Instance-Dependent Generalization Bounds via Optimal Transport

Songyan Hou * songyan.hou@ethz.ch
Department of Mathematics, ETH Zurich

Parnian Kassraie * pkassraie@ethz.ch
Department of Computer Science, ETH Zurich

Anastasis Kratsios * kratsioa@mcmaster.ca
Department of Mathematics, McMaster University

Jonas Rothfuss * jonas.rothfuss@inf.ethz.ch
Department of Computer Science, ETH Zurich

Andreas Krause krausea@ethz.ch
Department of Computer Science, ETH Zurich

Abstract

Existing generalization bounds fail to explain crucial factors that drive generalization of modern neural networks. Since such bounds often hold uniformly over all parameters, they suffer from over-parametrization, and fail to account for the fact that the set of parameters, considered during initialization and training, is much more restricted than the entire parameter space. As an alternative, we propose a novel *optimal transport* interpretation of the generalization problem. This allows us to derive *instance-dependent* generalization bounds that depend on the *local Lipschitz regularity* of the *learned prediction function* in the data space. Therefore, our bounds are agnostic to the parametrization of the model and work well when the number of training samples is much smaller than the number of parameters. With small modifications, our approach yields accelerated rates for data on *low-dimensional manifolds*, and guarantees under *distribution shifts*. We empirically analyze our generalization bounds for neural networks, showing that the bound values are meaningful and capture the effect of popular regularization methods during training.

Keywords: Generalization Bound, Instance-Dependent, Optimal Transport, Local Lipschitz Regularity

1. Introduction

A core challenge in machine learning is to generalize well beyond the training data. We want to choose a hypothesis $f \in \mathcal{F}$ that not only gives small training error but also yields good predictions for previously unseen data points. Accordingly, statistical learning theory aims to provide generalization guarantees and understand the factors that drive it. Generalization is typically described through the discrepancy between two key quantities: The empirical risk $\hat{\mathfrak{R}}(f)$, i.e., the prediction error of f on the training data and the expected risk $\mathfrak{R}(f)$, i.e., the expected error under the unknown data-distribution. A common type of guarantees are *uniform bounds* which control the generalization gap $\mathfrak{R}(f) - \hat{\mathfrak{R}}(f)$ with high probability, simultaneously for *all hypotheses* $f \in \mathcal{F}$ (e.g. Vapnik and Chervonenkis, 1971; Bartlett and

∗. Equal contribution, alphabetic order.

Mendelson, 2002). Such bounds include terms that quantify the complexity of the hypothesis f or hypothesis space $\mathcal{F}$. For neural networks (NNs), this complexity term grows rapidly with the number of parameters (e.g. Bartlett et al., 2017; Neyshabur et al., 2015b; Harvey et al., 2017). While the parameter space of NNs is vast, regular networks which are used in practice only seem to populate a small subset of the parameter space. This subset seemingly generalizes well, and depends on model structure, initialization scheme and optimization method in a complex manner. In addition, there are many NN parameter configurations that correspond to the same neural network mapping, artificially inflating the complexity of the parametric hypothesis space. Thus, such uniform bounds in the parameter space fail to explain the empirical generalization behavior of neural networks in the over-parameterized setting where the the number of training examples is much smaller than the number of parameters (Belkin et al., 2019).

Addressing this issue, we base our analysis on the geometric properties of the learned prediction function (i.e., hypothesis f) in the data domain. In particular, we partition the input domain into smaller neighborhoods, and locally characterize f via its *local Lipschitz constant* when the domain is restricted to each neighborhood. Using principles from *optimal transport*, we obtain a bound that depends on the instance f through its local Lipschitz constants, and is built on the following two key ideas. First, we view the generalization gap as the worst-case impact on the loss when probability mass is transported from the empirical measure to the true data distribution. The magnitude of this impact depends on the local regularity of f multiplied by the local transport cost which decreases w.h.p. with the number for samples. Second, unlike uniform bounds that hold with high probability simultaneously for all $f \in \mathcal{F}$, our analysis focuses one instance $f \in \mathcal{F}$. This approach is an alternative to the classic uniform bound, and allows us to forego arguments about the complexity of the hypothesis space which typically lead to extremely vacuous bounds.

Overall, the presented generalization bound (Theorem 4) has the following properties: 1) It is instance-dependent and thus can capture the combined effect of initialization, training method and model structure. 2) It characterizes f geometrically via its local Lipschitz regularity, therefore in contrast to parametric bounds, it does not suffer from over-parametrization. 3) It is tighter than bounds based on the global Lipschitz properties of f due to the fine-grained local analysis which takes into account changes in regularity of f throughout the domain. While our bounds generically hold for any machine learning model, we focus our exposition on neural network generalization, and empirically verify the mentioned properties through experiments. When applied to fully-connected ReLU networks, trained on simple regression and classification tasks, we observe that our result provides meaningful bound values in the same order of magnitude as the empirical risk, even for small sample sizes. We empirically show that, unlike the majority of prior works, the bound does not explode as the number of network parameters increases. Moreover, the value of the bound reflects the effect of regularization techniques applied *during* training, e.g. weight-decay, early-stopping and adversarial training.

Due to its transport-based derivation, our framework can be seamlessly adapted to obtain generalization certificates under distribution shifts or adversarial perturbations. Results mentioned above are corollaries of our core theorem, which is a optimal-transport-based concentration inequality for data-dependent locally regular functions. This theorem may be of independent interest and considers a spectrum of functions with different degrees of regularity, from non-smooth α-Hölder functions to smooth and s-time differentiable instances.

Outline The paper is structured as follows.

- Section 3 formalizes the problem setting and presents our main generalization bound (Theorem 4) together with an extension to when the data is known to be concentrated on a low-dimensional manifold (Proposition 5).
- Section 4 discusses of the key properties of our generalization bound which is instance-dependent (Section 4.1), non-parametric (Section 4.2), and localized (Section 4.3). Every section also presents corresponding experiments on neural networks.
- Section 5 considers instance-dependent generalization under distribution shifts, which is a natural corollary of our approach (Corollary 8).
- Section 6 focuses on our core result (Theorem 9). Section 6.1 highlights the key technical tools used for this theorem, and Section 6.2 outlines the proof methodology.

2. Related Work

Our work provides generalization bounds for learned prediction functions, contributing to the rich literature on generalization properties of learners. Classical generalization bounds often provide uniform or PAC-Bayesian guarantees on the learnability of a class of estimators, also referred to as the hypothesis space.

A popular class of uniform bounds depend on the combinatorial complexity of the hypothesis-space, e.g., expressed in form of the VC-dimension (Vapnik and Chervonenkis, 1971) or the Rademacher complexity (Koltchinskii, 2001; Bartlett and Mendelson, 2002). For neural networks, however, the hypothesis space is large and combinatorially explodes in size with the neural network width and depth, making the corresponding bounds loose (cf. Bartlett et al., 1998; Harvey et al., 2017; Bartlett et al., 2019; Sun et al., 2016). Bounds that utilize the parametric characterization of the network and depend on the norm of the neural network parameters similarly grow rapidly with the size of the neural network (Bartlett et al., 2017; Neyshabur et al., 2015b,a). Overall, these approaches hardly explain the empirical generalization behavior of neural networks in the over-parameterized setting, where the number of samples is much smaller than the number of parameters (Belkin et al., 2019). In fact, measures of neural network complexity based on the VC-dimension or parameter norm were found to be negatively correlated with the generalization gap for convolutional neural networks (Jiang et al., 2019b; Kuhn et al., 2021). In contrast, we present results which are instance-dependent and use the geometric properties, i.e. local regularity, of the learned prediction function f. This allows us to avoid the dependence on the combinatorial complexity of function classes as well as direct dependency on the parametrization of f.

Alternatively, PAC-Bayesian learning theory provides generalization bounds for random (Gibbs) learners (McAllester, 1998; Shawe-Taylor and Williamson, 1997; Catoni, 2007; Alquier et al., 2016; Mhammedi et al., 2019). Since PAC-Bayesian bounds do not trivially explode with the number of parameters of the model, they have gained increasing popularity in the context of neural networks (Langford and Caruana, 2001; Dziugaite and Roy, 2017;

Neyshabur et al., 2018; Zhou et al., 2018; Golowich et al., 2020). For instance, they have been related to the sharpness of minima, i.e. the robustness to perturbations in the weight space (Keskar et al., 2017; Neyshabur et al., 2017; Dziugaite and Roy, 2017), or the compressibility of a neural network (Zhou et al., 2018; Arora et al., 2018; Kuhn et al., 2021). Nonetheless, due to their inherent focus on a model's parameters, they still suffer from the standard pitfalls of the over-parameterized setting since these bounds become very loose once employed for larger networks. We argue that the generalization capability of a learner is directly influenced by the geometrical properties of the learned model in the data domain, rather than the number or values of its constructing parameters. Following this spirit, a small body of work uses the properties of the classification margin (Antos et al., 2002; Sokolic et al., 2017; Jiang et al., 2019a; Bartlett et al., 2017; Soudry et al., 2018; Gunasekar et al., 2018) to quantify generalization. A common flavor in such works, is that the generalization ability of neural networks relies crucially on the optimization procedure and can not be solely described by the hypothesis class. Following this logic, Dziugaite and Roy (2017, 2018) adjust the training procedure so that it minimizes the bounds, and thereby, attain non-vacuous PAC-Bayesian bounds. While the bound of Dziugaite and Roy (2018) depends on a instance-dependent prior, it considers generalization error with respect to a posterior distribution over neural network parameters. In contrast, we focus on the generalization properties of a single learning hypothesis (e.g., a single neural network) which is the result of training.

Our work is also connected to a body of literature that quantifies the local regularity of the learned prediction function. Examples of this are counting the number of linear regions of train neural neural networks, (Montufar et al., 2014), calculating the local Lipschitz constant of neural networks (Jordan and Dimakis, 2020; Herrera et al., 2020) or the local Rademacher complexity (Bartlett et al., 2005).

We also contribute to the literature of distributional robust optimization, since with little effort, our bounds can be extended into a distributional robustness certificate (see Section 5). Our bound suggests that locally Lipschitz estimators are more robust to distribution shifts, confirming recent results which control the global or local Lipschitz constants in order to achieve adversarially robust neural networks (Cisse et al., 2017; Salman et al., 2019; Cohen et al., 2019; Gouk et al., 2021; Anil et al., 2019; Muthukumar and Sulam, 2022). Similar to recent work on distributional robustness (Gao and Kleywegt, 2022; Kuhn et al., 2019; Cranko et al., 2021), we rely on a transport-based change of measure inequality. However, while this work only bounds the difference between expected risk under distribution shift and empirical validation error (i.e., a deviation bound), we present a stronger result which bounds the gap to the training error. We provide a more in-depth comparison in Section 5, once the notation is formally set.

3. Instance Dependent Bound on Generalization Error

We consider datapoints (x, y) where $x \in \mathcal{X}$ are observed input features and $y \in \mathcal{Y}$ are target values/labels. To formulate the learning problem, we assume that the data is generated via an *unknown* probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Given a dataset $\mathcal{D}^N = \{(x_i, y_i)\}_{i=1}^N$ of i.i.d. draws from μ, the goal of supervised learning is to find a function $\hat{f}^N$ which can accurately predict the targets. The quality of an estimator $\hat{f}^N$ is measured through a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Accordingly, we seek to attain a small *expected risk*, i.e., the expected loss under the data generating distribution

$$\mathfrak{R}(\hat{f}^N;\mu) \coloneqq \mathbb{E}_{(x,y)\sim\mu}\left[\ell(\hat{f}^N(x),y))\right].$$

Since μ is unknown, it is not possible to directly evaluate $\mathfrak{R}(\hat{f}^N;\mu)$ given the training data $\mathcal{D}^N$. However, based on the dataset, we can compute the *empirical risk*

$$\hat{\mathfrak{R}}(\hat{f}^N) \coloneqq \mathfrak{R}(\hat{f}^N;\mu^N) = \frac{1}{N}\sum_{n=1}^{N}\ell(\hat{f}^N(x_i),y_i).$$

which corresponds the expected loss under an empirical measure $\mu^N := \frac{1}{N}\sum_{i=1}^N \delta_{(x_i,y_i)}$ for $\mathcal{D}^N$. Often $\hat{\mathfrak{R}}(\hat{f}^N)$ is also referred to as training error. In this work, we aim to bound the *generalization gap* $\mathfrak{R}(\hat{f}^N;\mu) - \hat{\mathfrak{R}}(\hat{f}^N)$. Importantly, as $\hat{f}^N$ already depends on the data $\mathcal{D}^N$, $\hat{\mathfrak{R}}(\hat{f}^N)$ is a biased estimator of the expected risk $\mathfrak{R}(\hat{f}^N;\mu)$. Thus, standard results for the concentration of averages do not apply. Instead, to bound the generalization gap, we also need to take into consideration the learning hypothesis $\hat{f}^N$ and quantify how well it generalizes from the training data $\mathcal{D}^N$ to the general data distribution μ.

In the following, we introduce the basic assumptions and tools which form the foundation of our generalization bounds:

Assumption 1 *The domain $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, the d-dimensional Euclidean space, and $\mathcal{Y}$ is a compact subset of $\mathbb{R}$.*

The assumption that $\mathcal{X}$ and $\mathcal{Y}$ are compact are very common in statistical learning theory and imply that the bounded is the target values y are observed with a bounded noise. For instance, they are commonly used for uniform generalization bounds (e.g., Alon et al., 1997; Bartlett and Mendelson, 2002), PAC-Bayesian Bounds (e.g., McAllester, 1998; Catoni, 2007), and the more recent instance-dependent generalization bounds (e.g., Dziugaite and Roy, 2017; Neyshabur et al., 2018; Golowich et al., 2020).

In addition, we require geometric regularity assumptions on both the estimator and the loss function. For this purpose, we define the *local* Lipschitz constant of a function $g:\mathcal{X}\to\mathcal{Y}$ when restricted to $P\subset\mathcal{X}$ as,

$$\mathrm{Lip}(g|P) \coloneqq \sup_{\substack{x_1,x_2\in P\\ x_1\neq x_2}} \frac{|g(x_1)-g(x_2)|}{\|x_1-x_2\|}.$$

For $0\leq L<\infty$, we say that a function g is L-Lipschitz if the global Lipschitz constant $\mathrm{Lip}(g)\coloneqq \mathrm{Lip}(g|\mathcal{X})$ is bounded by L. We assume that the learned function $\hat{f}$ is Lipschitz continuous with Lipschitz constant $L_{\hat{f}}$:

Assumption 2 *There exists a constant $L_{\hat{f}}\geq 0$ such that the estimator $\hat{f}^N$ almost surely satisfies* $\mathrm{Lip}(\hat{f}^N)\leq L_{\hat{f}}$.

Lipschitz estimators are perhaps the most common class of estimators and include, Gaussian processes with non-smooth kernels, and neural networks with popular activation functions such as ReLUs, ELUs and tanh functions. In Section 6 we extend our result to α-Hölder and smooth estimators. We also require a Lipschitz loss function:

Assumption 3 *The loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is L_ℓ-Lipschitz.*

Examples of Lipschitz continuous loss functions for classification are logit, hinge or ramp loss (Hajek and Raginsky, 2019). A Lipschitz loss for regression is the Huber loss, which satisfies $L_\ell = 1$. This loss is commonly used for training neural networks (e.g Morales, 2020; Meyer, 2021), since compared to the squared error loss, it is more robust to outliers and large gradients that destabilize training.

We take a localized approach and instead of bounding the generalization error directly on the entire $\mathcal{X} \times \mathcal{Y}$ space, we first partition the space, and then compare the empirical and expected risk separately on each element of this partitioning. A partitioning $\boldsymbol{P}$ of size k is a collection $\{P_1, \ldots, P_i, \ldots, P_k\}$ subsets of $\mathcal{X} \times \mathcal{Y}$, where $P_i \cap P_j = \emptyset$ and $\cup_{i=1}^k P_i = \mathcal{X} \times \mathcal{Y}$, for every $1 \leq i < j \leq k$. Consequently, our analysis relies on two key localized notions: $\mathrm{Lip}(\hat{f}^N|P)$ the local Lipschitz constant of the estimator restricted to a part $P \in \boldsymbol{P}$, and $\mu|_P$ the data generating distribution restricted to P, defined via $\mu|_P(\cdot) \coloneqq \mu(\cdot \cap P)/\mu(P)$. The localized empirical distribution can be similarly defined as $\mu^N|_P(\cdot) \coloneqq \mu^N(\cdot \cap P)\mu^N(P)$. We note that $\mu^N(P) = N_P/N$ where $N_P \coloneqq |\{\mathcal{D}^N \cap P\}|$ counts the number of samples which fall into the set P. We are now ready to present our instance-dependent bound on the generalization error. This theorem is a corollary of our main result of Theorem 9, and Appendix A.1 presents its proof.

Theorem 4 (Generalization error of Lipschitz estimators) *Let $\hat{f}^N$ be a learned function which may depend on the dataset $\mathcal{D}^N$. Suppose Assumptions 1, 2, and 3 hold with some $L_\ell, L_{\hat{f}} > 0$. Then, for any data-independent partitioning $\boldsymbol{P}$ of $\mathcal{X} \times \mathcal{Y}$, we have*

$$\mathfrak{R}(\hat{f}^N; \mu) - \hat{\mathfrak{R}}(\hat{f}^N) \leq \mathrm{cost}_{\mathrm{transport}}(\boldsymbol{P}) + \mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}) + \mathrm{cost}_{\mathrm{partition}}(\boldsymbol{P})$$

with probability greater than $1 - \delta$, where

$$\begin{aligned}
\mathrm{cost}_{\mathrm{transport}}(\boldsymbol{P}) &\coloneqq \frac{C_{d+1,1} L_\ell}{N} \sum_{P \in \boldsymbol{P}} N_P^{\frac{d}{d+1}} \max\left\{1, \mathrm{Lip}(\hat{f}^N|P_{\mathcal{X}})\right\} \mathrm{diam}(P), \\
\mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}) &\coloneqq \sqrt{\frac{\ln(4/\delta)}{N}} L_\ell \max\{1, L_{\hat{f}}\} \max_{P \in \boldsymbol{P}} \mathrm{diam}(P), \\
\mathrm{cost}_{\mathrm{partition}}(\boldsymbol{P}) &\coloneqq \begin{cases} \|\ell\|_\infty \max\left\{\sqrt{\frac{2\ln(4/\delta)}{N}}, \sqrt{\frac{|\boldsymbol{P}|}{N}}\right\} & |\boldsymbol{P}| > 1 \\ 0 & |\boldsymbol{P}| = 1 \end{cases}
\end{aligned}$$

Here $P_{\mathcal{X}}$ denotes the projection of P onto $\mathcal{X}$, and the constant $C_{d+1,1}$ is recorded in Table 1.

The generalization gap is the discrepancy in calculating the expectation of the loss calculated with respect to the two distributions μ and μ^N. Intuitively, our bound is based on the cost from transporting probability mass from μ^N to μ. In particular, this cost accrues from how far do we have to transport probability mass on average and how much the loss can change in process. We perform this analysis transport-based analysis locally, by partitioning $\mathcal{X}$ and bounding the cost of changing the measure from $\mu|_P$ to $\mu^N|_P$ for every $P \in \boldsymbol{P}$. Since the dataset $\mathcal{D}^N$ is drawn at random, this cost is a random variable. The term $\mathrm{cost}_{\mathrm{transport}}$ upper bounds the expected value of this cost, and the term $\mathrm{err}_{\mathrm{transport}}$ controls the deviation error. The last term $\mathrm{cost}_{\mathrm{partition}}$ denotes the cost we pay for partitioning, and it is equal to zero if $\boldsymbol{P} = \{\mathcal{X} \times \mathcal{Y}\}$. The previous two terms account for transporting probability mass within parts of the domain. However, if $\mu(P) \neq \mu^N(P)$, mass also needs to be transported across parts. $\mathrm{cost}_{\mathrm{partition}}$ upper bounds the potential change in the risk due to this global transport of mass. Naturally, the more parts we have in our partitioning, the higher the $\mathrm{cost}_{\mathrm{partition}}$.

The error bound of Theorem 4 converges with $\mathcal{O}(N^{-1/(d+1)})$ which, for higher dimensional domains, implies relatively slow convergence. However, this rate is already an improvement upon Rademacher generalization bounds for Lipschitz estimators (see Section 4.1). We do not impose any constraints on μ other than having a compact support. Thus, our bound holds for any $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, also unfavorable edge-cases such as a uniform distribution over the domain. Without further assumptions, the optimal-transport cost (i.e. Wasserstein distance) of μ^N to μ inherently have an exponential dependence on the dimensionality of the domain.

