JMLR: Workshop and Conference Proceedings 1–20, 2018

Multi-task Kernel Learning Based on Probabilistic Lipschitzness

Anastasia Pentina* ANASTASIA.PENTINA@SDSC.ETHZ.CH
Swiss Data Science Center, ETH Zurich

Shai Ben-David SHAI@CS.UWATERLOO.CA
School of Computer Science, University of Waterloo

Editor: Editor's name

Abstract

In multi-task learning the learner is given data for a set of related learning tasks and aims to improve the overall learning performance by transferring information between them. A typical assumption exploited in this setting is that the tasks share a beneficial representation that can be learned form the joint training data of all tasks. This way, the training data of each task can be utilized to enhance the learning of other tasks in the set. Probabilistic Lipschitzness (PL) is a parameter that reflects one way in which some data representation can be beneficial for a classification learning task. In this work we propose to achieve multi-task learning by learning a kernel function relative to which each of the tasks in the set has a "high level" of probabilistic Lipschitzness. In order to be able to do that, we need to introduce a new variant of PL - one that allows reliable estimation of its value from finite size samples. We show that by having access to large amounts of training data *in total* (possibly the union of training sets for various tasks), the learner can identify a kernel function that would lead to fast learning rates per task when used for Nearest Neighbor classification or in a cluster-based active labeling procedure.

1. Introduction

Traditionally machine learning research concentrates on developing algorithms for solving individual learning problems. Significant progress has been achieved in this framework from both theoretical and practical perspectives. However, this approach clearly neglects important aspects of human learning. In particular, humans are often able to benefit from having access to more than one problem at a time and comparing and contrasting several learning concepts. This observation motivates the multi-task learning setting (Caruana, 1997), where the learner faces a set of learning tasks simultaneously and is able to transfer information between them in order to improve the overall performance.

Intuitively, for information transfer to be useful, the tasks of interest should be related in some way. A natural and commonly used approach to formalize task relatedness is through an assumption that there exists a data representation that is beneficial for solving all the given tasks. The corresponding methods differ by the type of representation they are searching for and by the measure of its quality and focus on the cases when the type of predictors used based on the obtained representation is predefined. One common approach is based on metric

* Part of work performed while at IST Austria

© 2018 A. Pentina & S. Ben-David.

learning, where the representation has a form of a learned Mahalanobis distance. The quality in this case depends on how well it satisfies the predefined constraints that typically resemble intuition behind nearest neighbor approach for classification and force objects from the same class to lie close to each other and objects from different classes to be far apart (Yang et al., 2013). Another group of the multi-task methods for representation transfer focuses on the case when every task is solved using a linear predictor and aims at learning a joint representation together with the task-specific weight vectors. In this case the quality of the representation is measured by the empirical error achieved by the linear predictors under this representation that may have a form of projection on a low-dimensional subspace (Argyriou et al., 2007a) or a dictionary of features that allows for sparse solutions for every task of interest (Argyriou et al., 2008). In this work we, instead, propose to measure the quality of representation by its Probabilistic Lipschitzness (Urner et al., 2011), which allows us to obtain representation that could be useful in various scenarios such as classification using Nearest Neighbors or active learning.

Benefits of representation transfer in multi-task learning from theoretical perspective are captured by demonstrating a reduction of the number of samples per task needed for reliable generalization, compared to those needed when every task is solved in isolation. This reduction is a consequence of reducing the hypothesis set, used for solving each individual task, from the initial one, described by the union of all considered representations, to the one determined by the learned common representation. Thus, it only affects constant factors of some of the complexity terms, while keeping the convergence rates $O(1/\sqrt{n})$, where n is the number of samples per task.

To be able to demonstrate faster learning rates, it is necessary to assume low level of noise in the labeling function. In particular, in this work we focus on the case of deterministic labels. However, even this stronger assumption is not sufficient (Ben-David and Urner, 2014). One way to overcome this problem is to add sufficient realizability assumption. However, this kind of condition cannot be influenced by selecting an appropriate feature representation - it either already holds for the initial hypothesis set that is a union over all considered representations, or it does not hold for any of them. Alternative is given by the notion of Probabilistic Lipschitzness (PL) (Urner et al., 2011) that is capable of quantifying "easiness" of data and has been shown to control the sample complexity of Nearest Neighbor and potential reductions in label complexity in active learning (Urner et al., 2013). The advantage of PL is that it directly depends on the choice of feature representation - by changing the metric one can transform an initially hard learning problem into an easier one (i.e. achieve faster learning rate).

Probabilistic Lipschitzness (Urner et al., 2011) captures the intuition that under good representations similar instances are likely to have similar labels. Consider for example a set of image classification tasks, say a collection of detecting tasks for images captured by a car camera; detecting if there is a person in sight, or a traffic sign, or some major obstacle to driving. In all of those tasks it is rare to encounter two images that look similar from a human perception point of view, and yet should be classified differently. In other words, the human perception has some implicit common image representation under which each of those tasks enjoys a high lever of Probabilistic Lipschitzness (PL). Our goal in this work is to formalize this intuition and propose an algorithmic paradigm that can be shown to take advantage of such relationships between learning tasks. Aiming to use PL for multi-task

learning requires a way of evaluating the PL values of a given classification task based on training samples. However, the versions of PL that have been proposed in previous works cannot be reliably estimated from finite samples.

The first contribution of this work is the introduction of a modification of the original notion of Probabilistic Lipschitzness that measures the probability of two points violating Lipschitzness. The advantage of the proposed measure is that it can be reliably estimated from finite data and thus it can be used for selecting a beneficial representation in the multi-task setting. In particular, we demonstrate that for a family of kernels with finite pseudodimension the learner can estimate the average multi-task PL for every kernel in the family reliably by having access to large *total* number of samples. Thus, the overhead associated with estimating PL for multiple kernel functions is spread across multiple tasks. Assuming that in a given kernel family there exists a kernel with respect to which the probability of two points having different labels decays sufficiently quickly as the kernel similarity between them tends to one, we demonstrate two consequences of our result in multi-task representation learning. First, we show how one could use it to select a kernel that would lead to faster learning rates of Nearest Neighbor applied to every task. Second, we demonstrate that it could also be used to identify a kernel with respect to which to run an active labeling procedure with provable label complexity reductions. In this case, any empirical risk minimization or regularized risk minimization method can be used later on for solving all tasks of interest and one obtains faster rates in terms of label complexity.

