Learning with metric losses

Dan Tsir Cohen DANTSIR@POST.BGU.AC.IL
Ben-Gurion University of the Negev

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Ben-Gurion University of the Negev

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We propose a practical algorithm for learning mappings between two metric spaces, $\mathcal{X}$ and $\mathcal{Y}$. Our procedure is strongly Bayes-consistent whenever $\mathcal{X}$ and $\mathcal{Y}$ are topologically separable and $\mathcal{Y}$ is "bounded in expectation" (our term; the separability assumption can be somewhat weakened). At this level of generality, ours is the first such learnability result for unbounded loss in the agnostic setting. Our technique is based on metric medoids (a variant of Fréchet means) and presents a significant departure from existing methods, which, as we demonstrate, fail to achieve Bayes-consistency on general instance- and label-space metrics. Our proofs introduce the technique of *semi-stable compression*, which may be of independent interest.

Keywords: regression; metric space; sample compression; Bayes-consistency

1. Introduction

Regression and multiclass classification fall under the rubric of supervised prediction from labeled examples. The chief difference between the two is that classification typically assumes the discrete metric on the label space (captured by the 0-1 loss), while regression (at least with absolute loss[1]) implicitly assumes the standard metric over the real-valued labels. In this paper, we study the considerably more general setting of learning with metric losses, where the labels reside in an arbitrary metric space. This setting subsumes both multiclass classification and real-valued regression under L_1 loss, and strictly generalizes these.

We consider the following fundamental learning problem: the instance space $\mathcal{X}$ is endowed with a metric ρ and the label space $\mathcal{Y}$ with a metric ℓ. The learner receives a training sample (X_i, Y_i), $i \in [n]$, drawn iid from an unknown distribution $\bar{\mu}$ on $\mathcal{X} \times \mathcal{Y}$. The learner's goal is to (efficiently) produce a hypothesis $f_n : \mathcal{X} \to \mathcal{Y}$, based on the labeled sample, so as to minimize the *risk* $R(f_n) := \mathbb{E}_{(X,Y)\sim\bar{\mu}}\, \ell(f_n(X), Y)$. In particular, we say that the learning procedure is strongly universally Bayes-consistent if, for every $\bar{\mu}$, we have that $R(f_n) \to R(f^*)$ almost surely as $n \to \infty$, where f^* is the minimizer of $R(\cdot)$ over all measurable $f : \mathcal{X} \to \mathcal{Y}$.

Our contribution. We propose a novel algorithm, MedNet, for learning in the metric-valued regression setting. To our knowledge, this is the first strong Bayes-consistency result for unbounded loss with agnostic noise. While inspired by the OptiNet algorithm of Hanneke et al. (2021), the extension from 0-1 loss to arbitrary metrics required non-trivial modifications to the learning procedure and the risk analysis; a detailed account of the similarities and innovations is provided in Section 2. We show that under quite general, natural conditions on the metric spaces $(\mathcal{X}, \rho)$ and

[1] Quadratic loss can be captured by the *inframetrics* (Fraigniaud et al., 2008) or *near-metrics* (Hanneke, 2021b).

$(\mathcal{Y}, \ell)$, our algorithm is strongly universally Bayes-consistent. The structural assumptions on $\mathcal{X}$ and $\mathcal{Y}$ are truly minimalistic: we require them to be separable metric spaces, and for $\mathcal{Y}$ to be *bounded in expectation*: $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(y_0, Y) < \infty$ for some $y_0 \in Y$. A byproduct of our analysis is the introduction of the *semi-stable compression* technique, which may be of independent interest.

Related work. The regression setting with labels residing in a metric space other than $\mathbb{R}$ is relatively uncommon; such works include Ferraty et al. (2011) and Biess et al. (2019), who study the Banach- and Hilbert-space valued cases, respectively, and mappings between Riemannian manifolds (Hein, 2009; Steinke et al., 2010); some more recent results are discussed below. Our work builds on Hanneke et al. (2021), who gave a complete characterization of the metric spaces $(\mathcal{X}, \rho)$ for which there exists a strong Universal Bayes-Consistent (UBC) learner, where $\mathcal{Y} = \mathbb{N}$ is endowed with the discrete metric. They also provided an algorithm, OptiNet, which achieves strong UBC whenever the latter is achievable by *any* learner (also provided therein is a comprehensive literature review of the strengths and limitations of previous metric-space methods, including k-NN). Györfi and Weiss (2021) followed up with a simplified algorithm, Proto-NN, which, in addition to enjoying all of the properties of OptiNet, is also strongly UBC in L_1 for unbounded real labels $\mathcal{Y} = \mathbb{R}$, as long as $\mathbb{E}\,|Y| < \infty$; our boundedness in expectation generalizes this condition.

Hanneke (2021b) introduced the very general paradigm of "learning whenever learning is possible," which extends the iid setting to essentially the broadest possible class of random processes. That work dealt mostly with the realizable (noiseless) case and bounded losses, though certain kinds of noise were considered in Section 9 therein. A series of recent preprints followed: Blanchard and Cosson (2021); Blanchard (2021); Hanneke (2021a); Blanchard et al. (2021). These also study sampling processes far more general than iid, but consider label-loss structures that are bounded, or noiseless, or both. Blanchard and Cosson (2021), for example, provide a reduction from the metric-valued regression problem to the binary classification problem for the realizable setting with bounded loss. Another aspect in which the above works are incomparable to ours is their use of *non-algorithmic* learning procedures: more in the spirit of an existence proof, these involve non-constructive operations such as enumerating elements of a σ-algebra.[2]

In the special case of the singleton $\mathcal{X} = \{x\}$ and a general metric space $(\mathcal{Y}, \ell)$, the consistency of various Fréchet means (which are naturally related to medoids) has been recently examined by Evans and Jaffe (2020); Schötz (2021). More tangentially related works include Morvant et al. (2012), who, in a PAC-Bayesian setting, gave multiclass risk bounds with a *confusion matrix* error structure, which is close in spirit to assuming a metric on the label set. The assumptions there are fairly restrictive (every label must appear at least a constant number of times in the sample), and no learning procedure or Bayes-consistency result was provided. On the algorithmic front, our procedure partitions the instance space $\mathcal{X}$ into Voronoi cells and chooses the label $y \in \mathcal{Y}$ for each cell based on a variant of the *medoid* principle. A number of loosely medoid-based learning algorithms have been proposed (der Laan et al., 2003; Gottlieb et al., 2016; Newling and Fleuret, 2017; Baharav and Tse, 2019), but our approach is distinct from all of these, in that we compute medoids in the *label* (rather than instance) space.

Finally, *stable compression* was a technique introduced by Bousquet et al. (2020) for the realizable case and extended to the agnostic case by Hanneke and Kontorovich (2021). We introduce a

[2] So as not to get sidetracked down with computability issues over continuous inputs, we only claim full algorithmic constructivity for countable $\mathcal{Y}$. When $\mathcal{Y}$ is merely separable, we assume access to an oracle for computing ε-nets over $\mathcal{Y}$. Such an oracle is easily constructed for, e.g., $\mathcal{Y} = \mathbb{R}^d$.

semi-stable variant for both cases by allowing additional side information; only the compression set (and not the side information) is required to satisfy the stability condition of Bousquet et al.

2. Main results and overview of techniques

Our main result is the existence of a strong Universal Bayes-Consistent (UBC) learner for metric-valued regression.

Theorem 1 *There exists a learning algorithm,* MedNet*, with the following property. Let $(\mathcal{X},\rho)$ and $(\mathcal{Y},\ell)$ be separable metric spaces endowed with a distribution $\bar{\mu}$ supported on the product Borel σ-algebra of $\mathcal{X}\times\mathcal{Y}$, such that $\mathcal{Y}$ is* bounded in expectation*: $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(y_0,Y)<\infty$ for some $y_0\in\mathcal{Y}$. Given a training sample $(X_i,Y_i)_{i\in[n]}$ drawn iid from $\bar{\mu}$ as input,* MedNet *outputs a predictor $f_n:\mathcal{X}\to\mathcal{Y}$ that is strongly universally Bayes-consistent: $R(f_n)\to R^*$ almost surely (under $\bar{\mu}$) as $n\to\infty$, where $R(f)=\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(f(X),Y)$ is the risk and R^* the Bayes-optimal risk (minimum risk achieved by any measurable f).*

The proof proceeds via a sequence of incremental results, culminating in Theorem 12, which is a restatement of Theorem 1. A few remarks regarding our assumptions on $(\mathcal{X},\rho)$ and $(\mathcal{Y},\ell)$ are in order. As per Hanneke et al. (2021), the assumption of separability may be weakened to *essential* separability (ES): this is the condition that the support of $\bar{\mu}$ is contained in a separable subspace. It was shown therein that the ES condition (on $\mathcal{X}$) is also necessary for *any* learner to succeed, and observed that in practice, any plausibly realistic metric space will be ES; in fact, the existence of non-ES metric spaces is widely believed to be independent of ZFC. For countable $\mathcal{Y}$, a variant of Theorem 1 — namely, Theorem 10 — holds for *any* bounded loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,L]$; no metric structure is necessary. Finally, a word about the computational efficiency of MedNet. The latter, conceptually, consists of two stages: (I) computing a γ-net on the finite training sample (residing in $\mathcal{X}$) and (II) for each Voronoi cell C induced on the sample by the γ-net, finding a *medoid* $y\in\mathcal{Y}$ that minimizes $\sum_{c\in C}\ell(c,y)$. Assuming black-box access to an evaluator for the metric ρ, the γ-net in stage (I) can indeed be efficiently constructed on a RAM machine (Gottlieb et al., 2014; Kpotufe and Verma, 2017). At stage (II), MedNet truncates $\mathcal{Y}$ adaptively to some finite $\mathcal{Y}'$ over which the medoid can always be computed in a runtime linear in $|\mathcal{Y}'|$. When additional structural information regarding $\mathcal{Y}$ is available, it may be leveraged to obtain more efficient medoid oracles.

To illustrate our significant departure from previous techniques, let us provide some simple examples where those techniques fail to be Bayes-consistent for various simple label metrics. Let $\mathcal{X}=\{0\}$ be the trivial singleton metric space and $\mathcal{Y}=\{a,b,c,o\}$ be the label space endowed with the metric $\ell(a,b)=\ell(b,c)=\ell(c,a)=1$; $\ell(o,a)=\ell(o,b)=\ell(o,c)=1/2$, and let the distribution $\bar{\mu}$ be such that $\mathbb{P}_{(X,Y)\sim\bar{\mu}}(Y=a)=\mathbb{P}_{(X,Y)\sim\bar{\mu}}(Y=b)=\mathbb{P}_{(X,Y)\sim\bar{\mu}}(Y=c)=1/3$. We observe that any *majority-vote* based method, such as k-NN, which takes a vote among the k nearest neighbors (Györfi et al., 2002), or OptiNet, which takes a vote within each Voronoi cell (Hanneke et al., 2021), or the memory-based techniques of Blanchard and Cosson (2021); Blanchard (2021); Blanchard et al. (2021), or the hybrid approach of Györfi and Weiss (2021) cannot achieve Bayes-consistency in this case — for the simple reason that they can only output the *observed* labels a,b,c (and hence, at best, achieve an asymptotic risk of $2/3$), while the Bayes-optimal predictor $f^*\equiv o$ achieves $R(f^*)=1/2$.

The necessity of predicting labels that never occurred in the sample required overcoming a subtle challenge not present in Hanneke et al. (2021). As in that work, we obtain finite-sample generalization bounds via a sample compression scheme. The latter selects a sub-sample $S_I = (X_i, Y_i)_{i \in I \subset [n]}$, based on which the predictor will be constructed. To mitigate the noise, we occasionally wish to relabel a point $X_i \in S_I$ with a label other than Y_i. One could do this using b bits of side information, but Hanneke et al. sidestepped this issue by doubling the compression set size: the first k pairs (X_i, Y_i) indicate which X's to use (their Y's are discarded) and the second k pairs indicate how to label those first k points X_i. This stratagem is not applicable when we wish to relabel an X with a $Y \in \mathcal{Y}$ not occurring in the sample. Instead, MedNet adaptively truncates $\mathcal{Y}$ to a finite $\mathcal{Y}_n$, whose elements can be described in $b(n) = \log_2 |\mathcal{Y}_n|$ bits of side information. This solves two problems simultaneously: the concentration inequalities we invoke require a bounded range, and our compression schemes require bounded side information. We introduce a semi-stable variant of the *stable* compression scheme (Bousquet et al., 2020; Hanneke and Kontorovich, 2021) to analyze the behavior of our truncated medoid.

Finally, we recall the family of techniques based on Lipchitz-extension, which has found applications in some metric-space learning problems. A binary classifier based on the McShane-Whitney extension theorem was shown to be Bayes-consistent (Kontorovich and Weiss, 2014); this technique was also applied by Gottlieb et al. (2017); Ashlagi et al. (2021) to real-valued regression. When $\mathcal{X}$ and $\mathcal{Y}$ are both Hilbert spaces, the Kirszbraun extension theorem likewise provides a basis for a regression algorithm (Biess et al., 2019). (While the latter three works do not prove Bayes-consistency, the finite-sample generalization bounds provided therein are likely straightforwardly adaptable to such a result via an appropriate regularization schedule.) Unfortunately, Lipschitz extension is not suitable for learning general metric-to-metric mappings. Indeed, this technique is limited to a small number of metric spaces with a special structure; besides the aforementioned cases, Naor and Sheffield (2012) established one for $\mathcal{X}$ a locally compact length space and $\mathcal{Y}$ a metric tree, remarking that "It is rare for a pair of metric spaces [...] to have the isometric extension property." As a concrete example, the spaces $(\mathcal{X}, \rho) = (\mathbb{R}^3, \|\cdot\|_1)$ and $(\mathcal{Y}, \ell) = (\mathbb{R}^2, \|\cdot\|_2)$ fail to have this property (Naor, 2015, Counterexample 2.4).

Open problem. The bounded in expectation (BIE) condition on $\mathcal{Y}$ is a natural generalization of the real-valued variant that $\mathbb{E}|Y| < \infty$ (or, more generally, $\mathbb{E}|Y|^p < \infty$ if L_p risk is being considered). These conditions, while sufficient for Bayes consistency, are clearly not always necessary. Consider, for example, $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, endowed with the standard metric, where the distribution $\bar{\mu}$ is such that the $\mathcal{X}$-marginal is Cauchy (i.e., has density $f(x) = [\pi(1 + x^2)]^{-1}$) and $X = Y$ almost surely. In this case, $\mathbb{E}|Y| = \infty$ and the more general BIE condition also fails. Yet the identity predictor $h(x) = x$ achieves the Bayes-optimal risk of 0, and various simple learning algorithms, including linear regression, achieve Bayes consistency (we conjecture that MedNet does as well). Problem: formulate a necessary and sufficient condition on the metric spaces $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \ell)$, and the joint distribution $\bar{\mu}$ such that MedNet (or some other learning algorithm) is strongly Bayes-consistent. A natural and optimistic candidate is the condition $R^* < \infty$.

3. Definitions and notation

For $n \in \mathbb{N} := \{1, 2, \ldots\}$, define $[n] := \{1, \ldots, n\}$; for any set $\mathcal{Z}$, we write $\mathcal{Z}^+ := \bigcup_{n=1}^{\infty} \mathcal{Z}^n$ and $\mathcal{Z}^{\le k} := \bigcup_{n=1}^{k} \mathcal{Z}^n$, where $|z|$ denotes the sequence length. For $A \in \mathcal{Z}^+$, we write $B \subset A$ to denote the subsequence relation. Our instance and label spaces are the metric spaces $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \ell)$, respectively, whose product Borel σ-algebra is equipped with the probability measure $\bar{\mu}$, whose $\mathcal{X}$-marginal will be denoted by μ and $\mathcal{Y}$-marginal by $\mu_{\mathcal{Y}}$. We say that $\mathcal{Y}$ is *bounded in expectation*

(BIE) if $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(y_0, Y) < \infty$ for some $y_0 \in Y$. Some of our results will also hold for countable $\mathcal{Y}$ equipped with an arbitrary (possibly non-metric) loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. We denote set cardinalities by $|\mathcal{Y}|$ and the diameter by

$$\|\mathcal{Y}\| \equiv \operatorname{diam}(\mathcal{Y}) := \sup_{y,y' \in \mathcal{Y}} \ell(y, y'); \tag{3.1}$$

the latter is also meaningful when ℓ is not a metric. For $x \in \mathcal{X}$ and $r > 0$, $B_r(x)$ denotes the open ball of radius r about x; an analogous definition holds when $(\mathcal{Y}, \ell)$ is a metric. Unless specified otherwise, $S_n = (X_i, Y_i)_{i\in[n]}$ is always sampled iid from $\bar{\mu}$. To any measurable mapping $f : \mathcal{X} \to \mathcal{Y}$ we associate the (true) risk $R(f) := \mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(f(X), Y)$ and the empirical risk $\widehat{R} : \mathcal{Y}^{\mathcal{X}} \times (\mathcal{X} \times \mathcal{Y})^{+} \to \mathbb{R}$ by

$$\widehat{R}(f; S) := |S|^{-1} \sum_{(x,y)\in S} \ell(f(x), y). \tag{3.2}$$

The Bayes-optimal risk is $R^* := \inf_f R(f)$, where the infimum is over all measurable $f : \mathcal{X} \to \mathcal{Y}$.