In many applications with high-dimensional data domains it has been postulated that the data lies on some low-dimensional manifold (Narayanan and Mitter, 2010; Fefferman et al., 2016). For instance, Pope et al. (2021) empirically demonstrate the validity of this assumption on popular image datasets. Under the assumption that the data lies on a $\tilde{d}$-dimensional manifold where $\tilde{d} \ll d$, we can improve the convergence rate. Proposition 5 shows that the generalization error would then *only* depend on the intrinsic dimension $\tilde{d}$. The proof is presented in Appendix A.2.

Proposition 5 (Fast rates for structured data) *Consider the setting and assumptions of Theorem 4. In addition, suppose μ is such that the data lies almost surely on a $\tilde{d}$-dimensional C^1-Riemannian manifold. Then for any $0 < \delta \leq 1$ and any data-independent partitioning $\boldsymbol{P}$ on $\mathcal{X} \times \mathcal{Y}$, there exists $C(\tilde{d})$ for which*

$$\mathfrak{R}(\hat{f}^N; \mu) - \hat{\mathfrak{R}}(\hat{f}^N) \leq \frac{C(\tilde{d}) L_\ell}{N} \sum_{P \in \boldsymbol{P}} N_P^{1-1/\tilde{d}} \max\left\{1, \mathrm{Lip}(\hat{f}^N | P_{\mathcal{X}})\right\} \mathrm{diam}(P) + \mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}) + \mathrm{cost}_{\mathrm{partition}}(\boldsymbol{P})$$

with probability $1-\delta$ where the terms $\mathrm{err}_{\mathrm{transport}}$ and $\mathrm{cost}_{\mathrm{partition}}$ are as defined in Theorem 4.

Bounds of Theorem 4 and Proposition 5 can be made tighter by directly considering $\mathrm{Lip}(\ell \circ \hat{f}^N | P)$ the local Lipschitz regularity of the *composition* of the loss and the prediction function. In fact, a direct instantiation of Theorem 9 would result a bound that depends on this quantity. However, for a clearer exposition, we split the Lipschitz constants of $\ell \circ \hat{f}^N$ using $\mathrm{Lip}(\ell \circ \hat{f}^N | P) \leq L_\ell \cdot \mathrm{Lip}(\hat{f}^N | P)$, and present the bounds in terms of the global Lipschitz constant of the loss. In our numerical experiments, we use the tighter variant based on $\mathrm{Lip}(\ell \circ \hat{f}^N | P)$.

4. Key Properties of the Generalization Bound: NN Perspective

Theorem 4 presents a generalization bound that is *instance-dependent*, *non-parametric* and *localized*. Further, it captures the post-training properties of the learned prediction function $\hat{f}^N$. In this section, we elaborate on these properties with a focus on neural networks. As an empirical running example, we consider two simple supervised learning tasks. We generate synthetic random datasets for 1D regression and 2D binary classification (Fig. 7), and train overparametrized fully-connected ReLU networks on them with stochastic gradient descent (SGD). We then evaluate the bound of Theorem 4 for the resulting estimator.[1] Details of the experiments are reserved for Appendix D. To calculate the local Lipschitz constant $\mathrm{Lip}(\hat{f}|P)$, we simply consider a fine grid of the domain and evaluate the gradient of the network over this mesh. This only requires light computations, since our toy examples are two-dimensional at most. For higher dimensional domains, Jordan and Dimakis (2020) and Fazlyab et al. (2019) propose scalable algorithms that approximate the local Lipschitz constant of a neural network. For the regression task, we use the Huber loss (Equation D.2). Since, the Huber loss has a Lipschitz constant of $L_\ell = 1$, Theorem 4 applies directly to the regression case. For the binary classification, we use the labels $\mathcal{Y} = \{-1, 1\}$ and aim to bound the expected classification error $\mathbb{P}(\hat{f}^N(X) \neq Y)$. The 0-1 classification error $\mathbf{1}(\hat{f}^N(x) \cdot y < 0)$ is not Lipschitz. However, following Hajek and Raginsky (2019), we use the ramp loss

$$\ell_\gamma(y_1, y_2) \coloneqq \min\left\{1, \left(1 - \tfrac{y_1 y_2}{\gamma}\right)_+\right\} \ , \text{ with } \gamma > 0 \ , \tag{1}$$

as a Lipschitz proxy and upper bound of the 0-1 loss. This allows us to obtain a corollary of Theorem 4 which upper bounds on the classification error:

Corollary 6 (Classification error bound) *Consider a compact input domain, and labels in $\mathcal{Y} = \{-1, 1\}$. Assume that the observation noise is i.i.d and may only flip the label. Let $\gamma > 0$, $\boldsymbol{P}$ be any partitioning of size k on $\mathcal{X}$, independent of the data $\mathcal{D}^N$. Then under Assumption 3, with probability greater than $1 - \delta$,*

$$\begin{aligned}\mathbb{P}(\hat{f}^N(X) \neq Y) \leq & \frac{1}{N}\sum_{i=1}^{N} \ell_\gamma(\hat{f}^N(X_i), Y_i) + \frac{2^{1/d} C_{d,1}}{\gamma} \sum_{P \in \boldsymbol{P}} \frac{N_P^{1-1/d}}{N} \mathrm{Lip}(\hat{f}^N|P)\mathrm{diam}(P) \\ & + \sqrt{\frac{\ln(4/\delta)}{N}} \frac{L_{\hat{f}}}{\gamma} \max_{P \in \boldsymbol{P}} \mathrm{diam}(P) + \sqrt{\frac{2}{N}} \max\left\{\sqrt{\ln(4/\delta)}, \sqrt{k}\right\}\end{aligned}$$

here $N_P = \left|\left\{(X, Y) \in \mathcal{D}^N \ \ s.t. \ \ X \in P\right\}\right|$.

The proof of Corollary 6 is given in Appendix A.3. Since for classification $\mathcal{Y}$ is only a finite set, we can marginally reduce the dimension dependence of the generic bound from $\mathcal{O}(N^{1-1/(d+1)})$ to $\mathcal{O}(N^{1-1/d})$.

4.1 Instance-Dependent vs. Uniform

Understanding generalization of overparametrized neural networks requires analyzing the *combination* of model architecture, initialization method, and training procedure. A trained network $\hat{f}^N$ inherits the joint effect of the three elements. Therefore, instance-dependent bounds, which are calculated for $\hat{f}^N$ post-hoc are more informative in describing the generalization behaviour of overparametrized networks than uniform bounds which only consider the complexity of the hypothesis space and thus neglect the effect of initialization and training procedure.

1. More precisely, we visualize the tighter variant of Theorem 4 which directly depends on $\mathrm{Lip}(\ell \circ \hat{f}^N|P)$, since splitting the constant as $L_\ell \cdot \mathrm{Lip}(\hat{f}^N|P)$ may loosen the bound.

Such bounds also allow the practitioner to predict the test error of the model once it is trained, and can be used as a certificate for model selection between a finite number of estimators.

Theorem 4 bounds the generalization error of an instance $\hat{f}^N$, as oppose to uniformly bounding the error for any f within a hypothesis class $\mathcal{F}$ (e.g., Blumer et al., 1989). We compare Theorem 4 with a classic uniform law on generalization of Lipschitz estimators, since our only assumption about the estimator is almost sure Lipschitz continuity. Let $\mathcal{F}_L$ denote the class of L-Lipschitz functions mapping $\mathcal{X}$ to $\mathcal{Y}$, and recall that $\mathcal{D}^N$ is an i.i.d. random sample of size N drawn according to the probability distribution μ. Under Assumption 1 and 3, the Rademacher generalization bound (Theorem 9, Bartlett and Mendelson, 2002) implies that, with probability $1-\delta$ there exists $C > 0$ for which *every* $f \in \mathcal{F}_L$ satisfies

$$\mathfrak{R}(f;\mu) - \hat{\mathfrak{R}}(f) \leq C L_\ell \left(\frac{(\mathrm{diam}(\mathcal{X})L)^d \, d^2 D^2}{N} \right)^{1/(d+2)} + \|\ell\|_\infty \sqrt{\frac{8 \log 2/\delta}{N}} \tag{2}$$

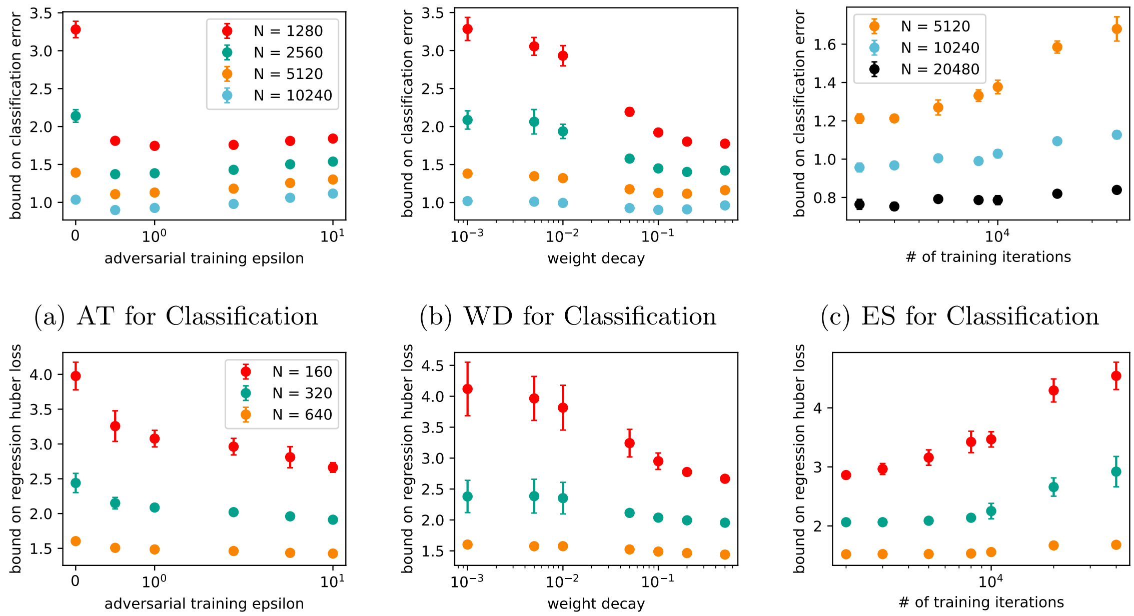
where $D \coloneqq \sup_{f \in \mathcal{F}_L} \|f\|_\infty$. In Appendix C we formalize this statement and provide a proof for completeness. The first term on the right-hand-side of (2), which corresponds to the Rademacher complexity of $\mathcal{F}_L$, dominates this bound. It rapidly grows for large high-dimensional domains, or for a large Lipschitz constant, and converges at a $O(N^{-1/(d+2)})$ rate. Theorem 4 only marginally improves upon this rate since it converges with $O(N^{-1/(d+1)})$. However, the value of the constants are significantly smaller. The term $\mathrm{cost}_{\mathrm{transport}}$ has the slowest decay with N, and its constant is proportional to $\mathrm{Lip}(\hat{f}^N|P)\mathrm{diam}(P)$. Consequently, for a typical estimator $\hat{f}^N$ the bound of Theorem 4 would be much tighter than (2).

Uniform bounds only reflect the properties of the function class, and are known to be vacuous for large hypothesis classes such as neural networks (Bartlett and Long, 2021; Golowich et al., 2020). We empirically evaluate our generalization bound when applied to neural networks trained on regression and classification tasks. Figures 2 and 5 show that, contrary to the majority of prior works which deem vacuous for overparametrized neural networks, our bound assumes values in the same order of magnitude as the expected error and becomes non-vacuous for ca. $N > 10000$ classification examples.

Furthermore, our instance-dependent bound reflects the positive effect of common regularization techniques on the generalization. It is known that certain training techniques, such as adversarial training, weight decay and early stopping, can lead stochastic gradient descent to solutions that generalize better. Fig. 1 illustrates the effect of these methods on our bound when applied to neural networks. As we can observe in Fig. 1, the bound improves once the aforementioned regularization techniques are employed during training. In particular, for smaller samples sizes, the change in the value of the bound suggests that all three methods of adversarial training, weight decay and early stopping produce networks that tend to generalize better. This observation matches the prior works of Xing et al. (2021), Krogh and Hertz (1991) and Li et al. (2020) respectively. This empirically supports our core idea that, since Theorem 4 directly depends on the learned neural network instance $\hat{f}^N$, it is able to capture the joint effect of model structure, initialization and training.

4.2 Geometric vs. Parametric Characterization of the Estimator

We characterize the estimator via its local Lipschitz regularity. A key idea in our work is that this local geometry has an immediate effect on the generalization ability of the network,



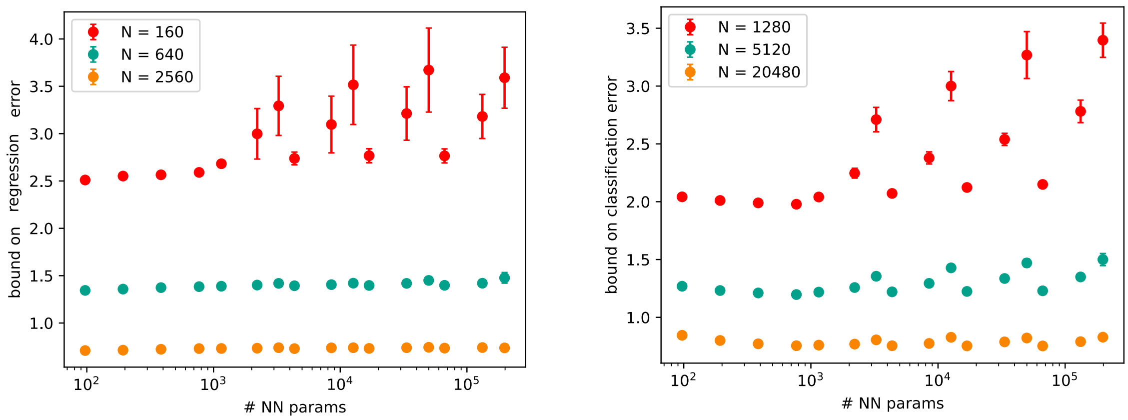
(a) AT for Classification (b) WD for Classification (c) ES for Classification

(d) AT for Regression (e) WD for Classification (f) ES for Regression

Figure 1: Effect of adversarial training (AT), weight decay (WD) and early stopping (ES). Generalization bounds of Theorem 4 and Corollary 6 suggest that these training techniques result in networks that generalize better. See Appendix D.3 for details of the plots.

compared to the network size or architecture. An alternative approach, are parametric bounds which consider the network structure. Such bounds are data-dependent and are often a function of the Frobenius norm of network's weights, i.e. $\|\boldsymbol{W}_j\|_F$ where $1 \leq j \leq l$ indexes the layer number. Examples are the rademacher-type bound of Neyshabur et al. (2015b), Bartlett et al. (2017) which follows a covering number argument, and Neyshabur et al. (2018) which takes a PAC-Bayesian approach. These norm-based bounds roughly grow with $\mathcal{O}(\mathrm{diam}(\mathcal{X})\mathrm{Poly}(d,h,l)\prod_{j=1}^{l}\|\boldsymbol{W}_j\|_F\sqrt{1/N})$, where h denotes the width of the network.[2] Golowich et al. (2020) improve prior results and present a bounds of the rate $\mathcal{O}(\mathrm{diam}(\mathcal{X})\prod_{j=1}^{l}\|\boldsymbol{W}_j\|_F\sqrt{l/N})$. While these bounds have a polynomial dependency on the input dimension d, they quickly become vacuous for larger networks to due their polynomial dependence on network size. Therefore such bounds fail to capture or explain the benefit of over-parametrization (Bartlett and Long, 2021), in contrary to the observation that large neural networks tend to generalize better in practice (Zhang et al., 2017). A key issue with parametric analysis is that there are combinatorially many parameter configurations, or in cases even entire sub-spaces that correspond to the same neural network mapping. This artificially inflates the hypothesis space compared to the set of neural network functions that is actually considered by the learning algorithm.

2. Not all the bounds have this polynomial dependency, e.g. the bound of Neyshabur et al. (2015b) depends on the network depth exponentially.



(a) Regression

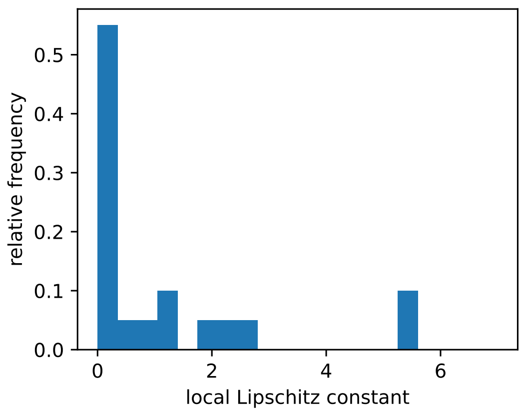
(b) Classification

Figure 2: Generalization error bound vs. number of neural network parameters. The neural network size has only a minor effect on the bound values. See Appendix D.4 for details.

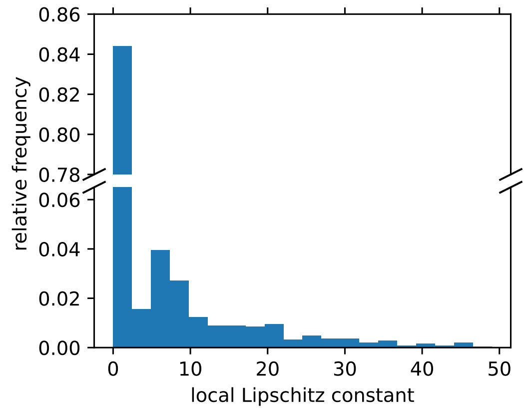
Our bound is based on the geometry of the learned function rather than its parameters, and it does not suffer from the described issue. We observe that increasing the network size has negligible effects on the local regularity of the learned prediction function. Thus, the generalization bound of Theorem 4 hardly grows with the neural network size. To demonstrate this empirically, we train networks of increasingly larger width and depth, and calculate the corresponding bounds. Fig. 2 shows that even increasing the number of neural network parameters by factor 1000 has only a minor effect on the value of the bound, in particular, when the dataset size is large. Our geometry-based approach seem to capture the empirically observed generalization behavior of over-parametrized neural networks better than previous bounds.

4.3 Localized vs. Global Analysis

In Theorem 4 we partition the domain into many subsets and locally bound the generalization risk when the domain is restricted to each subset. As a result, $\mathrm{cost}_{\mathrm{transport}}$, which is often the largest term of the bound and shrinks with N the slowest, is independent of the global Lipschitz constant and depends on $\hat{f}^N$ only through $\mathrm{Lip}(\hat{f}^N|P)$. When applied to neural networks, Theorem 4 tends to benefit from this localized analysis, as the local regularity of trained networks strongly varies across the domain. Figure 3 displays the distribution of local Lipschitz constants across parts $P \in \boldsymbol{P}$. We observe that $\mathrm{Lip}(\hat{f}^N|P)$ is fairly low in the majority of parts, while there exist a few parts with much higher local Lipschitz constant, which contributes to a large overall global constant. Therefore, we expect that our localize analysis to be more informative than a global argument which treats all parts uniformly and in turn, produces a bound depending on $\mathrm{diam}(\mathcal{X})\mathrm{Lip}(\hat{f}^N)$. Fig. 4 empirically supports this claim by visualizing the local generalization error restricted to each part P, which changes strongly across the neighborhoods. In particular, we observe that the local



(a) Regression



(b) Classification

Figure 3: Frequency of local Lipschitz constant values per part $P \in \boldsymbol{P}$ for fully-connected ReLU nets trained with SGD. The local Lipschitz constant is small in the majority of parts P, contrary to the large global constant. For training details see Appendix D.2.

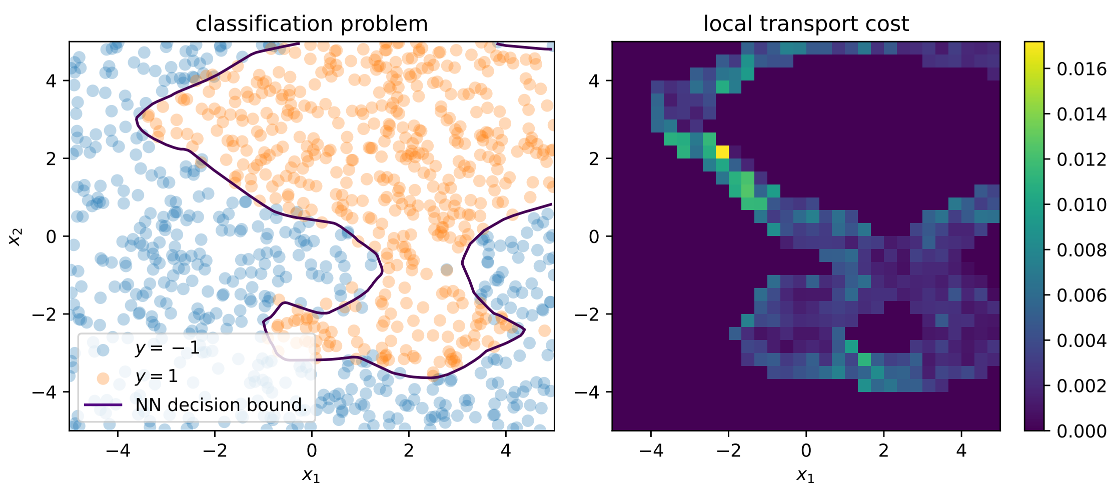


Figure 4: Local break down of $\mathrm{cost}_{\mathrm{transport}}$. Areas closer to the decision boundary of $\hat{f}^N$ contribute more to the bound. Each mosaic P in the heatmap shows the local transport cost $\mathrm{Lip}(\ell \circ \hat{f}^N|P)\mathrm{diam}(P)\sqrt{N_P}/N$.

generalization error is small in areas away from the decision boundary of $\hat{f}^N$, and that it achieves its maximum in an area where the estimator misclassifies the training data.