Related work. Multi-task methods based on feature learning typically rely on the assumption that there exists a common representation that leads to low approximation error for all tasks of interest and aim at inferring this representation from the data. In particular, in (Argyriou et al., 2007a, 2008) this assumption was explored in the case where every task is solved using a sparse combination of original features or their linear transformations. These methods were later extended to be able to handle different levels of relatedness between tasks (Argyriou et al., 2007b), disjoint (Zhang and Yeung, 2011) or overlapping (Kumar and Daumé III, 2012) groups of related tasks and exploit known unrelated tasks (Romera-Paredes et al., 2012). This paradigm was also applied to kernel methods, where the common representation is assumed to be described by a kernel function (Jebara, 2004, 2011; Zhou et al., 2010).

Potential benefits of representation learning in the multi-task setting have been also theoretically analyzed for cases when this representation takes form of a sparse combination of initial features (Maurer and Pontil, 2013; Lounici et al., 2009), their linear transformations (Maurer, 2006; Maurer et al., 2014, 2013) or a kernel function (Pentina and Ben-David, 2015). However, such theoretical studies focus on the cases when the learning method that is used in conjunction with the learned representation is predefined and typically is an empirical risk minimization. And the advantages of learning a representation based on multiple tasks takes a form of a reduction of the complexity terms corresponding to the overhead of learning that representation and in the limit of infinitely many tasks the corresponding guarantees typically are reduced to those for single-task learning with beneficial representation known in advance. Thus, the convergence rate with respect to the number of samples per task reduces only by some constant factor.

2. Probabilistic Lipschitzness

Let $\mathcal{X}$ be the domain set equipped with a distance measure dist and $\mathcal{Y} = \{0, 1\}$ be the label set. Throughout this paper we focus on *tasks* that are defined by a pair consisting of a probability distribution D over $\mathcal{X}$ and a deterministic labeling function $l : \mathcal{X} \to \mathcal{Y}$.

Many learning paradigms, such as Nearest Neighbor or geometrically defined classifiers, implicitly assume that there is some correlation between the geometry of the input space, the marginal distribution and the labels. The notion of Probabilistic Lipschitzness (PL), introduced in (Urner et al., 2011), quantifies this correlation. In particular, it relaxes the condition of Lipschitzness on the labeling rule and formalizes the intuition that under suitable feature representation the probability of two close points having different labels is small. We will refer to the original definition of PL (Urner et al., 2011) as *PL-Unary*:

Definition 1 (PL-Unary (Urner et al., 2011)) *The labeling function l satisfies ϕ_U-PL-Unary if for all $\lambda > 0$:*

$$\Pr_{x_1 \sim D}\left[\Pr_{x_2 \sim D}[l(x_1) \neq l(x_2) \text{ and } \text{dist}(x_1, x_2) < \lambda] > 0\right] \leq \phi_U(\lambda). \tag{1}$$

A slightly different version of this measure was used in (Kushagra and Ben-David, 2015), which we refer to as *PL-Conditional*:

Definition 2 (PL-Conditional (Kushagra and Ben-David, 2015)) *The labeling function l satisfies ϕ_C-PL-Conditional if for all $\lambda > 0$:*

$$\Pr_{x_1, x_2 \sim D}[l(x_1) \neq l(x_2) \mid \text{dist}(x_1, x_2) < \lambda] \leq \phi_C(\lambda). \tag{2}$$

Both PL-Unary and PL-Conditional have been used to demonstrate faster learning rates for nicer distributions. In particular, they have been shown to characterize the sample complexity of the Nearest Neighbor classifier. In addition, PL-Unary has also been used to quantify the label complexity savings in active learning paradigm (Urner et al., 2013). These results imply that the faster the decay of $\phi_U(\lambda)$ (or $\phi_C(\lambda)$) as $\lambda \to 0$, the nicer the distribution and the easier it is to learn the task of interest (in terms of the sample complexity).

A distinctive feature of PL-Unary and PL-Conditional is that they directly depend on the geometry of the input space. In particular they indicate that by changing the distance measure one could make the task of interest easier or harder to learn. Thus, it would be advantageous if one could automatically select a feature representation under which the correlation between the input space and the labeling function, as measured by PL, is strong and the task of interest becomes easier to learn. Of course, selecting such a good data representation is likely to be as difficult as learning a good classifier. However, in the multi-task setting, under the assumption that for a set of tasks there exists a representation under which they all enjoy good PL, the required training sample can be divided between the different tasks. As a result, the more tasks there are in the pool the less data per task it takes to improve the data representation. This is where one leverages the multi-task aspect of the learning process. Such selection process would naturally require comparison of $\phi_U(\lambda)$ (or $\phi_C(\lambda)$) for various feature representations. However, because the true marginal distribution D and labeling function f are not known to the learner, neither are $\phi_U(\lambda)$

and $\phi_C(\lambda)$. Thus, one would need to estimate them from the data. However, both these quantities are not reliably estimable from finite samples. Indeed, estimating PL-Unary would require determining for any given ball whether the minority label in it has non-zero mass, which would require large number of samples if that mass of the minority label is close to 0. In (Kushagra and Ben-David, 2015) the authors proposed to estimate PL-Conditional by representing the conditional probability as a ratio and estimate the nominator and the denominator separately. However, as the authors notice it would not be reliable for small values of λ, since for them both the denominator and the nominator are expected to be close to 0 and even small additive errors could result in large multiplicative errors. Thus, we propose a different version of PL which we refer to as *PL-Pairwise*:

Definition 3 (PL-Pairwise) *The labeling function l satisfies ϕ_P-PL-Pairwise if for all $\lambda > 0$:*

$$\Pr_{x_1,x_2\sim D}[l(x_1) \neq l(x_2) \text{ and } \operatorname{dist}(x_1,x_2) < \lambda] \leq \phi_P(\lambda). \tag{3}$$

As we show in the next sections, the advantage of PL-Pairwise over previously used versions of PL is that it can be reliably estimated from finite samples (Section 3) and used in the multi-task setting to identify a feature representation that leads to fast convergence rates when used in conjunction with the nearest neighbor classifier (Sections 4) or in active learning paradigm (Section 5).