We implicitly assume the existence of fixed measurable total orders on $\mathcal{X}$ and on $\mathcal{Y}$, whose existence is guaranteed by Hanneke et al. (2021, Proposition D.1), and refer to these orderings as *lexicographic*. For $A \subseteq \mathcal{X}$, denote its γ-envelope by $\mathrm{UB}_\gamma(A) := \cup_{x\in A} B_\gamma(x)$ and consider the γ-*missing mass* of S_n, defined as the following random variable:

$$\mathsf{mm}_\gamma(S_n) := \mu(\mathcal{X} \setminus \mathrm{UB}_\gamma(S_n)). \tag{3.3}$$

As in Hanneke et al. (2021), we denote, for any labeled sequence $S = (x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, and any $x \in \mathcal{X}$, the nearest neighbor of x with respect to S and its label by $X_{\mathsf{nn}}(x, S)$ and $Y_{\mathsf{nn}}(x, S)$, respectively:

$$(X_{\mathsf{nn}}(x, S), Y_{\mathsf{nn}}(x, S)) := \operatorname*{argmin}_{(x_i,y_i)\in S} \rho(x, x_i),$$

where ties are broken lexicographically. The 1-NN predictor induced by S is defined as $h_S(x) := Y_{\mathsf{nn}}(x, S)$. For any $m \in \mathbb{N}$, any sequence $\boldsymbol{X} = \{x_1, \ldots, x_m\} \in \mathcal{X}^m$ induces a *Voronoi partition* of $\mathcal{X}$, $\mathcal{V}(\boldsymbol{X}) := \{V_1(\boldsymbol{X}), \ldots, V_m(\boldsymbol{X})\}$, where each Voronoi cell is

$$V_i(\boldsymbol{X}) := \left\{ x \in \mathcal{X} : i = \operatorname*{argmin}_{1\le j\le m} \rho(x, x_j) \right\},$$

again breaking ties lexicographically. In particular, for $\boldsymbol{X} = \{X_i : (X_i, Y_i) \in S\}$, we have $h_S(x) = Y_i$ for all $x \in V_i(\boldsymbol{X})$. A 1-NN algorithm is a mapping from an i.i.d. labeled sample $S_n \sim \bar{\mu}^n$ to a labeled set $S'_n \subseteq \mathcal{X} \times \mathcal{Y}$, yielding the 1-NN predictor $h_{S'_n}$. For $A \subseteq \mathcal{X}$ and $\gamma > 0$, a γ-*net* of A is any *maximal set* $B \subseteq A$ in which all interpoint distances are at least γ. For a partition $\mathcal{A}$ of $B \subseteq \mathcal{X}$, we write $\|\mathcal{A}\| := \sup_{A\in\mathcal{A}} \|A\|$ (again, as in (3.1), $\|A\| := \operatorname{diam} A$). Given a labeled set $S_n = (x_i, y_i)_{i\in[n]}$, $d \in [n]$, and any $\boldsymbol{i} = \{i_1, \ldots, i_d\} \in [n]^d$, denote the sub-sample of S_n indexed by $\boldsymbol{i}$ by $S_n(\boldsymbol{i}) := \{(x_{i_1}, y_{i_1}), \ldots, (x_{i_d}, y_{i_d})\}$. Similarly, for a vector $\boldsymbol{y}' = \{y'_1, \ldots, y'_d\} \in \mathcal{Y}^d$, define $S_n(\boldsymbol{i}, \boldsymbol{y}') := \{(x_{i_1}, y'_1), \ldots, (x_{i_d}, y'_d)\}$, namely the sub-sample of S_n as determined by $\boldsymbol{i}$ where the labels are replaced with $\boldsymbol{y}'$. Lastly, for $\boldsymbol{i}, \boldsymbol{j} \in [n]^d$, we denote $S_n(\boldsymbol{i}; \boldsymbol{j}) := \{(x_{i_1}, y_{j_1}), \ldots, (x_{i_d}, y_{j_d})\}$.

We use standard order-of-magnitude notation throughout the paper; thus, for $f, g : \mathbb{N} \to [0, \infty)$ we write $f(n) \in O(g(n))$ to mean $\limsup_{n\to\infty} f(n)/g(n) < \infty$ and $f(n) \in o(g(n))$ to mean $\limsup_{n\to\infty} f(n)/g(n) = 0$. Likewise, $f(n) \in \Omega(g(n))$ means that $g(n) \in O(f(n))$. In accordance with common convention, we often use the less precise notation $f(n) = O(g(n))$, etc.

We say that a metric space $(\mathcal{X}, \rho)$ is *separable* if it contains a dense countable set. A metric probability space $(\mathcal{X}, \rho, \mu)$ is separable if there is a measurable $\mathcal{X}' \subseteq \mathcal{X}$ with $\mu(\mathcal{X}') = 1$ such that $(\mathcal{X}', \rho)$ is separable.

A *sample compression scheme* (κ, ψ) of size at most k using b bits of side-information consists of a *compression function* and a *reconstruction function*. The *compression function* κ maps every finite sample set to b bits plus a *compression set*, which is a subset of at most k labeled examples.

$$\kappa : (\mathcal{X} \times \mathcal{Y})^+ \to (\mathcal{X} \times \mathcal{Y})^{\leq k} \times \{0,1\}^b .$$

The *reconstruction function* ψ maps every possible compression set and b bits to a hypothesis:

$$\psi : (\mathcal{X} \times \mathcal{Y})^{\leq k} \times \{0,1\}^b \to \mathcal{Y}^{\mathcal{X}}.$$

4. Semi-stable compression

In this section, we expand the definition of *stable compression* and present our results for this notion. First, we split the compression function κ into its two components. For $S \in (\mathcal{X} \times \mathcal{Y})^+$, we write $\kappa(S) = (\kappa_{\mathsf{cs}}(S), \kappa_{\mathsf{si}}(S)) \in (\mathcal{X} \times \mathcal{Y})^{\leq k} \times \{0,1\}^b$; these are the *compression set* and the *side information*. We say that (κ, ψ) is *semi-stable* if the κ_{cs} component is stable in the sense of Bousquet et al. (2020): whenever $\kappa_{\mathsf{cs}}(S) \subseteq S' \subseteq S$, we have that

$$\psi(\kappa_{\mathsf{cs}}(S'), \kappa_{\mathsf{si}}(S)) = \psi(\kappa_{\mathsf{cs}}(S), \kappa_{\mathsf{si}}(S)) = \psi(\kappa(S)).$$

We denote by $|\kappa_{\mathsf{cs}}(\cdot)|$ and $|\kappa_{\mathsf{si}}(\cdot)|$ the sizes of the compression set and side information (in bits), respectively.

Theorem 2 (proof deferred to Section C.3.4) *Suppose that $\mathcal{X}$ is an instance space and $\mathcal{Y}$ a label space with a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, L]$, and (κ, ψ) is semi-stable compression scheme. For any distribution $\bar{\mu}$ over $\mathcal{X} \times \mathcal{Y}$, any $n \in \mathbb{N}$, and any $\delta \in (0, 1)$, for $S_n \sim \bar{\mu}^n$ we have that*

$$\begin{aligned} R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) \leq & \left(20\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 20\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 15\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \right) \widehat{R}(\psi(\kappa(S_n)); S_n) \\ & + (6L + 18)\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 8L\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + (2L + 12)\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} \\ & + 7L\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + (3L + 10)\frac{\ln(\frac{4e^2}{\delta})}{n} + 6L\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \end{aligned}$$

holds with probability at least $1 - \delta$.

(α, k, b)**-semi-stable-compression**

Let $\mathcal{X}$, $\mathcal{Y}$, ℓ, and S_n be as in the statement of Theorem 2. For $k \leq n$, $b \in \mathbb{N}$, and $\alpha \geq 0$, we say that $(S_n', h_{S_n'})$ is an (α, k, b)-*semi-stable-compression* of S_n if there exist $\boldsymbol{i} \in [n]^k$ and $\mathbf{Y} \in \mathcal{Y}^k$ such that:

1. $h_{S_n'}$ and $S_n' = S_n(\boldsymbol{i}, \mathbf{Y})$ are a result of a semi-stable compression scheme of size k with at most b bits of side information. Thus, $S_n' = \kappa_{\mathsf{si}}(S_n)$ and $h_{S_n'} = \psi(\kappa_{\mathsf{cs}}(S), \kappa_{\mathsf{si}}(S))$.

2. $\widehat{R}(h_{S_n'}; S_n) \leq \alpha$.

Lemma 3 (proof in Section B.2) *Let $\mathcal{X}$, $\mathcal{Y}$, ℓ, L, and S_n be as in Theorem 2. For $k \leq n$, define*

$$\begin{aligned} Q(n, \alpha, k, b, \delta, L) := Q_n(\alpha, k, b, \delta, L) := &\left(20\sqrt{\frac{k}{n}} + 20\sqrt{\frac{b}{n}} + 15\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} + 1\right)\alpha \\ &+ (6L+18)\frac{k}{n} + 8L\sqrt{\frac{k}{n}} + (2L+12)\frac{b}{n} + 7L\sqrt{\frac{b}{n}} \\ &+ (3L+10)\frac{\ln(\frac{4e^2}{\delta})}{n} + 6L\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}. \end{aligned} \tag{4.1}$$

Then the function Q satisfies the following properties:

Q1. *For any $n \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over $S_n \sim \bar{\mu}^n$, for all $\alpha \in [0, L]$, $k \in [n]$, $b \in \mathbb{N}$: If $(S_n', h_{S_n'})$ is an (α, k, b)-semi-stable-compression of S_n, then*

$$R(h_{S_n'}) \leq Q_n(\alpha, k, b, \delta, L).$$

Q2. *For any fixed $n \in \mathbb{N}$ and $\delta \in (0,1)$, Q is monotonically increasing in α and in k.*

Q3$''$. *There is a sequence $\{\delta_n\}_{n=1}^{\infty}$, $\delta_n \in (0,1)$ such that $\sum_{n=1}^{\infty} \delta_n < \infty$, and for any $k_n \in o(n)$ we have that*

$$\lim_{n\to\infty} \sup_{\alpha \in [0,L]} (Q_n(\alpha, k_n, b, \delta_n, L) - \alpha) = 0.$$

5. Metric approximations

Our proof technique involves performing several distinct truncations, approximating a potentially unbounded quantity by a finite one. In this section, we adapt a variant of this method from Hanneke et al. (2021) for the 0-1 loss to arbitrary bounded losses. Here, $(\mathcal{X}, \rho)$ is assumed to be a separable metric space, and $\mathcal{Y}$ a *countable* label space with a loss function $\ell : \mathcal{Y}^2 \to [0, L]$. Let $\mathcal{V} = \{V_1, \dots\}$ be a countable partition of $\mathcal{X}$, and define the function $I_{\mathcal{V}} : \mathcal{X} \to \mathcal{V}$ such that $I_{\mathcal{V}}(x)$ is the unique $V \in \mathcal{V}$ for which $x \in V$. For any measurable set $\emptyset \neq E \subseteq \mathcal{X}$ define the true medoid label $y^*(E)$ by

$$y^*(E) = \operatorname*{argmin}_{y \in \mathcal{Y}} \int_{X \in E} \ell(y, Y) \mathrm{d}\bar{\mu}, \tag{5.1}$$

where ties are broken lexicographically according to fixed total order on $\mathcal{Y}$. Given $\mathcal{V}$ and a measurable set $W \subseteq \mathcal{X}$, define the true medoid predictor $h^*_{\mathcal{V},W} : \mathcal{X} \to \mathcal{Y}$ given by

$$h^*_{\mathcal{V},W}(x) = y^*(I_{\mathcal{V}}(x) \cap W). \tag{5.2}$$

Lemma 4 (proof in Section B.3) *Let $\bar{\mu}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$-marginal μ, where $\mathcal{X}$ is a metric probability space, and $\mathcal{Y}$ a* countable *label space with a loss function ℓ such that $L := \|\mathcal{Y}\| < \infty$. For any $\nu > 0$, there exists a diameter $\beta = \beta(\nu) > 0$ such that for any countable measurable partition $\mathcal{V} = \{V_1, \ldots\}$ of $\mathcal{X}$ and any measurable set $W \subseteq \mathcal{X}$ satisfying*

(i) $\mu(\mathcal{X} \setminus W) \leq \nu$

(ii) $\sup_{V \in \mathcal{V}} \|V \cap W\| \leq \beta$,

the true medoid predictor $h^*_{\mathcal{V},W}$ defined in (5.2) satisfies

$$R(h^*_{\mathcal{V},W}) \leq R^* + 9L\nu.$$

The proof of Lemma 4 is similar to that of Hanneke et al. (2021, Lemma 3.6), with novel arguments to handle the general loss function setting. Next, we state two results from Hanneke et al. (2021):

Lemma 5 (variant of Lemma 3.7, Hanneke et al. (2021)) *Let $(\mathcal{X}, \rho, \mu)$ be a separable metric probability space. For $S_n \sim \mu^n$, let $\boldsymbol{X}(\gamma)$ be any γ-net of S_n. Then, for any $\gamma > 0$, there exists a function $t_\gamma : \mathbb{N} \to \mathbb{R}_+$ in $o(n)$ such that $\mathbb{P}\left[\sup_{\gamma\text{-nets } \boldsymbol{X}(\gamma)} |\boldsymbol{X}(\gamma)| \geq t_\gamma(n)\right] \leq 1/n^2$.*

Lemma 6 (Lemma 3.8, Hanneke et al. (2021)) *Let $(\mathcal{X}, \rho, \mu)$ be a separable metric probability space, $\gamma > 0$ be fixed, and the γ-missing mass mm_γ defined as in (3.3). Then there exists a function $u_\gamma : \mathbb{N} \to \mathbb{R}_+$ in $o(1)$, such that $\mathbb{P}\left[\mathsf{mm}_\gamma(S_n) \geq u_\gamma(n) + t\right] \leq \exp\left(-nt^2\right)$ for $S_n \sim \mu^n$ and $t > 0$.*

6. Algorithms and analysis: finite $\mathcal{Y}$

In this section, we give the most basic version of our algorithm, denoted $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$, for the case where $(\mathcal{X}, \rho)$ is a separable metric and $\mathcal{Y}$ is a *finite* set equipped with an *arbitrary* (not necessarily metric) loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$. This rudimentary setting provides the basis for extension to more general settings, in the sequel.

The input is the sample S_n; the set of instances in the sample is denoted by $\boldsymbol{X}_n = \{X_1, \ldots, X_n\}$. The algorithm defines a set Γ of all $\binom{n}{2}$ scales $\gamma > 0$ which are interpoint distances in $\boldsymbol{X}_n$, and the additional scale $\gamma = \infty$. For each scale in Γ, the algorithm constructs a γ-net of $\boldsymbol{X}_n$. Denote the constructed γ-net by

$$\boldsymbol{X}(\gamma) := \{X_{i_1}, \ldots, X_{i_M}\}, \tag{6.1}$$

where

$$M \equiv M_n(\gamma) := |\boldsymbol{X}(\gamma)| \tag{6.2}$$

denotes its size and $\boldsymbol{i} \equiv \boldsymbol{i}(\gamma) := \{i_1, \ldots, i_M\} \in [n]^M$ denotes the indices selected from S_n for this γ-net.

For each γ-net, Algorithm 1 finds the *empirical medoid labels* in the Voronoi cells defined by the partition $\mathcal{V}(\boldsymbol{X}(\gamma)) = \{V_1(\boldsymbol{X}(\gamma)), \ldots, V_M(\boldsymbol{X}(\gamma))\}$. These labels are denoted by $\boldsymbol{Y}'(\gamma) \in \mathcal{Y}^M$. Formally, for $i \in [M]$,

$$Y_i'(\gamma) := \underset{y \in \mathcal{Y}}{\operatorname{argmin}} \sum_{j \in [n] : X_j \in V_i} \ell(y, Y_j). \tag{6.3}$$

As always, ties are broken lexicographically. The output of $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$ is a labeled set $S_n'(\gamma) := S_n(\boldsymbol{i}(\gamma), \boldsymbol{Y}'(\gamma))$ for every candidate scale $\gamma \in \Gamma$. The algorithm then selects a single scale $\gamma^* \equiv \gamma_n^*$ from Γ, and outputs the hypothesis that it induces, $h_{S_n'(\gamma^*)}$. The choice of γ^* is executed by minimizing a generalization error bound, denoted Q, which upper-bounds $R(h_{S_n'(\gamma)})$ with high probability.

```
Algorithm 1: $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$
Assumptions: $(\mathcal{X}, \rho)$ is a separable metric space, $\mathcal{Y}$ a *finite* label space with a loss function $\ell$.
             Define $L := \|\mathcal{Y}\| = \max_{y,y' \in \mathcal{Y}} \ell(y, y')$ and $b := \log_2 |\mathcal{Y}|$.
Input        : Sample $S_n = (X_i, Y_i)_{i \in [n]}$, confidence $\delta_n \in (0,1)$
Output       : predictor $h : \mathcal{X} \to \mathcal{Y}$
Let $\Gamma := (\{\rho(X_i, X_j) : i, j \in [n]\} \cup \{\infty\}) \setminus \{0\}$
for $\gamma \in \Gamma$ do
    Let $\boldsymbol{X}(\gamma)$ be a $\gamma$-net of $\{X_1, \ldots, X_n\}$
    Let $M_n(\gamma) := |\boldsymbol{X}(\gamma)|$
    For each $i \in [M_n(\gamma)]$, let $Y_i'(\gamma)$ be the *empirical medoid label* of $V_i(\boldsymbol{X}(\gamma))$ as in (6.3)
    Set $S_n'(\gamma) := (\boldsymbol{X}(\gamma), \boldsymbol{Y}'(\gamma))$
    Set $h_{S_n'(\gamma)} := x \mapsto Y_{\mathsf{nn}}(x, S_n'(\gamma))$.
    Set $\alpha_n(\gamma) := \widehat{R}(h_{S_n'(\gamma)}; S_n)$
end
Find $\gamma_n^* \in \operatorname{argmin}_{\gamma \in \Gamma} Q_n(\alpha_n(\gamma), M_n(\gamma), b, \delta, L)$, where $Q_n$ is defined in (4.1)
Set $S_n' := S_n'(\gamma_n^*)$
return $h = h_{S_n'}$
```

Bayes Consistency of $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$

The Bayes consistency result of Hanneke et al. (2021) was for the 0-1 loss. Their approach was also compression-based, but did not leverage the stability property, had no need for side information, and did not have to truncate potentially unbounded losses. Our main technical innovation was constructing MedNet (formally defined in Section A.1) as a semi-stable compression scheme with side-information, and then invoking it with an appropriate truncation schedule for infinite and unbounded $\mathcal{Y}$.

The first order of business is to verify that $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$ indeed furnishes a semi-stable compression scheme for any fixed γ:

Lemma 7 (proof in Section B.1) *Let $(\mathcal{X}, \rho)$ be a separable metric space, and $\mathcal{Y}$ a* finite *label space with a loss function ℓ. For any fixed scale $\gamma \in \Gamma$, the procedure in Algorithm 1 generating $h_{S'_n(\gamma)}$ is a semi-stable compression scheme.*

The following key technical lemma is a generalization of Hanneke et al. (2021, Lemma 3.5) from 0-1 loss to the general loss setting.