As we partition the data domain into ever finer parts, two forces are at play: The diameter of each part P shrinks and the local Lipschitz constant may get smaller, making $\mathrm{cost}_{\mathrm{transport}}$ and $\mathrm{err}_{\mathrm{transport}}$ in Theorem 4 shrink. At the same time, as parts become smaller, mismatches in probability mass of μ and μ^N become more pronounced so that we have to account for more transport of mass across the partitions, increasing the $\mathrm{cost}_{\mathrm{partition}}$ term. Hence, by making partitioning finer, we trade off $\mathrm{cost}_{\mathrm{transport}}$ and $\mathrm{err}_{\mathrm{transport}}$ against $\mathrm{cost}_{\mathrm{partition}}$. Figure 5 displays our bounds in response to an increasing partition size. We can empirically observe the the trade-off as the bound values initially decrease, and, as the partitioning becomes much fines, increase again. Since the marginal gains from a finer partitioning decrease, there is typically a sweet-spot, i.e., a degree partitioning that leads to the tightest bounds.

Without partitioning, or equivalently by considering only one *global* partition $\boldsymbol{P}_{\mathrm{global}} \coloneqq \{\mathcal{X} \times \mathcal{Y}\}$, we obtain the *global* counterpart of Theorem 4:

$$\mathfrak{R}(\hat{f}^N; \mu) - \hat{\mathfrak{R}}(\hat{f}^N) \leq \mathrm{cost}_{\mathrm{transport}}(\boldsymbol{P}_{\mathrm{global}}) + \mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}_{\mathrm{global}}). \tag{3}$$

Fig. 5 also visualizes this global bound, and empirically confirms that partitioning often has a significant advantage over the global analysis. In particular, for the classification task, the global bound (8-20 times larger) is vacuous while partitioning allows to achieve a non-vacuous guarantees. Crucially, there exist estimators for which the generalization error of Theorem 4 is bounded and vanishes as $N \to \infty$, while the global inequality (3) diverges. In Proposition 7, we construct such an estimator. Perhaps surprisingly, we prove that the global bound diverges already for a shallow ReLU network with exactly one neuron defined over $\mathcal{X} = [0, 1]$, while a localized analysis with a partition with size of order $|\boldsymbol{P}| = \mathcal{O}(N^{0.6})$ gives a vanishing error bound. The proof is presented in Appendix A.4.

Proposition 7 (Partitioning can help) *Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = [0, 1]$. For any probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, there is an increasing sequence of Lipschitz constants $\{L_{f,N}\}_{N=1}^{\infty}$, a sequence of partitions $\{\boldsymbol{P}_N\}_{N=1}^{\infty}$, and a sequence of ReLU feedforward networks with one neuron $\{f_N\}_{N=1}^{\infty}$, such that the local bound of Theorem 4 for the tuple $(L_{f,N}, \boldsymbol{P}_N, f_N)$ converges:*

$$\lim_{N\to\infty} \mathrm{cost}_{\mathrm{transport}}(\boldsymbol{P}_N) + \mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}_N) + \mathrm{cost}_{\mathrm{partition}}(\boldsymbol{P}_N) \to 0$$

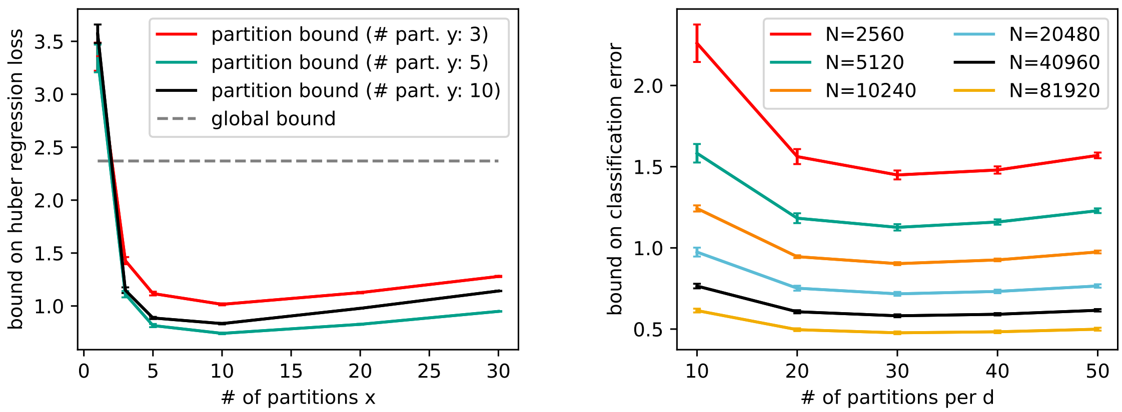
while the global bound of Equation (3) diverges:

$$\lim_{N\to\infty} \mathrm{cost}_{\mathrm{transport}}\left(\boldsymbol{P}_{\mathrm{global}}\right) + \mathrm{err}_{\mathrm{transport}}\left(\boldsymbol{P}_{\mathrm{global}}\right) \to \infty.$$

where $\boldsymbol{P}_{\mathrm{global}} = \{[0, 1] \times [0, 1]\}$, and the terms $\mathrm{cost}_{\mathrm{transport}}, \mathrm{err}_{\mathrm{transport}}$ and $\mathrm{cost}_{\mathrm{partition}}$ are defined as in Theorem 4.

5. Generalization under Distribution Shifts

A desirable characteristic of an estimator $\hat{f}^N$ is robustness to changes in the data generating distribution μ between train and test time. A change in μ may be due to covariate-shifts, adversarial attacks, or small changes in data-generating process over time. In safety-critical applications such as perception systems in self-deriving cars or models for medical diagnosis, it is crucial that we can certify performance of $\hat{f}^N$ under changes in the distribution. In this section, we employ our framework to obtain an instance-dependent generalization bound under distribution shift. In particular, we bound the risk calculated with respect to a shifted distribution μ^{adv}, by the *training* error of the estimator on a dataset of size N which is sampled from data generating distribution μ.



(a) Regression. All curves correspond to $N = 2560$ samples.

(b) Classification. Depending on N, the global bound is 8-20 times larger (see Table 2).

Figure 5: Generalization error bound values for a varying number of partitions. Left: Bound on the Huber regression loss for differing numbers of partitions in $\mathcal{X}$ and $\mathcal{Y}$, together with the global Lipschitz bound in Eq. 3. Right: Bound on the classification error for differing numbers partitions per dimension of $\mathcal{X}$. For more details see Appendix D.5.

In addition to the transport of mass from μ^N to μ which is at the core of Theorem 4, our optimal transport based approach allows us seamlessly consider the additional change of measure from μ to μ^{adv}. In particular, we use the Wassertstein-1 distance $\mathcal{W}_1(\mu, \mu^{\text{adv}})$ to quantify the amount of distribution shift. This is consistent with previous work on robustness certificates which are often given for distributions within an ϵ-Wasserstein ball centered in the data generating distribution (Sinha et al., 2018; Lee and Raginsky, 2018; Blanchet and Murthy, 2019; Levine and Feizi, 2020; Gao and Kleywegt, 2022). The Wasserstein-1 distance is defined as the minimum ℓ_1-cost of transporting probability mass from ν_1 to ν_2, that

$$\mathcal{W}_1(\nu_1, \nu_2) \coloneqq \inf_{\gamma \in \Gamma(\nu_1, \nu_2)} \int_{\mathcal{Z} \times \mathcal{Z}} \left\| x - x' \right\|_1 d(\gamma(x, x'))$$

where $\Gamma(\nu_1, \nu_2) \subset \mathcal{P}(\mathcal{Z} \times \mathcal{Z})$ denotes the set of all couplings between ν_1 and ν_2, in other words, the set of joint distributions who's marginals are ν_1 and ν_2.

Corollary 8 presents our instance-dependent generalization bound under distribution shift. For simplicity, we present this result in the global case of $\boldsymbol{P}_{\text{global}} = \{\mathcal{X} \times \mathcal{Y}\}$, i.e., without partitioning. However the analysis can also be carried out locally with partitioning analogous to Theorem 4. The proof is given in Appendix A.5.

Corollary 8 (Locally Lipschitz estimators are robust to distribution shift) *Let $\mu^{\text{adv}} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Under Assumptions 1, 2, and 3 with probability greater than $1 - \delta$, we have*

$$\begin{aligned} \mathfrak{R}(\hat{f}^N; \mu^{\text{adv}}) - \hat{\mathfrak{R}}(\hat{f}^N) \leq\ & \text{cost}_{\text{transport}}\left(\boldsymbol{P}_{\text{global}}\right) + \text{err}_{\text{transport}}(\boldsymbol{P}_{\text{global}}) \\ & + L_\ell \max\left(1, \text{Lip}(\hat{f}^N)\right) \mathcal{W}_1\left(\mu, \mu^{\text{adv}}\right) \end{aligned}$$

where $\boldsymbol{P}$ denotes a data-independent partition on $\mathcal{X}$, and $(\text{cost}_{\text{transport}}, \text{err}_{\text{transport}})$ are identical to that of Theorem 4.

The bound of Corollary 8 has the same generalization terms as the global bound in Equation (3), plus an additional term which accounts for the potentially negative impact of the distribution shift by multiplying $\mathcal{W}_1\left(\mu, \mu^{\text{adv}}\right)$ with the Lipschitz constants of the ℓ and $\hat{f}^N$. In particular, if $\mathcal{W}_1(\mu, \mu^{\text{adv}}) \leq \epsilon$, then the corollary implies that in the worst case, the estimator suffers from a $\epsilon L_\ell \text{Lip}(\hat{f}^N)$ increase in risk when evaluated on the perturbed distribution.

There are many connections between Corollary 8 and prior work on robust estimators. Gao (2022) verifies distributionally robust learnability of $\mathcal{F}_L$ the class of Lipschitz functions through a uniform bound. We expect this bound to be vacuous if calculated empirically, since it has large terms depending on the complexity of the function class, e.g. via Rademacher complexity or metric entropy. Kuhn et al. (2019) and Cranko et al. (2021) bound the difference between the finite-sample validation error of a fixed (data-independent) estimator f, and $\mathfrak{R}(f; \mu^{\text{adv}})$. Both works use this robustness certificate to develop methods for distributionally robust optimization. Perhaps, closest to our result, is Mohajerin Esfahani and Kuhn (2018) which gives a generalization bound for a robust estimator defined via

$$\hat{f}_\epsilon^N = \underset{f \in \mathcal{F}_L}{\arg\min} \max_{\mathcal{W}_1(\mu^N, \mu^{\text{adv}}) \leq \epsilon} \mathcal{R}(f; \mu^{\text{adv}}).$$

In contrast, Corollary 8 holds for any data-dependent Lipschitz estimator. In practice, we have access to the training error $\hat{\mathfrak{R}}(\hat{f}^N)$ and aim to verify how this performance generalizes to unseen data generated from a shifted distribution ($\mathfrak{R}(f; \mu^{\text{adv}})$). Therefore, instance dependent generalization bounds such as Corollary 8 or Mohajerin Esfahani and Kuhn (2018) are of more practical relevance, compared to uniform (Gao, 2022) or deviation bounds (Kuhn et al., 2019). Corollary 8 further suggests that (locally) lipschitz estimators tend to be more robust towards distribution shifts, and contributes to prior results connecting distributional or adversarial robustness to Lipschitzness (Cisse et al., 2017; Finlay et al., 2018; Sinha et al., 2018; Anil et al., 2019). Corollary 8 does not depend on the number of model parameters. Hence, gives a powerful guarantee when applied to over-parametrized neural networks, in particular when the data lies on a low-dimensional manifold. Therefore, it acts as an advocate for training methods which effectively regularize the Lipschitz constant of the network, e.g., (Bartlett et al., 2017; Cisse et al., 2017; Anil et al., 2019; Sagawa et al., 2020).

6. Main Result

Our main result is a transport based concentration inequality, which states that the empirical mean of a sample-dependent function concentrates around its expectation, if the function satisfies some degree of regularity. Let $\mathcal{Z} \subset \mathbb{R}^{d_\mathcal{Z}}$ denote the domain, and consider functions $g : \mathcal{Z} \to \mathbb{R}$. We work with two classes of regular functions, smooth and non-smooth. The $\mathcal{C}^s$-smooth class identifies functions that admit all partial derivatives up to order s, for an $s \in \mathbb{N}_+$. The smoothness of a $\mathcal{C}^s$-smooth function g, when restricted to $P \subset \mathcal{Z}$, is quantified by

$$\|g\|_{s:P} := \max_{|\beta| \leq s} \max_{z \in P} \left| \frac{\partial^{|\beta|} g(z)}{\partial_{\beta_1}, \ldots, \partial_{\beta_{d_\mathcal{Z}}}} \right|,$$

where $\beta \in \mathbb{N}^{d_\mathcal{Z}}$ is a multi-index and $|\beta| = \beta_1 + \cdots + \beta_{d_\mathcal{Z}}$. Further, when $P = \mathcal{Z}$, we simplify this notation to $\|g\|_s$. In machine learning applications, examples of smooth estimators include Gaussian processes with smooth kernels, physics-informed neural networks (Raissi et al., 2019), Neural-ODE solvers (Chen et al., 2018), invertible neural networks Hyndman and Kratsios (2021), and feedforward networks with smooth activation functions (De Ryck et al., 2021). For the non-smooth category, i.e. functions which are not differentiable, we use the α-Hölder property as a geometric notion for quantifying regularity. More formally, for $0 < \alpha \leq 1$, a function g is α-Hölder if

$$\mathrm{Lip}_\alpha(g|P) \coloneqq \sup_{\substack{x_1, x_2 \in P \\ x_1 \neq x_2}} \frac{\|g(x_1) - g(x_2)\|}{\|x_1 - x_2\|^\alpha}.$$

is finite. For $\alpha = 1$ this recovers the case of Lipschitz functions, which are discussed in the previous sections. Setting $\alpha \in (0, 1)$ produces rougher models which are typically used for predicting from long time-series (Morrill et al., 2021) or rough paths of Neural-SDE models (Cuchiero et al., 2020). Theorem 9 formalizes our main result. In Section 6.2, we sketch the proof for Theorem 9 for the non-smooth case to highlight the main techniques. The complete proof can be found in Appendix A.6.

Theorem 9 *Set $0 < \delta \leq 1$, and $N \in \mathbb{N}$. Let $\mathcal{Z} \subseteq \mathbb{R}^{d_\mathcal{Z}}$ be a compact set, $\mu \in \mathcal{P}(\mathcal{Z})$ be a probability measure, and $\boldsymbol{P}$ a data-independent partitioning of any size $k \in \mathbb{N}$ on $\mathcal{Z}$. Suppose $g^N : \mathcal{Z} \mapsto \mathbb{R}$ is a real-valued random function that may depend on $Z_1, \ldots, Z_N$ which are samples drawn independently from μ. For $\alpha \in (0, 1]$, define*

$$\mathrm{err}(\alpha) \coloneqq \sqrt{\frac{\ln(4/\delta)}{N}} L \max_{P \in \boldsymbol{P}} \mathrm{diam}(P)^\alpha + \frac{\|g^N\|_\infty}{\sqrt{N}} \max\{\sqrt{2\ln(4/\delta)}, \sqrt{k}\}.$$

*(i) **Non-Smooth:** Set $0 < \alpha \leq 1$, and let $\mathcal{F}_{L,\alpha} = \{g \in \mathcal{C}(\mathcal{Z}, \mathbb{R}) \colon \mathrm{Lip}_\alpha(g) \leq L\}$. Suppose $g^N \in \mathcal{F}_{L,\alpha}$ almost surely. Then with probability greater than $1 - \delta$, we have*

$$\mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \leq C_{d_\mathcal{Z},\alpha} \sum_{P \in \boldsymbol{P}} \frac{N_P}{N} \mathrm{rate}_{d_\mathcal{Z},\alpha}(N_P) \mathrm{diam}(P) \mathrm{Lip}_\alpha(g^N|P) + \mathrm{err}(\alpha)$$

where $\mathrm{rate}_{d_\mathcal{Z},\alpha}$ and $C_{d_\mathcal{Z},\alpha}$ depend only on the Hölder coefficient α and on the dimension $d_\mathcal{Z}$. The explicit expressions are recorded in Table 1.

*(ii) **Smooth:** Set $s \geq 1$, and let $\mathcal{F}_{L,s} = \{g \in \mathcal{C}^s(\mathcal{Z}, \mathbb{R}) \colon \|g\|_s \leq L\}$. Suppose $g^N \in \mathcal{F}_{L,s}$ almost surely. Then there exists constant $C_{d_\mathcal{Z},s} > 0$ which with probability greater than $1 - \delta$ satisfies*

$$\mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \leq C_{d_\mathcal{Z},s} \sum_{P \in \boldsymbol{P}} \frac{N_P}{N} \mathrm{rate}_{d_\mathcal{Z},s}(N_P) \mathrm{diam}(P) \|g^N\|_{s:P} + \mathrm{err}(1)$$

where $\mathrm{rate}_{d_\mathcal{Z},s}$ depends only on s and on $d_\mathcal{Z}$ and is recorded in Table 1.

Regularity	**Dimension**	**Rate** ($\text{rate}_{d_{\mathcal{Z}},\alpha}$, or $\text{rate}_{d_{\mathcal{Z}},s}$)	**Constant** ($C_{d_{\mathcal{Z}},\alpha}$ or $C_{d_{\mathcal{Z}},s}$)
α-Hölder	$d_{\mathcal{Z}} < 2\alpha$	$N_P^{-1/2}$	$C_{d_{\mathcal{Z}},\alpha} = \frac{d_{\mathcal{Z}}^{\alpha/2} 2^{d_{\mathcal{Z}}/2-2\alpha}}{1-2^{d_{\mathcal{Z}}/2-\alpha}}$
	$d_{\mathcal{Z}} = 2\alpha$	$\left(\alpha 2^{\alpha+2} + \log_2(N_P)\right) N_P^{-1/2}$	$C_{d_{\mathcal{Z}},\alpha} = \frac{d_{\mathcal{Z}}^{\alpha/2}}{\alpha 2^{\alpha+1}}$
	$d_{\mathcal{Z}} > 2\alpha$	$N_P^{-\alpha/d_{\mathcal{Z}}}$	$C_{d_{\mathcal{Z}},\alpha} = 2\left(\frac{\frac{d_{\mathcal{Z}}}{2}-\alpha}{2\alpha(1-2^{\alpha-d_{\mathcal{Z}}/2})}\right)^{2\alpha/d_{\mathcal{Z}}} \left(1+\frac{\alpha}{2^{\alpha}(\frac{d_{\mathcal{Z}}}{2}-\alpha)}\right) d_{\mathcal{Z}}^{\alpha/2}$
s-Smooth	$s > \frac{d_{\mathcal{Z}}}{2}$	$N_P^{-1/2}$	$\exists C_{d_{\mathcal{Z}},s} > 0$
	$s = \frac{d_{\mathcal{Z}}}{2}$	$\left(\log(N_P)+1)N_P^{-1/2}\right)$	$\exists C_{d_{\mathcal{Z}},s} > 0$
	$s < \frac{d_{\mathcal{Z}}}{2}$	$N_P^{-s/d_{\mathcal{Z}}}$	$\exists C_{d_{\mathcal{Z}},s} > 0$

Table 1: Rates and constants for Theorem 9

The generalization bounds of Sections 3, 4 and 5 all follow from Theorem 9 in the non-smooth case with $\alpha = 1$, so that the α-Hölder regularity coincides with Lipschitzness. Since we are concerned with the loss of a machine learning estimator we use $g^N(Z) = g^N(X,Y) = \ell(\hat{f}^N(X), Y)$ to obtain the risk bounds.