3. Estimating PL-Pairwise

In the multi-task setting the learner is given a collection of T training sets $S_1,\ldots,S_T$, where each $S_t = \{(x_1^t, l_t(x_1^t)),\ldots,(x_n^t, l_t(x_n^t))\}$ consists of n examples sampled i.i.d. from the task-specific data distribution D_t and labeled by the ground truth labeling function l_t. We consider the setting where the learner is given a family of kernel functions $\mathcal{K}$ and its goal is to identify the best kernel in this family. As a quality measure of kernel we will use the following multi-task version of PL-Pairwise:

Definition 4 *A set of tasks $\langle D_1,l_1\rangle,\ldots,\langle D_T,l_T\rangle$ satisfies Φ-MT-PL-Pairwise with respect to a kernel function K if for all $\lambda > 0$:*

$$\frac{1}{T}\sum_{t=1}^{T}\Pr_{x,x'\sim D_t}[l_t(x) \neq l_t(x') \wedge \|x-x'\|_K < \lambda] = \Phi(\lambda,K). \tag{4}$$

The following theorem shows that this quantity can be reliably estimated from a collection of finite training sets. In particular, note that the quality of estimation ($\alpha(k)$) vanishes as $\frac{1}{\sqrt{Tn}}$, therefore what matters is the *total size* of the training data. This indicates the advantage of having access to multiple tasks - as the number of tasks T grows, the amount of training examples per task n, sufficient for reliable estimation, decreases.

Theorem 5 *For any tasks $\langle D_1,l_1\rangle,\ldots,\langle D_T,l_T\rangle$, any set $\mathcal{K}$ of B^2-bounded kernels with pseudodimension p and any $\delta > 0$, with probability at least $1-2\delta$ over $\mathbf{S} = S_1\ldots,S_T$ the following holds uniformly for all $K \in \mathcal{K}$ and all $k = 1,2,\ldots$:*

$$\Phi(2^{-k},K) - \alpha(k) \leq \widehat{\Phi}_{\mathbf{S}}(2^{-(k-1)},K) \leq \Phi(3\cdot 2^{-k},K) + \alpha(k). \tag{5}$$

where

$$\alpha(k) = \sqrt{\frac{32p}{Tn}\log\left(\frac{30eT^2n^2B^2\cdot 2^{2k}}{p}\right) + \frac{32}{Tn}\log\frac{2}{\delta} + \frac{32k}{Tn}} = \tilde{O}\left(\sqrt{\frac{pk}{Tn}}\right) \tag{6}$$

$$\widehat{\Phi}_{\mathbf{S}}(\lambda, K) = \frac{1}{T}\sum_{t=1}^{T}\frac{2}{n(n-1)}\sum_{1\le i<j\le n}[\![y_i^t \neq y_j^t \wedge \|x_i^t - x_j^t\|_K < \lambda]\!] \tag{7}$$

Proof First, we utilize the standard 3-step procedure for some fixed λ.
Step 1. Symmetrization. Define:

$$Q = \{\mathbf{S} \in (\mathcal{X}\times\mathcal{Y})^{(T,n)} :\ \exists K :\ \Phi(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \epsilon\}$$
$$R = \{(\mathbf{S}, \tilde{\mathbf{S}}) \in (\mathcal{X}\times\mathcal{Y})^{(T,n)} \times (\mathcal{X}\times\mathcal{Y})^{(T,n)} :\ \exists K :\ \widehat{\Phi}_{\tilde{\mathbf{S}}}(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \epsilon/2\}$$

By triangle inequality:

$$\Phi(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \epsilon \wedge \Phi(\lambda, K) \le \widehat{\Phi}_{\tilde{\mathbf{S}}}(\lambda, K) + \frac{\epsilon}{2} \Rightarrow \widehat{\Phi}_{\tilde{\mathbf{S}}}(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \frac{\epsilon}{2}. \tag{8}$$

Therefore:

$$\Pr(R) \ge \Pr\{\exists K :\ \Phi(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \epsilon \wedge \Phi(\lambda, K) \le \widehat{\Phi}_{\tilde{\mathbf{S}}}(\lambda, K) + \epsilon/2\} \tag{9}$$

For $\mathbf{S} \in Q$ fix some K such that $\Phi(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \epsilon$. Then, by McDiarmid's inequality $\Pr\{\Phi(\lambda, K) > \widehat{\Phi}_{\tilde{\mathbf{S}}}(\lambda, K) + \epsilon/2\} \le 0.5$ if $Tn\epsilon^2 > 8\log 2$. Consequently $\Pr(Q) \le 2\Pr(R)$.
Step 2. Permutations. Define Γ to be a set of permutations on pairs $\{(1,1), \ldots, (T, 2n)\}$ such that $\{\sigma(i,j), \sigma(i, n+j)\} = \{(i,j), (i, n+j)\}$ for every $i = 1, \ldots, T$ and $j = 1, \ldots, n$. Then:

$$\Pr(R) \le \max_{(\mathbf{S},\tilde{\mathbf{S}})} \Pr_{\sigma}\{\sigma((\mathbf{S}, \tilde{\mathbf{S}})) \in R\}. \tag{10}$$

Step 3. Reduction to a finite class. Fix $(\mathbf{S}, \tilde{\mathbf{S}})$ such that there exists a kernel K:

$$\widehat{\Phi}_{\tilde{\mathbf{S}}}(\lambda, K) \ge \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \epsilon/2. \tag{11}$$