Lemma 8 (proof in Section B.4) *Let $\bar{\mu}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a metric probability space, and $\mathcal{Y}$ a* countable *label space endowed with a loss function $\ell \le L < \infty$. Let t_γ as in Lemma 5. Then there exist functions $\varepsilon \mapsto \gamma(\varepsilon)$ and $\varepsilon \mapsto \gamma := \nu(\varepsilon) \in (0, \frac{\varepsilon}{176L})$ such that for each $\varepsilon, b > 0$ there is an $N_0(\nu(\varepsilon), b, \delta_n, t_\gamma)$ such that for all $n \ge N_0$, and all $d \in [t_\gamma(n)]$,*

$$p_d := \mathbb{P}\left[Q_n(\alpha_n(\gamma), M_n(\gamma), b, \delta_n, L) > R^* + \varepsilon \;\wedge\; \mathsf{mm}_\gamma(S_n) \le \frac{\varepsilon}{18L} \;\wedge\; M_n(\gamma) = d\right] \le e^{-\frac{n\varepsilon^2}{32}} + e^{-\frac{1}{2}n\nu^2},$$

where $M_n(\gamma)$ is defined in (6.2).

The main result of this section is

Theorem 9 *Let $(\mathcal{X}, \rho)$ be a separable metric space, and $\mathcal{Y}$ a* finite *label space with a loss function ℓ. Then there exists a choice of $\delta_{n\in\mathbb{N}}$ such that the sequence of hypotheses h_n computed by $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}(S_n, \delta_n)$ is strongly Bayes consistent: $\mathbb{P}[\lim_{n\to\infty} R(h_n) = R^*] = 1$.*

Proof Recall that $L := \|\mathcal{Y}\| = \max_{y,y' \in \mathcal{Y}} \ell(y, y')$ and let $b := \log_2 |\mathcal{Y}|$. Let Q be the generalization bound in (4.1) and set the input confidence δ for input size n to δ_n as stipulated by **Q3**$''$.

Given a sample $S_n \sim \bar{\mu}^n$, we abbreviate the optimal empirical error $\alpha_n^* = \alpha(\gamma_n^*)$ and the optimal compression size $M_n^* = M(\gamma_n^*)$ as computed by Algorithm 1. By Lemma 7, the labeled set $S'_n(\gamma_n^*)$ computed by Algorithm 1 is an (α_n^*, M_n^*, b)-semi-stable compression of the sample S_n. For brevity we denote $Q_n(\alpha, k) := Q_n(\alpha, k, b, \delta_n, L)$. To prove the Theorem, we first follow the standard technique, used also in Hanneke et al. (2021), of decomposing the excess risk into two terms:

$$R(h_{S'_n(\gamma_n^*)}) - R^* \quad = \quad \left(R(h_{S'_n(\gamma_n^*)}) - Q_n(\alpha_n^*, M_n^*)\right) + \left(Q_n(\alpha_n^*, M_n^*) - R^*\right) =: T_{\mathrm{I}}(n) + T_{\mathrm{II}}(n)$$

and arguing that each term decays to zero almost surely. For $T_{\mathrm{I}}(n)$ we have, similarly to Hanneke et al., that Property **Q1** from Lemma 3 implies that for any $n > 0$,

$$\mathbb{P}\left[R(h_{S'_n(\gamma_n^*)}) - Q_n(\alpha_n^*, M_n^*) > 0\right] \le \delta_n. \tag{6.4}$$

Applying Borel-Cantelli to the fact that $\sum \delta_n < \infty$ yields $\limsup_{n\to\infty} T_{\mathrm{I}}(n) \le 0$ almost surely. The main departure from the proof in Hanneke et al. is in establishing $\limsup_{n\to\infty} T_{\mathrm{II}}(n) \le 0$ almost surely. We will argue that there exist $N = N(\varepsilon) > 0$, $\gamma = \gamma(\varepsilon) > 0$, $\nu = \nu(\varepsilon) > 0$, and universal constants $c, C > 0$ such that $\forall n \ge N$,

$$\mathbb{P}[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon] \le Cne^{-cn\varepsilon^2} + ne^{-\nu^2 n/2} + 1/n^2. \tag{6.5}$$

For any $\gamma > 0$ (even if $\gamma \notin \Gamma$), Algorithm 1 finds a γ_n^* such that

$$Q_n(\alpha_n^*, M_n^*) \quad = \quad \min_{\gamma' \in \Gamma} Q_n(\alpha_n(\gamma'), M_n(\gamma')) \;\le\; Q_n(\alpha_n(\gamma), M_n(\gamma)).$$

The bound in (6.5) thus implies that $\forall n \geq N$,

$$\mathbb{P}[Q_n(\alpha_n^*, M_n^*) > R^* + \varepsilon] \leq Cne^{-cn\varepsilon^2} + ne^{-\nu^2 n/2} + 1/n^2. \tag{6.6}$$

By the Borel-Cantelli lemma, this implies that almost surely, $\limsup_{n\to\infty} T_{\mathrm{II}}(n) = \limsup_{n\to\infty}(Q_n(\alpha_n^*, M_n^*) - R^*) \leq 0$. Since $\forall n, T_{\mathrm{I}}(n) + T_{\mathrm{II}}(n) \geq 0$, this implies $\lim_{n\to\infty} T_{\mathrm{II}}(n) = 0$ almost surely, thus completing the proof.

It remains to prove (6.5), the 0-1 loss analog of which was proved in Hanneke et al., Eq. (3.4). That argument does not hold for general losses, and we present the novel argument below. We bound the left-hand side of (6.5) using a function $n \mapsto t_\gamma(n) \in o(n)$, used to upper bound the compression size; the latter is furnished by Lemma 5.

$$\begin{aligned}
&\mathbb{P}[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon] \\
\leq\ & \mathbb{P}\Big[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon \ \wedge\ \mathsf{mm}_\gamma(S_n) \leq \frac{\varepsilon}{18L} \ \wedge\ M_n(\gamma) \leq t_\gamma(n)\Big] \\
& + \mathbb{P}[\mathsf{mm}_\gamma(S_n) > \frac{\varepsilon}{18L}] + \mathbb{P}[M_n(\gamma) > t_\gamma(n)] =: P_{\mathrm{I}} + P_{\mathrm{II}} + P_{\mathrm{III}}.
\end{aligned} \tag{6.7}$$

We estimate P_{I} via a union bound:

$$\begin{aligned}
&\mathbb{P}\Big[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon \ \wedge\ \mathsf{mm}_\gamma(S_n) \leq \frac{\varepsilon}{18L} \ \wedge\ M_n(\gamma) \leq t_\gamma(n)\Big] \\
\leq & \sum_{d=1}^{t_\gamma(n)} \mathbb{P}\Big[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon \ \wedge\ \mathsf{mm}_\gamma(S_n) \leq \frac{\varepsilon}{18L} \ \wedge\ M_n(\gamma) = d\Big].
\end{aligned}$$

Thus, it suffices to bound each term in the summation separately. Applying Lemma 8 and summing, we have, for n sufficiently large that $t_\gamma(n) \leq n$,

$$P_{\mathrm{I}} \leq \sum_{d=1}^{t_\gamma(n)} p_d \ \leq\ t_\gamma(n)(e^{-\frac{n\varepsilon^2}{32}} + e^{-\frac{1}{2}n\nu^2}) \leq n(e^{-\frac{n\varepsilon^2}{32}} + e^{-\frac{1}{2}n\nu^2}). \tag{6.8}$$

Now, using the function t_γ, we note that $P_{\mathrm{III}} \leq 1/n^2$ thanks to Lemma 5. A bound on P_{II}, which bounds the γ-missing-mass $\mathsf{mm}_\gamma(S_n)$, is furnished by Lemma 6. Taking n sufficiently large so that $u_\gamma(n)$, as furnished by Lemma 6, satisfies $u_\gamma(n) \leq \varepsilon/36L$, and invoking Lemma 6 with $t = \varepsilon/36L$, we have $P_{\mathrm{II}} = \mathbb{P}[\mathsf{mm}_\gamma(S_n) > \varepsilon/18L] \leq e^{-\frac{n\varepsilon^2}{1296L^2}}$. Plugging this, (6.8), and $P_{\mathrm{III}} \leq 1/n^2$ into (6.7) yields (6.5), which completes the proof. ■

7. Extensions

7.1. Countable $\mathcal{Y}$ with finite diameter: $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$

In this section we describe an extension of $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$, denoted $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$, which is strongly Bayes-consistent for countably infinite $\mathcal{Y}$, but still with a finite diameter. A modification of $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$ is required because the latter uses a compression scheme with $b = \log_2 |\mathcal{Y}|$ bits of side information.

Our variant is formally presented in Algorithm 3 and operates as follows. We fix in advance a specific sequence $b_n \in \mathbb{N}$, to be specified in the sequel. The family of γ-nets over the input sample is generated exactly as in $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$. For each γ-net, $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ (presented in Section A.2) computes the *truncated empirical medoid labels* in the Voronoi cells defined by the partition $\mathcal{V}(\boldsymbol{X}(\gamma)) = \{V_1(\boldsymbol{X}(\gamma)), \dots, V_M(\boldsymbol{X}(\gamma))\}$. These labels are denoted by $\boldsymbol{Y}'(\gamma) \in \mathsf{pref}(\mathcal{Y}, b_n)^M$. Formally, for $i \in [M]$,

$$Y_i' := \underset{y \in \mathsf{pref}(\mathcal{Y}, b_n)}{\operatorname{argmin}} \sum_{j \in [n]: X_j \in V_i} \ell(y, Y_j), \tag{7.1}$$

where $\mathsf{pref}(\mathcal{Y}, b) := \left\{ y \in \mathcal{Y}' : \omega(y) \le 2^b \right\}$ for $b \in \mathbb{N}$, for some fixed canonical injection $\omega : \mathcal{Y} \to \mathbb{N}$. In words, $\mathsf{pref}(\mathcal{Y}, b_n)$ is a sample-dependent, cardinality-based truncation of the label space. Other than the truncation, $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ behaves exactly as $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$.

Theorem 10 (proof in Section B.6) *Let $(\mathcal{X}, \rho)$ be a separable metric space, and $\mathcal{Y}$ a* countable *label space with a loss function $\ell \le L < \infty$. Then there is a choice of $\delta_{n \in \mathbb{N}}$ and truncation schedule $b_{n \in \mathbb{N}}$ such that the sequence of hypotheses h_n computed by $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}(S_n, \delta_n, b_n)$ is strongly Bayes consistent: $\mathbb{P}[\lim_{n\to\infty} R(h_n) = R^*] = 1$.*

7.2. Countable metric space $(\mathcal{Y}, \ell)$ with unbounded diameter: $\mathsf{MedNet}^{\aleph_0}$

In this section, we extend $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ to the case where $(\mathcal{Y}, \ell)$ is a countable metric space. That is, the loss ℓ is now assumed to be a metric, but the boundedness condition $\|\mathcal{Y}\| < \infty$ is relaxed to boundlessness-in-expectation (BIE): $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\, \ell(y_0, Y) < \infty$ for some $y_0 \in \mathcal{Y}$. Boundedness was used in the analysis of $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ in order to invoke a distribution-free concentration inequality (Hoeffding's). The present extension, denoted $\mathsf{MedNet}^{\aleph_0}$, invokes $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ as a subroutine with an appropriately diameter-truncated label space. The latter is defined as follows. Fix a $y_0 \in \mathcal{Y}$ that is a witness of the BIE property[3]. For $y \in \mathcal{Y}$ and $L > 0$, define $[\![\mathcal{Y}]\!]_L := B(y_0, L)$ and the *diameter-truncation* operation

$$y \wedge L := \underset{\hat{y} \in [\![\mathcal{Y}]\!]_L}{\operatorname{argmin}}\, \ell(y, \hat{y}). \tag{7.2}$$

In words, $y \wedge L$ is the closest $\hat{y}$ to y in the L-ball about y_0.

$\mathsf{MedNet}^{\aleph_0}$ is formally presented in Algorithm 2 and operates as follows. The cardinality- and diameter-truncation schedules $b_{n\in\mathbb{N}}$ and $L_{n\in\mathbb{N}}$ are fixed in advance; the former as any $b_n \in o(n)$ and the latter specified in the sequel. Next, the labels Y_i of the input sample are truncated to $[\![Y_i]\!] := Y_i \wedge L_n$; this is a substantive difference from the cardinality-based truncation in $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$, which does not modify the sample labels.

Theorem 11 (proof in Section B.7) *Let $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \ell)$ be metric spaces, separable and countable, respectively, equipped with a product distribution $\bar{\mu}$ such that BIE holds for $\mathcal{Y}$. Then there is a choice of $\delta_{n\in\mathbb{N}}$ and truncation schedules $b_{n\in\mathbb{N}}$, $L_{n\in\mathbb{N}}$ such that the sequence of hypotheses h_n computed by $\mathsf{MedNet}^{\aleph_0}(S_n, \delta_n, b_n, L_n)$ is strongly Bayes consistent: $\mathbb{P}[\lim_{n\to\infty} R(h_n) = R^*] = 1$.*

The only remaining extension to render the proof of Theorem 1 complete is from countable to *separable* $(\mathcal{Y}, \ell)$; this straightforward step is carried out in Section A.1.

[3] Lemma 18 shows that if BIE holds then *every* $y' \in \mathcal{Y}$ is such a witness, and in particular, we may always choose y_0 as the "first" element under the canonical ordering.

```
Algorithm 2: $\mathsf{MedNet}^{\aleph_0}$
Assumptions: $(\mathcal{X}, \rho)$ is a separable metric space, $(\mathcal{Y}, \ell)$ a BIE countable metric space
Input      : Sample $S_n = (X_i, Y_i)_{i \in [n]}$, $\delta_n \in (0, 1)$, $b_n \in \mathbb{N}$, $L_n > 0$
Output     : predictor $h : \mathcal{X} \to \mathcal{Y}$
Set $[\![S_n]\!] := \{(X_i, [\![Y_i]\!]) : i \in [n]\}$, where $[\![Y_i]\!] := Y_i \wedge L_n$
Set $h_n := \mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\| < \infty}([\![S_n]\!], \delta_n, b_n)$ in truncated label space $[\![\mathcal{Y}]\!]_{L_n}$
                                        (i.e., $\mathcal{Y}$ in (6.3) is replaced with $[\![\mathcal{Y}]\!]_{L_n}$)
return $h = h_n$
```

Acknowledgements. This research was partially supported by the Israel Science Foundation (grant No. 1602/19), an Amazon Research Award, and the Ben-Gurion University Data Science Research Center.

References

Yair Ashlagi, Lee-Ad Gottlieb, and Aryeh Kontorovich. Functions with average smoothness: structure, algorithms, and learning. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 186–236. PMLR, 2021. URL http://proceedings.mlr.press/v134/ashlagi21a.html.

Tavor Z. Baharav and David Tse. Ultra fast medoid identification via correlated sequential halving. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c4de8ced6214345614d33fb0b16a8ac

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361 – 1385, 2015. doi: 10.3150/14-BEJ605. URL https://doi.org/10.3150/14-BEJ605.

Armin Biess, Aryeh Kontorovich, Yury Makarychev, and Hanan Zaichyk. Regression via Kirszbraun extension with applications to imitation learning. *CoRR*, abs/1905.11930, 2019. URL http://arxiv.org/abs/1905.11930.

Moïse Blanchard. Universal strong and weak online learning with bounded loss, preprint. 2021.

Moïse Blanchard and Romain Cosson. Universal online learning with bounded loss: Reduction to binary classification. 2021.

Moïse Blanchard, Romain Cosson, and Steve Hanneke. Universal online learning with unbounded losses: Memory is all you need. 2021.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609. PMLR, 2020. URL http://proceedings.mlr.press/v125/bousquet20a.html.

Mark Van der Laan, Katherine Pollard, and Jennifer Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003. doi: 10.1080/0094965031000136012. URL https://doi.org/10.1080/0094965031000136012.

Steven N. Evans and Adam Q. Jaffe. Strong laws of large numbers for Fréchet means, 2020.

Frédéric Ferraty, Ali Laksaci, Amel Tadj, and Philippe Vieu. Kernel regression with functional response. *Electronic Journal of Statistics*, 5(none):159 – 171, 2011. doi: 10.1214/11-EJS600. URL https://doi.org/10.1214/11-EJS600.

Pierre Fraigniaud, Emmanuelle Lebhar, and Laurent Viennot. The inframetric model for the internet. In *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, pages 1085–1093, 2008. doi: 10.1109/INFOCOM.2008.163.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data (extended abstract COLT 2010). *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014. doi: 10.1109/TIT.2014.2339840. URL http://dx.doi.org/10.1109/TIT.2014.2339840.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction (extended abstract: ALT 2013). *Theoretical Computer Science*, pages 105–118, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate Lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag New York, Inc., 2002.

László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151):1–25, 2021. URL http://jmlr.org/papers/v22/20-1081.html.

Steve Hanneke. Universally consistent online learning with arbitrarily dependent responses, preprint. 2021a.

Steve Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research*, 22(130):1–116, 2021b. URL http://jmlr.org/papers/v22/17-298.html.

Steve Hanneke and Aryeh Kontorovich. Stable sample compression schemes: New applications and an optimal SVM margin bound. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 697–721. PMLR, 2021. URL http://proceedings.mlr.press/v132/hanneke21a.html.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129 – 2150, 2021. doi: 10.1214/20-AOS2029. URL https://doi.org/10.1214/20-AOS2029.

Matthias Hein. Robust nonparametric regression with metric-space valued output. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 718–726, 2009.

Aryeh Kontorovich and Roi Weiss. A Bayes consistent 1-NN classifier. In *Artificial Intelligence and Statistics (AISTATS 2015)*, 2014.

Samory Kpotufe and Nakul Verma. Time-accuracy tradeoffs in kernel prediction: Controlling prediction quality. *J. Mach. Learn. Res.*, 18:44:1–44:29, 2017. URL http://jmlr.org/papers/v18/16-538.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization, 2009.

Emilie Morvant, Sokol Koço, and Liva Ralaivola. PAC-bayesian generalization bound on confusion matrix for multi-class classification. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/434.pdf.

Assaf Naor. Metric embeddings and Lipschitz extensions, 2015.

Assaf Naor and Scott Sheffield. Absolutely minimal Lipschitz extension of tree-valued mappings. *Mathematische Annalen*, 354(3):1049–1078, November 2012. ISSN 0025-5831. doi: 10.1007/s00208-011-0753-1.

James Newling and François Fleuret. A sub-quadratic exact medoid algorithm. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 185–193. PMLR, 2017. URL http://proceedings.mlr.press/v54/newling17a.html.

David Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

Mark J. Schervish. *Theory of statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1995. ISBN 0-387-94546-6. doi: 10.1007/978-1-4612-4250-5. URL https://doi.org/10.1007/978-1-4612-4250-5.