6.1 Transport-based Change of Measure Inequality

A key ingredient of Theorem 9 is a change of measure inequality, which we borrow from the optimal transport literature. We elaborate on this inequality as it is of independent interest. Change of measure inequalities upper bound the expectation of a function $g : \mathcal{Z} \to \mathbb{R}$ with respect to a measure $\nu_1 \in \mathcal{P}(\mathcal{Z})$, by the moments of g with respect to the another measure $\nu_2 \in \mathcal{P}(\mathcal{Z})$ and the divergence between the two measures $d(\nu_1, \nu_2)$. Such inequalities are in the core of PAC-Bayesian Bounds (Ambroladze et al., 2006; Seldin et al., 2012; Bégin et al., 2016; Mhammedi et al., 2019), and therefore there has been effort on developing tight change of measure inequalities that rely on various notions divergences such as the Kullback-Leibler or the f-divergence (Donsker and Varadhan, 1975; Picard-Weibel and Guedj, 2022; Ohnishi and Honorio, 2021; Alquier and Guedj, 2018). In this work we use a transport-based change of measure inequality which relies on the Wasserstein distance the between two probability measures.

Lemma 10 (Kantorovich and Rubinstein (1958)) *Let $\mathcal{Z}$ be a complete and separable metric space and $\nu_1, \nu_2 \in \mathcal{P}(\mathcal{Z})$. Let $g : \mathcal{Z} \to \mathbb{R}$ be a Lipschitz function. Then*

$$\int_{x\in\mathcal{Z}} g(x)\,\nu_1(dx) \le \mathrm{Lip}(g)\mathcal{W}_1(\nu_1,\nu_2) + \int_{x\in\mathcal{Z}} g(x)\,\nu_2(dx).$$

The proof is given in Appendix B.1 for completeness. We use this lemma with $\nu_1 = \mu$ and $\nu_2 = \mu^N$, to redistribute the data according to the empirical distribution μ^N, since the data generating distribution μ is unknown. This inequality allows us to directly benefit from the fact that in expectation, $\mathcal{W}_1(\mu, \mu^N)$ converges to zero at a $\Theta(N^{-1/d})$ rate, as more samples N are provided. Lemma B.5 in the appendix formalizes this claim, and presents a concentration inequality on the Wasserstein distance between μ and μ^N. This Lemma is a direct consequence of Kloeckner (2020). We believe that the change of measure inequality of Lemma 10, in combination with the concentration inequality of Lemma B.5 are of independent interest, in particular for analyses that consider data generating and empirical measures.

6.2 Proof outline for Theorem 9

Consider a partition $\boldsymbol{P}$ of the space $\mathcal{Z}$. We first apply the Kantorovich-Rubinstein duality theorem (Kantorovich and Rubinstein, 1958) on each part $P \in \boldsymbol{P}$ and prove a *local* variant of the change of measure inequality in Lemma 10. We then analyze how the empirical mean of g^N concentrates around $\mathbb{E}[g^N(z)]$ when the domain is restricted to a set $P \subset \mathcal{Z}$. We elaborate on the the three key steps of this proof.

Step 1: Local change of measure. By Lemma 10, the difference of expectation between two measures is dominated by integral probability metric. Thus, in order to capture a local difference, we apply Lemma 10 locally by restricting the domain to P and using $\nu_1 = \mu|_P$ and $\nu_2 = \mu^N|_P$. Taking a sum over all $P \in \boldsymbol{P}$ results in

$$\int_{z\in\mathcal{Z}} g^N(z)\,\mu(dz) \leq \int_{z\in\mathcal{Z}} g^N(z)\,\mu^N(dz) + \overbrace{\sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\mathcal{W}_\alpha(\mu|_P,\mu^N|_P)}^{\mathrm{I}} + \underbrace{\sum_{P\in\boldsymbol{P}}\left(1-\frac{\mu^N(P)}{\mu(P)}\right)\int_{z\in P} g^N(z)\,\mu(dz)}_{\mathrm{II}}.$$

The term (I) indicates the transport cost, and (II) appears due to the potential mismatch of measures over the parts of the partitioning of $\mathcal{Z}$. We bound (I) and (II) separately.

Step 2: Bounding (I). The Wasserstein distance between a probability distribution μ and its empirical counterpart μ^N is bounded in expectation and concentrates as more samples are taken into account. More formally, we show that due to Kloeckner (2020, Theorem 2.1), for all $\epsilon > 0$, and $N \in \mathbb{N}$ it holds that

$$\mathbb{P}\left(\left|\mathcal{W}_\alpha(\mu,\mu^N) - \mathbb{E}\Big[\mathcal{W}_\alpha(\mu,\mu^N)\Big]\right| \geq \epsilon\right) \leq 2e^{-\frac{2N\epsilon^2}{\mathrm{diam}(\mathcal{Z})^{2\alpha}}},$$

and

$$\mathbb{E}\Big[\mathcal{W}_\alpha(\mu,\mu^N)\Big] \leq C_{d_\mathcal{Z},\alpha}\,\mathrm{diam}(\mathcal{Z})\,\mathrm{rate}_{d_\mathcal{Z},\alpha}(N)$$

where the rates and the constants are as in Table 1. We apply the above inequalities locally, by considering $\mathcal{W}_\alpha(\mu|_P,\mu^N|_P)$ and sum over all $P \in \boldsymbol{P}$. We show that this summation is a weighted sum of independent sub-Gaussian random variables which we control via Lemma B.11. This treatment results in an upper bound for (I).

Step 3: Bounding (II). We interpret this term as the penalty we face for partitioning. It is zero if the probability mass of the data generating and empirical measures match across the partitioning, i.e. $\mu(P) = \mu^N(P)$ for all $P \in \boldsymbol{P}$, or if the analysis is carried out globally, i.e. $\boldsymbol{P} = \{\mathcal{Z}\}$. Considering discretized measures $\tilde{\mu}, \tilde{\mu}^N \in \mathcal{P}(\boldsymbol{P})$, which satisfy $\tilde{\mu}(\{P\}) = \mu(P)$ and $\tilde{\mu}^N(\{P\}) = \mu^N(P)$, we use that

$$\sum_{P\in\boldsymbol{P}}\left(1-\frac{\mu^N(P)}{\mu(P)}\right)\int_{z\in P} g^N(z)\,\mu(dz) \leq \|g\|_\infty \mathrm{TV}(\tilde{\mu},\tilde{\mu}^N)$$

where $\mathrm{TV}(\tilde{\mu}, \tilde{\mu}^N)$ denotes the total variation distance between $\tilde{\mu}^N$ and $\tilde{\mu}$. Since the $\mathrm{TV}(\tilde{\mu}, \tilde{\mu}^N)$ concentrates around zero and we can bound term II with high probability. Combining these three steps concludes the proof.

7. Conclusion

This paper presents novel instance-dependent generalization bound for locally regular estimators. We empirically and theoretically demonstrated the benefits of an instance-dependent non-parametric bound, and the effectiveness of a localized treatment of the risk. In particular, we showed that the instance-dependent bound remains relatively tight for over-parametrized models and captures a number of neural network generalization phenomena. In contrast, existing uniform or data-dependent parametric bounds tend to explode for large neural networks, and fail to explain their good generalization behavior.

Key observations made in this work could be relevant for future work that aims to improve learning algorithms. For example, our result suggest that the *local* regularity of a model plays a crucial role in its generalization ability and robustness to distribution shifts. This might of interest for developing robust and regularized training techniques. Finally, our optimal transport based approach constitutes a novel avenue towards theoretically analyzing generalization in machine learning. We introduce the necessary technical tools and proof methodology, hoping that future work can further explore this avenue and improve our results.

Acknowledgments

This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 815943 and the European Research Council (ERC) Starting Grant 852821—SWING. Anastasis Kratsios and Songyan Hou were additionally supported by the ETH Zürich foundation. Jonas Rothfuss was supported by an Apple Scholars in AI/ML fellowship. Anastasis was further funded by their McMaster University Startup Grant.

A. Proofs

A.1 Proof of the Main Generalization Bound (Theorem 4)

Proof of Theorem 4: Let $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$, $d_{\mathcal{Z}} = d+1$ and $g^N(x,y) = \ell(\hat{f}^N(x), y)$. Notice that $\nabla_y g^N = \nabla_{y_2} \ell \leq L_\ell$ and $\nabla_x g^N = \nabla_{y_1} \ell \cdot \nabla_x \hat{f}^N \leq L_\ell \cdot \mathrm{Lip}(\hat{f}^N)$. So $\mathrm{Lip}(g^N|P) \leq L_\ell \max\{1, \mathrm{Lip}(\hat{f}^N|P_{\mathcal{X}})\} \leq L_\ell \max\{1, L_f\}$ for all $P \in \boldsymbol{P}$. Now we can apply Theorem 9 with $\alpha = 1$, $\mathcal{F}_{L,1} = \{g \in \mathcal{C}(\mathcal{Z}, \mathbb{R})\colon \mathrm{Lip}(g) \leq L = L_\ell \max\{1, L_{\hat{f}}\}\}$. Then for all partition $\boldsymbol{P}$ with $|\boldsymbol{P}| \leq k$, for all $0 < \delta \leq 1$ and $N \in \mathbb{N}$, there exists an explicit constant $C_{d_{\mathcal{Z}},1} > 0$ s.t. with probability greater than $1 - \delta$

$$
\begin{aligned}
& \mathfrak{R}(\hat{f}^N; \mu) - \hat{\mathfrak{R}}(\hat{f}^N) \\
= \; & \mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \\
\leq \; & C_{d_{\mathcal{Z}},\alpha} \sum_{P \in \boldsymbol{P}} \frac{N_P}{N} \mathrm{rate}_{d_{\mathcal{Z}},\alpha}(N_P) \mathrm{diam}(P) \mathrm{Lip}(g^N|P) + \epsilon \\
\leq \; & C_{d_{\mathcal{Z}},\alpha} \sum_{P \in \boldsymbol{P}} \frac{N_P}{N} \mathrm{rate}_{d_{\mathcal{Z}},\alpha}(N_P) \mathrm{diam}(P) L_\ell \max\{1, \mathrm{Lip}(\hat{f}^N|P_{\mathcal{X}})\} + \epsilon \\
= \; & \frac{C_{d+1,1} L_\ell}{N} \sum_{P \in \boldsymbol{P}} N_P^{\frac{d}{d+1}} \max\Big\{1, \mathrm{Lip}(\hat{f}^N|P_{\mathcal{X}})\Big\} \mathrm{diam}(P) + \epsilon \\
= \; & \mathrm{cost}_{\mathrm{transport}} + \epsilon,
\end{aligned}
\tag{A.1}
$$

where

$$
\begin{aligned}
\epsilon &:= \sqrt{\frac{\ln(4/\delta)}{N}} L_\ell \max\{1, L_{\hat{f}}\} \max_{P \in \boldsymbol{P}} \mathrm{diam}(P) + \frac{\|\ell\|_\infty}{\sqrt{N}} \max\Big\{\sqrt{2\ln(4/\delta)}, \sqrt{k}\Big\} \\
&= \mathrm{err}_{\mathrm{transport}} + \mathrm{cost}_{\mathrm{partition}}.
\end{aligned}
\tag{A.2}
$$

■

A.2 Proof of Generalization Bound on Manifold Domain (Proposition 5)

We start by proving the manifold extension of our main result in Theorem 9. Then present the proof of Theorem 5 as a corollary of this theorem.

Theorem A.1 (Concentration of measure on a compact manifold) *Set $0 < \delta \leq 1$, and $N \in \mathbb{N}$. Let $\mathcal{Z}$ be a $d_{\mathcal{Z}}$-dimensional compact class C^1 Riemannian manifold. Let μ be a Borel probability measure on $\mathcal{Z}$, and $\boldsymbol{P}$ a partition of size k on $\mathcal{Z}$. Suppose g^N is a real-valued random function on $\mathcal{Z}$ depending on $Z_1, \ldots, Z_N$. Let $\mathcal{F}_L = \{g \in \mathcal{C}(\mathcal{Z}, \mathbb{R})\colon \mathrm{Lip}(g) \leq L\}$. Suppose $g^N \in \mathcal{F}_L$ almost surely. Then with probability greater than $1 - \delta$*

$$
\mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \leq C_{\mathcal{Z}} \sum_{P \in \boldsymbol{P}} \frac{N_P^{1-1/d_{\mathcal{Z}}}}{N} \mathrm{diam}(P) \mathrm{Lip}(g^N|P) + \mathrm{err}
\tag{A.3}
$$

where $C_{\mathcal{Z}} > 0$ is a constant depending on $d_{\mathcal{Z}}$ and

$$\mathrm{err} := \sqrt{\frac{\ln(4/\delta)}{N}} L \max_{P \in \boldsymbol{P}} \mathrm{diam}(P) + \frac{\|g^N\|_\infty}{\sqrt{N}} \max\{\sqrt{2\ln(4/\delta)}, \sqrt{k}\}.$$

Proof of Theorem A.1. All the steps are similar to the proof of Theorem 9, and we only invoke a different Wasserstein concentration lemma. Deploying Lemma B.7 we find that there exists $C_{\mathcal{Z}} > 0$ for which

$$\textstyle\sum_{P \in \boldsymbol{P}} \mu^N(P) \mathrm{Lip}(g^N|P)\, \mathbb{E}\big[\mathcal{W}_1(\mu|_P, \mu^N|_P)|N_P\big] \le \; C_{\mathcal{Z}} \sum_{P \in \boldsymbol{P}} \mu^N(P) \mathrm{Lip}(g^N|P) \mathrm{diam}(P) N_P^{-1/d_{\mathcal{Z}}}$$

and that

$$\mathbb{P}\Big(\Big|\mathcal{W}_1(\mu|_P, \mu^N|_P) - \mathbb{E}\big[\mathcal{W}_1(\mu|_P, \mu^N|_P)|N_P\big]\Big| \ge \epsilon \,\Big|\, N_P\Big) \le 2e^{-\frac{2N_P \epsilon^2}{\mathrm{diam}(P)^2}}.$$

By defining X_P similar to proof of Theorem 9 and invoking Lemma B.11, we get that there exists $C_{\mathcal{Z}} > 0$ such that the following holds with probability $1 - \delta_1$

$$\begin{aligned}\mathcal{B}^N \le C_{\mathcal{Z}} \sum_{P \in \boldsymbol{P}} & \mu^N(P) \mathrm{Lip}(g^N|P) \mathrm{diam}(P) N_P^{-1/d_{\mathcal{Z}}} \\ &+ L \max_{P \in \boldsymbol{P}} \mathrm{diam}(P) \left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2}.\end{aligned}$$

Terms $\mathcal{E}^N$ and $\mathcal{R}^N$ are identical to proof of Theorem 9. Therefore, plugging in everything we get,

$$\mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \le C_{\mathcal{Z}} \sum_{P \in \boldsymbol{P}} \frac{N_P^{1-1/d_{\mathcal{Z}}}}{N} \mathrm{diam}(P) \mathrm{Lip}(g^N|P) + \mathrm{err}.$$

■

Proof of Proposition 5 The proof is nearly identical to Theorem 4, however, here we invoke Theorem A.1 instead of Theorem 9. For completeness we repeat some of the steps. Let $\mathcal{Z}$ be the manifold which denotes the support of μ, then $d_{\mathcal{Z}} = \tilde{d}$. Let $g^N(x, y) = \ell(\hat{f}^N(x), y)$. We then apply Theorem A.1 with $\mathcal{F}_L := \{g : \mathcal{Z} \to \mathbb{R} \colon \mathrm{Lip}(g) \le L = L_\ell \max\{1, L_{\hat{f}}\}\}$. Then for all partition $\boldsymbol{P}$ with $|\boldsymbol{P}| \le k$, for all $0 < \delta \le 1$, $N \in \mathbb{N}$, there exists an $C_{\mathcal{Z}} > 0$ s.t. with probability greater than $1 - \delta$

$$\begin{aligned}\mathfrak{R}(\hat{f}^N; \mu) - \hat{\mathfrak{R}}(\hat{f}^N) =\; & \mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \\ \le\; & C_{\mathcal{Z}} \sum_{P \in \boldsymbol{P}} \frac{N_P^{1-1/d_{\mathcal{Z}}}}{N} \mathrm{diam}(P) \mathrm{Lip}(g^N|P) + \mathrm{err} \\ =\; & \frac{C(\tilde{d}) L_\ell}{N} \sum_{P \in \boldsymbol{P}} N_P^{1-1/\tilde{d}} \max\Big\{1, \mathrm{Lip}(\hat{f}^N|P_{\mathcal{X}})\Big\} \mathrm{diam}(P) \\ & + \mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}) + \mathrm{cost}_{\mathrm{partition}}(\boldsymbol{P})\end{aligned}$$

where the last line is obtained by using the definition of $\mathrm{err}_{\mathrm{transport}}$ and $\mathrm{cost}_{\mathrm{partition}}$ (as defined in Theorem 4). In addition, to transparently show the dependency of $C_{\mathcal{Z}}$ on $\tilde{d}$, we have renamed the constant. ■

A.3 Proof of the Classification Error Bound (Corollary 6)

Proof of Corollary 6 Let $g^N\big((x,y)\big) \coloneqq \ell_\gamma(\hat{f}^N(x), y)$, where $y \in \{-1, 1\}$ and $x \in \mathcal{X} \subset \mathbb{R}^d$. Consider a partition $\boldsymbol{P} \in \mathcal{X}$ of size k. For all $P \in \boldsymbol{P}$, we may define two sets

$$P_{(-)} \coloneqq \{(x, -1), \forall x \in P\} \qquad P_{(+)} \coloneqq \{(x, +1), \forall x \in P\}.$$

Note that $\mathrm{diam}(P_{(-)}) = \mathrm{diam}(P_{(+)}) = \mathrm{diam}(P)$. We construct the $\boldsymbol{P}_\pm$ partition on $\mathcal{X} \times \mathcal{Y}$ as

$$\boldsymbol{P}_\pm = \{P_{(-)} \mid P \in \boldsymbol{P}\} \cup \{P_{(+)} \mid P \in \boldsymbol{P}\}.$$

Note that $|\boldsymbol{P}_\pm| = 2|\boldsymbol{P}|$. For some $P_{(+)} \in \boldsymbol{P}_\pm$, we calculate $\mathrm{Lip}(g|P_{(+)})$,

$$\mathrm{Lip}(g|P_{(+)}) = \mathrm{Lip}\left(\ell_\gamma\left(\hat{f}^N(\cdot), +1\right)\Big|P\right) \le \frac{1}{\gamma}\mathrm{Lip}(\hat{f}^N|P)$$

and similarly for any $P_{(-)} \in \boldsymbol{P}_\pm$. Now invoking Theorem 9, for g and $\boldsymbol{P}_\pm$ we get, with probability greater than $1-\delta$

$$\begin{aligned}
\mathfrak{R}_\gamma(\hat{f}^N; \mu) - \hat{\mathfrak{R}}_\gamma(\hat{f}^N) &\le C_{d,1} \sum_{P_\pm \in \boldsymbol{P}_\pm} \frac{N_{P_\pm}^{1-1/d}}{N} \mathrm{diam}(P_\pm)\mathrm{Lip}(g^N|P_\pm) + \mathrm{err} \\
&= C_{d,1} \sum_{P \in \boldsymbol{P}} \Bigg[\frac{N_{P_{(+)}}^{1-1/d}}{N} \mathrm{diam}\left(P_{(+)}\right)\mathrm{Lip}(g^N|P_{(+)}) \\
&\qquad + \frac{N_{P_{(-)}}^{1-1/d}}{N}\mathrm{diam}\left((P_{(-)}\right)\mathrm{Lip}(g^N|P_{(-)})\Bigg] + \mathrm{err} \\
&\le C_{d,1} \sum_{P \in \boldsymbol{P}} \frac{N_{P_{(+)}}^{1-1/d} + N_{P_{(-)}}^{1-1/d}}{N} \mathrm{diam}(P)\frac{\mathrm{Lip}(\hat{f}^N|P)}{\gamma} + \mathrm{err} \\
&\le 2^{1/d} C_{d,1} \sum_{P \in \boldsymbol{P}} \frac{N_P^{1-1/d}}{N}\mathrm{diam}(P)\frac{\mathrm{Lip}(\hat{f}^N|P)}{\gamma} + \mathrm{err}
\end{aligned}$$

where $N_P = N_{P_{(+)}} + N_{P_{(-)}}$ and

$$\begin{aligned}
\mathrm{err} &= \sqrt{\frac{\ln(4/\delta)}{N}}\mathrm{Lip}(g^N) \max_{P_\pm \in \boldsymbol{P}_\pm} \mathrm{diam}(P_\pm) + \frac{1}{\sqrt{N}}\max\{\sqrt{2\ln(4/\delta)}, \sqrt{|\boldsymbol{P}_\pm|}\} \\
&= \sqrt{\frac{\ln(4/\delta)}{N}}\frac{\mathrm{Lip}(\hat{f}^N)}{\gamma}\max_{P \in \boldsymbol{P}} \mathrm{diam}(P) + \frac{1}{\sqrt{N}}\max\{\sqrt{2\ln(4/\delta)}, \sqrt{2k}\}.
\end{aligned}$$