Let $\tilde{\mathcal{K}}(3\lambda^2/2)$ be a subset of kernel family $\mathcal{K}$ such that for every $K \in \mathcal{K}$ there exists a $\tilde{K} \in \tilde{\mathcal{K}}(3\lambda^2/2)$ such that for every $x_i^t, x_j^t \in \mathbf{S} \cup \tilde{\mathbf{S}}$:

$$\|x_i^t - x_j^t\|_{\tilde{K}}^2 - 3\lambda^2/2 \le \|x_i^t - x_j^t\|_K^2 \le \|x_i^t - x_j^t\|_{\tilde{K}}^2 + 3\lambda^2/2. \tag{12}$$

Then:

$$\|x_i^t - x_j^t\|_{\tilde{K}} \le \sqrt{2.5}\lambda \Rightarrow \|x_i^t - x_j^t\|_K \le 2\lambda; \qquad \|x_i^t - x_j^t\|_K \le \lambda \Rightarrow \|x_i^t - x_j^t\|_{\tilde{K}} \le \sqrt{2.5}\lambda$$

Therefore:

$$\widehat{\Phi}_{\tilde{\mathbf{S}}}(\sqrt{2.5}\lambda, \tilde{K}) \ge \widehat{\Phi}_{\mathbf{S}}(\sqrt{2.5}\lambda, \tilde{K}) + \epsilon/2. \tag{13}$$

As a result, using McDiarmid's inequality:

$$\Pr(R) \le \max_{(\mathbf{S},\tilde{\mathbf{S}})} \Pr_{\sigma}\{\sigma((\mathbf{S}, \tilde{\mathbf{S}})) \in R\} \le |\tilde{\mathcal{K}}(3\lambda^2/2)| \exp\left(-\frac{2\epsilon^2/4}{Tn(4/Tn)^2}\right)$$

According to Lemma 13 in Appendix A:

$$|\tilde{\mathcal{K}}(3\lambda^2/2)| \leq \mathcal{N}_{2Tn}^{\|\cdot\|^2}(\mathcal{K}, 3\lambda^2/2) \leq \left(\frac{30eT^2n^2B^2}{p\lambda^2}\right)^p. \tag{14}$$

By re-writing the above in terms of δ we obtain that for a fixed λ w.p. $1-\delta$ for all $K \in \mathcal{K}$:

$$\Phi_T(\lambda, K) \leq \widehat{\Phi}_{\mathbf{S}}(2\lambda, K) + \sqrt{\frac{32p}{Tn}\log\left(\frac{30eT^2n^2B^2}{p\lambda^2}\right) + \frac{32}{Tn}\log\frac{2}{\delta}}. \tag{15}$$

Step 4. Union bound. Now we take a union bound over all λ of form 2^{-k} by setting the corresponding δ to $\delta/2^k$. As a result we obtain that w.p. at least $1-\delta$ uniformly for all $K \in \mathcal{K}$ and $k = 1, 2, \ldots$:

$$\Phi_T(2^{-k}, K) \leq \widehat{\Phi}_{\mathbf{S}}(2^{-(k-1)}, K) + \sqrt{\frac{32p}{Tn}\log\left(\frac{30eT^2n^2B^2 \cdot 2^{2k}}{p}\right) + \frac{32}{Tn}\log\frac{2}{\delta} + \frac{32k}{Tn}}. \tag{16}$$

By repeating the same procedure in the opposite direction, we obtain the statement of Theorem 5. ■

4. Application to Nearest Neighbor

Both PL-Unary and PL-Conditional have been shown to characterize the sample complexity of the Nearest Neighbor classifier (Urner et al., 2011; Kushagra and Ben-David, 2015). The following theorem shows that similar result can be obtained using PL-Pairwise (proof can be found in Appendix B.1).

Theorem 6 *Let $\mathcal{X}$ be the input domain equipped with a distance measure d. Assume that the tasks $\langle D_1, l_1\rangle, \ldots, \langle D_T, l_T\rangle$ satisfy MT-PL-Pairwise with function $\Phi(\lambda)$ (with respect to the distance* dist*). Let $S_1, \ldots, S_T$ be T training sets, where every S_i consists of n points sampled i.i.d. according to the task-specific marginal distribution D_i and labeled by l_i. Then the following holds:*

$$\mathop{\mathbf{E}}_{S_1,\ldots,S_T} \frac{1}{T}\sum_{t=1}^{T} \mathrm{er}_t(NN(S_t)) \leq \min_{\lambda}\left(\frac{r(\lambda)}{ne} + n\Phi(\lambda)\right), \tag{17}$$

where $\mathrm{er}_t(NN(S_t))$ is the expected error of the nearest neighbor classifier on task $\langle D_t, l_t\rangle$ obtained based on the training set S_t and $r(\lambda)$ is the number of sets of diameter at most λ needed to cover the input space.

This result shows that $\Phi(\lambda)$ can be used to characterize how easy (on average) it is to learn given tasks using Nearest Neighbor. In general, the faster the decay of $\Phi(\lambda)$ with $\lambda \to 0$, the fewer samples per task n are sufficient to guarantee low average expected error. In order to see how Theorem 6 compares to the analogous results for PL-Unary (Urner and Ben-David, 2013) and PL-Conditional (Kushagra and Ben-David, 2015) (fully stated in Appendix B.2), consider the case of a single task, $T = 1$, and $\mathcal{X} \subset [0,1]^d$ with Euclidean metric. In this case $r(\lambda)$ behaves as λ^{-d}.

Exponential Lipschitzness First suppose that PL-Pairwise is satisfied with $\phi_P(\lambda) = e^{-1/\lambda}$. In this case Theorem 6 leads to sample complexity $O\left(\frac{1}{\epsilon\delta}\log^d\frac{1}{\epsilon\delta}\right)$. With the same condition on PL-Conditional (Theorem 15) gives the same complexity bound. The same condition on PL-Unary (Theorem 14) results in $O\left(\frac{1}{\epsilon\delta}\log^d\frac{1}{\epsilon}\right)$. So we see that under exponentially decaying PL Theorem 6 provides essentially the same sample complexity guarantees for Nearest Neighbor, as the previously known measures.