Christof Schötz. Strong laws of large numbers for generalizations of Fréchet mean sets, 2021.

Florian Steinke, Matthias Hein, and Bernhard Schölkopf. Nonparametric regression between general riemannian manifolds. *SIAM J. Imaging Sci.*, 3(3):527–563, 2010. doi: 10.1137/080744189. URL https://doi.org/10.1137/080744189.

Appendix A. Deferred results

A.1. Separable metric space $(\mathcal{Y}, \ell)$ with unbounded diameter

The extension from countable to separable $(\mathcal{Y}, \ell)$ — implemented by the final, subscript-free version of MedNet— is quite straightforward. The approximation arguments we invoke are standard, and hence we only give a sketch of the proof. In Section E, we give a countable discretization $\mathcal{Y}_\varepsilon \subseteq \mathcal{Y}$, with a corresponding discretized version $\bar{\mu}_\varepsilon$ of $\bar{\mu}$ and the induced Bayes-optimal risk R^*_ε on the discretized space. Theorem 20 guarantees that $R^*_\varepsilon \to R^*$ as $\varepsilon \to 0$.

As discussed in the Introduction, we assume access to an oracle that takes $\varepsilon > 0$ as input and returns a (necessarily at most countable, due to separability) ε-net $\mathcal{Y}_\varepsilon$ of $\mathcal{Y}$. Given this oracle, MedNet operates as follows. First, a sequence $\varepsilon_n \downarrow 0$ is fixed. For each $n \in \mathbb{N}$, the sample S_n is drawn and the ε-net $\mathcal{Y}_n := \mathcal{Y}_{\varepsilon_n}$ is constructed. Next, each label Y_i in S_n is projected onto $\mathcal{Y}_n$ — i.e., replaced by $Y'_i \in \mathcal{Y}_n$ that is closest to Y_i. The resulting modified sample S'_n is then fed into $\mathsf{MedNet}^{\aleph_0}$ with the additional arguments δ_n, b_n, L_n as in Theorem 11. The latter shows that almost surely, the the constructed predictor's risk minus $R^*_{\varepsilon_n}$ decays to zero.[4] This, coupled with Theorem 20, implies Theorem 1:

Theorem 12 *Let $(\mathcal{X}, \rho)$ and be $(\mathcal{Y}, \ell)$ separable metric spaces equipped with a product distribution $\bar{\mu}$ such that BIE holds for $\mathcal{Y}$. For any $\varepsilon_n \downarrow 0$, let $\mathcal{Y}_n$ be a sequence of ε_n-nets as above. Discretize each sample $S_n \sim \bar{\mu}^n$ to S'_n with labels in $\mathcal{Y}_n$, as above. Then there is a choice of $\delta_{n\in\mathbb{N}}$ and truncation schedules $b_{n\in\mathbb{N}}$, $L_{n\in\mathbb{N}}$ such that the sequence of hypotheses h_n computed by* $\mathsf{MedNet}^{\aleph_0}(S'_n, \delta_n, b_n, L_n)$ *is strongly Bayes consistent:* $\mathbb{P}[\lim_{n\to\infty} R(h_n) = R^*] = 1$.

[4] Formally, Theorem 11 proves convergence on a fixed label space $\mathcal{Y}$, but a standard diagonal argument lets us apply it to the sequence $\mathcal{Y}_n$ and conclude the aforementioned claim.

A.2. The $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ algorithm

Algorithm 3: $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$

```
Assumptions: Let $(\mathcal{X}, \rho)$ be a separable metric space, and $\mathcal{Y}$ a countable label space with a loss
             function $\ell$ such that $L := \|\mathcal{Y}\| < \infty$.
Input       : Sample $S_n = (X_i, Y_i)_{i \in [n]}$, confidence $\delta_n \in (0, 1)$, side-information size $b_n \in \mathbb{N}$
Output      : predictor $h : \mathcal{X} \to \mathcal{Y}$
Let $\Gamma := (\{\rho(X_i, X_j) : i, j \in [n]\} \cup \{\infty\}) \setminus \{0\}$
for $\gamma \in \Gamma$ do
    Let $\boldsymbol{X}(\gamma)$ be a $\gamma$-net of $\{X_1, \dots, X_n\}$
    Let $M_n(\gamma) := |\boldsymbol{X}(\gamma)|$
    For each $i \in [M_n(\gamma)]$, let $Y'_i(\gamma)$ be the truncated empirical medoid label of $V_i(\boldsymbol{X}(\gamma))$ as in
      (7.1)
    Set $S'_n(\gamma) := (\boldsymbol{X}(\gamma), \boldsymbol{Y}'(\gamma))$
    Set $h_{S'_n(\gamma)} := x \mapsto Y_{\mathsf{nn}}(x, S'_n(\gamma))$.
    Set $\alpha_n(\gamma) := \widehat{R}(h_{S'_n(\gamma)}; S_n)$
end
Find $\gamma^*_n \in \operatorname{argmin}_{\gamma \in \Gamma} Q_n(\alpha_n(\gamma), M_n(\gamma), b_n, \delta, L)$, where $Q_n$ is defined in (4.1)
Set $S'_n := S'_n(\gamma^*_n)$
return $h = h_{S'_n}$
```

Appendix B. Auxiliary Proofs

B.1. Proof of Lemma 7

Lemma *Let $(\mathcal{X}, \rho)$ be a separable metric space, and $\mathcal{Y}$ a* finite *label space with a loss function ℓ. For any fixed scale $\gamma \in \Gamma$, the procedure in Algorithm 1 generating $h_{S'_n(\gamma)}$ is a semi-stable compression scheme.*

Proof Fix a $\gamma \in \Gamma$. Define $b := \log_2 |\mathcal{Y}|$ and the map $\mathsf{bits} : \mathcal{Y} \to \{0,1\}^b$ as one that converts the lexicographic index of $y \in \mathcal{Y}$ to its unique b-bit binary representation. Our compression function:

$$(\mathcal{X} \times \mathcal{Y})^+ \to (\mathcal{X} \times \mathcal{Y})^+ \times \{0,1\}^b.$$

Recall our notation $\boldsymbol{i}_\gamma$ as the γ-net indices calculated and selected for a sample $S_n \sim (\mathcal{X} \times \mathcal{Y})^n$, and $\mathbf{Y}'(\boldsymbol{i}_\gamma)$ the empirical medoid labels of a γ-net $\boldsymbol{X}(\boldsymbol{i}_\gamma)$, as defined in (6.3). Let κ be such that $\kappa(S_n) = (\kappa_{\mathsf{cs}}(S_n), \kappa_{\mathsf{si}}(S_n))$, where

$$\kappa_{\mathsf{cs}}(S_n) = S_n(\boldsymbol{i}_\gamma) \in (\mathcal{X} \times \mathcal{Y})^{|\boldsymbol{i}_{\gamma^*}|}$$

$$\kappa_{\mathsf{si}}(S_n) = \left\{\mathsf{bits}(Y') : Y' \in \mathbf{Y}'(\boldsymbol{i}_\gamma)\right\} \in \left(\{0,1\}^b\right)^{|\boldsymbol{i}_{\gamma^*}|}.$$

In words, κ_{cs} compresses the sample S_n to a specific γ-net keeping original labels, while κ_{si} calculates the respective empirical limited medoid labels of the resulting sub-sample. As for the reconstruction function

$$\psi : (\mathcal{X} \times \mathcal{Y})^+ \times \{0,1\}^b \to \mathcal{Y}^{\mathcal{X}},$$

it is defined as

$$\psi\left(S_n(\boldsymbol{i}_\gamma), \mathbf{Y}'(\boldsymbol{i}_\gamma)\right) = h_{S_n(\boldsymbol{i}_\gamma, \mathbf{Y}'(\boldsymbol{i}_\gamma))};$$

in words, we take the 1-nearest-neighbor rule predictor of the sub-sample $S_n(\boldsymbol{i}_\gamma)$ labeled by $\mathbf{Y}'(\boldsymbol{i}_\gamma)$.

It remains to argue that our compression scheme is semi-stable. Indeed, since the scale γ is fixed and the net is constructed in a deterministic fashion, for any S' satisfying $\kappa_{\mathsf{cs}}(S_n) \subseteq S' \subseteq S_n$, the γ-net computed by the algorithm will be the same. Therefore $\kappa_{\mathsf{cs}}(S') = \kappa_{\mathsf{cs}}(S_n)$ and the definition follows. ■

B.2. Proof of Lemma 3

Lemma *Let $\mathcal{X}$, $\mathcal{Y}$, ℓ, L, and S_n be as in Theorem 2. For $k \le n$, define*

$$\begin{aligned} Q(n,\alpha,k,b,\delta,L) := Q_n(\alpha,k,b,\delta,L) := &\left(20\sqrt{\frac{k}{n}} + 20\sqrt{\frac{b}{n}} + 15\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} + 1\right)\alpha \\ &+ (6L+18)\frac{k}{n} + 8L\sqrt{\frac{k}{n}} + (2L+12)\frac{b}{n} + 7L\sqrt{\frac{b}{n}} \\ &+ (3L+10)\frac{\ln(\frac{4e^2}{\delta})}{n} + 6L\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}. \end{aligned}$$

Then the function Q satisfies the following properties:

Q1. *For any $n \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$ over $S_n \sim \bar{\mu}^n$, for all $\alpha \in [0,L]$, $k \in [n]$, $b \in \mathbb{N}$: If $(S'_n, h_{S'_n})$ is an (α,k,b)-semi-stable-compression of S_n, then*

$$R(h_{S'_n}) \le Q_n(\alpha,k,b,\delta,L).$$

Q2. *For any fixed $n \in \mathbb{N}$ and $\delta \in (0,1)$, Q is monotonically increasing in α and in k.*

Q3$''$. *There is a sequence $\{\delta_n\}_{n=1}^{\infty}$, $\delta_n \in (0,1)$ such that $\sum_{n=1}^{\infty}\delta_n < \infty$, and for any $k_n \in o(n)$ we have that*

$$\lim_{n\to\infty} \sup_{\alpha\in[0,L]} \left(Q_n(\alpha,k_n,b,\delta_n,L) - \alpha\right) = 0.$$

Proof Let $\mathcal{X}$ be an instance space, and $\mathcal{Y}$ a label space with a loss function ℓ such that $L := \|\mathcal{Y}\| < \infty$. Starting from **Q1**, let $(S'_n, h_{S'_n})$ be an (α,k,b)-semi-stable-compression of S_n. Thus, Q satisfies property **Q1** by Theorem 2.

Furthermore, property **Q2** (monotonicity in α and in k) can also be easily verified from the definition in (4.1).

To establish **Q3$''$**, an inspection of $Q_n(\alpha,k_n,b,\delta_n,L) - \alpha$ shows that since $k_n \in o(n)$, only the terms containing $\frac{\ln(\frac{4e^2}{\delta_n})}{n}$ are not *obviously* decaying to zero. To ensure the latter, any choice of δ_n with $-\log\delta_n \in o(n)$ suffices. The additional constraint δ_n must satisfy is $\sum_{n=1}^{\infty}\delta_n < \infty$; one such choice $\delta_n = e^{-\sqrt{n}}$. ■

B.3. Proof of Lemma 4

Lemma *Let $\bar{\mu}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$-marginal μ, where $\mathcal{X}$ is a metric probability space, and $\mathcal{Y}$ a* countable *label space with a loss function ℓ such that $L := \|\mathcal{Y}\| < \infty$. For any $\nu > 0$, there exists a diameter $\beta = \beta(\nu) > 0$ such that for any countable measurable partition $\mathcal{V} = \{V_1, \dots\}$ of $\mathcal{X}$ and any measurable set $W \subseteq \mathcal{X}$ satisfying*

(i) $\mu(\mathcal{X} \setminus W) \leq \nu$

(ii) $\sup_{V \in \mathcal{V}} \|V \cap W\| \leq \beta$,

the true medoid predictor $h^*_{\mathcal{V},W}$ defined in (5.2) satisfies

$$R(h^*_{\mathcal{V},W}) \leq R^* + 9L\nu.$$

Proof We begin similarly to the proof of Hanneke et al. (2021, Lemma 3.6). Let $\eta_y : \mathcal{X} \to [0,1]$ be the conditional probability function for label $y \in \mathcal{Y}$,

$$\eta_y(x) = \mathbb{P}(Y = y \,|\, X = x),$$

and let $\zeta_y : \mathcal{X} \to [0, L]$ be the expected loss function for the label $y \in \mathcal{Y}$:

$$\zeta_y(x) = \mathop{\mathbb{E}}_{\bar{\mu}}[\ell(y, Y) \,|\, X = x] = \int_{y' \in \mathcal{Y}} \ell(y, y')\eta_{y'}(x) \mathrm{d}\mu$$

which is measurable by Schervish (1995, Corollary B.22).

Define $\tilde{\eta}_y : \mathcal{X} \to [0,1]$ as η_y's conditional expectation function with respect to $(\mathcal{V}, W)$: For x such that $I_{\mathcal{V}}(x) \cap W \neq \emptyset$,

$$\tilde{\eta}_y(x) = \mathbb{P}(Y = y \,|\, X \in I_{\mathcal{V}}(x) \cap W) = \frac{\int_{I_{\mathcal{V}}(x) \cap W} \eta_y(z) \,\mathrm{d}\mu(z)}{\mu(I_{\mathcal{V}}(x) \cap W)}.$$

Otherwise, if $I_{\mathcal{V}}(x) \cap W = \emptyset$, define $\tilde{\eta}_y(x) = \mathbf{1}[y \text{ is lexicographically first}]$. Note that $(\tilde{\eta}_y)_{y \in \mathcal{Y}}$ are piecewise constant on the cells of the restricted partition $\mathcal{V} \cap W$. Accordingly, define $\tilde{\zeta}_y \to [0, L]$:

$$\tilde{\zeta}_y(x) = \mathop{\mathbb{E}}_{\bar{\mu}}[\ell(y, Y) \,|\, X \in I_{\mathcal{V}}(x) \cap W] = \sum_{y' \in \mathcal{Y}} \ell(y, y')\tilde{\eta}_{y'}(x).$$

In the proof of Hanneke et al., there is no appearance of ζ_y or $\tilde{\zeta}_y$; there, the much simpler conditional *error probabilities* suffice. Note likewise that their local majority vote classifier has been replaced here by the local medoid. By definition, the Bayes-optimal predictor h^* and the true medoid predictor $h^*_{\mathcal{V},W}$ satisfy

$$\begin{aligned} h^*(x) &= \operatorname*{argmin}_{y \in \mathcal{Y}} \zeta_y(x), \\ h^*_{\mathcal{V},W}(x) &= \operatorname*{argmin}_{y \in \mathcal{Y}} \tilde{\zeta}_y(x). \end{aligned}$$

It follows that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*_{\mathcal{V},W}(X),Y) \,|\, X=x\right] - \mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*(X),Y) \,|\, X=x\right] \\
&\qquad = \zeta_{h^*_{\mathcal{V},W}(x)}(x) - \zeta_{h^*(x)}(x) \\
&\qquad = \zeta_{h^*_{\mathcal{V},W}(x)}(x) - \tilde{\zeta}_{h^*_{\mathcal{V},W}(x)}(x) + \tilde{\zeta}_{h^*_{\mathcal{V},W}(x)}(x) - \tilde{\zeta}_{h^*(x)}(x) \\
&\qquad + \tilde{\zeta}_{h^*(x)}(x) - \zeta_{h^*(x)}(x) \\
&\qquad \le \zeta_{h^*_{\mathcal{V},W}(x)}(x) - \tilde{\zeta}_{h^*_{\mathcal{V},W}(x)}(x) + \tilde{\zeta}_{h^*(x)}(x) - \zeta_{h^*(x)}(x) \\
&\qquad \le 2 \max_{y' \in \left\{h^*_{\mathcal{V},W}(x), h^*(x)\right\}} \left|\zeta_{y'}(x) - \tilde{\zeta}_{y'}(x)\right| \\
&\qquad = 2 \max_{y' \in \left\{h^*_{\mathcal{V},W}(x), h^*(x)\right\}} \left|\sum_{y\in\mathcal{Y}} \ell(y',y)\left(\eta_y(x) - \tilde{\eta}_y(x)\right)\right| \\
&\qquad \le 2L \left|\sum_{y\in\mathcal{Y}} \left(\eta_y(x) - \tilde{\eta}_y(x)\right)\right| \\
&\qquad \le 2L \sum_{y\in\mathcal{Y}} \left|\eta_y(x) - \tilde{\eta}_y(x)\right|.
\end{aligned}$$

By condition (i) in the lemma statement, $\mu(\mathcal{X}\setminus W) \le \nu$. Thus,

$$\begin{aligned}
R(h^*_{\mathcal{V},W}) - R^* &= \mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*_{\mathcal{V},W}(X),Y)\right] - \mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*(X),Y)\right] \\
&= \int_{\mathcal{X}\setminus W} \left(\mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*_{\mathcal{V},W}(X),Y) \,|\, X=x\right] - \mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*(X),Y) \,|\, X=x\right]\right) \mathrm{d}\mu(x) \\
&\quad + \int_{W} \left(\mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*_{\mathcal{V},W}(X),Y) \,|\, X=x\right] - \mathop{\mathbb{E}}_{\bar{\mu}}\left[\ell(h^*(X),Y) \,|\, X=x\right]\right) \mathrm{d}\mu(x) \\
&\le L\cdot\mu(\mathcal{X}\setminus W) + 2L\int_W \sum_{y\in\mathcal{Y}} \left|\eta_y(x) - \tilde{\eta}_y(x)\right| \mathrm{d}\mu(x) \\
&\le L\nu + 2L\sum_{y\in\mathcal{Y}} \int_W \left|\eta_y(x) - \tilde{\eta}_y(x)\right| \mathrm{d}\mu(x).
\end{aligned}$$

Let $\mathcal{Y}_\nu \subseteq \mathcal{Y}$ be a finite set of labels such that $\mathbb{P}[Y \in \mathcal{Y}_\nu] \ge 1-\nu$. Then

$$\begin{aligned}
\sum_{y\notin\mathcal{Y}_\nu} \int_W \left|\eta_y(x) - \tilde{\eta}_y(x)\right| \mathrm{d}\mu(x) &\le \sum_{y\notin\mathcal{Y}_\nu} \int_W \eta_y(x)\mathrm{d}\mu(x) \\
&= \sum_{y\notin\mathcal{Y}_\nu} \int_W \mathbb{P}_{\bar{\mu}}(Y=y \,|\, X=x)\mathrm{d}\mu(x) \\
&= \sum_{y\notin\mathcal{Y}_\nu} \mathbb{P}_{\bar{\mu}}(Y=y, X\in W) \\
&\le \sum_{y\notin\mathcal{Y}_\nu} \mathbb{P}(Y=y) \\
&= \mathbb{P}(y\notin\mathcal{Y}_\nu) \le \nu.
\end{aligned}$$

We conclude:

$$R(h^*_{\mathcal{V},W}) - R^* \leq 3L\nu + 2L \sum_{y\in\mathcal{Y}_\nu} \int_W |\eta_y(x) - \tilde{\eta}_y(x)| \mathrm{d}\mu(x). \tag{B.1}$$

To bound the integrals in (B.1), we invoke a result from the proof of Lemma 3.6 in Hanneke et al. (2021), which showed that

$$\sum_{y\in\mathcal{Y}_\nu} \int_W |\eta_y(x) - \tilde{\eta}_y(x)| \,\mathrm{d}\mu(x) \leq \sum_{y\in\mathcal{Y}_\nu} \frac{3\nu}{|\mathcal{Y}_\nu|} = 3\nu.$$

Applying this bound to (B.1), we conclude $R(h^*_{\mathcal{V},W}) - R^* \leq 9L\nu$. ■

B.4. Proof of Lemma 8

Lemma *Let $\bar{\mu}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a metric probability space, and $\mathcal{Y}$ a* countable *label space endowed with a loss function $\ell \leq L < \infty$. Let t_γ as in Lemma 5. Then there exist functions $\varepsilon \mapsto \gamma(\varepsilon)$ and $\varepsilon \mapsto \gamma := \nu(\varepsilon) \in (0, \frac{\varepsilon}{176L})$ such that for each $\varepsilon, b > 0$ there is an $N_0(\nu(\varepsilon), b, \delta_n, t_\gamma)$ such that for all $n \geq N_0$, and all $d \in [t_\gamma(n)]$,*

$$p_d := \mathbb{P}\left[Q_n(\alpha_n(\gamma), M_n(\gamma), b, \delta_n, L) > R^* + \varepsilon \ \wedge\ \mathsf{mm}_\gamma(S_n) \leq \frac{\varepsilon}{18L} \ \wedge\ M_n(\gamma) = d\right] \leq e^{-\frac{n\varepsilon^2}{32}} + e^{-\frac{1}{2}n\nu^2},$$

where $M_n(\gamma)$ is defined in (6.2).