The ramp loss acts as a Lipschitz proxy of the zero-one loss and allows us to analyze the classification error defined via

$$\mathfrak{R}_{01}(\hat{f}^N;\mu) := \mathbb{E}_{(X,Y)\sim\mu}\ell_{01}\left(\hat{f}^N(X), Y\right) = \mathbb{P}(\hat{f}^N(X) \neq Y).$$

Here $\ell_{01}(y_1, y_2) := \mathbf{1}_{[y_1 y_2 \leq 0]}$ is the zero-one loss which is not Lipschitz itself. Finally, we note that $\ell_\gamma \geq \ell_{01}$, and therefore

$$\mathfrak{R}_\gamma(\hat{f}^N;\mu) \geq \mathfrak{R}_{01}(\hat{f}^N;\mu) = \mathbb{P}(\hat{f}^N(X) \neq Y)$$

which implies,

$$\mathbb{P}(\hat{f}^N(X) \neq Y) \leq \frac{1}{N}\sum_{i=1}^{N} \ell_\gamma(\hat{f}^N(X_i), Y_i) + 2C_{d,1}\sum_{P\in\boldsymbol{P}} \frac{N_P^{1-1/d}}{N}\operatorname{diam}(P)\frac{\operatorname{Lip}(\hat{f}^N|P)}{\gamma} + \text{err}$$

with probability greater than $1-\delta$. ■

A.4 Proof of Proposition 7 on Partitioning

Proof of Proposition 7: Let σ be ReLU activation function and, for every $N \in \mathbb{N}$, consider the feedforward neural network f_N with one layer and one neuron defined by

$$f_N(x) = \frac{\sqrt{N}}{\log_2(\log_2(N))} \cdot \sigma(x - 1 + \frac{1}{N}).$$

We directly compute the following "global quantities" associated to f_N:

$$\|f_N\|_\infty \leq 1 \text{ and } \operatorname{Lip}(f_N) = \frac{\sqrt{N}}{\log_2(\log_2(N))}.$$

Set $L_{f,N} \stackrel{\text{def.}}{=} \sqrt{N}$ and define each $\mathcal{F}_N$ to be the set of Lipschitz functions from $[0,1]$ to

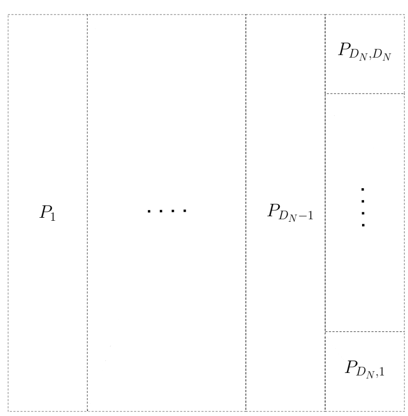


Figure 6: Partition $\boldsymbol{P}_N$ used in proof of Proposition 7

itself with Lipschitz constant at-most $L_{f,N}$, as in Assumption 2. We build a partition $\boldsymbol{P}_N$ of $[0,1] \times [0,1]$ as follows. Let $\Delta_N = N^{-0.6}$ and $D_N = \lceil \frac{1}{\Delta_N} \rceil = \mathcal{O}(N^{0.6})$.

$$\begin{aligned}\boldsymbol{P}_N &= \{B_1 \times \mathcal{Y}, \dots, B_{D_N - 1} \times \mathcal{Y}, B_{D_N} \times B_1, \dots, B_{D_N} \times B_{D_N}\} \\ &= \{P_1, \dots, P_n, \dots, P_{D_N - 1}, P_{D_N,1}, \dots, P_{D_N,n}, \dots, P_{D_N,D_N}\}\end{aligned}$$

where the sets $B_1, \dots, B_{D_N}$ subdivide $[0,1]$ into D_N intervals as defined by

$$B_n := \begin{cases} \left[\frac{n-1}{D_N}, \frac{n}{D_N}\right) : & n = 1, \dots, D_N - 1 \\ \left[\frac{D_N - 1}{D_N}, 1\right] : & n = D_N \end{cases}$$

Fig. 6 illustrates $\boldsymbol{P}_N$. This partition implies the following estimates on local Lipschitz constant of f_N

$$\mathrm{Lip}\Big(f_N \Big| P_n\Big) = 0 \quad \text{and} \quad \mathrm{Lip}\Big(f_N \Big| P_{D_N,n}\Big) = \frac{\sqrt{N}}{\log_2(\log_2(N))}, \qquad \forall n \in \mathbb{N}_{<D_N}.$$

Furthermore, we compute the following partition related quantities

$$\max_{P \in \boldsymbol{P}_N} \mathrm{diam}(P) \leq \sqrt{2} \text{ and } |\boldsymbol{P}_N| = 2D_N - 1 = \mathcal{O}(N^{0.6}).$$

From the above quantities, we compute the "localized bound" of Theorem 4, by calculating the terms $\mathrm{cost}_{\mathrm{partition}}, \mathrm{cost}_{\mathrm{transport}}, \mathrm{err}_{\mathrm{transport}}$.

$$\begin{aligned}\mathrm{cost}_{\mathrm{transport}} &= \frac{C_{d+1,1} L_\ell}{N} \sum_{P \in \boldsymbol{P}} N_P^{\frac{d}{d+1}} \max\{1, \mathrm{Lip}(f_N | P)\} \,\mathrm{diam}(P) \\ &= C_{2,1} \sum_{P \in \boldsymbol{P}_N} \frac{\big(8 + \log_2(N_P)\big) N_P^{1/2}}{N} \mathrm{diam}(P) L_\ell \max\{1, \mathrm{Lip}(f_N | P)\} \\ &= C_{2,1} L_\ell \sum_{n=1}^{N} \frac{\big(8 + \log_2(N_{P_{D_N,n}})\big) N_{P_{D_N,n}}^{1/2}}{N} \frac{\sqrt{N}}{\log_2(\log_2(N))} \frac{\sqrt{2}}{D_N} \\ &\leq \sqrt{2} C_{2,1} L_\ell \sum_{j=1}^{N} \frac{\big(8 + \log_2(1)\big)}{N} \frac{\sqrt{N}}{\log_2(\log_2(N))} \frac{1}{N^{0.6}} \qquad \text{(by Jensen's inequality)} \\ &= \frac{8\sqrt{2} C_{2,1} L_\ell}{\log_2(\log_2(N)) N^{0.1}}.\end{aligned}$$

For the other two terms,

$$\begin{aligned}\mathrm{err}_{\mathrm{transport}} &\leq \sqrt{\frac{\ln(4/\delta)}{N}} L_\ell \max\{1, L_{\hat{f}}\} \max_{P \in \boldsymbol{P}} \sqrt{\mathrm{diam}(P)^2 + 4B_y^2} \\ &\leq \sqrt{2} L_\ell \frac{\sqrt{N}}{\log_2(\log_2(N))} \frac{\sqrt{\ln(4/\delta)}}{N^{1/2}} \\ \mathrm{cost}_{\mathrm{partition}} &= \frac{B_\ell}{\sqrt{N}} \max\left\{\sqrt{2\ln(4/\delta)}, \sqrt{k_N}\right\} \\ &\leq \frac{\|\ell\|_\infty}{N^{1/2}} \max\{\sqrt{2\ln(4/\delta)}, \sqrt{\mathcal{O}(N^{0.6})}\}\end{aligned}$$

Therefore, in the $N \to \infty$ limit,

$$\lim_{N\to\infty} \mathrm{cost}_{\mathrm{transport}} + \mathrm{err}_{\mathrm{transport}} + \mathrm{cost}_{\mathrm{partition}} = 0$$

and the "local bound" of Theorem 4 converges. In contrast, upon inspecting the "global bound" of Equation (3), we note that it is bounded below by the following quantity

$$\begin{aligned} \mathrm{cost}_{\mathrm{transport}} &= C_{2,1}\frac{\big(8+\log_2(N)\big)}{N^{1/2}}\mathrm{diam}(\mathcal{X}\times\mathcal{Y})L_\ell \max\{1, \mathrm{Lip}(f_N)\} \\ &= \sqrt{2}C_{2,1}L_\ell\frac{\big(8+\log_2(N)\big)}{N^{1/2}}\sqrt{N} \\ &= \sqrt{2}C_{2,1}L_\ell\big(8+\log_2(N)\big). \end{aligned} \tag{A.4}$$

We conclude that the "global bound" diverges as N approaches infinity, since the quantity in (A.4) does; i.e. $\lim_{N\to\infty} \sqrt{2}C_{2,1}L_\ell\big(8+\log_2(N)\big) = \infty$. ■

A.5 Proof of Corollary 8 on Robustness to Distribution Shifts

Proof of Corollary 8. By Kantorovich Duality (Villani, 2009, Theorem 5.10), we have

$$\mathfrak{R}(\hat{f}^N; \mu^{\mathrm{adv}}) \le \mathfrak{R}(\hat{f}^N; \mu) + L_\ell \max\Big(1, \mathrm{Lip}(\hat{f}^N)\Big)\, \mathcal{W}\left(\mu, \mu^{\mathrm{adv}}\right). \tag{A.5}$$

By Theorem 4, we have for all $0 < \delta \le 1$ with probability greater than $1-\delta$,

$$\mathfrak{R}(\hat{f}^N; \mu^{\mathrm{adv}}) - \hat{\mathfrak{R}}(\hat{f}^N) \le \mathrm{cost}_{\mathrm{transport}}\left(\boldsymbol{P}_{\mathrm{global}}\right) + \mathrm{err}_{\mathrm{transport}}(\boldsymbol{P}_{\mathrm{global}}). \tag{A.6}$$

Combining (A.5) and (A.6), we complete the proof of Corollary 8. ■

A.6 Proof of Theorem 9

Proof of Theorem 9. We first consider the non-smooth case. That is, we consider the case where g^N is almost surely α-Hölder with $0 < \alpha \le 1$.

Step 1 (Change of measure)**.** By Lemma B.2, we deduce that

$$\begin{aligned} \int_{z\in\mathcal{Z}} g^N(z)\,\mu(dz) \le \int_{z\in\mathcal{Z}} g^N(z)\,\nu(dz) &+ \sum_{P\in\boldsymbol{P}} \nu(P)\mathrm{Lip}_\alpha(g^N|P)\mathcal{W}_\alpha(\mu|_P, \nu|_P) \\ &+ \sum_{P\in\boldsymbol{P}}\left(1-\frac{\nu(P)}{\mu(P)}\right)\int_{z\in P} g^N(z)\,\mu(dz). \end{aligned} \tag{A.7}$$

Setting $\nu = \mu^N$, we obtain an estimation of the expectation of g^N under μ since

$$\begin{aligned} \int_{z\in\mathcal{Z}} g^N(z)\,\mu(dz) \le \int_{z\in\mathcal{Z}} g^N(z)\,\mu^N(dz) &+ \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) \\ &+ \sum_{P\in\boldsymbol{P}}\left(1-\frac{\mu^N(P)}{\mu(P)}\right)\int_{z\in P} g^N(z)\,\mu(dz). \end{aligned}$$

We simplify notations by defining for each $P \in \boldsymbol{P}$, the following three abbreviations:

$$\begin{aligned}
\mathcal{D}_P &\stackrel{\text{def.}}{=} W_\alpha(\mu|_P, \mu^N|_P), \\
\mathcal{B}^N &\stackrel{\text{def.}}{=} \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\mathcal{W}_\alpha(\mu|_P, \mu^N|_P), \\
\mathcal{R}^N &\stackrel{\text{def.}}{=} \sum_{P\in\boldsymbol{P}} \left(1 - \frac{\mu^N(P)}{\mu(P)}\right) \int_{z\in P} g^N(z)\,\mu(dz).
\end{aligned}$$

With these notational short-hands, we concisely rewrite (A.7) as

$$\int_{z\in\mathcal{Z}} g^N(z)\,\mu(dz) \le \int_{z\in\mathcal{Z}} g^N(z)\,\mu^N(dz) + \mathcal{B}^N + \mathcal{R}^N. \tag{A.8}$$

In order to control our upper-bound in (A.8), we must control the terms $\mathcal{B}^N$ and $\mathcal{R}^N$; which we now do.

Step 2 (Integral probability metric concentration)**.** First we control the term $\mathcal{B}^N$. Notice that

$$\begin{aligned}
\mathcal{B}^N &= \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) \\
&= \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\,\mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big] \\
&\quad + \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\left(\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) - \mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big]\right) \\
&\le \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\,\mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big] \\
&\quad + L\sum_{P\in\boldsymbol{P}} \mu^N(P)\Big|\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) - \mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big]\Big|.
\end{aligned}$$

Deploying Lemma B.5, we find that

$$\begin{aligned}
&\sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\,\mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big] \\
\le\;& \sum_{P\in\boldsymbol{P}} \mu^N(P)\mathrm{Lip}_\alpha(g^N|P)\,C_{d_\mathcal{Z},\alpha}\,\mathrm{diam}(P)\,\mathrm{rate}_{d_\mathcal{Z},\alpha}(N_P).
\end{aligned} \tag{A.9}$$

and that

$$\mathbb{P}\Big(\Big|\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) - \mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big]\Big| \ge \epsilon \,\Big|\, N_P\Big) \le 2e^{-\frac{2N_P\,\epsilon^2}{\mathrm{diam}(P)^{2\alpha}}}.$$

Synchronizing our notation with that of Lemma B.11 we set $C_P = 2$, $\sigma_P^2 = \mathrm{diam}(P)^{2\alpha}/4N_P$, $\alpha_P = L\mu^N(P)$ and

$$X_P = \Big|\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) - \mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big]\Big|.$$

for $P \in \boldsymbol{P}$. Apply Lemma B.11 while conditioning on N_P we have that for every $\epsilon > 0$ and each $N \in \mathbb{N}$

$$\mathbb{P}\Big[\big|\sum_{P\in\boldsymbol{P}} \alpha_P X_P\big| \geq \epsilon \,\big|\, N_P\Big] \leq 2e^{-\frac{\epsilon^2}{8\tilde{\sigma}^2}},$$

where $\tilde{\sigma}^2$ is can be bounded above as follows

$$\begin{aligned}
\tilde{\sigma}^2 =& \sum_{P\in\boldsymbol{P}} C_P^2 \alpha_P^2 \sigma_P^2 \\
=& \sum_{P\in\boldsymbol{P}} 4\mu^N(P)^2 L^2 \frac{\operatorname{diam}(P)^{2\alpha}}{4N_P} \\
=& \sum_{P\in\boldsymbol{P}} \frac{N_P}{N^2} L^2 \operatorname{diam}(P)^{2\alpha} \\
\leq& \frac{L^2}{N} \max_{P\in\boldsymbol{P}} \operatorname{diam}(P)^{2\alpha}.
\end{aligned}$$

Therefore, we deduce the following concentration inequality

$$\begin{aligned}
\mathbb{P}\Big(\big|\sum_{P\in\boldsymbol{P}} \alpha_P X_P\big| \geq \epsilon\Big) &= \mathbb{E}\Big[\mathbb{P}\Big(\big|\sum_{P\in\boldsymbol{P}} \alpha_P X_P\big| \geq \epsilon \,\big|\, N_P\Big)\Big] \\
&\leq \mathbb{E}\Big[2\exp\Big\{-\frac{N\epsilon^2}{L^2 \max_{P\in\boldsymbol{P}} \operatorname{diam}(P)^{2\alpha}}\Big\}\Big] \\
&= 2\exp\Big\{-\frac{N\epsilon^2}{L^2 \max_{P\in\boldsymbol{P}} \operatorname{diam}(P)^{2\alpha}}\Big\}
\end{aligned}$$

Fix our "threshold probability" $0 < \delta_1 \leq 1$. With probability $1-\delta_1$ it holds that

$$L \textstyle\sum_{P\in\boldsymbol{P}} \mu^N(P)\Big|\mathcal{W}_\alpha(\mu|_P, \mu^N|_P) - \mathbb{E}\big[\mathcal{W}_\alpha(\mu|_P, \mu^N|_P)|N_P\big]\Big| \leq L \max_{P\in\boldsymbol{P}} \operatorname{diam}(P)^\alpha \left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2}.$$

Combining it with (A.9), we conclude that the following holds with probability $1-\delta_1$

$$\begin{aligned}
\mathcal{B}^N \leq& \sum_{P\in\boldsymbol{P}} \mu^N(P) \operatorname{Lip}_\alpha(g^N|P)\, C_{d_{\mathcal{Z}},\alpha} \operatorname{diam}(P) \operatorname{rate}_{d_{\mathcal{Z}},\alpha}(N_P) \\
&+ L \max_{P\in\boldsymbol{P}} \operatorname{diam}(P)^\alpha \left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2}.
\end{aligned} \tag{A.10}$$

Step 3: (Global concentration). It remains to estimate the term $\mathcal{R}^N$. Let $\delta_2 > 0$, by Lemma B.13, we know that the following holds with probability $1-\delta_2$

$$\mathcal{R}_N \leq \frac{\|g^N\|_\infty}{N^{1/2}} \max\{\sqrt{2\ln(2/\delta_2)}, \sqrt{k}\}. \tag{A.11}$$

Combining (A.8), (A.10) and (A.11), we have with probability greater than $(1-\delta_1)(1-\delta_2)$

$$\mathbb{E}\big[g^N(Z)\big] - \frac{1}{N}\sum_{n=1}^N g^N(Z_n) \leq \sum_{P\in\boldsymbol{P}} \mu^N(P) \operatorname{Lip}_\alpha(g^N|P)\, C_{d_{\mathcal{Z}},\alpha} \operatorname{diam}(P) \operatorname{rate}_{d_{\mathcal{Z}},\alpha}(N_P) + \epsilon,$$

where the term ϵ is given by

$$\epsilon \stackrel{\text{def.}}{=} L \max_{P \in \boldsymbol{P}} \operatorname{diam}(P)^{\alpha} \left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2} + \frac{\|g^N\|_\infty}{N^{1/2}} \max\{\sqrt{2\ln(2/\delta_2)}, \sqrt{k}\}.$$

Let $\delta \in (0,1]$. Set $\delta_1 \stackrel{\text{def.}}{=} \delta_2 \stackrel{\text{def.}}{=} \delta/2$. We now have with probability greater than $1-\delta$ that

$$\epsilon \stackrel{\text{def.}}{=} L \max_{P \in \boldsymbol{P}} \operatorname{diam}(P) \left(\frac{\ln(4/\delta)}{N}\right)^{1/2} + \frac{\|g^N\|_\infty}{N^{1/2}} \max\{\sqrt{2\ln(4/\delta)}, \sqrt{k}\}.$$

We now turn our attention to the proof of the smooth case; which is similar modulo some a few changes at key points in its proof.

Step 1 (Change of measure)**.** By Applying Lemma B.4 we have

$$\begin{aligned} \int_{z \in \mathcal{Z}} g^N(z)\, \mu(dz) \leq \int_{z \in \mathcal{Z}} g^N(z)\, \nu(dz) &+ \sum_{P \in \boldsymbol{P}} \nu(P) \|g^N\|_{s:P} \mathcal{W}_{\mathcal{C}^s}(\mu|_P, \nu|_P) \\ &+ \sum_{P \in \boldsymbol{P}} \left(1 - \frac{\nu(P)}{\mu(P)}\right) \int_{z \in P} g^N(z)\, \mu(dz). \end{aligned}$$

Setting $\nu \stackrel{\text{def.}}{=} \mu^N$ in the above equation, we estimate the mean of g^N with respect to μ

$$\begin{aligned} \int_{z \in \mathcal{Z}} g^N(z)\, \mu(dz) \leq \int_{z \in \mathcal{Z}} g^N(z)\, \mu^N(dz) &+ \sum_{P \in \boldsymbol{P}} \mu^N(P) \|g^N\|_{s:P} \mathcal{W}_{\mathcal{C}^s}(\mu|_P, \mu^N|_P) \\ &+ \sum_{P \in \boldsymbol{P}} \left(1 - \frac{\mu^N(P)}{\mu(P)}\right) \int_{z \in P} g^N(z)\, \mu(dz). \end{aligned} \tag{A.12}$$

As before, we simplify our notation. For each for all $P \in \boldsymbol{P}$ we abbreviate

$$\begin{aligned} \mathcal{D}_P &\stackrel{\text{def.}}{=} \mathcal{W}_{\mathcal{C}^s}(\mu|_P, \mu^N|_P) \\ \mathcal{B}^N &\stackrel{\text{def.}}{=} \sum_{P \in \boldsymbol{P}} \mu^N(P) \|g^N\|_{s:P} \mathcal{W}_{\mathcal{C}^s}(\mu|_P, \mu^N|_P) \\ \mathcal{R}^N &\stackrel{\text{def.}}{=} \sum_{P \in \boldsymbol{P}} \left(1 - \frac{\mu^N(P)}{\mu(P)}\right) \int_{z \in P} g^N(z)\, \mu(dz). \end{aligned}$$

Therefore, (A.12) can be succinctly written as

$$\int_{z \in \mathcal{Z}} g^N(z)\, \mu(dz) \leq \int_{z \in \mathcal{Z}} g^N(z)\, \mu^N(dz) + \mathcal{B}^N + \mathcal{R}^N. \tag{A.13}$$

As in the non-smooth case, we need only bound the terms $\mathcal{B}^N$ and $\mathcal{R}^N$ in order to control the left-hand side of (A.13).