Polynomial Lipschitzness Suppose now that the labeling function satisfies PL-Pairwise with $\phi_P(\lambda) = \lambda^m$ for some $m \in \mathbb{N}$, $m \geq d$. In this case Theorem 6 leads to sample complexity of $O\left(\left(\frac{1}{\epsilon\delta}\right)^{\frac{m+d}{m-d}}\right)$. Under the same assumption but on PL-Unary (Theorem 14) one obtains $O\left(\frac{1}{\delta}\left(\frac{1}{\epsilon}\right)^{\frac{d+m}{m}}\right)$. Finally, the same condition on PL-Conditional (Theorem 15) leads to $O\left(\left(\frac{1}{\epsilon\delta}\right)^{\frac{m+d}{m}}\right)$. Thus, it might appear like the result of Theorem 6 is weaker than those based on PL-Unary and PL-Conditional. However, it is based on potentially weaker assumption, because PL-Pairwise might be decaying regardless of the labeling function just because having two λ-close points might be unlikely. In fact, as the following theorem shows, these results are incomparable (proof can be found in Appendix B.2.1):

Theorem 7 *Let the domain be $\mathcal{X} \subset [-1, 1]$ and the labeling function be $l(x) = \operatorname{sign}(x)$. Then there exist distributions D_1 and D_2 such that guarantees provided by Theorem 6 based on PL-Pairwise are stronger than those based on PL-Unary and PL-Conditional for D_1 and weaker for D_2.*

Now we show how Theorem 6, combined with Theorem 5 can be used to select a kernel for using it later with the nearest neighbor classification and obtaining faster learning rates.

4.1. Selecting a kernel for Nearest Neighbor

Assume that for all kernels K $r(K, \lambda)$ behaves as λ^{-d} and that there exists a kernel $K^* \in \mathcal{K}$ such that $\Phi(K, \lambda) = \lambda^m$ for some $m > d$. For fixed ϵ and δ select a kernel $\widehat{K}$ for which $\hat{\lambda}$ satisfying

$$\widehat{\Phi}(2\lambda, K)r(\lambda) < \epsilon^2\delta^2/2 \tag{18}$$

is the largest. For $\lambda^* \sim (\epsilon\delta)^{\frac{2}{m-d}}$ and $Tn \sim \tilde{O}\left(\left(\frac{1}{\epsilon\delta}\right)^{\frac{2m}{m-d}}\right)$:

$$r(\lambda^*)\alpha(-\log\lambda^*) < \epsilon^2\delta^2/4; \qquad r(\lambda^*)\Phi(3\lambda^*, K^*) < \epsilon^2\delta^2/4$$

Thus, according to Theorem 5:

$$r(\lambda^*)\widehat{\Phi}(2\lambda^*, K^*) \leq r(\lambda^*)(\Phi(3\lambda^*, K^*) + \alpha(-\log\lambda^*)) < \epsilon^2\delta^2/2. \tag{19}$$

Therefore, due to the choice of $\widehat{K}$, $\widehat{\lambda} \geq \lambda^*$. Thus, using Theorem 5 again:

$$r(\hat{\lambda})\Phi(\hat{\lambda}, \hat{K}) \leq r(\hat{\lambda})(\hat{\Phi}(2\hat{\lambda}, \hat{H}) + \alpha(-\log\hat{\lambda})) \leq \epsilon^2\delta^2/2 + r(\lambda)\alpha(-\log\lambda^*) < \epsilon^2\delta^2, \tag{20}$$

which means that (ϵ, δ)-multi-task-guarantees for Nearest Neighbor using kernel $\widehat{K}$ are implied with the number of sample per task n satisfying:

$$n \sim \sqrt{\frac{r(\widehat{\lambda})}{\Phi(\widehat{\lambda}, \widehat{K})}} \sim \left(\frac{1}{\epsilon\delta}\right)^{\frac{m+d}{m-d}}.$$

Thus one can identify a beneficial kernel and learn every given task using Nearest Neighbor with access to $T \sim \tilde{O}\left(\frac{1}{\epsilon\delta}\right)$ tasks and $n \sim O\left(\left(\frac{1}{\epsilon\delta}\right)^{\frac{m+d}{m-d}}\right)$ labeled samples per task. This example shows how overhead associated with estimation of $\Phi(\lambda, K)$ and identification of a beneficial kernel function spreads across the tasks and by having access to sufficiently many tasks one can recover the sample guarantees in terms of number of samples per task, as if that beneficial kernel was known in advance.

5. Application in active learning

The usefulness of the notion of Probabilistic Lipschitzness in cluster-based active learning was demonstrated in (Urner et al., 2013). The authors proposed a labeling procedure, called *PLAL*, that takes an unlabeled training set, queries labels of some of the sample points and returns a, possibly erroneous, full labeling of the sample. It starts with partitioning the input space into large regions and automatically assigns labels to those that seem to be label-homogeneous after a few queries, all the remaining clusters are refined and the procedure repeats recursively. The authors showed that it is possible to control the amount of incorrect labels that PLAL assigns and therefore this algorithm can be safely used as a pre-procedure to various learning methods, such as empirical risk or regularized empirical risk minimization. Moreover, the amount of queries that PLAL makes, and thus the potential savings in terms of label complexity that it offers, can be characterized using PL-Unary. As in the case of the sample complexity of Nearest Neighbor, the faster the decay of $\phi_U(\lambda)$, the fewer queries PLAL needs to produce a full labeling of the initial set. Thus, by selecting a data representation, under which this decay is fast, one could obtain significant savings in terms of label complexity. In the following we show that the number of queries that a refined version of PLAL, proposed in (Kpotufe et al., 2015), makes can be also characterized using PL-Pairwise.