Proof We begin the proof similarly to Hanneke et al. (2021, Lemma 3.5) and then diverge in order to extend their 0-1 loss to the general loss setting. Let $\boldsymbol{i} = \boldsymbol{i}(\gamma) \in [n]^d$ be the set of indices in the net $\boldsymbol{X} = \boldsymbol{X}(\gamma)$ selected by the algorithm. Let $\boldsymbol{Y}^* \in \mathcal{Y}^d$ be the true medoid labels with respect to the restricted partition $\mathcal{V}(\boldsymbol{X}) \cap \mathrm{UB}_{2\gamma}(\boldsymbol{X})$,

$$(\boldsymbol{Y}^*)_j = y^*(V_j \cap \mathrm{UB}_{2\gamma}(\boldsymbol{X})), \qquad j \in [d]. \tag{B.2}$$

We pair $\boldsymbol{X}$ with the labels $\boldsymbol{Y}^*$ to obtain the labeled set

$$S_n(\boldsymbol{i}, *) := S_n(\boldsymbol{i}, \boldsymbol{Y}^*) = (\boldsymbol{X}, \boldsymbol{Y}^*) \in (\mathcal{X} \times \mathcal{Y})^d. \tag{B.3}$$

Note that conditioned on $\boldsymbol{X}$, $S_n(\boldsymbol{i}, *)$ does not depend on the rest of S_n.

The induced 1-NN predictor $h_{S_n(\boldsymbol{i},*)}(x)$ can be expressed as $h^*_{\mathcal{V},W}(x) = y^*(I_{\mathcal{V}}(x) \cap W)$ with $\mathcal{V} = \mathcal{V}(\boldsymbol{X})$ and $W = \mathrm{UB}_{2\gamma}(\boldsymbol{X})$ (see (5.2) for the definition of $h^*_{\mathcal{V},W}$). We now show that

$$\mathsf{mm}_\gamma(\boldsymbol{X}_n) \leq \frac{\varepsilon}{18L} \qquad \Longrightarrow \qquad R(h_{S_n(\boldsymbol{i},*)}) \leq R^* + \varepsilon/2, \tag{B.4}$$

by showing that under the assumption $\mathsf{mm}_\gamma(\boldsymbol{X}_n) \leq \frac{\varepsilon}{18L}$, the conditions of Lemma 4 hold for $\mathcal{V}, W$ as defined above. To this end, we bound the diameter of the partition $\mathcal{V} \cap W = \mathcal{V} \cap \mathrm{UB}_{2\gamma}(\boldsymbol{X})$, and the measure of the missing mass $\mu(\mathcal{X} \setminus W) = \mathsf{mm}_{2\gamma}(\boldsymbol{X})$ under the assumption.

To bound the diameter of the partition $\mathcal{V} \cap \mathrm{UB}_{2\gamma}(\boldsymbol{X})$, let $x \in V_j \cap \mathrm{UB}_{2\gamma}(\boldsymbol{X})$. Note that V_j is the Voronoi cell centered at $x_{i_j} \in \boldsymbol{X}$. Then $\rho(x, x_{i_j}) = \min_{i\in\boldsymbol{i}} \rho(x, x_i)$ and, since $x \in \mathrm{UB}_{2\gamma}(\boldsymbol{X})$, $\min_{i\in\boldsymbol{i}} \rho(x, x_i) \leq 2\gamma$. Thus,

$$\|\mathcal{V} \cap W\| = \max_j \|V_j \cap \mathrm{UB}_{2\gamma}(\boldsymbol{X})\| \leq 4\gamma.$$

To bound $\mathsf{mm}_{2\gamma}(\boldsymbol{X})$ under the assumption $\mathsf{mm}_{\gamma}(\boldsymbol{X}_n) \leq \frac{\varepsilon}{18L}$, observe that for all $z \in \mathrm{UB}_{\gamma}(\boldsymbol{X}_n)$, there is some $i \in [n]$ such that $z \in B_{\gamma}(x_i)$. For this i, there is some $j \in \boldsymbol{i}$ such that $x_i \in B_{\gamma}(x_j)$, since $\boldsymbol{X}$ is a γ-net of $\boldsymbol{X}_n$. Therefore $z \in B_{2\gamma}(x_j)$. Thus, $z \in \mathrm{UB}_{2\gamma}(\boldsymbol{X})$. It follows that $\mathrm{UB}_{\gamma}(\boldsymbol{X}_n) \subseteq \mathrm{UB}_{2\gamma}(\boldsymbol{X})$, thus $\mathsf{mm}_{2\gamma}(\boldsymbol{X}) \leq \mathsf{mm}_{\gamma}(\boldsymbol{X}_n)$. Under the assumption, we thus have $\mathsf{mm}_{2\gamma}(\boldsymbol{X}) \leq \frac{\varepsilon}{18L}$. Hence, by the choice of $\gamma = \gamma(\varepsilon)$ in the statement of the lemma, Lemma 4 implies (B.4).

To bound $Q_n(\alpha_n(\gamma), M_n(\gamma))$, we consider the relationship between the hypothetical true medoid predictor $h_{S_n(\boldsymbol{i},*)}$ and the actual predictor returned by the algorithm, $h_{S_n(\boldsymbol{i},\boldsymbol{Y'})}$. Firstly, for any $\nu \in (0,1)$, there exists a finite $\mathcal{Y}'_\nu \subseteq \mathcal{Y}$ such that $\mathbb{P}[Y \in \mathcal{Y}'_\nu] \geq 1-\nu$. Therefore, since b is non-decreasing in n, there exists an $N_1(\nu)$ large enough such that for any $n \geq N_1(\nu)$ $\mathcal{Y}'_\nu \subseteq \mathsf{pref}(\mathcal{Y}, b)$. Fix such a ν specifically such that $\nu < \frac{\varepsilon}{176L}$. Thus we have that $\mathbb{P}[Y \in \mathsf{pref}(\mathcal{Y}, b)] \geq 1-\nu$.

For brevity we denote

$$Q_n(\alpha, k) := Q_n(\alpha, k, b, \delta_n, L).$$

Let us split our cases:

$$\begin{aligned} p_d &= \mathbb{P}\big[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon \ \wedge \ \mathsf{mm}_\gamma(\boldsymbol{X}_n) \leq \frac{\varepsilon}{18L} \ \wedge \ M_n(\gamma) = d \\ &\qquad \wedge \ \widehat{R}(h_{S_n(\boldsymbol{i},\boldsymbol{Y'})}; S_n) \leq \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n) + 2\nu L\big] + \\ &\quad + \ \mathbb{P}\big[Q_n(\alpha_n(\gamma), M_n(\gamma)) > R^* + \varepsilon \ \wedge \ \mathsf{mm}_\gamma(\boldsymbol{X}_n) \leq \frac{\varepsilon}{18L} \ \wedge \ M_n(\gamma) = d \\ &\qquad \wedge \ \widehat{R}(h_{S_n(\boldsymbol{i},\boldsymbol{Y'})}; S_n) > \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n) + 2\nu L\big] \\ &:= (p_d)_1 + (p_d)_2. \end{aligned} \tag{B.5}$$

Now, let $\mathbf{Y}^*$ be the best medoids possible for the sample, from the entire label space $\mathcal{Y}$:

$$\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n) = \min_{\mathbf{Y} \in \mathcal{Y}^d} \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y})}; S_n).$$

This means:

$$\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n) \leq \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')}; S_n) \qquad \text{and} \qquad \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n) \leq \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n).$$

Specifically, we note that since

$$\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')}; S_n) = \min_{\mathbf{Y} \in \mathsf{pref}(\mathcal{Y}, b_n)^d} \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y})}; S_n),$$

we have that

$$\begin{aligned} \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')}; S_n) - \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n) &= \frac{1}{n} \sum_{(X,Y) \in S_n, Y \notin \mathsf{pref}(\mathcal{Y}, b_n)} \ell(h_{S_n(\boldsymbol{i},\mathbf{Y}')}(X), Y) - \ell(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}(X), Y) \\ &\leq \frac{1}{n} L \left|\{(X,Y) \in S_n, Y \notin \mathsf{pref}(\mathcal{Y}, b_n)\}\right| \end{aligned}$$

$$\Rightarrow \mathbb{E}\left[\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')}; S_n) - \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n)\right] \leq \frac{1}{n} L \cdot \nu n = \nu L.$$

Hence, we observe:

$$\mathbb{E}\left[\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')}; S_n) - \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n)\right] \leq \nu L \qquad \text{and} \qquad \mathbb{E}\left[\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}^*)}; S_n) - \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n)\right] \leq 0,$$

whence

$$\mathbb{E}\left[\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')};S_n)-\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n)\right]\le \nu L.$$

Next, Hoeffding's inequality implies that for any $t>0$:

$$\mathbb{P}\left(\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')};S_n)-\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n)>\nu L+t\right)<\exp\left(-\frac{nt^2}{2L^2}\right).$$

Taking $t=\nu L$ we get:

$$(p_d)_2\le \mathbb{P}\left(\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')};S_n)-\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n)>2\nu L\right)<\exp\left(-\frac{1}{2}n\nu^2\right).$$

Next, we examine the case $\widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')};S_n)-\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n)\le 2\nu L$, and use the monotonicity Property **Q2** of Q:

$$\begin{aligned}
Q_n(\alpha_n(\gamma),M_n(\gamma)) &\le Q_n\left(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n)+2\nu L,M_n(\gamma)\right)\\
&\le Q_n\left(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),M_n(\gamma)\right)+2\nu L\left(20\sqrt{\frac{M_n(\gamma)}{n}}+20\sqrt{\frac{b}{n}}+15\sqrt{\frac{\ln(\frac{4e^2}{\delta_n})}{n}}+1\right)\\
&\le Q_n\left(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),M_n(\gamma)\right)+2\nu L\left(21+20\sqrt{\frac{b}{n}}+15\sqrt{\frac{\ln(\frac{4e^2}{\delta_n})}{n}}\right)\\
&:= Q_n\left(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),M_n(\gamma)\right)+2\nu L\cdot F_n(b,\delta_n).
\end{aligned}$$

Examining, $F_n(b,\delta_n)$ we note (as shown in the proof of Lemma 3) that for n sufficiently large (larger than some $N_2(b,\delta_n)$), we have $F_n(b,\delta_n)\le 22$. Thus,

$$Q_n(\alpha_n(\gamma),M_n(\gamma))\le Q_n\left(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),M_n(\gamma)\right)+44\nu L. \tag{B.6}$$

Combining (B.4) and (B.6),

$$\begin{aligned}
&\Big\{Q_n(\alpha_n(\gamma),M_n(\gamma))>R^*+\varepsilon \;\wedge\; \mathsf{mm}_\gamma(\boldsymbol{X}_n)\le\frac{\varepsilon}{18L} \;\wedge\; M_n(\gamma)=d\\
&\quad\wedge\; \widehat{R}(h_{S_n(\boldsymbol{i},\mathbf{Y}')};S_n)\le \widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n)\Big\}\\
&\qquad\Longrightarrow\; \left\{Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),d)+44\nu L>R(h_{S_n(\boldsymbol{i},*)})+\frac{\varepsilon}{2}\;\wedge\;|\boldsymbol{i}|=d\right\}.
\end{aligned}$$

Thus, for all $d\le t_\gamma$,

$$\begin{aligned}
(p_d)_1 &\le \mathbb{P}\left[Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),d)>R(h_{S_n(\boldsymbol{i},*)})+\frac{\varepsilon}{2}-44\nu L\;\wedge\;|\boldsymbol{i}|=d\right]\\
&\le \mathbb{P}\Big[\exists\boldsymbol{i}\in[n]^d: Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)};S_n),d)\\
&\qquad\qquad > R(h_{S_n(\boldsymbol{i},*)})+\frac{\varepsilon}{2}-44\nu L\Big].
\end{aligned}\tag{B.7}$$

To bound the last expression, let $\boldsymbol{i} \in [n]^d$ and denote

$$r_{d,n} = \sup_{\alpha \in (0,L)} (Q_n(\alpha, d) - \alpha).$$

We therefore have

$$Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n), d) \leq \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n) + r_{d,n}.$$

Let $\boldsymbol{i}' = \{1, \ldots, n\} \setminus \boldsymbol{i}$ and note that

$$\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n) \leq \frac{n-d}{n} \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n(\boldsymbol{i}')) + \frac{d}{n}.$$

Combining the two inequalities above, we get

$$Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n), d) \leq \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n(\boldsymbol{i}')) + \frac{d}{n} + r_{d,n}.$$

Recalling $t_\gamma \in o(n)$, by Property **Q3**$''$,

$$\lim_{n \to \infty} \frac{t_\gamma}{n} + r_{t_\gamma,n} = 0.$$

In addition, by **Q2**, we have $r_{d,n} \leq r_{t_\gamma,n}$ for all $d \leq t_\gamma$. Hence, using our $\nu < \frac{\varepsilon}{176L}$, we take n sufficiently large (larger than some $N_3(t_\gamma)$), so that for all $d \leq t_\gamma$,

$$\frac{d}{n} + r_{d,n} \leq \frac{t_\gamma}{n} + r_{t_\gamma,n} \leq \frac{\varepsilon}{4} - 44\nu L,$$

and thus

$$Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n), d) \leq \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n(\boldsymbol{i}')) + \frac{\varepsilon}{4} - 44\nu L.$$

Therefore, for any $n \geq N_0(\nu, b, \delta_n, t_\gamma) := \max\{N_1, N_2, N_3\}$,

$$\begin{aligned} & Q_n(\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n), d) > R(h_{S_n(\boldsymbol{i},*)}) + \frac{\varepsilon}{2} - 44\nu L \\ \Longrightarrow \quad & \widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n(\boldsymbol{i}')) > R(h_{S_n(\boldsymbol{i},*)}) + \frac{\varepsilon}{4}. \end{aligned}$$

Now,

$$\begin{aligned} & \mathbb{P}\left[\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n(\boldsymbol{i}')) > R(h_{S_n(\boldsymbol{i},*)}) + \frac{\varepsilon}{4}\right] \qquad \text{(B.8)} \\ = & \mathop{\mathbb{E}}_{S_n(\boldsymbol{i})}\left[\mathbb{P}_{S_n(\boldsymbol{i}') \mid S_n(\boldsymbol{i})}\left[\widehat{R}(h_{S_n(\boldsymbol{i},*)}; S_n(\boldsymbol{i}')) > R(h_{S_n(\boldsymbol{i},*)}) + \frac{\varepsilon}{4}\right]\right]. \end{aligned}$$

Since $\mathbb{P}_{S_n(\boldsymbol{i}') \mid S_n(\boldsymbol{i})}$ is a product distribution, by Hoeffding's inequality we have that (B.8) is bounded above by $e^{-2(n-d)(\frac{\varepsilon}{4})^2}$. Since $h_{S_n(\boldsymbol{i},*)}$ is invariant to permutations of $\boldsymbol{i}$'s entries, bounding (B.7) by a union bound over $\boldsymbol{i}$ yields

$$(p_d)_1 \leq \binom{n}{d} e^{-2(n-d)(\frac{\varepsilon}{4})^2} \leq e^{d \log(\frac{en}{d}) - 2(n-d)(\frac{\varepsilon}{4})^2},$$

where we used $\binom{n}{d} \leq \left(\frac{en}{d}\right)^d$. Selecting n large enough so that for all $d \leq t_\gamma$ we have $d \log(en/d) \leq (n-d)(\frac{\varepsilon}{4})^2$ and $d \leq n/4$. Combining this with (B.7) proves the lemma. ■

B.5. Proof of Lemma 5

Lemma *Let $(\mathcal{X}, \rho, \mu)$ be a separable metric probability space. For $S_n \sim \mu^n$, let $\boldsymbol{X}(\gamma)$ be any γ-net of S_n. Then, for any $\gamma > 0$, there exists a function $t_\gamma : \mathbb{N} \to \mathbb{R}_+$ in $o(n)$ such that* $\mathbb{P}\left[\sup_{\gamma\text{-nets } \boldsymbol{X}(\gamma)} |\boldsymbol{X}(\gamma)| \geq t_\gamma(n)\right] \leq 1/n^2$.