Step 2 (Integral probability metric concentration)**.** We again first control the term $\mathcal{B}^N$.

Observe that

$$\begin{aligned}
\mathcal{B}^N &= \sum_{P\in\boldsymbol{P}} \mu^N(P)\|g^N\|_{s:P}\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P) \\
&= \sum_{P\in\boldsymbol{P}} \mu^N(P)\|g^N\|_{s:P}\,\mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big] \\
&\quad+ \sum_{P\in\boldsymbol{P}} \mu^N(P)\|g^N\|_{s:P}\Big(\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P) - \mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big]\Big) \\
&\le \sum_{P\in\boldsymbol{P}} \mu^N(P)\|g^N\|_{s:P}\,\mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big] \\
&\quad+ L\sum_{P\in\boldsymbol{P}} \mu^N(P)\Big|\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P) - \mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big]\Big|
\end{aligned}$$

Applying Lemma B.6 we have both that

$$\begin{aligned}
&\sum_{P\in\boldsymbol{P}} \mu^N(P)\|g^N\|_{s:P}\,\mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big] \\
\le\;& \sum_{P\in\boldsymbol{P}} \mu^N(P)\|g^N\|_{s:P}\,C_{d_{\mathcal{Z}},s}\operatorname{diam}(P)\operatorname{rate}_{d_{\mathcal{Z}},s}(N_P).
\end{aligned} \tag{A.14}$$

and that

$$\mathbb{P}\Big(\Big|\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P) - \mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big]\Big| \ge \epsilon \,\Big|\, N_P\Big) \le 2e^{-\frac{2N_P\epsilon^2}{\operatorname{diam}(P)^2}}.$$

Synchronizing notation with Lemma B.11 we denote $C_P = 1$, $\sigma_P^2 = \operatorname{diam}(P)^2/4N_P$ and $\alpha_P = L\mu^N(P)$,

$$X_P = \Big|\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P) - \mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big]\Big|$$

for $P \in \boldsymbol{P}$. Applying Lemma B.11 while conditioning on N_P, we have for all $\epsilon > 0$ and $N \in \mathbb{N}$,

$$\mathbb{P}\Big[\big|\sum_{P\in\boldsymbol{P}} \alpha_P X_P\big| \ge \epsilon \,\big|\, N_P\Big] \le 2e^{-\frac{\epsilon^2}{8\tilde{\sigma}^2}},$$

where, similarly to the smooth case, $\tilde{\sigma}^2$ is bounded above by

$$\begin{aligned}
\tilde{\sigma}^2 &= \sum_{P\in\boldsymbol{P}} C_P^2\alpha_P^2\sigma_P^2 \\
&= \sum_{P\in\boldsymbol{P}} 4\mu^N(P)^2L^2\frac{\operatorname{diam}(P)^2}{4N_P} \\
&= \sum_{P\in\boldsymbol{P}} \frac{N_P}{N^2}L^2\operatorname{diam}(P)^2 \\
&\le \frac{L^2}{N}\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)^2.
\end{aligned}$$

Therefore, we arive at the fact that

$$\begin{aligned}\mathbb{P}\Big(\big|\sum_{P\in\boldsymbol{P}}\alpha_P X_P\big|\geq\epsilon\Big)&=\mathbb{E}\Big[\mathbb{P}\Big(\big|\sum_{P\in\boldsymbol{P}}\alpha_P X_P\big|\geq\epsilon\,\big|\,N_P\Big)\Big]\\&\leq\mathbb{E}\Big[2\exp\Big\{-\frac{N\epsilon^2}{L^2\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)^2}\Big\}\Big]\\&=2\exp\Big\{-\frac{N\epsilon^2}{L^2\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)^2}\Big\}.\end{aligned}$$

Let $0<\delta_1\leq 1$ be our "threshold probability". Thus, with probability $1-\delta_1$ it holds that

$$L\sum_{P\in\boldsymbol{P}}\mu^N(P)\Big|\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)-\mathbb{E}\big[\mathcal{W}_{\mathcal{C}^s}(\mu|_P,\mu^N|_P)|N_P\big]\Big|\leq L\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)\left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2}.$$

Combine this with (A.14) we conclude that with probability $1-\delta_1$

$$\mathcal{B}^N\leq\sum_{P\in\boldsymbol{P}}\mu^N(P)\|g^N\|_{s:P}\,C_{d_{\mathcal{Z}},s}\operatorname{diam}(P)\operatorname{rate}_{d_{\mathcal{Z}},s}(N_P)+L\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)\left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2}. \tag{A.15}$$

Step 3: (Global concentration)**.** As with the smooth case, it only now remains to control the term $\mathcal{R}^N$. Set $0<\delta_2\leq 1$. Then, by Lemma B.13, we have that the following holds with probability at-least $1-\delta_2$

$$\mathcal{R}_N\leq\frac{\|g^N\|_\infty}{N^{1/2}}\max\{\sqrt{2\ln(2/\delta_2)},\sqrt{k}\}. \tag{A.16}$$

Combining (A.13), (A.15) and (A.16), we have with probability greater than $(1-\delta_1)(1-\delta_2)$

$$\mathbb{E}\big[g^N(Z)\big]-\frac{1}{N}\sum_{n=1}^N g^N(Z_n)\leq\sum_{P\in\boldsymbol{P}}\mu^N(P)\|g^N\|_{s:P}\,C_{d_{\mathcal{Z}},s}\operatorname{diam}(P)\operatorname{rate}_{d_{\mathcal{Z}},s}(N_P)+\epsilon,$$

where the "error term" ϵ is given by

$$\epsilon\overset{\text{def.}}{=}L\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)\left(\frac{\ln(2/\delta_1)}{N}\right)^{1/2}+\frac{\|g^N\|_\infty}{N^{1/2}}\max\{\sqrt{2\ln(2/\delta_2)},\sqrt{k}\}.$$

Fix $\delta\in(0,1]$. Set $\delta_1\overset{\text{def.}}{=}\delta_2\overset{\text{def.}}{=}\delta/2$. Thus, from our analysis we arrive at the conclusion that

$$\epsilon=L\max_{P\in\boldsymbol{P}}\operatorname{diam}(P)\left(\frac{\ln(4/\delta)}{N}\right)^{1/2}+\frac{\|g^N\|_\infty}{N^{1/2}}\max\{\sqrt{2\ln(4/\delta)},\sqrt{k}\},$$

holds with probability at-least $1-\delta$. This concludes our proof. ■

B. Helper Lemmas

B.1 Change of Measure Helper Inequalities

Lemma B.1 (Change of Measure of Hölder Function) *Let $\mathcal{Z}$ be a complete and separable metric space and $\mu, \nu \in \mathcal{P}_1(\mathcal{Z})$. Let $g : \mathcal{Z} \to \mathbb{R}$ be a α-Hölder function with $\alpha \in (0, 1]$. Then*

$$\int_{z\in\mathcal{Z}} g(z)\,\mu(dz) \le \operatorname{Lip}_\alpha(g)\mathcal{W}_\alpha(\mu,\nu) + \int_{z\in\mathcal{Z}} g(z)\,\nu(dz).$$

Proof of Lemma B.1. Since $\mathcal{Z}$ is complete and separable then Kantorovich Duality (Villani, 2009, Theorem 5.10) with the transport-cost function $c(x_1, x_2) = \|x_1 - x_2\|$ implies that

$$\mathcal{W}_\alpha(\mu,\nu) \ge \sup_{f\in\mathcal{C}(\mathcal{Z}),\,\operatorname{Lip}_\alpha(f)\le 1} \int f(z)\,\mu(dz) - \int f(z)\,\nu(dz).$$

If g is constant, then Lemma B.1 holds trivially. Otherwise, $\operatorname{Lip}_\alpha(g) > 0$ and we let $\tilde{g} \overset{\text{def.}}{=} \operatorname{Lip}_\alpha(g)^{-1} g$. Then $\operatorname{Lip}_\alpha(\tilde{g}) \le 1$ and we have

$$\begin{aligned}
\operatorname{Lip}_\alpha(g)^{-1} \int g\,\mu(dz) = & \int \tilde{g}(z)\,\mu(dz) \\
\le & \mathcal{W}_\alpha(\mu,\nu) + \int \tilde{g}(z)\,\nu(dz) \\
= & \mathcal{W}_\alpha(\mu,\nu) + \operatorname{Lip}_\alpha(g)^{-1} \int g(z)\,\nu(dz).
\end{aligned}$$

Multiplying across by $\operatorname{Lip}_\alpha(g) > 0$ yields the desired result. ■

Lemma B.2 (Local Change of Measure of Hölder Function) *Let $\mathcal{Z}$ a subset of $\mathbb{R}^{d_{\mathcal{Z}}}$ and $\mu, \nu \in \mathcal{P}_1(\mathcal{Z})$. Let $g : \mathcal{Z} \to \mathbb{R}$ be locally α-Hölder with $\alpha \in (0, 1]$. Then*

$$\begin{aligned}
\int_{z\in\mathcal{Z}} g(z)\,\mu(dz) \le \int_{z\in\mathcal{Z}} g(z)\,\nu(dz) & + \sum_{P\in\boldsymbol{P}} \nu(P)\operatorname{Lip}_\alpha(g|P)\mathcal{W}_\alpha(\mu|_P, \nu|_P) \\
& + \sum_{P\in\boldsymbol{P}} \left(1 - \frac{\nu(P)}{\mu(P)}\right) \int_{z\in P} g(z)\,\mu(dz).
\end{aligned}$$

Proof of Lemma B.2. Let for all $P \in \boldsymbol{P}$ that

$$\mu_P(\cdot) = \mu(\cdot \cap P), \quad \nu_P(\cdot) = \nu(\cdot \cap P), \quad \tilde{\mu}_P = \frac{\nu(P)}{\mu(P)}\mu_P.$$

Then we apply Lemma B.1 to $\tilde{\mu}_P$ and ν_P for all $P \in \boldsymbol{P}$ and have

$$\begin{aligned}
\int_{z\in\mathcal{Z}} -g(z)\nu(dz) = & \sum_{P\in\boldsymbol{P}} \int_{z\in\mathcal{Z}} -g(z)\nu_P(dz) \\
\le & \sum_{P\in\boldsymbol{P}} \Big(\operatorname{Lip}_\alpha(g|P)\mathcal{W}_\alpha\big(\nu_P, \tilde{\mu}_P\big) + \int_{z\in\mathcal{Z}} -g(z)\tilde{\mu}_P(dz)\Big).
\end{aligned}$$

By adding $\int_{z\in\mathcal{Z}} g(z)\mu(dz)$ on both sides and rearranging terms we have that

$$\begin{aligned}
\int_{z\in\mathcal{Z}} g(z)\mu(dz) \le& \int_{z\in\mathcal{Z}} g(z)\nu(dz) + \sum_{P\in\boldsymbol{P}} \operatorname{Lip}_\alpha(g|P)\mathcal{W}_\alpha\big(\nu_P, \tilde{\mu}_P\big) \\
&+ \int_{z\in\mathcal{Z}} g(z)\mu_P(dz) - \int_{z\in\mathcal{Z}} g(z)\frac{\nu(P)}{\mu(P)}\mu_P(dz) \\
\le& \int_{z\in\mathcal{Z}} g(z)\nu(dz) + \sum_{P\in\boldsymbol{P}} \operatorname{Lip}_\alpha(g|P)\nu(P)\mathcal{W}_\alpha\Big(\frac{\nu_P}{\nu(P)}, \frac{\mu_P}{\mu(P)}\Big) \\
&+ \int_{z\in\mathcal{Z}} g(z)\mu_P(dz) - \int_{z\in\mathcal{Z}} g(z)\frac{\nu(P)}{\mu(P)}\mu_P(dz) \\
\le& \int_{z\in\mathcal{Z}} g(z)\nu(dz) + \sum_{P\in\boldsymbol{P}} \operatorname{Lip}_\alpha(g|P)\nu(P)\mathcal{W}_\alpha\Big(\nu|_P, \mu|_P\Big) \\
&+ \Big(1 - \frac{\nu(P)}{\mu(P)}\Big)\int_{z\in\mathcal{Z}} g(z)\mu_P(dz) \\
\le& \int_{z\in\mathcal{Z}} g(z)\mu^N(dz) + \sum_{P\in\boldsymbol{P}} \operatorname{Lip}_\alpha(g|P)\mu^N(P)\mathcal{W}_\alpha\Big((\mu|_P)^{N_P}, \mu|_P\Big) \\
&+ \Big(1 - \frac{\mu^N(P)}{\mu(P)}\Big)\int_{z\in\mathcal{Z}} g(z)\mu_P(dz).
\end{aligned}$$

■

Lemma B.3 (Change of Measure of Smooth Function) *Let $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, $d_\mathcal{Z} \in \mathbb{N}$ and $\mu, \nu \in \mathcal{P}_1(\mathcal{Z})$. Let $g \in \mathcal{C}^s(\mathcal{Z})$ with $s \in \mathbb{N}$. Then*

$$\int_{z\in\mathcal{Z}} g(z)\,\mu(dz) \le \|g\|_s \mathcal{W}_{\mathcal{C}^s}(\mu,\nu) + \int_{z\in\mathcal{Z}} g(z)\,\nu(dz).$$

Proof of Lemma B.3. The proof is similar with the proof of Lemma B.1. Recall the definition of $\mathcal{W}_{\mathcal{C}^s}$ that

$$\mathcal{W}_{\mathcal{C}^s}(\mu,\nu) \overset{\text{def.}}{=} \sup_{f\in\mathcal{C}(\mathcal{Z}),\,\|f\|_s\le 1} \int f(z)\,\mu(dz) - \int f(z)\,\nu(dz).$$

If g is constant, then Lemma B.3 holds trivially. Otherwise, $\|g\|_s > 0$ and we Let $\tilde{g} \overset{\text{def.}}{=} \|g\|_s^{-1} g$. Then $\|g\|_s \le 1$ and we have

$$\begin{aligned}
\|g\|_s^{-1} \int g\,\mu(dz) =& \int \tilde{g}(z)\,\mu(dz) \\
\le& \mathcal{W}_\alpha(\mu,\nu) + \int \tilde{g}(z)\,\nu(dz) \\
=& \mathcal{W}_\alpha(\mu,\nu) + \|g\|_s^{-1} \int g(z)\,\nu(dz).
\end{aligned}$$

Multiplying across by $\|g\|_s > 0$ yields the desired result. ■

Lemma B.4 (Local Change of Measure of Smooth Function) *Let $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, $d_\mathcal{Z} \in \mathbb{N}$ and $\mu, \nu \in \mathcal{P}_1(\mathcal{Z})$. Let $g \in \mathcal{C}^s(\mathcal{Z})$ with $s \in \mathbb{N}$. Then*

$$\int_{z\in\mathcal{Z}} g(z)\,\mu(dz) \leq \int_{z\in\mathcal{Z}} g(z)\,\nu(dz) + \sum_{P\in\boldsymbol{P}} \mu(P)\|g\|_{s:P}\mathcal{W}_{\mathcal{C}^s}(\mu|_P, \nu|_P) + \sum_{P\in\boldsymbol{P}} \left(1 - \frac{\nu(P)}{\mu(P)}\right) \int_{z\in P} g(z)\,\mu(dz).$$

Proof of Lemma B.4. The proof is similar with the proof of Lemma B.2. Let for all $P \in \boldsymbol{P}$ that

$$\mu_P(\cdot) = \mu(\cdot \cap P), \quad \nu_P(\cdot) = \nu(\cdot \cap P), \quad \tilde{\mu}_P = \frac{\nu(P)}{\mu(P)}\mu_P.$$

Then we apply Lemma B.1 to $\tilde{\mu}_P$ and ν_P for all $P \in \boldsymbol{P}$ and have

$$\begin{aligned}\int_{z\in\mathcal{Z}} -g(z)\nu(dz) &= \sum_{P\in\boldsymbol{P}} \int_{z\in\mathcal{Z}} -g(z)\nu_P(dz) \\ &\leq \sum_{P\in\boldsymbol{P}} \Big(\|g\|_{s:P}\mathcal{W}_{\mathcal{C}^s}(\nu_P, \tilde{\mu}_P) + \int_{z\in\mathcal{Z}} -g(z)\tilde{\mu}_P(dz)\Big).\end{aligned}$$

By adding $\int_{z\in\mathcal{Z}} g(z)\mu(dz)$ on both sides and rearranging terms we have that

$$\begin{aligned}\int_{z\in\mathcal{Z}} g(z)\mu(dz) &\leq \int_{z\in\mathcal{Z}} g(z)\nu(dz) + \sum_{P\in\boldsymbol{P}} \|g\|_{s:P}\mathcal{W}_{\mathcal{C}^s}(\nu_P, \tilde{\mu}_P) \\ &\quad + \int_{z\in\mathcal{Z}} g(z)\mu_P(dz) - \int_{z\in\mathcal{Z}} g(z)\frac{\nu(P)}{\mu(P)}\mu_P(dz) \\ &\leq \int_{z\in\mathcal{Z}} g(z)\nu(dz) + \sum_{P\in\boldsymbol{P}} \nu(P)\|g\|_{s:P}\mathcal{W}_{\mathcal{C}^s}\Big(\frac{\nu_P}{\nu(P)}, \frac{\mu_P}{\mu(P)}\Big) \\ &\quad + \int_{z\in\mathcal{Z}} g(z)\mu_P(dz) - \int_{z\in\mathcal{Z}} g(z)\frac{\nu(P)}{\mu(P)}\mu_P(dz) \\ &\leq \int_{z\in\mathcal{Z}} g(z)\nu(dz) + \sum_{P\in\boldsymbol{P}} \nu(P)\|g\|_{s:P}\mathcal{W}_{\mathcal{C}^s}\Big(\nu|_P, \mu|_P\Big) \\ &\quad + \Big(1 - \frac{\nu(P)}{\mu(P)}\Big) \int_{z\in\mathcal{Z}} g(z)\mu_P(dz) \\ &\leq \int_{z\in\mathcal{Z}} g(z)\mu^N(dz) + \sum_{P\in\boldsymbol{P}} \mu^N(P)\|g\|_{s:P}\mathcal{W}_{\mathcal{C}^s}\Big((\mu|_P)^{N_P}, \mu|_P\Big) \\ &\quad + \Big(1 - \frac{\mu^N(P)}{\mu(P)}\Big) \int_{z\in\mathcal{Z}} g(z)\mu_P(dz).\end{aligned}$$

■

B.2 Helper Wasserstein Concentration Inequalities

Lemma B.5 (Concentration of Hölder Wasserstein Metric) *Let $\mathcal{Z}$ a compact subset of $\mathbb{R}^{d_\mathcal{Z}}$ and $\mu \in \mathcal{P}_1(\mathcal{Z})$. Then for all $\epsilon > 0$, $N \in \mathbb{N}$*

$$\mathbb{P}\left(\left|\mathcal{W}_\alpha(\mu, \mu^N) - \mathbb{E}\Big[\mathcal{W}_\alpha(\mu, \mu^N)\Big]\right| \geq \epsilon\right) \leq 2e^{-\frac{2N\epsilon^2}{\operatorname{diam}(\mathcal{Z})^{2\alpha}}}$$

where $C_{d_\mathcal{Z},\alpha}$ is given in Table 1 and

$$\mathbb{E}\Big[\mathcal{W}_\alpha(\mu, \mu^N)\Big] \leq C_{d_\mathcal{Z},\alpha} \operatorname{diam}(\mathcal{Z}) \operatorname{rate}_{d_\mathcal{Z},\alpha}(N)$$

with $\operatorname{rate}_{d_\mathcal{Z},\alpha}(N)$ also given in Table 1.