The algorithm from (Kpotufe et al., 2015) (for completeness the pseudo-code is provided in Appendix C) takes as input accuracy and confidence parameters ϵ and δ, an unlabeled training set and a hierarchical partition of the input space:

Definition 8 *A hierarchical partition $P = \{P_l, l \in \mathbb{N}\}$ is a collection of partitions of the domain $\mathcal{X}$. Formally, for every $l \in \mathbb{N}$ T_l is a collection of disjoint sets $\mathcal{C}$ such that:*

- *for every $C \in \mathcal{C}$ $diam(C) = \sup_{x,x' \in C} \mathrm{dist}(x, x') \leq 2^{-l}$*
- *$\mathcal{X} \subset \cup_{C \in \mathcal{C}} C$*
- *every $C \in P_l$ has a parent C' in P_{l-1} such that $C \subset C'$*

The hierarchical partition P has tree-growth rate $\kappa \geq 1$ if for every l $|P_l| \leq 2^{\kappa l}$.

The following theorem provides the guarantees for this method in terms of PL-Pairwise:

Theorem 9 *Let $\mathcal{X}$ be the input domain equipped with a distance measure d with diameter at most 1. Let $0 < \epsilon, \delta < 1/2$ and $\tau = 1/2 - \epsilon/162$. Assume that the tasks $\langle D_1, l_1\rangle, \ldots, \langle D_T, l_T\rangle$ satisfy MT-PL-Pairwise with function $\Phi(\lambda)$ (with respect to the distance d). Let $X_1, \ldots, X_T$ be T training sets, where every X_i consists of n points sampled i.i.d. according to the task-specific marginal distribution D_i. Assume that $n \geq 81(16V_P \log(2n) + \log(8T/\delta))/\epsilon^2$, where V_P is the Vapnik-Chervonenkis dimension of the class of all clusters of P. Then the following holds:*

- *running algorithm from (Kpotufe et al., 2015) on each of X_t with parameters ϵ and δ/T produces a full labeling of X_t with at most ϵ-fraction of labels being incorrect with probability at least $1 - \delta$*
- *the expected number of queries that the algorithm makes, averaged across all tasks, is at most:*

$$\min_l \left(2^{\kappa l} \cdot 2\kappa l n\epsilon + 7\delta n + n \cdot \sqrt{\frac{2 \cdot 2^{\kappa l}\Phi(2^{-l})}{1 - 4\tau_\epsilon^2}} \right). \tag{21}$$

The first statement of the above theorem follows directly from (Kpotufe et al., 2015). The proof of the second statement can be found in Appendix C.1.

To better see the implications of this result and how it compares to the result of Kpotufe et al. (2015) , consider the case of one task, $T = 1$, and suppose the the data distribution satisfies polynomial Lipschitzness: $\phi_P(\lambda) = \lambda^m$. Then the expected number of queries, as guaranteed by Theorem 9, is at most:

$$\tilde{O}\left(\left(\frac{1}{\epsilon} \right)^{\frac{4\kappa+m}{\kappa+m}} \right). \tag{22}$$

Analogously to the case of the nearest neighbor classification, it leads to the conclusion that as the distribution gets nicer, i.e. $m \to \infty$, the query complexity of the labeling procedure reduces to $\tilde{O}\left(\frac{1}{\epsilon}\right)$. Under the same assumption on PL-Unary, from Theorem 11 in (Kpotufe et al., 2015) one obtains the label complexity of:

$$C2^{\kappa\alpha/(\kappa+\alpha)} \cdot \epsilon^{\alpha/(\kappa+\alpha)} \log(1/\epsilon) \cdot n = \tilde{O}\left(\left(\frac{1}{\epsilon} \right)^{\frac{2\kappa+\alpha}{\kappa+\alpha}} \right) \tag{23}$$

While it may seem like, as in the case of Nearest Neighbors, the guarantees provided by Theorem 9 are weaker than (23), both results lead to the same conclusion: as the distribution gets nicer, i.e. $\alpha \to \infty$, the query complexity of the labeling procedure reduces to $\tilde{O}\left(\frac{1}{\epsilon}\right)$. Moreover, in fact, the following theorem shows that they are incomparable (the proof can be found in Appendix C.2):

Theorem 10 *Let the domain be $\mathcal{X} \subset [-1, 1]$ and the labeling function be $l(x) = \text{sign}(x)$. Then there exist distributions D_1 and D_2 such that guarantees provided by Theorem 9 based on PL-Pairwise are stronger than those based on PL-Unary for D_1 and weaker for D_2.*

Now we show that in combination with Theorem 5, we obtain a method for selecting a kernel in a given kernel family, such that when used in algorithm from Kpotufe et al. (2015), it would lead to label complexity saving, while providing a full labeling of initially unlabeled samples which then can be safely used by any (regularized) empirical risk minimization method.

5.1. Selecting a kernel for active learning

Suppose that for all kernels in $\mathcal{K}$ there exists a hierarchical partition with tree-growth rate κ and that there exists a kernel K^* for which $\Phi(\lambda, K^*) = \lambda^m$ for some $m > \kappa$. Consider two cases.

Case 1: m is known. Let $\hat{K}$ be a kernel for which $\hat{\lambda}$ satisfying:

$$\frac{\hat{\Phi}(2\lambda, \hat{K})}{\lambda^\kappa} \leq \epsilon^{\frac{3(m-\kappa)}{m+\kappa}} \tag{24}$$

is the largest. For $\lambda^* \sim \epsilon^{\frac{3}{m+k}}$ and $Tn \sim \left(\frac{1}{\epsilon}\right)^{\frac{6m}{m+\kappa}}$ according to Theorem 5:

$$\frac{\hat{\Phi}(2\lambda^*, K^*)}{(\lambda^*)^\kappa} \leq \frac{\Phi(2\lambda^*, K^*) + \alpha(-\log \lambda^*)}{(\lambda^*)^\kappa} \leq \epsilon^{\frac{3(m-\kappa)}{m+\kappa}}$$

Thus, $\hat{\lambda} \geq \lambda^*$. Moreover, again by Theorem 5:

$$\frac{\Phi(\hat{\lambda}, \hat{K})}{\hat{\lambda}^\kappa} \leq \frac{\hat{\Phi}(2\hat{\lambda}, \hat{K}) + \alpha(-\log \hat{\lambda})}{\hat{\lambda}^\kappa} \leq 2\epsilon^{\frac{3(m-\kappa)}{m+\kappa}}. \tag{25}$$

Therefore, the expected average number of queries that algorithm from Kpotufe et al. (2015) based on kernel $\hat{K}$ will perform on T tasks is at most:

$$\left(\frac{\epsilon}{\hat{\lambda}^\kappa} + \sqrt{\frac{\Phi(\hat{\lambda}, \hat{K})}{\hat{\lambda}^\kappa \epsilon}}\right) \cdot n' \leq \left(\frac{\epsilon}{(\lambda^*)^\kappa} + \epsilon^{\frac{m-2\kappa}{m+\kappa}}\right) \cdot n' \sim \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{m+4\kappa}{m+\kappa}}\right), \tag{26}$$

which for large m is less than $\tilde{O}\left(\frac{1}{\epsilon^2}\right)$ and recovers the guarantees for label complexity when using a kernel with $\phi_P(\lambda) = \lambda^m$.