Proof Almost identical to Hanneke et al. (2021, Lemma 3.7) — the former has a factor of 2 multiplying $|\boldsymbol{X}(\gamma)|$ — and hence omitted. ■

B.6. Proof of Theorem 10

Theorem *Let $(\mathcal{X}, \rho)$ be a separable metric space, and $\mathcal{Y}$ a* countable *label space with a loss function $\ell \leq L < \infty$. Then there is a choice of $\delta_{n\in\mathbb{N}}$ and truncation schedule $b_{n\in\mathbb{N}}$ such that the sequence of hypotheses h_n computed by $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}(S_n, \delta_n, b_n)$ is strongly Bayes consistent:* $\mathbb{P}[\lim_{n\to\infty} R(h_n) = R^*] = 1$.

Proof The claim will follow if we succeed in showing how the lemmas and theorems invoked in proving Bayes consistency of $\mathsf{MedNet}_{|\mathcal{Y}|<\infty}$ (Theorem 9) are applicable in the present setting.

In Lemma 7, which established that the procedure is a semi-stable compression scheme a globally constant $b = \log_2 |\mathcal{Y}|$ was used. The claim remains perfectly true if the size of the label space happens to depend on the sample size, which is precisely how the result is being invoked in the present setting.

Next, we argue that the Q_n bound in Lemma 3 remains valid for sufficiently slowly growing b_n, and our chosen rate of $o(n)$ suffices. The remaining lemmas 4 and 8 can be used freely since they allow for countable label spaces. Lemmas 5 and 6 likewise do not require any modifications. Thus, we have shown how a straightforward modification of the proof of Theorem 9 also proves the theorem in question. ■

B.7. Proof of Theorem 11

Theorem *Let $(\mathcal{X}, \rho)$ and $(\mathcal{Y}, \ell)$ be metric spaces, separable and countable, respectively, equipped with a product distribution $\bar{\mu}$ such that BIE holds for $\mathcal{Y}$. Then there is a choice of $\delta_{n\in\mathbb{N}}$ and truncation schedules $b_{n\in\mathbb{N}}$, $L_{n\in\mathbb{N}}$ such that the sequence of hypotheses h_n computed by $\mathsf{MedNet}^{\aleph_0}(S_n, \delta_n, b_n, L_n)$ is strongly Bayes consistent:* $\mathbb{P}[\lim_{n\to\infty} R(h_n) = R^*] = 1$.

Proof Let Q be the generalization bound as defined in Lemma 3, and set the input confidence δ for input size n to δ_n as stipulated by **Q3**$''$. Choose any $b_n \in o(n)$, and L_n such that $L_n^2 k_n, L_n^2 b_n, L_n^2 \ln(\frac{4e^2}{\delta_n}) \in o(n)$. Similarly, we choose $k_n \in o(n)$ arbitrarily (cf. Lemma 3). Given a sample $S_n \sim \bar{\mu}^n$, we abbreviate the optimal empirical error $[\![\alpha_n^*]\!] = \alpha(\gamma_n^*)$ and the optimal compression size $[\![M_n^*]\!] = M(\gamma_n^*)$ as computed by Algorithm 1 on our truncated sample. Let $[\![h_n]\!]$ be the output of $\mathsf{MedNet}^{\aleph_0}$,

$$[\![f_n^*]\!] := \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{Y}\wedge L_n} R(f),$$

be the Bayes-optimal predictor for the truncated label space and

$$[\![R_n^*]\!] := R([\![f_n^*]\!]).$$

be the optimal risk in the *truncated setting* on our modified sample. For brevity we denote

$$Q_n(\alpha, k) := Q_n(\alpha, k, b, \delta_n, L_n).$$

To prove Theorem 11, we first follow the standard technique, used also in Theorem 9, of decomposing the excess error over the Bayes error into two terms:

$$\begin{aligned} R([\![h_n]\!]) - [\![R(f_n^*)]\!] &= \big(R([\![h_n]\!]) - Q_n(\alpha_n^*, M_n^*)\big) + \big(Q_n(\alpha_n^*, M_n^*) - [\![R(f_n^*)]\!]\big) \\ &=: T_{\mathrm{I}}(n) + T_{\mathrm{II}}(n). \end{aligned}$$

We now show that each term decays to zero almost surely. Regarding, $T_{\mathrm{I}}(n)$ we wish to invoke Lemmas 7 and 3 as in the proof of Theorem 9. The argument of Lemma 7, which shows that $\mathsf{MedNet}^{\aleph_0}_{\|\mathcal{Y}\|<\infty}$ is a semi-stable compression scheme, applies verbatim to $\mathsf{MedNet}^{\aleph_0}$.

As for Lemma 3, properties **Q1** and **Q2** trivially continue to hold, while to ensure **Q3′′**, we constrain the choice of the diameter truncation schedule L_n to satisfy $L_n^2 k_n, L_n^2 b_n, L_n^2 \ln(\frac{4e^2}{\delta_n}) \in o(n)$, where $k_n \in o(n)$ is as in Lemma 3. Having verified the applicability of Lemmas 7 and 3, we invoke property **Q1** from Lemma 3:

$$\mathbb{P}[R([\![h_n]\!]) - Q_n(\alpha_n^*, M_n^*) > 0] \le \delta_n, \qquad n \ge 1.$$

Since δ_n was chosen as furnished by Lemma 3(**Q3′′**), the Borel-Cantelli lemma implies that $\limsup_{n\to\infty} T_{\mathrm{I}}(n) \le 0$ with probability 1.

We now proceed to argue that the generalization bound $Q_n(\alpha_n^*, M_n^*)$ approaches the truncated optimal Bayes error $[\![R_n^*]\!]$, which will establish $\limsup_{n\to\infty} T_{\mathrm{II}}(n) \le 0$ almost surely. Since Lemmas 4, 8, 5 and 6 do not rely on L_n being fixed, the are applicable to our setting. Thus, the argument from the proof of Theorem 9 applies here as well, and thus $\limsup_{n\to\infty} T_{\mathrm{II}}(n) \le 0$ almost surely. It follows that $\lim_{n\to\infty} R([\![h_n]\!]) - [\![R_n^*]\!] = 0$. It remains to exploit the BIE property of $\mathcal{Y}$ and invoke Theorem 19 to conclude that $\lim_{n\to\infty} [\![R(f_n^*)]\!] - R^* = 0$, whence

$$\lim_{n\to\infty} R([\![h_n]\!]) - R^* = 0.$$

■

Appendix C. Compression Scheme Theorems

In this section we introduce a series of new theorems regarding semi-stable compression schemes, each leading to the next, culminating in Theorem 2 which is used to bound the generalization of MedNet.

C.1. Definitions

C.2. Setting

Let $\mathcal{X}$ be an instance space, and $\mathcal{Y}$ of finite diameter, and $\bar{\mu}$ a distribution supported on the product Borel σ-algebra of $\mathcal{X} \times \mathcal{Y}$, such that $\mathcal{Y}$ is bounded by some $L > 0$: $\forall y_1, y_2 \in \mathcal{Y} : \ell(y_1, y_2) \le L$.

C.3. Theorems & Lemmas

C.3.1. SEMI-STABLE AGNOSTIC SAMPLE COMPRESSION SCHEME OF GIVEN SIZE

Theorem 13 *For any $k, b \in \mathbb{N} \cup \{0\}$, let (κ, ψ) be any semi-stable compression scheme of size at most k using at most b bits of side-information. For any distribution $\bar{\mu}$ over $\mathcal{X} \times \mathcal{Y}$, any $n \in \mathbb{N}$ with $n > 2k$, and any $\delta \in (0,1)$, for $S_n \sim \bar{\mu}^n$, with probability at least $1-\delta$*

$$\left|R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n)\right| \leq \sqrt{\frac{4L^2}{n-2k}\left(k\ln(4) + \ln(4/\delta)\right)} + \sqrt{\frac{L^2}{n-2k} b\ln(2)}.$$

Proof If $k = 0$, the result trivially follows from Hoeffding's inequality, so let us suppose $k \geq 1$. As in the proof of Theorem 5 in Bousquet et al. (2020), fix any $T_n \in [n-1]$ and let $\mathcal{I}_n$ be any family of subsets of $[n]$ with the properties that each $I \in \mathcal{I}_n$ has $|I| \leq n - T_n$, and for every $i_1, \ldots, i_k \in [n]$ there exists $I \in \mathcal{I}_n$ such that $\{i_1, \ldots, i_k\} \subseteq I$. In particular, Bousquet et al. construct a family $\mathcal{I}_n$ satisfying the properties above with $T_n = k\lfloor \frac{n}{2k} \rfloor$, and with $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$: namely, let $D_1, \ldots, D_{2k}$ be any partition of $[n]$ with each $|D_i| \in \{\lfloor \frac{n}{2k} \rfloor, \lceil \frac{n}{2k} \rceil)\}$, and define $\mathcal{I}_n = \{\bigcup\{D_j : j \in \mathcal{J}\} : \mathcal{J} \subseteq [2k], |\mathcal{J}| = k\}$; that is, $\mathcal{I}_n$ contains all unions of exactly k of the $2k$ sets D_j.

Let there be a sample $S_n \sim \bar{\mu}^n$. Recall our notation that for any $I \in \mathcal{I}_n$ we have $S(I) \subseteq S_n$. For any $I \in \mathcal{I}_n$, define $\bar{I} := [n] \setminus I$. Let $\sigma : [n] \to [n]$ be a uniformly random permutation of $[n]$, and for any $I = (i_1, \ldots, i_\ell) \subseteq [n]$ define $\boldsymbol{\sigma}(I) := (\sigma(i_1), \ldots, \sigma(i_\ell))$.

Next, for any $I \subseteq [n]$ and any $\boldsymbol{b} \in \{0,1\}^b$, let $\hat{h}_{I,\boldsymbol{b}} := \psi\left(\kappa_{\mathsf{cs}}\left(S\left(I\right)\right), \boldsymbol{b}\right)$. Now, since $S(\bar{I})$ is independent of $S(I)$, Hoeffding's Inequality (applied under the conditional distribution given $S(I)$) and the law of total probability imply that with probability $1 - \frac{\delta}{2|\mathcal{I}_n| \cdot 2^b}$:

$$\begin{aligned}\left|R\left(\hat{h}_{I,\boldsymbol{b}}\right) - \widehat{R}\left(\hat{h}_{I,\boldsymbol{b}}; S\left(\bar{I}\right)\right)\right| &\leq \sqrt{\frac{L^2 \ln(\frac{4|\mathcal{I}_n| \cdot 2^b}{\delta})}{2(n-|I|)}} \\ &= \sqrt{\frac{L^2\left(\ln(|\mathcal{I}_n|) + b\ln(2) + \ln(4/\delta)\right)}{2(n-|I|)}}.\end{aligned}$$

Applying this under the conditional distribution given σ, together with the union bound and the law of total probability, we have that with probability at least $1 - \frac{\delta}{2}$, every $I \in \mathcal{I}_n$ and every $\boldsymbol{b} \in \{0,1\}^b$ has

$$\left|R\left(\hat{h}_{\boldsymbol{\sigma}^{-1}(I),\boldsymbol{b}}\right) - \widehat{R}\left(\hat{h}_{\boldsymbol{\sigma}^{-1}(I),\boldsymbol{b}}; S\left(\overline{\boldsymbol{\sigma}^{-1}\left(I\right)}\right)\right)\right| \leq \sqrt{\frac{L^2\left(\ln(|\mathcal{I}_n|) + b\ln(2) + \ln(4/\delta)\right)}{2(n-|I|)}}.$$

In particular, let $\boldsymbol{i}^*$ be the indices such that

$$\kappa_{\mathsf{cs}}(S_n) = \{(X_i, Y_i) \in S_n : i \in \boldsymbol{i}^*\} = S_n(\boldsymbol{i}^*).$$

Now, by property (ii) of $\mathcal{I}_n$ there must exist $I^* \in \mathcal{I}_n$ such that

$$\boldsymbol{\sigma}(\boldsymbol{i}^*) \subseteq I^*,$$

which means:

$$\begin{aligned}&\Rightarrow \quad \boldsymbol{i}^* \subseteq \boldsymbol{\sigma}^{-1}(I^*)\\ &\Rightarrow \quad \kappa_{\mathsf{cs}}(S_n) = S_n(\boldsymbol{i}^*) \subseteq S_n(\boldsymbol{\sigma}^{-1}(I^*)).\end{aligned}$$

Due to the semi-stability property of (κ, ψ) and since $S_n(\boldsymbol{\sigma}^{-1}(I^*)) \subseteq S_n$, this implies

$$\Rightarrow \psi(\kappa_{\mathsf{cs}}(S_n(\boldsymbol{\sigma}^{-1}(I^*))), \kappa_{\mathsf{si}}(S_n)) = \psi(\kappa_{\mathsf{cs}}(S_n), \kappa_{\mathsf{si}}(S_n)) = \psi(\kappa\,(S_n)).$$

Thus, on the above event of probability at least $1 - \frac{\delta}{2}$ we get, for $I = I^*$ and $\boldsymbol{b} = \kappa_{\mathsf{si}}(S_n)$,

$$\left|R(\psi(\kappa\,(S_n))) - \widehat{R}(\psi(\kappa\,(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)}))\right| \le \sqrt{\frac{L^2\,(\ln(|\mathcal{I}_n|) + b\ln(2) + \ln(4/\delta))}{2(n - |I^*|)}}.$$

Furthermore, by property (i) of I_n we have that $n - |I^*| \ge T_n$, and so

$$\left|R(\psi(\kappa\,(S_n))) - \widehat{R}(\psi(\kappa\,(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)}))\right| \le \sqrt{\frac{L^2\,(\ln(|\mathcal{I}_n|) + b\ln(2) + \ln(4/\delta))}{2T_n}}. \tag{C.1}$$

Next, we want to relate $\widehat{R}(\psi(\kappa\,(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)}))$ to $\widehat{R}(\psi(\kappa\,(S_n)); S_n)$. Let $\hat{h} := \psi(\kappa\,(S_n))$. For each $i \in [n]$, let $\ell_i := \ell(\hat{h}(X_i), Y_i)$. For any $I \in \mathcal{I}_n$, by Hoeffding's inequality without replacement (Bardenet and Maillard, 2015) applied under the conditional distribution given S_n, together with the law of total probability, with probability at least $1 - \frac{\delta}{2|\mathcal{I}_n|}$ it holds that

$$\left|\frac{1}{n - |I|}\sum_{i \in \overline{\boldsymbol{\sigma}^{-1}(I)}} \ell_i - \widehat{R}(\psi(\kappa\,(S_n)); S_n)\right| \le \sqrt{\frac{L^2\ln(4|\mathcal{I}_n|/\delta)}{2(n - |I|)}}.$$

By the union bound, this holds simultaneously for all $I \in \mathcal{I}_n$ with probability at least $1 - \frac{\delta}{2}$. In particular, taking $I = I^*$, and recalling that $n - |I^*| \ge T_n$, on this event we have that

$$\left|\widehat{R}(\psi(\kappa\,(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)})) - \widehat{R}(\psi(\kappa\,(S_n)); S_n)\right| \le \sqrt{\frac{L^2\ln(4|\mathcal{I}_n|/\delta)}{2T_n}} \tag{C.2}$$

By the union bound, the above two events (each of probability at least $1 - \frac{\delta}{2}$) hold simultaneously with probability at least $1 - \delta$, in which case (C.1) and (C.2) together imply

$$\begin{aligned}\left|R(\psi(\kappa\,(S_n))) - \widehat{R}(\psi(\kappa\,(S_n)); S_n)\right| &\le \left|R(\psi(\kappa\,(S_n))) - \widehat{R}(\psi(\kappa\,(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)}))\right| +\\ &\quad + \left|\widehat{R}(\psi(\kappa\,(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)})) - \widehat{R}(\psi(\kappa\,(S_n)); S_n)\right|\\ &\le \sqrt{\frac{L^2\,(\ln(|\mathcal{I}_n|) + b\ln(2) + \ln(4/\delta))}{2T_n}} + \sqrt{\frac{L^2\ln(4|\mathcal{I}_n|/\delta)}{2T_n}}\\ &\le \sqrt{\frac{4L^2\,(\ln(|\mathcal{I}_n|) + \ln(4/\delta))}{2T_n}} + \sqrt{\frac{L^2 b\ln(2)}{2T_n}}\end{aligned}$$

The theorem now immediately follows from plugging the aforementioned family $\mathcal{I}_n$ from Bousquet et al. (2020), having $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$ and $T_n = k\left\lfloor\frac{n}{2k}\right\rfloor > \frac{n-2k}{2}$. ■

C.3.2. ACCOUNTING FOR REALIZABLE CASE

Lemma 14 *Let $Z_1, \dots, Z_n$ be i.i.d random variables with values in $[0, L]$ for some $L > 0$, and let $\delta > 0$. Define $\bar{Z} := \frac{1}{n}\sum_{i=1}^n Z_i$. Then with probability at least $1 - \delta$ we have:*

$$\mathbb{E}[\bar{Z}] - \bar{Z} \leq \bar{Z} \cdot \sqrt{\frac{2\ln(4/\delta)}{n-1}} + L\sqrt{\frac{2\ln(4/\delta)}{n-1}} + \frac{7L\ln(4/\delta)}{3(n-1)}.$$

Proof Let $Z'_i := \frac{Z_i}{L} \in [0, 1]$ for all $i \in \{1, \dots, n\}$. Using the empirical Bernstein inequality stated in Maurer and Pontil (2009, Theorem 4), we get

$$\begin{aligned} \mathbb{E}[\bar{Z}'] - \bar{Z}' &\leq \sqrt{\frac{2\widehat{\mathrm{Var}}(\boldsymbol{Z'})\ln(4/\delta)}{n}} + \frac{7\ln(4/\delta)}{3(n-1)} \\ \Rightarrow \mathbb{E}[\bar{Z}] - \bar{Z} &\leq L\sqrt{\frac{2\widehat{\mathrm{Var}}(\boldsymbol{Z})\ln(4/\delta)}{L^2 n}} + \frac{7L\ln(4/\delta)}{3(n-1)} \qquad \text{(C.3)} \\ &= \sqrt{\frac{2\widehat{\mathrm{Var}}(\boldsymbol{Z})\ln(4/\delta)}{n}} + \frac{7L\ln(4/\delta)}{3(n-1)}, \end{aligned}$$

where $\widehat{\mathrm{Var}}(\boldsymbol{Z})$ is defined to be:

$$\widehat{\mathrm{Var}}(\boldsymbol{Z}) := \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (Z_i - Z_j)^2.$$