Proof of Lemma B.5. In the proof of Lemma B.5, we consider two different norms on the cube $[0,1]^{d_\mathcal{Z}}$ in order to apply (Kloeckner, 2020, Theorem 2.1). The first is the *Euclidean norm* $\|u\|_2^2 := \sum_{i=1}^{d_\mathcal{Z}} u_i^2$ and the second is the ∞-*norm* defined by $\|u\|_\infty := \max_{i=1,\dots,d_\mathcal{Z}} |u_i|$. When needed from the context, we emphasize implicitly used when defining the Wasserstein distance by $\mathcal{W}_{\alpha:2}$ and $\mathcal{W}_{\alpha:\infty}$ for the Euclidean and ∞ norms, respectively. By (Kloeckner, 2020, Theorem 2.1), we have for $\mathcal{Z} = [0,1]^{d_\mathcal{Z}}$ and $N \in \mathbb{N}$

$$\mathbb{E}\Big[\mathcal{W}_{\alpha:\infty}(\mu, \mu^N)\Big] \leq d_\mathcal{Z}^{-\alpha/2} C_{d_\mathcal{Z},\alpha} \operatorname{rate}_{d_\mathcal{Z},\alpha}(N).$$

By a simple fact that $\mathcal{W}_{\alpha:2} \leq d_\mathcal{Z}^{\alpha/2} \mathcal{W}_{\alpha:\infty}$, we have

$$\mathbb{E}\Big[\mathcal{W}_\alpha(\mu, \mu^N)\Big] = \mathbb{E}\Big[\mathcal{W}_{\alpha:2}(\mu, \mu^N)\Big] \leq C_{d_\mathcal{Z},\alpha} \operatorname{rate}_{d_\mathcal{Z},\alpha}(N).$$

We scale $[0,1]^{d_\mathcal{Z}}$ with $\operatorname{diam}(\mathcal{Z})$ to conclude that for general $\mathcal{Z} \subset \mathbb{R}^{d_\mathcal{Z}}$

$$\mathbb{E}\Big[\mathcal{W}_\alpha(\mu, \mu^N)\Big] \leq C_{d_\mathcal{Z},\alpha} \operatorname{diam}(\mathcal{Z}) \operatorname{rate}_{d_\mathcal{Z},\alpha}(N).$$

Now we define $f\colon \mathcal{Z}^N \to \mathbb{R}$ s.t.

$$f_N(z_1, \dots, z_N) \stackrel{\text{def.}}{=} \mathcal{W}_\alpha\Big(\frac{1}{N}\sum_{n=1}^N \delta_{z_n}, \mu\Big).$$

For every $i = 1, \dots, N$ and every $(z_1, \dots, z_N)$, $(z'_1, \dots, z'_N) \in \mathcal{Z}^N$ that differs only in the i-th coordinate, we have

$$|f(z_1, \dots, z_N) - f(z'_1, \dots, z'_N)| \leq \mathcal{W}_\alpha\Big(\frac{1}{N}\sum_{n=1}^N \delta_{z_n}, \frac{1}{N}\sum_{n=1}^N \delta_{z'_n}\Big) \leq \frac{\operatorname{diam}(\mathcal{Z})^\alpha}{N}.$$

Therefore, with $c = \frac{\operatorname{diam}(\mathcal{Z})^\alpha}{N}$, f has $(c, \dots, c)$-bounded differences property i.e. Lipschitz w.r.t. Hamming distance. Applying Lemma B.12 (the McDiarmid's inequality) with f proves that for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\mathcal{W}_\alpha(\mu, \mu^N) - \mathbb{E}\Big[\mathcal{W}_\alpha(\mu, \mu^N)\Big]\right| \geq \epsilon\right) \leq 2e^{-\frac{2N\epsilon^2}{\operatorname{diam}(\mathcal{Z})^{2\alpha}}}.$$

■

Lemma B.6 (Concentration of Smooth Wasserstein Metric) *Let $\mathcal{Z} = \mathbb{R}^{d_\mathcal{Z}}$, $d_\mathcal{Z} \in \mathbb{N}$ and $\mu, \nu \in \mathcal{P}_1(\mathcal{Z})$. Let $g \in \mathcal{C}^s(\mathcal{Z})$ with $s \in \mathbb{N}$. Then there exist constant $C_{d_\mathcal{Z},s} > 0$ s.t. for all $\epsilon > 0$, $N \in \mathbb{N}$*

$$\mathbb{P}\left(\left|\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N) - \mathbb{E}\Big[\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N)\Big]\right| \geq \epsilon\right) \leq 2e^{-\frac{2N\epsilon^2}{\operatorname{diam}(\mathcal{Z})^2}},$$

and

$$\mathbb{E}\Big[\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N)\Big] \leq C_{d_\mathcal{Z},s} \operatorname{diam}(\mathcal{Z}) \operatorname{rate}_{d_\mathcal{Z},s}(N).$$

Proof of Lemma B.6. The proof is similar to the proof of Lemma B.5. By (Kloeckner, 2020, Theorem 1.4), we have for $\mathcal{Z} = [0,1]^{d_\mathcal{Z}}$ and $N \in \mathbb{N}$

$$\mathbb{E}\Big[\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N)\Big] \leq C_{d_\mathcal{Z},s} \operatorname{rate}_{d_\mathcal{Z},s}(N).$$

Next, we scale $[0,1]^{d_\mathcal{Z}}$ with $\operatorname{diam}(\mathcal{Z})$ to conclude that for general $\mathcal{Z} \subset \mathbb{R}^{d_\mathcal{Z}}$

$$\mathbb{E}\Big[\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N)\Big] \leq C_{d_\mathcal{Z},s} \operatorname{diam}(\mathcal{Z}) \operatorname{rate}_{d_\mathcal{Z},s}(N).$$

Now we define $f\colon \mathcal{Z}^N \to \mathbb{R}$ s.t.

$$f_N(z_1, \dots, z_N) \overset{\text{def.}}{=} \mathcal{W}_{\mathcal{C}^s}\Big(\frac{1}{N}\sum_{n=1}^N \delta_{z_n}, \mu\Big).$$

For every $i = 1, \dots, N$ and every $(z_1, \dots, z_N)$, $(z'_1, \dots, z'_N) \in \mathcal{Z}^N$ that differs only in the i-th coordinate, we have

$$|f(z_1, \dots, z_N) - f(z'_1, \dots, z'_N)| \leq \mathcal{W}_{\mathcal{C}^s}\Big(\frac{1}{N}\sum_{n=1}^N \delta_{z_n}, \frac{1}{N}\sum_{n=1}^N \delta_{z'_n}\Big) \leq \frac{\operatorname{diam}(\mathcal{Z})}{N}.$$

Therefore, with $c = \frac{\operatorname{diam}(\mathcal{Z})}{N}$, f has $(c, \dots, c)$-bounded differences property i.e. Lipschitz w.r.t. Hamming distance. Applying Lemma B.12 (the McDiarmid's inequality) with f proves that for all $\epsilon > 0$

$$\mathbb{P}\left(\left|\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N) - \mathbb{E}\Big[\mathcal{W}_{\mathcal{C}^s}(\mu, \mu^N)\Big]\right| \geq \epsilon\right) \leq 2e^{-\frac{2N\epsilon^2}{\operatorname{diam}(\mathcal{Z})^2}}.$$

■

Lemma B.7 (Concentration of Wasserstein Metric on a Manifold) *Let $\mathcal{Z}$ be a $d_\mathcal{Z}$-dimensional compact class C^1 Riemannian manifold. Let μ be a Borel probability measure on $\mathcal{Z}$, and let μ^N denote the corresponding empirical distribution based on a sample of size N. Then exist for every $\epsilon > 0$ and $N \in \mathbb{N}$,*

$$\mathbb{P}\left(\left|\mathcal{W}_1\left(\mu^N, \mu\right) - \mathbb{E}\left[\mathcal{W}_1\left(\mu^N, \mu\right)\right]\right| \geq \epsilon\right) \leq 2e^{\frac{-2N\epsilon^2}{\operatorname{diam}(\mathcal{Z})^2}}, \tag{B.1}$$

and there exists constant $C_\mathcal{Z} > 0$ such that

$$\mathbb{E}\left[\mathcal{W}_1\left(\mu^N, \mu\right)\right] \leq C_\mathcal{Z} \cdot \operatorname{diam}(\mathcal{Z}) N^{-1/d_\mathcal{Z}}. \tag{B.2}$$

Proof of Lemma B.7. We recall that a $d_{\mathcal{Z}}$-dimensional class C^1-Riemannian manifold is $d_{\mathcal{Z}}$-dimensional topological manifold which is locally C^1-diffeomorphic to an open subset of $\mathbb{R}^{d_{\mathcal{Z}}}$. We first show that $\mathcal{Z}$ has Assouad dimension $d_{\mathcal{Z}}$ see (Robinson, 2011, Definitions 9.1 and 9.5). Then, we deduce the desired concentration inequality for metric spaces of Assouad dimension $d_{\mathcal{Z}}$. The for compact Riemannian $\mathcal{Z}$ then follows.

Step 1 - Computing $\mathcal{Z}$'s Metric (Assouad) Dimension Since $\mathcal{Z}$ is a $d_{\mathcal{Z}}$-dimensional manifold then, there exists smooth charts $\{(U_k,\phi_k)\}_{k=1}^K$ where $\mathcal{Z} = \cup_{k\le K} U_k$, $K \in \mathbb{N}\cup\{\infty\}$, and for $k=1,\dots,K$, $\phi_k : U_k \to B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1)$ is a (class C^1) diffeomorphism and each U_k is an open and bounded subset of $\mathcal{Z}$ and such that

$$\mathcal{Z} = \bigcup_{k=1}^{K} \phi_k^{-1}\Big[B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)\Big].$$

Since $\mathcal{Z}$ is compact and $\{U_k\}_{k\le K}$ is an open cover thereof then, we may without loss of generality assume that K is finite.

Applying (Robinson, 2011, Lemma 9.6 (iii)) we deduce that both $B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)$ and $B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1)$ have Assouad dimension $d_{\mathcal{Z}}$. By (Robinson, 2011, Lemma 9.6 (i)) we deduce that the closed Euclidean ball $\overline{B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)} = \{u \in \mathbb{R}^{d_{\mathcal{Z}}} : \|u\| \le 1/2\}$ must have Assouad dimension $d_{\mathcal{Z}}$. Since each ϕ_k is a diffeomorphism onto its image then $\phi_k^{-1} : B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1) \to U_k$ and ϕ_k are both locally Lipschitz. Thus, each ϕ_k is bi-Lipschitz when restricted to the compact set $\overline{B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)}$. Consequentially, (Robinson, 2011, Lemma 9.6 (v)) implies that each $\phi_k^{-1}[\overline{B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)}]$ has Assouad dimension $d_{\mathcal{Z}}$. Since K is finite and $U_1,\dots,U_K$ all have Assouad dimension $d_{\mathcal{Z}}$ and since $\{\phi_k^{-1}[B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)]\}_{k=1}^K$ is a cover of $\mathcal{Z}$ (since $\{\phi_k^{-1}[\overline{B_{\mathbb{R}^{d_{\mathcal{Z}}}}(0,1/2)}]\}_{k=1}^K$ is a cover of $\mathcal{Z}$) then (Robinson, 2011, Lemma 9.6 (ii)) implies that $\mathcal{Z}$ has Assouad dimension $d_{\mathcal{Z}}$.

Step 2 - The Concentration Inequality The assumption that $\mathcal{Z}$ has finite Assouad dimension $d_{\mathcal{Z}}$ is equivalent to the existence of a constant $K_{\mathcal{Z}}$ satisfying: for every $r > 0$

$$\mathcal{N}_{\mathcal{Z}}^{cov}(r) \le K_{\mathcal{Z}}\Big(\frac{\operatorname{diam}(\mathcal{Z})}{r}\Big)^{d_{\mathcal{Z}}} \tag{B.3}$$

Therefore, $\mathcal{Z}$ satisfies the Assumption made in (Boissard and Le Gouic, 2014, Equation (2)); hence, we may apply (Boissard and Le Gouic, 2014, Corollary 1.2) to conclude that:

$$\mathbb{E}\left[\mathcal{W}_1\left(\mu^N,\mu\right)\right] \le c\cdot K_{\mathcal{Z}}^{1/d_{\mathcal{Z}}}\left(\frac{2}{d_{\mathcal{Z}}-2}\right)^{2/d_{\mathcal{Z}}} \operatorname{diam}(\mathcal{Z}) N^{-1/d_{\mathcal{Z}}}; \tag{B.4}$$

for some constant $0 \le c \le \frac{2^6}{3}$. Let $C_{\mathcal{Z}} \overset{\text{def.}}{=} c\cdot K_{\mathcal{Z}}$ and we prove (B.2). Next, since $\operatorname{diam}(\mathcal{Z}) < \infty$ and μ is a Borel measure on the polish space $\mathcal{Z}$ then, (Weed and Bach, 2019, Proposition 20) applies; hence for every $\epsilon > 0$ we have the estimate

$$\mathbb{P}\left(\left|\mathcal{W}_1\left(\mu^N,\mu\right) - \mathbb{E}\left[\mathcal{W}_1\left(\mu^N,\mu\right)\right]\right| \ge \epsilon\right) \le 2e^{\frac{-2N\epsilon^2}{\operatorname{diam}(\mathcal{Z})^2}}.$$

■

Remark B.8 (Acceleration of Rates in Lemma B.7 Under Additional Regularity) *If $\mathcal{Z}$ is a submanifold of Euclidean space with finite-reach*[3] *and if μ has a density with respect to the volume measure on $\mathcal{Z}$ then a variant of Lemma B.7 with a faster concentration rate can be derived using the results of Block et al. (2021) instead of Boissard and Le Gouic (2014).*

B.3 Helper Sub-Gaussian Concentration Inequalities

Definition B.9 (Sub-Gaussian distribution) *A centered random variable X is called sub-Gaussian if there exists $C > 0$ and $\sigma > 0$ s.t. for all $x > 0$ that*

$$\mathbb{P}[|X| \geq x] \leq Ce^{-\frac{x^2}{2\sigma^2}},$$

denoted by $X \sim \mathrm{subG}(C, \sigma^2)$.

Lemma B.10 *Let $C, \sigma > 0$, $\tilde{\sigma} = \sigma \max\{C, 1\}$, and X be a centered random variable. Then each statement bellow implies the next:*

1. $X \sim \mathrm{subG}(C, \sigma^2)$.
2. $\mathbb{P}[|X| \geq x] \leq Ce^{-\frac{x^2}{2\sigma^2}}$ *for all* $x > 0$.
3. $\mathbb{E}[|X|^k] \leq (2\tilde{\sigma}^2)^{\frac{k}{2}} \left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}\right)$ *for all* $k \in \mathbb{N}_{\geq 2}$.
4. $\mathbb{E}[\exp(tX)] \leq e^{4\tilde{\sigma}^2 t^2}$ *for all* $t \in \mathbb{R}$.
5. $X \sim \mathrm{subG}(2, 4\tilde{\sigma}^2)$.

Proof of Lemma B.10. $(i) \Rightarrow (ii)$ by definition. $(ii) \Rightarrow (iii)$ is true by the following estimate:

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}[|X|^k \geq t]dt \\
&= \int_0^\infty \mathbb{P}[|X| \geq t^{\frac{1}{k}}]dt \\
&\leq \int_0^\infty Ce^{-\frac{t^{2/k}}{2\sigma^2}}dt \\
&\leq \frac{Ck(2\sigma^2)^{\frac{k}{2}}}{2} \int_0^\infty e^{-u}u^{\frac{k}{2}-1}du \\
&\leq C\left(\frac{k}{2}\right)(2\sigma^2)^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right) \\
&\leq \max\{1, C^2\}(2\sigma^2)^{\frac{k}{2}}\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}\right) \\
&\leq (2\max\{1, C^2\}\sigma^2)^{\frac{k}{2}}\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}\right) \\
&\leq (2\tilde{\sigma}^2)^{\frac{k}{2}}\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}\right).
\end{aligned}$$

3. The *reach* of a submanifold $\mathcal{Z}$ of a Euclidean space is the largest radius $r \geq 0$ for which each point in the Euclidean space whose Euclidean distance from $\mathcal{Z}$ is at-most r has a unique projection onto $\mathcal{Z}$; see Genovese et al. (2012) for further details.

$(iii) \Rightarrow (iv)$ is true because for all $t \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[e^{tX}] &\leq 1 + \sum_{k=2}^{\infty} \frac{t^k \mathbb{E}[|X|^k]}{k!} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{(2\tilde{\sigma}^2 t^2)^k 2k\Gamma(k)}{(2k)!} + \sum_{k=1}^{\infty} \frac{(2\tilde{\sigma}^2 t^2)^{k+1/2}(2k+1)\Gamma(k+1/2)}{(2k+1)!} \\
&\leq 1 + (2 + \sqrt{2\tilde{\sigma}^2 t^2}) \sum_{k=1}^{\infty} \frac{(2\tilde{\sigma}^2 t^2)^k k!}{(2k)!} \\
&\leq 1 + (1 + \sqrt{\frac{\tilde{\sigma}^2 t^2}{2}}) \sum_{k=1}^{\infty} \frac{(2\tilde{\sigma}^2 t^2)^k}{k!} \\
&\leq e^{2\tilde{\sigma}^2 t^2} + \sqrt{\frac{\tilde{\sigma}^2 t^2}{2}} (e^{2\tilde{\sigma}^2 t^2} - 1) \\
&\leq e^{4\tilde{\sigma}^2 t^2}.
\end{aligned}$$

$(iv) \Rightarrow (v)$: for all $x > 0$ and $t > 0$

$$\begin{aligned}
\mathbb{P}(X > x) &= \mathbb{P}(e^{tX} > e^{tx}) \\
&\leq \frac{\mathbb{E}[e^{tX}]}{e^{tx}} \\
&\leq e^{4\tilde{\sigma}^2 t^2 - tx}.
\end{aligned}$$

Therefore we have that

$$\mathbb{P}(X \geq x) \leq e^{-\frac{x^2}{8\tilde{\sigma}^2}}.$$

Similarly we can prove that

$$\mathbb{P}(X \leq -x) \leq e^{-\frac{x^2}{8\tilde{\sigma}^2}}.$$

Therefore we conclude that

$$\mathbb{P}(|X| \geq x) \leq 2e^{-\frac{x^2}{8\tilde{\sigma}^2}},$$

that is $X \sim \mathrm{subG}(2, 4\tilde{\sigma}^2)$.

■

Lemma B.11 *Let $X_1, \ldots, X_n$ be independent with $X_i \sim \mathrm{subG}(C, \sigma_i^2)$, and let $\alpha_i \geq 0$, $\tilde{\sigma}_i = \sigma_i \max\{C, 1\}$, for all $i = 1, \ldots, n$. Then we have*

$$\sum_{i=1}^{n} \alpha_i X_i \sim \mathrm{subG}(2, 4\sum_{i=1}^{n} \alpha_i^2 \tilde{\sigma}_i{}^2),$$

that is, for all $x > 0$,

$$\mathbb{P}\Big[\big|\sum_{i=1}^{n} \alpha_i X_i\big| \geq x\Big] \leq 2e^{-\frac{x^2}{8\tilde{\sigma}^2}},$$

where $\tilde{\sigma}^2 = \sum_{i=1}^{n} \alpha_i^2 \tilde{\sigma}_i{}^2$.