Case 2: m is unknown. Let $\hat{K}$ be a kernel for which $\hat{\lambda}$ satisfying:

$$\frac{\hat{\Phi}(2\lambda, \hat{K})}{\lambda^\kappa} \leq \epsilon^3 \tag{27}$$

is the largest. For $\lambda^* \sim \epsilon^{\frac{3}{m-k}}$ and $Tn \sim \left(\frac{1}{\epsilon}\right)^{\frac{6m}{m-\kappa}}$ according to Theorem 5:

$$\frac{\hat{\Phi}(2\lambda^*, K^*)}{(\lambda^*)^\kappa} \leq \frac{\Phi(2\lambda^*, K^*) + \alpha(-\log \lambda^*)}{(\lambda^*)^\kappa} \leq \epsilon^3$$

Thus, $\hat{\lambda} \geq \lambda^*$. By Theorem 5:

$$\frac{\Phi(\hat{\lambda}, \hat{K})}{\hat{\lambda}^\kappa} \leq \frac{\hat{\Phi}(2\hat{\lambda}, \hat{K}) + \alpha(-\log \hat{\lambda})}{\hat{\lambda}^\kappa} \leq 2\epsilon^3. \tag{28}$$

Therefore, the expected average number of queries that the algorithm from Kpotufe et al. (2015) based on kernel $\hat{K}$ will perform on T tasks is at most:

$$\left(\frac{\epsilon}{\hat{\lambda}^{\kappa}}+\sqrt{\frac{\Phi(\hat{\lambda},\hat{K})}{\hat{\lambda}^{\kappa}\epsilon}}\right)\cdot n' \leq \left(\frac{\epsilon}{(\lambda^*)^{\kappa}}+\epsilon\right)\cdot n' \sim \tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{m+2\kappa}{m-\kappa}}\right), \tag{29}$$

which again for large m leads to improvement over querying all $n' \sim O\left(\frac{1}{\epsilon^2}\right)$ labels.

Thus, in both cases, under an assumption of existence of a beneficial kernel (with sufficiently quickly decaying $\Phi(K,\lambda)$), using Theorem 5 one is able to identify a kernel that would lead to siginifcant label complexity reductions when used in active labeling procedure from Kpotufe et al. (2015) and thus, would lead to label complexity savings regardless of which risk minimization method is used later on for solving all tasks of interest.

6. Conclusion

In this work we have shown an alternative approach to multi-task representation learning that uses the notion of Pairwise-PL as a quality measure of kernel functions. We have shown that PL-Pairwise, in contrast to previously proposed versions of PL, can be reliably estimated from finite data. In particular, we demonstrated how to select a representation (a kernel function) that would lead to faster learning rates it terms of number of samples per task when used for the nearest neighbor classification or in active learning paradigm. However, our analysis of these two applications is limited by the assumption that the covering number of the input space in case of NN classification and tree-width of hierarchical partition in active learning behave equally for all kernels in the family. In practice, this assumption might not be satisfied and one would need to take it into account. In this work we focused only on statistical aspects of learning a kernel based on PL-Pairwise. Thus an important future direction is to develop an efficient algorithmic solution for the proposed approach.

Acknowledgments

This work was in parts funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

References

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2007a.

A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In *Conference on Neural Information Processing Systems (NIPS)*, 2007b.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3), 2008.

S. Ben-David and R. Urner. The sample complexity of agnostic learning under deterministic labels. In *Conference on Learning Theory (COLT)*, 2014.

R. Caruana. Multitask learning. *Machine Learning*, 1997.

T. Jebara. Multi-task feature and kernel selection for SVMs. In *International Conference on Machine Learning (ICML)*, 2004.

T. Jebara. Multitask sparsity via maximum entropy discrimination. *Journal of Machine Learning Research (JMLR)*, 2011.

S. Kpotufe, R. Urner, and S. Ben-David. Hierarchical label queries with data-dependent partitions. In *Conference on Learning Theory (COLT)*, 2015.

A. Kumar and H. Daumé III. Learning task grouping and overlap in multi-task learning. In *International Conference on Machine Learning (ICML)*, 2012.

S. Kushagra and S. Ben-David. Information preserving dimensionality reduction. In *International Conference on Algorithmic Learning Theory (ALT)*, 2015.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. van De Geer. Taking advantage of sparsity in multi-task learning. In *Conference on Learning Theory (COLT)*, 2009.

A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research (JMLR)*, 2006.

A. Maurer and M. Pontil. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory (COLT)*, 2013.

A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *International Conference on Machine Learning (ICML)*, 2013.

A. Maurer, M. Pontil, and B. Romera-Paredes. An inequality with applications to structured sparsity and multitask dictionary learning. In *Conference on Learning Theory (COLT)*, 2014.

A. Pentina and S. Ben-David. Multi-task and lifelong learning of kernels. In *International Conference on Algorithmic Learning Theory (ALT)*, 2015.

B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

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

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Conference on Learning Theory (COLT)*, 2006.

R. Urner and S. Ben-David. Probabilistic Lipschitzness: A niceness assumption for deterministic labels. In *Learning Faster from Easy Data - Workshop @ NIPS*, 2013.

R. Urner, S. Shalev-Shwartz, and S. Ben-David. Access to unlabeled data can speed up prediction time. In *International Conference on Machine Learning (ICML)*, 2011.