Observe that

$$\begin{aligned} \widehat{\mathrm{Var}}(\boldsymbol{Z}) &= \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (Z_i - Z_j)^2 \\ &= \frac{1}{n(n-1)} \cdot n \sum_{i=1}^n (Z_i - \bar{Z})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n Z_i^2 - 2\bar{Z} \sum_{i=1}^n Z_i + n\bar{Z}^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n Z_i^2 - 2n\bar{Z}^2 + n\bar{Z}^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n Z_i^2 - n\bar{Z}^2 \right) \\ &\leq \frac{1}{n-1} \left(nL^2 - n\bar{Z}^2 \right) \\ &= \frac{n}{n-1} \left(L^2 - \bar{Z}^2 \right). \end{aligned}$$

Now plugging this back into (C.3):

$$\begin{aligned}
\mathbb{E}[\bar{Z}] - \bar{Z} &\leq \sqrt{\frac{2(L^2 - \bar{Z}^2)\ln(4/\delta)}{n-1}} + \frac{7L\ln(4/\delta)}{3(n-1)} \\
&\leq \sqrt{\frac{2(L^2 + \bar{Z}^2)\ln(4/\delta)}{n-1}} + \frac{7L\ln(4/\delta)}{3(n-1)} \\
&\leq \bar{Z}\sqrt{\frac{2\ln(4/\delta)}{n-1}} + L\sqrt{\frac{2\ln(4/\delta)}{n-1}} + \frac{7L\ln(4/\delta)}{3(n-1)},
\end{aligned}$$

where the last inequality used $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y > 0$. ■

Lemma 15 *Let $S_n = \{(X_i, Y_i)\}_{i=1}^n \sim (\mathcal{X} \times \mathcal{Y})^n$ be a* given *sample. Let $\hat{h} : \mathcal{X} \to \mathcal{Y}$ be a predictor, and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ be a bounded loss function by some $L > 0$. Let $I \subseteq [n]$ be a random variable* sampled without replacement *from $[n]$. Then for any $\delta \in (0,1)$, with confidence at least $1 - \delta$:*

$$\left| \frac{1}{|I|} \sum_{i \in I} \ell\left(\hat{h}(X_i), Y_i\right) - \widehat{R}(\hat{h}; S_n) \right| \leq \widehat{R}(\hat{h}; S_n)\sqrt{\frac{2\ln(2/\delta)}{|I|}} + L\sqrt{\frac{2\ln(2/\delta)}{|I|}} + \frac{2L\ln(2/\delta)}{3|I|}.$$

Proof Let $\ell_i := \ell(\hat{h}(X_i), Y_i)$ for $i \in [n]$. Treating $\ell_1, \ldots, \ell_n$ as a given finite population, and $\{\ell_i\}_{i \in I}$ as a random sample drawn without replacement from it, we can use a version of Bernstein's inequality (Bardenet and Maillard, 2015),

$$\mathbb{P}\left(\left| \frac{1}{|I|} \sum_{i \in I} \ell_i - \mu \right| \geq \varepsilon \right) \leq 2\exp\left(-\frac{|I|\varepsilon^2}{2\sigma^2 + \frac{2L}{3}\varepsilon} \right) \qquad \varepsilon > 0,$$

where we have defined:

$$\begin{aligned}
\mu &:= \frac{1}{n}\sum_{i=1}^n \ell_i = \widehat{R}(\hat{h}; S_n) && \text{(population mean)} \\
\sigma^2 &:= \frac{1}{n}\sum_{i=1}^n (\ell_i - \mu)^2 && \text{(population variance).}
\end{aligned}$$

For any $\delta \in (0,1)$, we get that with confidence at least $1 - \delta$:

$$\left| \frac{1}{|I|} \sum_{i \in I} \ell_i - \mu \right| \leq \frac{2L}{3|I|}\ln(2/\delta) + \sqrt{\frac{2\sigma^2\ln(2/\delta)}{|I|}}.$$

Now similarly to the analysis in the proof of Lemma 14, we see that

$$\sigma^2 \leq L^2 + \widehat{R}^2(\hat{h}; S_n)$$

using that and that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y > 0$ we get the statement of the lemma. ■

Theorem 16 *For any $k \in \mathbb{N}$, $b \in \mathbb{N} \cup \{0\}$, let (κ, ψ) be any semi-stable compression scheme of size at most k using at most b bits of side-information. For any distribution $\bar{\mu}$ over $\mathcal{X} \times \mathcal{Y}$, any $n \in \mathbb{N}$ with $n > 4k+4$, and any $\delta \in (0,1)$, for $S_n \sim \bar{\mu}^n$, with probability at least $1-\delta$*

$$\begin{aligned} R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) \leq \widehat{R}(\psi(\kappa(S_n)); S_n) &\left(5\sqrt{\frac{8\left(\ln(\frac{4}{\delta}) + k\ln 4\right)}{n}} + 4\sqrt{\frac{8b\ln 2}{n}} \right) \\ &+ 2L\sqrt{\frac{8\left(\ln(\frac{4}{\delta}) + k\ln 4\right)}{n}} + \frac{(28+8L)\left(\ln(\frac{4}{\delta}) + k\ln 4\right)}{3n} \\ &+ L\sqrt{\frac{8b\ln 2}{n}} + \frac{28b\ln 2}{3n}. \end{aligned}$$

Proof Similarly to Hanneke and Kontorovich (2021), this proof follows similar arguments as Theorem 13, except using Lemma 14 in place of Hoeffding's inequality in both places in the proof where such inequalities are used.

Let $\mathcal{I}_n$ and T_n be as in the proof of Theorem 13, and let $[n] = \{1, \ldots, n\}$ for and $n \in \mathbb{N}$.

Let there be a sample $S_n = \{(X_i, Y_i)\}_{i=1}^n \sim \bar{\mu}^n$. As in Theorem 13, for any $I \in \mathcal{I}_n$ we have $S(I) = \{(X_i, Y_i) : i \in I\} \subseteq S_n$, for any $I \in \mathcal{I}_n$, define $\bar{I} := [n] \setminus I$. Let $\sigma : [n] \to [n]$ be a uniform random permutation of $[n]$, and for any $I = (i_1, \ldots, i_\ell) \subseteq [n]$ define $\boldsymbol{\sigma}(I) := (\sigma(i_1), \ldots, \sigma(i_\ell))$.

For any $I \subseteq [n]$ and any $\boldsymbol{b} \in \{0,1\}^b$, let $\hat{h}_{I,\boldsymbol{b}} := \psi\left(\kappa_{\textsf{cs}}\left(S\left(I\right)\right), \boldsymbol{b}\right)$. Now, since $S(\bar{I})$ is independent of $S(I)$, Lemma 14 (applied under the conditional distribution given $S(I)$) and the law of total probability imply that with probability $1 - \frac{\delta}{2|\mathcal{I}_n| \cdot 2^b}$:

$$\begin{aligned} R\left(\hat{h}_{I,\boldsymbol{b}}\right) - \widehat{R}\left(\hat{h}_{I,\boldsymbol{b}}; S\left(\bar{I}\right)\right) \leq \widehat{R}\left(\hat{h}_{I,\boldsymbol{b}}; S\left(\bar{I}\right)\right) &\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{n-|I|-1}} \\ &+ L\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{n-|I|-1}} + \frac{7\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{3(n-|I|-1)}. \end{aligned}$$

Applying this under the conditional distribution given σ, together with the union bound and the law of total probability, we have that with probability at least $1-\frac{\delta}{2}$, every $I \in \mathcal{I}_n$ and every $\boldsymbol{b} \in \{0,1\}^b$ has

$$\begin{aligned} R\left(\hat{h}_{\boldsymbol{\sigma}^{-1}(I),\boldsymbol{b}}\right) - \widehat{R}\left(\hat{h}_{\boldsymbol{\sigma}^{-1}(I),\boldsymbol{b}}; S\left(\overline{\boldsymbol{\sigma}^{-1}\left(I\right)}\right)\right) \leq \widehat{R}\left(\hat{h}_{\boldsymbol{\sigma}^{-1}(I),\boldsymbol{b}}; S\left(\overline{\boldsymbol{\sigma}^{-1}\left(I\right)}\right)\right) &\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{n-|I|-1}} \\ &+ L\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{n-|I|-1}} + \frac{7\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{3(n-|I|-1)}. \end{aligned}$$

In particular, let $\boldsymbol{i}^*$ be the indices such that

$$\kappa_{\textsf{cs}}(S_n) = \{(X_i, Y_i) \in S_n : i \in \boldsymbol{i}^*\} = S_n(\boldsymbol{i}^*)$$

Now, by property (ii) of $\mathcal{I}_n$ there must exist $I^* \in \mathcal{I}_n$ such that

$$\boldsymbol{\sigma}(\boldsymbol{i}^*) \subseteq I^*,$$

which means:

$$\begin{aligned}&\Rightarrow \quad \boldsymbol{i}^* \subseteq \boldsymbol{\sigma}^{-1}(I^*)\\ &\Rightarrow \quad \kappa_{\mathsf{cs}}(S_n) = S_n(\boldsymbol{i}^*) \subseteq S_n(\boldsymbol{\sigma}^{-1}(I^*)).\end{aligned}$$

Due to the semi-stability property of (κ, ψ) and since $S_n(\boldsymbol{\sigma}^{-1}(I^*)) \subseteq S_n$, this implies

$$\Rightarrow \psi(\kappa_{\mathsf{cs}}(S_n(\boldsymbol{\sigma}^{-1}(I^*))), \kappa_{\mathsf{si}}(S_n)) = \psi(\kappa_{\mathsf{cs}}(S_n), \kappa_{\mathsf{si}}(S_n)) = \psi(\kappa(S_n)).$$

Thus, on the above event of probability at least $1 - \frac{\delta}{2}$ we can get for $I = I^*$ and $\boldsymbol{b} = \kappa_{\mathsf{si}}(S_n)$,

$$\begin{aligned}R(\psi(\kappa(S_n))) - \widehat{R}\left(\psi(\kappa(S_n)); S\left(\overline{\boldsymbol{\sigma}^{-1}(I^*)}\right)\right) &\leq \widehat{R}\left(\psi(\kappa(S_n)); S\left(\overline{\boldsymbol{\sigma}^{-1}(I^*)}\right)\right)\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{n - |I^*| - 1}}\\ &\quad + L\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{n - |I^*| - 1}} + \frac{7\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{3(n - |I^*| - 1)}.\end{aligned}$$

Furthermore, by property (i) of I_n we have that $n - |I^*| \geq T_n$, and that $\widehat{R}\left(\psi(\kappa(S_n)); S\left(\overline{\boldsymbol{\sigma}^{-1}(I^*)}\right)\right) \leq \frac{n}{T_n}\widehat{R}(\psi(\kappa(S_n)); S_n)$, so

$$\begin{aligned}R(\psi(\kappa(S_n))) - \widehat{R}\left(\psi(\kappa(S_n)); S\left(\overline{\boldsymbol{\sigma}^{-1}(I^*)}\right)\right) &\leq \frac{n}{T_n}\widehat{R}(\psi(\kappa(S_n)); S_n)\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{T_n - 1}}\\ &\quad + L\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{T_n - 1}} + \frac{7\ln(\frac{4|\mathcal{I}_n|\cdot 2^b}{\delta})}{3(T_n - 1)}.\end{aligned} \tag{C.4}$$

Now, given S_n, we define $\ell_i := \ell(\psi(\kappa(S_n))(X_i), Y_i)$ for $i \in [n]$. Now for any $I \in \mathcal{I}_n$, under the conditional distribution given S_n we apply Lemma 15 and see that with probability at least $1 - \frac{\delta}{2|\mathcal{I}_n|}$:

$$\begin{aligned}\left|\frac{1}{n - |I|}\sum_{i \in \overline{\boldsymbol{\sigma}^{-1}(I)}} \ell_i - \widehat{R}(\psi(\kappa(S_n)); S_n)\right| &\leq \widehat{R}(\psi(\kappa(S_n)); S_n)\sqrt{\frac{2\ln(4|\mathcal{I}_n|/\delta)}{n - |I|}}\\ &\quad + L\sqrt{\frac{2\ln(4|\mathcal{I}_n|/\delta)}{n - |I|}} + \frac{2L\ln(4|\mathcal{I}_n|/\delta)}{3(n - |I|)}.\end{aligned}$$

By the union bound, this holds simultaneously for all $I \in \mathcal{I}_n$ with probability at least $1 - \frac{\delta}{2}$. In particular, taking $I = I^*$, and recalling that $n - |I^*| \geq T_n$, on this event we have that

$$\begin{aligned}\left|\widehat{R}(\psi(\kappa(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}(I^*)})) - \widehat{R}(\psi(\kappa(S_n)); S_n)\right| &\leq \widehat{R}(\psi(\kappa(S_n)); S_n)\sqrt{\frac{2\ln(4|\mathcal{I}_n|/\delta)}{T_n}}\\ &\quad + L\sqrt{\frac{2\ln(4|\mathcal{I}_n|/\delta)}{T_n}} + \frac{2L\ln(4|\mathcal{I}_n|/\delta)}{3T_n}.\end{aligned} \tag{C.5}$$

By the union bound, the two events represented by (C.4) and (C.5) hold simultaneously with probability at least $1-\delta$, in which case together we get:

$$
\begin{aligned}
R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) &\leq R\left(\psi(\kappa\left(S_n\right))\right) - \widehat{R}\left(\psi(\kappa\left(S_n\right)); S\left(\overline{\boldsymbol{\sigma}^{-1}\left(I^*\right)}\right)\right) \\
&\quad + \left|\widehat{R}(\psi(\kappa(S_n)); S(\overline{\boldsymbol{\sigma}^{-1}\left(I^*\right)})) - \widehat{R}(\psi(\kappa(S_n)); S_n)\right| \\
&\leq \widehat{R}(\psi(\kappa(S_n)); S_n)\left(\left(1+\frac{n}{T_n}\right)\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|}{\delta})}{T_n-1}} + \frac{n}{T_n}\sqrt{\frac{2b\ln 2}{T_n-1}}\right) \\
&\quad + 2L\sqrt{\frac{2\ln(\frac{4|\mathcal{I}_n|}{\delta})}{T_n-1}} + \frac{(7+2L)\ln(\frac{4|\mathcal{I}_n|}{\delta})}{3(T_n-1)} + L\sqrt{\frac{2b\ln 2}{T_n-1}} + \frac{7b\ln 2}{3(T_n-1)}.
\end{aligned}
$$

The theorem now follows from plugging the aforementioned family $\mathcal{I}_n$ from Bousquet et al. (2020), with $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$ and $T_n = k\left\lfloor\frac{n}{2k}\right\rfloor > \frac{n-2k}{2}$ — with the modification thatsince $n > 4k+4$ we have $\frac{n-2k}{2} - 1 > \frac{n}{4}$, meaning $T_n > T_n - 1 > \frac{n}{4}$:

$$
\begin{aligned}
R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) &\leq \widehat{R}(\psi(\kappa(S_n)); S_n)\left(5\sqrt{\frac{8\left(\ln(\frac{4}{\delta}) + k\ln 4\right)}{n}} + 4\sqrt{\frac{8b\ln 2}{n}}\right) \\
&\quad + 2L\sqrt{\frac{8\left(\ln(\frac{4}{\delta}) + k\ln 4\right)}{n}} + \frac{(28+8L)\left(\ln(\frac{4}{\delta}) + k\ln 4\right)}{3n} \\
&\quad + L\sqrt{\frac{8b\ln 2}{n}} + \frac{28b\ln 2}{3n}.
\end{aligned}
$$

■

C.3.3. SEMI-STABLE COMPRESSION SCHEME OF BOUNDED SAMPLE-DEPENDENT SIZE

Theorem 17 *Let (κ, ψ) be any semi-stable compression scheme with side-information. For any distribution $\bar{\mu}$ over $\mathcal{X}\times\mathcal{Y}$, any $n\in\mathbb{N}$, and any $\delta\in(0,1)$, for $S_n \sim \bar{\mu}^n$, with probability at least $1-\delta$, if $|\kappa_{\mathsf{cs}}(S_n)| < \frac{n}{4} - 1$ then*

$$
\begin{aligned}
R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) &\leq \widehat{R}(\psi(\kappa(S_n)); S_n)\left(5\sqrt{\frac{8T_{\kappa,S_n}}{n}} + 4\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}}\right) \\
&\quad + 2L\sqrt{\frac{8T_{\kappa,S_n}}{n}} + \frac{(28+8L)T_{\kappa,S_n}}{3n} \\
&\quad + L\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} + \frac{28|\kappa_{\mathsf{si}}(S_n)|\ln 2}{3n},
\end{aligned}
$$

where

$$
T_{\kappa,S_n} := \ln\left(\frac{4(|\kappa_{\mathsf{cs}}(S_n)|+1)(|\kappa_{\mathsf{cs}}(S_n)|+2)(|\kappa_{\mathsf{si}}(S_n)|+1)(|\kappa_{\mathsf{si}}(S_n)|+2)}{\delta}\right) + |\kappa_{\mathsf{cs}}(S_n)|\ln 4.
$$

Proof

Let (κ, ψ) be any semi-stable compression scheme with side-information. Now for each $k \in \mathbb{N} \cup \{0\}$ and $b \in \mathbb{N} \cup \{0\}$, let $(\kappa_{k,b}, \psi)$ be a compression scheme such that, for any $S_n \sim \mu^n$, if $|\kappa_{\mathsf{cs}}(S)| \leq k$ and $|\kappa_{\mathsf{si}}(S)| \leq b$, then $\kappa_{k,b}(S) = \kappa(S)$, and otherwise $(\kappa_{\mathsf{cs}})_{k,b}(S_n) = \emptyset$ and $(\kappa_{\mathsf{si}})_{k,b}(S_n) = \emptyset$. In particular, note that $|(\kappa_{\mathsf{cs}})_{k,b}(S)| \leq k$ and $|(\kappa_{\mathsf{si}})_{k,b}(S)| \leq b$ always. Thus, for each k and b, Theorem 16 implies that for any $n > 4k + 4$

$$\left| R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) \right| \leq \widehat{R}(\psi(\kappa(S_n)); S_n) \left(5\sqrt{\frac{8t_{k,b}}{n}} + 4\sqrt{\frac{8b\ln 2}{n}} \right) + 2L\sqrt{\frac{8t_{k,b}}{n}} + \frac{(28+8L)t_{k,b}}{3n} + L\sqrt{\frac{8b\ln 2}{n}} + \frac{28b\ln 2}{3n}$$

holds with probability at least $1 - \frac{\delta}{(k+1)(k+2)(b+1)(b+2)}$, where

$$t_{k,b} := \ln\left(\frac{4(k+1)(k+2)(b+1)(b+2)}{\delta}\right) + k\ln 4.$$

By the union bound, the above claim holds simultaneously for all $k \in \mathbb{N} \cup \{0\}$ and $b \in \mathbb{N} \cup \{0\}$ with probability at least $1 - \sum_{k,b} \frac{\delta}{(k+1)(k+2)(b+1)(b+2)} = 1 - \delta$. Finally, note that there necessarily exists some $k \in \mathbb{N} \cup \{0\}$ and $b \in \mathbb{N} \cup \{0\}$ for which $|\kappa_{\mathsf{cs}}(S)| = k$ and $|\kappa_{\mathsf{si}}(S)| = b$, in which case $\psi(\kappa(S)) = \psi(\kappa_{k,b}(S))$ for these k and b. This completes the proof. ■

C.3.4. SEMI-STABLE COMPRESSION SCHEME OF ANY SAMPLE-DEPENDENT SIZE

Theorem 2. *Suppose that $\mathcal{X}$ is an instance space and $\mathcal{Y}$ a label space with a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, L]$, and (κ, ψ) is semi-stable compression scheme. For any distribution $\bar{\mu}$ over $\mathcal{X} \times \mathcal{Y}$, any $n \in \mathbb{N}$, and any $\delta \in (0, 1)$, for $S_n \sim \bar{\mu}^n$ we have that*

$$R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) \leq \left(20\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 20\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 15\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \right) \widehat{R}(\psi(\kappa(S_n)); S_n) + (6L+18)\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 8L\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + (2L+12)\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7L\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + (3L+10)\frac{\ln(\frac{4e^2}{\delta})}{n} + 6L\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}$$

holds with probability at least $1 - \delta$.