Proof of Lemma B.11. By Lemma B.10, for all $i = 1, \ldots, n$ and $t \in \mathbb{R}$ we have that

$$\mathbb{E}[\exp(tX_i)] \leq e^{4\tilde{\sigma}_i^2 t^2}.$$

Then, by independence, we obtain

$$\mathbb{E}[\exp(t\sum_{i=1}^n \alpha_i X_i)] = \prod_{i=1}^n \mathbb{E}[\exp(t\alpha_i X_i)] \leq \exp\left(4\sum_{i=1}^n \alpha_i^2 \tilde{\sigma}_i^2 t^2\right).$$

Then, by Lemma B.10 we conclude that

$$\sum_{i=1}^n \alpha_i X_i \sim \operatorname{subG}(2, 4\sum_{i=1}^n \alpha_i^2 \tilde{\sigma_i}^2).$$

■

Lemma B.12 (McDiarmid's inequality) *Let $X_1, \cdots, X_N$ be independent random variables, where X_i has range $\mathcal{X}_i$. Let $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \to \mathbb{R}$ be any function with the $(c_1, \ldots, c_N)$-bounded differences property: for every $i = 1, \ldots, N$ and every $(x_1, \ldots, x_N)$, $(x'_1, \ldots, x'_N) \in \mathcal{X}_1 \times \ldots \mathcal{X}_N$ that differs only in the i-th coordinate, we have*

$$|f(x_1, \ldots, x_N) - f(x'_1, \ldots, x'_N)| \leq c_i.$$

Then for all $t > 0$ we have that

$$\mathbb{P}\Big[f(X_1, \ldots, X_N) - \mathbb{E}[f(X_1, \ldots, X_N)] \geq t\Big] \leq e^{\frac{-2t^2}{\sum_{i=1}^N c_i^2}},$$

and

$$\mathbb{P}\Big[\big|f(X_1, \ldots, X_N) - \mathbb{E}[f(X_1, \ldots, X_N)]\big| \geq t\Big] \leq 2e^{\frac{-2t^2}{\sum_{i=1}^N c_i^2}}.$$

Proof of Lemma B.12. See McDiarmid et al. (1989). ■

Lemma B.13 (Concentration of Global Mismatch) *Let $\mathcal{Z}$ a compact metric space and $\mu \in \mathcal{P}_1(\mathcal{Z})$. Let $g^N\colon \mathcal{Z} \to \mathbb{R}$ continuous and $\boldsymbol{P}$ a finite partition of $\mathcal{Z}$. Then for all $\delta > 0$, with probability $1 - \delta$*

$$\sum_{P \in \boldsymbol{P}} \left(1 - \frac{\mu^N(P)}{\mu(P)}\right) \int_{z \in P} g(z)\, \mu(dz) \leq \frac{\|g^N\|_\infty}{N^{1/2}} \max\{\sqrt{2\ln(2/\delta)}, \sqrt{k}\}.$$

Proof of Lemma B.13. We notice that

$$\begin{aligned}\mathcal{R}^N &\overset{\text{def.}}{=} \sum_{P \in \boldsymbol{P}} \left(1 - \frac{\mu^N(P)}{\mu(P)}\right) \int_{z \in P} g^N(z)\, \mu(dz) \\ &\leq \|g^N\|_\infty \sum_{P \in \boldsymbol{P}} \big|\mu^N(P) - \mu(P)\big|.\end{aligned}$$

Let $\tilde{\mu}$ be a discrete distribution on $\boldsymbol{P}$ s.t. $\tilde{\mu}(P) = \mu(P)$ and ν^N the empirical measure of ν with N samples. Then we have

$$\sum_{P \in \boldsymbol{P}} \left|\mu^N(P) - \mu(P)\right| = \mathrm{TV}(\nu, \nu^N),$$

where $\mathrm{TV}(\cdot,\cdot)$ denote the total variation distance. Therefore, by the empirical estimation under total variation distance (Canonne, 2020, Theorem 1), for all $\epsilon > 0$, $N \geq \max\{\frac{|\boldsymbol{P}|}{\epsilon^2}, \frac{2}{\epsilon^2}\log(2/\delta)\}$

$$\mathrm{TV}(\nu, \nu^N) \leq \epsilon.$$

Thus, we have with probability $1-\delta$

$$\mathcal{R}_N \leq \frac{\|g^N\|_\infty}{N^{1/2}} \max\{\sqrt{2\ln(2/\delta)}, \sqrt{k}\}.$$

■

C. Uniform Rademacher Generalization Bound

In this section we present the Rademacher generalization bound of Equation (2) with more rigor. Consider $\mathcal{F}_L$ the class of Lipschitz functions mapping $\mathcal{X}$ to $\mathcal{Y}$, with Lipschitz constant of at most L, and let $\tilde{\mathcal{F}}_L = \{\ell \circ f \,:\, f \in \mathcal{F}_L\}$. Under assumptions 1 and 3, for any random sample $\mathcal{D}^N$ of size N, and $0 < \delta < 1$ Bartlett and Mendelson (Theorem 8, 2002) states that with probability greater than $1-\delta$

$$\sup_{f \in \mathcal{F}_L} \left\{\mathfrak{R}(f;\mu) - \hat{\mathfrak{R}}(f)\right\} \leq 2\hat{\mathcal{R}}_N\left(\tilde{\mathcal{F}}_L\right) + \|\ell\|_\infty \sqrt{\frac{8\log 2/\delta}{N}} \tag{C.1}$$

where $\hat{\mathcal{R}}_N\left(\tilde{\mathcal{F}}_L\right)$ is the empirical Rademacher complexity of $\tilde{\mathcal{F}}_L$ which is defined via

$$\hat{\mathcal{R}}_N(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{N}\sum_{i=1}^N \epsilon_i h(X_i, Y_i)$$

with $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ being an i.i.d. vector of Rademacher random variables. By contraction of Rademacher complexity (Theorem 12, Bartlett and Mendelson, 2002), since the Loss is L_ℓ-Lipschitz we get

$$\hat{\mathcal{R}}_N\left(\tilde{\mathcal{F}}_L\right) \leq 2L_\ell \hat{\mathcal{R}}_N\left(\mathcal{F}_L\right). \tag{C.2}$$

In the next lemma, we bound the Rademacher complexity of the class of L-lipschitz functions, defined on a d-dimensional domain.

Lemma C.1 *There exists C_1 and C_2 such that,*

$$\mathcal{R}_N(\mathcal{F}_L) \leq C_1\left(\frac{(dD)^2(BL)^d}{N}\right)^{1/(d+2)} + C_2 D\left(\frac{1}{N}\left(\frac{BL}{dD}\right)^d\right)^{1/(d+2)}.$$

where $D := \sup_{f\in\mathcal{F}_L}\|f\|_\infty$ and $\mathrm{diam}(\mathcal{X}) \leq B$.

By this lemma, and due to Equations (C.1) and (C.2), there exists C for which with probability greater than $1-\delta$,

$$\mathfrak{R}(f;\mu) - \hat{\mathfrak{R}}(f) \leq C L_\ell \left(\frac{(dD)^2 (BL)^d}{N}\right)^{1/(d+2)} + \|\ell\|_\infty \sqrt{\frac{8\log 2/\delta}{N}}$$

implying that the generalization error vanishes at a $\mathcal{O}(N^{-1/(2+d)})$ rate. This concludes the derivation of Equation (2).

Proof of Lemma C.1. We start by bounding the Metric Entropy of the function class, and then applying a discretization bound. Without a loss of generality, we may assume that $\mathbf{0}$ is included in $\mathcal{X}\times\mathcal{Y}$. Since $\mathcal{X}$ and $\mathcal{Y}$ are compact, then there exists B such that $\mathcal{X}\subset[0,B]^d$. With a treatment similar to Wainwright (2019, Example 5.10), we can show that the metric entropy of $\mathcal{F}_L$ is bounded as

$$\log \mathcal{N}(\delta, \mathcal{F}_L, \|\cdot\|_\infty) = \Theta\left(\left(\frac{BL}{\delta}\right)^d\right).$$

The 1-step discretization bound (Wainwright, 2019, Proposition 5.17) implies that

$$\mathcal{R}_N(\mathcal{F}_L) \leq \frac{1}{\sqrt{N}} \inf_{\delta>0}\left(\delta\sqrt{N} + 2\sqrt{D^2 \log \mathcal{N}(\delta, \mathcal{F}_L, \|\cdot\|_\infty)}\right)$$

where D^2 is used to upper bound the N-norm $\sup_{f\in\mathcal{F}_L}\sum_{i=1}^N f^2(X_i)/N$. By solving for δ and plugging in the optimal value we get that there exists constants C_1 and C_2 for which

$$\mathcal{R}_N(\mathcal{F}_L) \leq C_1 \left(\frac{(dD)^2(BL)^d}{N}\right)^{1/(d+2)} + C_2 D\left(\frac{1}{N}\left(\frac{BL}{dD}\right)^d\right)^{1/(d+2)}.$$

■

D. Experiment Details

We include the details of the experiments in Section 4. For visualizing the bounds in all experiments, we have used $\mathrm{Lip}(\ell\circ\hat{f}^N|P)$ since splitting the constant as $L_\ell \mathrm{Lip}(\hat{f}^N|P)$ may loosen the bound, in particular for the classification experiments. All experiments are repeated for multiple random seeds, in each run the following are randomized: the learning problem (i.e. the training and test sets), the network initialization, the training algorithm.

D.1 Task Descriptions

We generate random datasets for two toy regression and classification tasks.

Regression problem. For our empirical evaluations of neural network regression, we use the simplistic problem of regressing on noisy observation of a modified logistic function. Formally, our target function $f^\star:\mathcal{X}\mapsto\mathcal{Y}$ with $\mathcal{X}=[-5,5]\subset\mathbb{R}$ and $\mathcal{Y}=[-1,2]\subset\mathbb{R}$ is defined as

$$f^\star(x) = \frac{1}{1+\exp(5(x+2))}. \tag{D.1}$$

The inputs $x \in \mathcal{X}$ are sampled i.i.d from a uniform distribution $\mathcal{U}(-5, 5)$ and the corresponding regression labels follow as $y = f^\star(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$ is i.i.d. Gaussian observation noise. Figure 7a illustrates $f^\star$ together with 20 noisy observations. The loss function we use for regression is the Huber loss,

$$\ell(y, \hat{f}(x)) := \begin{cases} (y - \hat{f}(x))^2 & \text{if } |y - \hat{f}(x)| < 1 \\ |y - \hat{f}(x)| & \text{otherwise} \end{cases} \tag{D.2}$$

which behaves like the mean squared error (MSE) for small and like the mean absolute error for large regression residuals. Training with the Huber loss is equivalent to training with the MSE plus gradient clipping and thus a common choice of practitioners to prevent large regression residuals from destabilizing the neural network training.

Classification problem. We also consider a binary classification with $\mathcal{X} = [-5, 5]^2 \subset \mathbb{R}^2$ and $\mathcal{Y} = \{-1, 1\}$. The input features $x \in \mathcal{X}$ are sampled i.i.d. from a uniform distribution over $\mathcal{X}$. The labels are sampled i.i.d from the Bernoulli distribution $\mathcal{B}(\sigma(f^{\text{logit}}(x_1, x_2)))$ where $\sigma(z) = 1/(1 + \exp(-z))$ is the logistic function and

$$f^{\text{logit}}(x_1, x_2) = 10\sqrt{(x_1 - 2)^2 + (x_2 - 2)^2} - \frac{1}{4}\sin(2x_1) + \frac{3}{2}\cos(x_2) \; . \tag{D.3}$$

This binary classification problem is illustrated in Figure 7b. During training, we use the negative cross-entropy error,

$$\ell(x_1, x_2, y) = (1 - y) f^{\text{logit}}(x_1, x_2) - \log(\sigma(f^{\text{logit}}(x_1, x_2))) \tag{D.4}$$

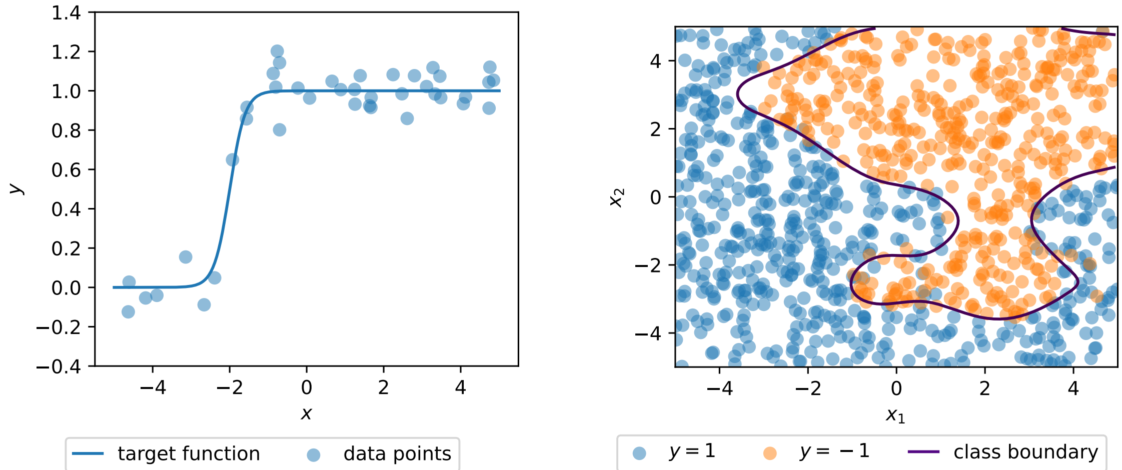
which is commonly used for training neural network classifiers. We visualize the bound of Corollary 6. To calculate the bound we consider the ramp lost ℓ_γ as defined in Section 4, with $\gamma = 5$.

D.2 Details on the Neural Network Training

In our empirical evaluations in Section 4, we use fully-connected neural networks with leaky ReLU activation functions,

$$\rho(z) = \begin{cases} z & z \geq 0, \\ z/10 & \text{otherwise.} \end{cases}$$

We train the neural network by stochastic gradient descent with the AdamW (Loshchilov and Hutter, 2019) optimizer which combines the adaptive learning rate method Adam with weight decay. Unless stated otherwise, we set the weight decay parameter to 0 (i.e., no weight decay), use an initial learning rate of 0.05 and decay the learning rate every 1000 iterations by 0.85. By default, we train for 20000 iterations with a mini-batch size of 8 in case of regression and 16 in case of classification. In the experiments where we do not vary the neural network size, we use $l = 3$ hidden layers with $w = 64$ neurons each.



(a) Regression

(b) Classification

Figure 7: One instance of the random datasets generated for neural network experiments

D.3 Details of the Training Techniques Experiment (Fig. 1)

The experiment is repeated for 10 random seeds and the error bounds show the standard error. Across the regression bounds, we use a partition of size 25, i.e. a square mesh with 5 intervals along the x and y dimension. As for the classification plots, we use a partition of size 1800. We construct is as the union of two 30×30 meshes in $\mathcal{X}$, one located at $y = 1$ and the other $y = -1$. For a more formal definition, see Appendix A.3. We visualize the plots for different sizes of training set N, the legend in Fig. 1 shows the values.

For adversarial training, we use perturbed samples $x^{\text{adv}} = x + \epsilon \nabla_x l(x, y)$ during stochastic gradient descent. How strongly the adversary perturbs the training inputs is controlled by ϵ, i.e., the higher ϵ, the higher the regularization effect. The x-axis of Fig. 1a and Fig. 1d corresponds to this parameter. For training with weight-decay we effectively use the loss function $\ell_{\text{new}}(\boldsymbol{W}) = \ell_{\text{original}}(\boldsymbol{W}) + \lambda \|\boldsymbol{W}\|_{\text{F}}^2$ where $\boldsymbol{W}$ denotes the network weights and λ is the weight-decay constant which is down on the x-axis of Fig. 1b and Fig. 1e. Lastly, Fig. 1c and Fig. 1f show the effect of early stopping, where the x-axis corresponds to the number of gradient descent iterations.

D.4 Details of the Network Size Experiment (Fig. 2)

For this experiment, we pick the depth of the network as $l \in \{1, 2, 3, 4\}$ and the width of the network as $w \in \{32, 64, 128, 256\}$. The experiment is repeated for 10 random seeds and the error bounds show the standard error. For the regression plot, we use a partition of size 50, i.e. a 10×5 mesh with 10 intervals along the $\mathcal{X}$ and 5 along the $\mathcal{Y}$ dimension. As for the classification plots, we use a partition of size 5000. A partition is constructed as union of two 50×50 meshes in $\mathcal{X}$, one located at $y = 1$ and the other $y = -1$. We visualize the plots for different sizes of training set N, the legend in Fig. 1 shows the values.

N	Global bound
2560	29.283 ± 3.708
5120	16.067 ± 0.990
10240	10.714 ± 0.386
20480	7.935 ± 0.252
40960	5.702 ± 0.258
81920	4.182 ± 0.258

Table 2: Values of global bound in Eq. 3 for the trained classification network, corresponding to Figure 2b. Listed are the means and standard error across 10 seeds.

D.5 Details of the Partitioning Experiment (Fig. 5)

The experiment is repeated for 10 random seeds and the error bounds show the standard error. For the regression bounds, we consider partitions that divide the space into a uniform $M_{\mathcal{X}} \times M_{\mathcal{Y}}$ mesh. The legend of Fig. 5a shows $M_{\mathcal{Y}}$, i.e. the number of parts made in $\mathcal{Y}$, and the horizontal axis shows $M_{\mathcal{X}}$ the number of parts along $\mathcal{X}$. The regression curves are all for a dataset size of $N = 2560$.

For the classification plot, we consider partitions of size $2M^2$, where M is shown on the horizontal axis of the plot. A partition is constructed as union of two $M \times M$ meshes in $\mathcal{X}$, one located at $y = 1$ and the other $y = -1$. For a more formal definition, see Appendix A.3. We visualize the plots for different sizes of training set N, the legend in Fig. 5b shows the values.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 1997.

P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 2018.

P. Alquier, J. Ridgway, N. Chopin, and Y. W. Teh. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016.

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter pac-bayes bounds. *Advances in neural information processing systems*, 2006.

C. Anil, J. Lucas, and R. Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, 2019.

A. Antos, B. Kégl, T. Linder, and G. Lugosi. Data-dependent margin-based generalization bounds for classification. *Journal of Machine Learning Research*, 2002.

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

P. Bartlett, V. Maiorov, and R. Meir. Almost linear vc dimension bounds for piecewise polynomial networks. *Advances in Neural Information Processing Systems*, 1998.

P. L. Bartlett and P. M. Long. Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 2021.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 2005.

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 2017.

P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 2019.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, 2016.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 2019.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 2019.

A. Block, Z. Jia, Y. Polyanskiy, and A. Rakhlin. Intrinsic dimension estimation. *arXiv preprint arXiv:2106.04018*, 2021.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 1989.

E. Boissard and T. Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. Henri Poincaré Probab. Stat.*, 2014.

C. L. Canonne. A short note on learning discrete distributions. *arXiv preprint*, 2020.

O. Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *Lecture notes - Monograph Series*, 2007.

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 2018.

M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, 2017.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

Z. Cranko, Z. Shi, X. Zhang, R. Nock, and S. Kornblith. Generalised lipschitz regularisation equals distributional robustness. In *International Conference on Machine Learning*, 2021.

C. Cuchiero, W. Khosrawi, and J. Teichmann. A generative adversarial network approach to calibration of local stochastic volatility models. *Risks*, 2020.

T. De Ryck, S. Lanthaler, and S. Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 2021.

M. Donsker and Varadhan. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 1975.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence*, 2017.

G. K. Dziugaite and D. M. Roy. Data-dependent pac-bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, 2018.

M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 2016.

C. Finlay, J. Calder, B. Abbasi, and A. Oberman. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint*, 2018.

R. Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 2022.

R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 2022.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40(2):941–963, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS994. URL https://doi.org/10.1214/12-AOS994.

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 2020.

H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 2021.

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 2018.

B. Hajek and M. Raginsky. Ece 543: Statistical learning theory, 2019.

N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Conference on Learning Theory*, 2017.

C. Herrera, F. Krach, and J. Teichmann. Estimating full lipschitz constants of deep neural networks. *arXiv preprint*, 2020.

C. Hyndman and A. Kratsios. NEU: A meta-algorithm for universal uap-invariant feature representation. *Journal of Machine Learning Research*, 2021.

Y. Jiang, D. Krishnan, H. Mobahi, and S. Bengio. Predicting the generalization gap in deep networks with margin distributions. In *International Conference on Learning Representations*, 2019a.

Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. *arXiv preprint*, 2019b.

M. Jordan and A. G. Dimakis. Exactly computing the local lipschitz constant of relu networks. *Advances in Neural Information Processing Systems*, 2020.

L. Kantorovich and G. Rubinstein. On a space of totally additive functions, vestn. *Vestnik Leningrad. Univ*, 1958.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Machine Learning*, 2017.

B. R. Kloeckner. Empirical measures: regularity is a counter-curse to dimensionality. *ESAIM Probab. Stat.*, 2020.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 2001.

A. Krogh and J. Hertz. A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 1991.

D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, 2019.

L. Kuhn, C. Lyle, A. N. Gomez, J. Rothfuss, and Y. Gal. Robustness to pruning predicts generalization in deep neural networks. *arXiv preprint*, 2021.

J. Langford and R. Caruana. (not) bounding the true error. In *Advances in Neural Information Processing Systems*, 2001.

J. Lee and M. Raginsky. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 2018.

A. Levine and S. Feizi. Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, 2020.

M. Li, M. Soltanolkotabi, and S. Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

D. A. McAllester. Some pac-bayesian theorems. In *Machine Learning*, pages 230–234. ACM Press, 1998.

C. McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 1989.

G. P. Meyer. An alternative probabilistic interpretation of the huber loss. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5261–5269, 2021.

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes un-expected bernstein inequality. *Advances in Neural Information Processing Systems*, 2019.

P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 2018.

G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 2014.

M. Morales. *Grokking deep reinforcement learning*. Manning Publications, 2020.

J. Morrill, C. Salvi, P. Kidger, and J. Foster. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

R. Muthukumar and J. Sulam. Adversarial robustness of sparse local lipschitz predictors. *arXiv preprint*, 2022.

H. Narayanan and S. Mitter. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23, 2010.

B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 2015a.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, 2015b.

B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2017.

B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Y. Ohnishi and J. Honorio. Novel change of measure inequalities with applications to pac-bayesian bounds and monte carlo estimation. In *International Conference on Artificial Intelligence and Statistics*, 2021.

A. Picard-Weibel and B. Guedj. On change of measure inequalities for f-divergences. *arXiv preprint*, 2022.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

M. Raissi, P. Perdikaris, and G. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 2019.

J. C. Robinson. *Dimensions, embeddings, and attractors*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2011.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, 2019.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 2012.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, 1997.

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

J. Sokolic, R. Giryes, G. Sapiro, and M. R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 2017.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 2018.

S. Sun, W. Chen, L. Wang, X. Liu, and T.-Y. Liu. On the depth of deep neural networks: A theoretical view. In *AAAI Conference on Artificial Intelligence*, 2016.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 1971.

C. Villani. *Optimal transport: old and new*. Springer, 2009.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 2019.

Y. Xing, Q. Song, and G. Cheng. On the generalization properties of adversarial training. In *International Conference on Artificial Intelligence and Statistics*, 2021.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

W. Zhou, V. Veitch, M. Austern, R. P. Adams, and P. Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2018.