Proof Let (κ, ψ) be any semi-stable compression scheme with side-information. We will observe the RHS of Theorem 17. Since $\ln(x^2) < \frac{x}{2}$ for any $x \geq \sqrt{3}$, firstly we have:

$$\begin{aligned}
T_{\kappa,S_n} &:= \ln\left(\frac{4(|\kappa_{\mathsf{cs}}(S_n)|+1)(|\kappa_{\mathsf{cs}}(S_n)|+2)(|\kappa_{\mathsf{si}}(S_n)|+1)(|\kappa_{\mathsf{si}}(S_n)|+2)}{\delta}\right) + |\kappa_{\mathsf{cs}}(S_n)|\ln 4 \\
&\leq \ln(\frac{4}{\delta}) + |\kappa_{\mathsf{cs}}(S_n)|\ln 4 + \frac{|\kappa_{\mathsf{cs}}(S_n)|+2}{2} + \frac{|\kappa_{\mathsf{si}}(S_n)|+2}{2} \\
&= |\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e}) + \frac{1}{2}|\kappa_{\mathsf{si}}(S_n)| + \ln(\frac{4e^2}{\delta})
\end{aligned}$$

Next, let us abbreviate the right-hand side of the bound in Theorem 17:

$$\begin{aligned}
\widehat{R}(\psi(\kappa(S_n)); S_n)\left(5\sqrt{\frac{8T_{\kappa,S_n}}{n}} + 4\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}}\right) + 2L\sqrt{\frac{8T_{\kappa,S_n}}{n}} + \frac{(28+8L)T_{\kappa,S_n}}{3n} \\
+L\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} + \frac{28|\kappa_{\mathsf{si}}(S_n)|\ln 2}{3n} \\
:= A_{\mathrm{I}}\widehat{R}(\psi(\kappa(S_n)); S_n) + A_{\mathrm{II}},
\end{aligned}$$

$$\begin{aligned}
A_{\mathrm{I}} &:= 5\sqrt{\frac{8T_{\kappa,S_n}}{n}} + 4\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} \\
&\leq 5\sqrt{\frac{8(|\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e}) + \frac{1}{2}|\kappa_{\mathsf{si}}(S_n)| + \ln(\frac{4e^2}{\delta}))}{n}} + 4\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} \\
&\leq 5\sqrt{\frac{8|\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e})}{n}} + 5\sqrt{\frac{4|\kappa_{\mathsf{si}}(S_n)|}{n}} + 5\sqrt{\frac{8\ln(\frac{4e^2}{\delta})}{n}} + 4\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} \\
&= 5\sqrt{8\ln(4\sqrt{e})}\cdot\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + (10+4\sqrt{8\ln 2})\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 5\sqrt{8}\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \\
&\leq 20\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 20\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 15\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \\
&:= B_{\mathrm{I}},
\end{aligned}$$

$$
\begin{aligned}
A_{\mathrm{II}} &:= 2L\sqrt{\frac{8T_{\kappa,S_n}}{n}} + \frac{(28+8L)T_{\kappa,S_n}}{3n} + L\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} + \frac{28|\kappa_{\mathsf{si}}(S_n)|\ln 2}{3n} \\
&\leq 2L\sqrt{\frac{8\left(|\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e}) + \frac{1}{2}|\kappa_{\mathsf{si}}(S_n)| + \ln(\frac{4e^2}{\delta})\right)}{n}} \\
&\quad + \frac{(28+8L)\left(|\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e}) + \frac{1}{2}|\kappa_{\mathsf{si}}(S_n)| + \ln(\frac{4e^2}{\delta})\right)}{3n} \\
&\quad + L\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} + \frac{28|\kappa_{\mathsf{si}}(S_n)|\ln 2}{3n} \\
&\leq 2L\sqrt{\frac{8|\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e})}{n}} + 2L\sqrt{\frac{4|\kappa_{\mathsf{si}}(S_n)|}{n}} + 2L\sqrt{\frac{8\ln(\frac{4e^2}{\delta})}{n}} \\
&\quad + \frac{(28+8L)\left(|\kappa_{\mathsf{cs}}(S_n)|\ln(4\sqrt{e}) + \frac{1}{2}|\kappa_{\mathsf{si}}(S_n)| + \ln(\frac{4e^2}{\delta})\right)}{3n} \\
&\quad + L\sqrt{\frac{8|\kappa_{\mathsf{si}}(S_n)|\ln 2}{n}} + \frac{28|\kappa_{\mathsf{si}}(S_n)|\ln 2}{3n} \\
&= \frac{(28+8L)\ln(4\sqrt{e})}{3}\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 2L\sqrt{8\ln(4\sqrt{e})}\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + \\
&\quad + \frac{14+4L+28\ln 2}{3}\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + (4+\sqrt{8\ln 2})L\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} \\
&\quad + \frac{28+8L}{3}\frac{\ln(\frac{4e^2}{\delta})}{n} + 2\sqrt{8}L\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \\
&\leq (6L+18)\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 8L\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + \\
&\quad + (2L+12)\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7L\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} \\
&\quad + (3L+10)\frac{\ln(\frac{4e^2}{\delta})}{n} + 6L\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \\
&= \left(6\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 8\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 2\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 3\frac{\ln(\frac{4e^2}{\delta})}{n} + 6\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}\right)L \\
&\quad + 18\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 12\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 10\frac{\ln(\frac{4e^2}{\delta})}{n} \\
&:= B_{\mathrm{II}}.
\end{aligned}
$$

Let $\bar{\mu}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$, and let $n \in \mathbb{N}$, and $\delta \in (0,1)$. From Theorem 17 we know that for $S_n \sim \bar{\mu}^n$, with probability at least $1-\delta$, if $|\kappa_{\mathsf{cs}}(S_n)| < \frac{n}{4} - 1$ then

$$
\begin{aligned}
R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n) &\leq A_{\mathrm{I}}\widehat{R}(\psi(\kappa(S_n)); S_n) + A_{\mathrm{II}} \\
&\leq B_{\mathrm{I}}\widehat{R}(\psi(\kappa(S_n)); S_n) + B_{\mathrm{II}}.
\end{aligned} \tag{C.6}
$$

Now, for B_{I} we note that even if $|\kappa_{\mathsf{cs}}(S_n)| \geq \frac{n}{4} - 1$ we have:

$$\begin{aligned} B_{\mathrm{I}} &:= 20\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 20\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 15\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}} \\ &\geq 0 \geq -1 \end{aligned}$$

and for B_{II}, we observe firstly that if $n \leq 4$:

$$\begin{aligned} B_{\mathrm{II}} := &\left(6\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 8\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 2\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 3\frac{\ln(\frac{4e^2}{\delta})}{n} + 6\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}\right) L \\ &+ 18\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 12\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 10\frac{\ln(\frac{4e^2}{\delta})}{n} \\ \geq &\left(\frac{3}{2} + 8\sqrt{\frac{1}{4}} + 2\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 3\frac{\ln(\frac{4e^2}{\delta})}{n} + 6\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}\right) L \\ &+ \frac{9}{2} + 12\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 10\frac{\ln(\frac{4e^2}{\delta})}{n} \\ \geq &\, L \end{aligned}$$

and now even if $n \geq 5$ and $|\kappa_{\mathsf{cs}}(S_n)| \geq \frac{n}{4} - 1$:

$$\begin{aligned} B_{\mathrm{II}} := &\left(6\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 8\sqrt{\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n}} + 2\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 3\frac{\ln(\frac{4e^2}{\delta})}{n} + 6\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}\right) L \\ &+ 18\frac{|\kappa_{\mathsf{cs}}(S_n)|}{n} + 12\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 10\frac{\ln(\frac{4e^2}{\delta})}{n} \\ \geq &\left(6\left(\frac{1}{4} - \frac{1}{n}\right) + 8\sqrt{\frac{1}{4} - \frac{1}{n}} + 2\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 3\frac{\ln(\frac{4e^2}{\delta})}{n} + 6\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}\right) L \\ &+ 18\left(\frac{1}{4} - \frac{1}{n}\right) + 12\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 10\frac{\ln(\frac{4e^2}{\delta})}{n} \\ \geq &\left(\frac{6}{20} + 8\sqrt{\frac{1}{20}} + 2\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 7\sqrt{\frac{|\kappa_{\mathsf{si}}(S_n)|}{n}} + 3\frac{\ln(\frac{4e^2}{\delta})}{n} + 6\sqrt{\frac{\ln(\frac{4e^2}{\delta})}{n}}\right) L \\ &+ \frac{18}{20} + 12\frac{|\kappa_{\mathsf{si}}(S_n)|}{n} + 10\frac{\ln(\frac{4e^2}{\delta})}{n} \\ \geq &\, L. \end{aligned}$$

Thus, even if $|\kappa_{\mathsf{cs}}(S_n)| \geq \frac{n}{4} - 1$, we have

$$\begin{aligned} B_{\mathrm{I}}\widehat{R}(\psi(\kappa(S_n)); S_n) + B_{\mathrm{II}} &\geq L - \widehat{R}(\psi(\kappa(S_n)); S_n) \\ &\geq R(\psi(\kappa(S_n))) - \widehat{R}(\psi(\kappa(S_n)); S_n). \end{aligned} \tag{C.7}$$

Finally, the theorem follows from (C.6) and (C.7). ■

Appendix D. Diameter-truncating BIE $\mathcal{Y}$

Recall the bounded-in-expectation (BIE) condition: $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(y_0, Y) < \infty$ for some $y_0 \in \mathcal{Y}$. Under BIE, the trivial $f_0(x) \equiv y_0$ achieves $R(f_0) < \infty$, and a fortiori, $R^* = \inf_f R(f) < \infty$.

Lemma 18 *If there exists* some $y_0 \in \mathcal{Y}$ *for which* $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(y_0, Y) < \infty$, *then this holds for* all $y \in \mathcal{Y}$.

Proof Suppose that $\mathbb{E}_{(X,Y)\sim\bar{\mu}}\,\ell(y_0, Y) < \infty$ for some $y_0 \in \mathcal{Y}$. Then, by the triangle inequality, for any other $y' \in \mathcal{Y}$, we have

$$\mathop{\mathbb{E}}_{(X,Y)\sim\bar{\mu}} \ell(y', Y) \le \mathop{\mathbb{E}}_{(X,Y)\sim\bar{\mu}} [\ell(y', y_0) + \ell(y_0, Y)] = \ell(y', y_0) + \mathop{\mathbb{E}}_{(X,Y)\sim\bar{\mu}} \ell(y_0, Y) < \infty.$$

■

Thus, the choice of $y_0 \in \mathcal{Y}$ is immaterial; let us fix one such element once and for all. For any sequence $L_n \uparrow \infty$, let $[\![\mathcal{Y}]\!]_n := B(y_0, L_n)$ denote the "L_n-truncated" space.

We observe that

$$f^*(x) := \mathop{\mathrm{argmin}}_{y' \in \mathcal{Y}} \mathbb{E}[\ell(y', Y) \,|\, X = x]$$

(where ties in $\mathcal{Y}$ are broken lexicographically) achieves $R(f^*) = R^*$, since it is a pointwise minimizer of the non-negative risk integrand. Let us also define a truncated version:

$$f_n^*(x) := \mathop{\mathrm{argmin}}_{\hat{y} \in [\![\mathcal{Y}]\!]_n} \mathbb{E}[\ell(\hat{y}, Y) \,|\, X = x].$$

Since $y_0 \in [\![\mathcal{Y}]\!]_n$, we have that

$$g_n(x) := \mathbb{E}[\ell(f_n^*(x), Y | X = x)] \le \mathbb{E}[\ell(y_0, Y | X = x)] =: h(x).$$

While one or both of g_n, h may be infinite for some x, BIE implies that h is integrable. Next, we claim that

$$\lim_{n\to\infty} g_n(x) = g(x) := \mathbb{E}[\ell(f^*(x), Y | X = x)], \qquad x \in \mathcal{X}.$$

Indeed, this follows from a stronger property: there is a function $N : \mathcal{X} \to \mathbb{N}$ such that $f_n^*(x) = f^*(x)$ for all x and all $n \ge N(x)$; this is immediate by construction and because $L_n \uparrow \infty$. Applying Lebesgue's Dominated Convergence Theorem to the sequence $g_n \le h$ yields

Theorem 19 *If $\mathcal{Y}$ is BIE, f^* is a Bayes-optimal predictor and f_n^* is its truncation as defined above, then*

$$\lim_{n\to\infty} R(f_n^*) = R(f^*) = R^*.$$

Appendix E. Discretizing separable $(\mathcal{Y}, \ell)$

For any separable $(\mathcal{Y}, \ell)$ and $\varepsilon > 0$, any ε-net $\mathcal{Y}_\varepsilon \subseteq \mathcal{Y}$ is always countable. Define $\pi_\varepsilon(y)$ as the closest element to y in $\mathcal{Y}_\varepsilon$, breaking ties lexicographically. Then the Voronoi cell $V(y)$ about each $y \in \mathcal{Y}_\varepsilon$ is given by $V(y) = \{y' \in \mathcal{Y} : \pi_\varepsilon(y') = y\}$. Any probability measure $\bar{\mu}$ on the product Borel σ-algebra on $\mathcal{X} \times \mathcal{Y}$ induces the product measure $\bar{\mu}_\varepsilon$ on $\mathcal{X} \times \mathcal{Y}_\varepsilon$ as follows. By Pollard (2002, Appendix F, Theorem 1), the measure $\bar{\mu}$ admits the disintegration intro $\mu^{\mathcal{Y}} \otimes \Lambda$, where $\mu^{\mathcal{Y}}$ is the $\mathcal{Y}$-marginal of $\bar{\mu}$ and $\Lambda(y, E) = \mathbb{P}_{(X,Y)\sim\bar{\mu}}(X \in E \,|\, Y = y)$ is the conditional kernel. Define the measure $\mu^{\mathcal{Y}}_\varepsilon$ on $\mathcal{Y}_\varepsilon$ by $\mu^{\mathcal{Y}}_\varepsilon(y) = \mu^{\mathcal{Y}}(V(y))$ and the product measure $\bar{\mu}_\varepsilon = \mu^{\mathcal{Y}}_\varepsilon \otimes \Lambda$ on $\mathcal{X} \times \mathcal{Y}_\varepsilon$. Let R^*, R^*_ε be the Bayes-optimal risk under $\bar{\mu}$ and $\bar{\mu}_\varepsilon$, respectively.

Theorem 20

$$\lim_{\varepsilon \to 0} R^*_\varepsilon = R^*.$$

Proof Appealing to a standard truncation argument, we assume without loss of generality that $R^* < \infty$. The product metric $\rho \oplus \ell$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, given by $\rho \oplus \ell((x, y), (x', y')) = \rho(x, x') + \ell(y, y')$ renders $(\mathcal{Z}, \rho \oplus \ell, \bar{\mu})$ and $(\mathcal{Z}_\varepsilon, \rho \oplus \ell, \bar{\mu}_\varepsilon)$ metric probability spaces. Let $h^* : \mathcal{X} \to \mathcal{Y}$ be the Bayes-optimal predictor for $(\mathcal{Z}, \rho \oplus \ell, \bar{\mu})$, and define $f^* : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ by $f^*(x, y) = \ell(h^*(x), y)$. Then

$$R^* = \int_{\mathcal{X}\times\mathcal{Y}} f^*(x, y) \mathrm{d}\bar{\mu}(x, y).$$

Since we assumed $R^* < \infty$, we have that $f^* \in L_1(\bar{\mu})$ and hence, by Hanneke et al. (2021, Lemma A.1) f^* may be approximated in L_1 by Lipschitz functions: for all $\eta > 0$, there is a $\Delta < \infty$ and a Δ-Lipschitz $\tilde{f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $\int_{\mathcal{X}\times\mathcal{Y}} |f^*(x, y) - \tilde{f}(x, y)| \mathrm{d}\bar{\mu}(x, y) < \eta$. Thus, there is no loss of generality in assuming f^* to be Δ-Lipschitz:

$$|f^*(x, y) - f^*(x', y')| \le \Delta(\rho(x, x') + \ell(y, y')).$$

Define the natural projection of $h \in \mathcal{Y}^{\mathcal{X}}$ onto $h_\varepsilon \in \mathcal{Y}_\varepsilon$, via $h_\varepsilon(x) := \pi_\varepsilon(h(x))$. Then by the Lipschitz property, $|R(h_\varepsilon) - R(h)| \le \Delta\varepsilon$. ■