R. Urner, S. Wulff, and S. Ben-David. PLAL: Cluster-based active learning. In *Conference on Learning Theory (COLT)*, 2013.

P. Yang, K. Huang, and C.-L. Liu. Geometry preserving multi-task metric learning. *Machine Learning*, 2013.

Y. Zhang and D.-Y. Yeung. Multi-task learning in heterogeneous feature spaces. In *Conference on Artificial Intelligence (AAAI)*, 2011.

Y. Zhou, R. Jin, and S. C. H. Hoi. Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Appendix A. Supplementary results

For a sample $\mathbf{x} = \{x_1, \ldots, x_n\}$ define:

$$D^{\mathbf{x}}_{\infty}(K, \tilde{K}) = \max_{i,j} |K(x_i, x_j) - \tilde{K}(x_i, x_j)|. \tag{30}$$

Definition 11 *The uniform ℓ_∞ kernel covering number $\mathcal{N}_{n,\infty}(\mathcal{K}, \epsilon)$ of a kernel family $\mathcal{K}$ is given by considering all possible samples $\mathbf{x}$ of size n:*

$$\mathcal{N}_{n,\infty}(\mathcal{K}, \epsilon) = \sup_{\mathbf{x}} \mathcal{N}_{D^{\mathbf{x}}_{\infty}}(\mathcal{K}, \epsilon). \tag{31}$$

Lemma 12 (Lemma 3 in (Srebro and Ben-David, 2006)) *For any kernel family $\mathcal{K}$ bounded by B^2 with pseudodimension p:*

$$\mathcal{N}_{n,\infty}(\mathcal{K}, \epsilon) \leq \left(\frac{en^2B^2}{\epsilon p}\right)^p. \tag{32}$$

Define:

$$D^{\mathbf{x}}_{\infty,\|\cdot\|^2}(K, \tilde{K}) = \max_{i,j} \left| \|x_i - x_j\|^2_K - \|x_i - x_j\|^2_{\tilde{K}} \right|, \tag{33}$$

$$\mathcal{N}^{\|\cdot\|^2}_{n,\infty}(\mathcal{K}, \epsilon) = \sup_{\mathbf{x}} \mathcal{N}_{D^{\mathbf{x}}_{\infty,\|\cdot\|^2}}(\mathcal{K}, \epsilon). \tag{34}$$

By definition:

$$\|x - x'\|^2_K = K(x, x) - 2K(x, x') + K(x', x'). \tag{35}$$

Therefore:

$$D^{\mathbf{x}}_{\infty,\|\cdot\|^2}(K, \tilde{K}) \leq 4D^{\mathbf{x}}_{\infty}(K, \tilde{K}) \tag{36}$$

and, consequently:

Lemma 13 *For any kernel family $\mathcal{K}$ bounded by B^2 with pseudodimension p:*

$$\mathcal{N}^{\|\cdot\|^2}_{n,\infty}(\mathcal{K}, \epsilon) \leq \left(\frac{4en^2B^2}{\epsilon p}\right)^p. \tag{37}$$

Appendix B. Nearest Neighbor

B.1. Proof of Theorem 6

Partition the domain $\mathcal{X}$ with $r(\lambda)$ sets with diameter λ for some $\lambda > 0$. For any $x \in \mathcal{X}$ denote by $C(x)$ (any of) the set it belongs to and by $NN_S(x)$ its nearest neighbor in the set S. Then:

$$\mathrm{er}_D(NN(S)) = \Pr_{x\sim D}[\![l(NN_S(x)) \neq l(x)]\!] \leq \Pr_{x\sim D}[S \cap C(x) = \emptyset] + \tag{38}$$

$$\Pr_{x\sim D}[NN_S(x) \in C(x) \wedge l(x) \neq l(NN_S(x))]. \tag{39}$$

By Lemma 19.2 in (Shalev-Shwartz and Ben-David, 2014):

$$\mathop{\mathbf{E}}_{S} \Pr_{x\sim D}[S \cap C(x) = \emptyset] \leq \frac{r(\lambda)}{ne}. \tag{40}$$

To analyze the second term we use the definition of ϕ:

$$\mathop{\mathbf{E}}_{S} \Pr_{x\sim D}[NN_S(x) \in C(x) \wedge l(x) \neq l(NN_S(x))] =$$

$$\mathop{\mathbf{E}}_{S} \mathop{\mathbf{E}}_{x}[\![NN_S(x) \in C(x) \wedge l(x) \neq l(NN_S(x)) \wedge \mathrm{dist}(x, NN_S(x)) \leq \lambda] = \tag{41}$$

$$\Pr_{S,x}[l(x) \neq NN_S(x) \wedge \mathrm{dist}(x, NN_S(x)) \leq \lambda] \leq \tag{42}$$

$$\Pr_{S,x}[\exists i \in [0,n] : \ \mathrm{dist}(x_i, x) \leq \lambda \wedge l(x_i) \neq l(x)] \leq \tag{43}$$

$$n \Pr_{x_1,x}[\mathrm{dist}(x_1, x) \leq \lambda \wedge l(x_1) \neq l(x)] \leq \tag{44}$$

$$n\phi(\lambda). \tag{45}$$

By combining (40) and (45) we obtain the statement of the theorem.

B.2. Comparison to previously known results

In terms of PL-Unary and PL-Conditional the sample complexity of Nearest Neighbor can be characterized as follows:

Theorem 14 (Urner and Ben-David (2013)) *Let the domain be $\mathcal{X} \subset [0,1]^d$. Assume that the labeling function l is deterministic and satisfies PL-Unary with function ϕ_U. Then the sample complexity of Nearest Neighbor is upper bounded by:*

$$\Pr_S[\mathrm{er}_D(NN(S)) > \epsilon] \leq \frac{2}{\epsilon ne}\left(\frac{\sqrt{d}}{\phi_U^{-1}(\epsilon/2)}\right)^d \tag{46}$$

Theorem 15 (Kushagra and Ben-David (2015)) *Let the domain be $\mathcal{X} \subset [0,1]^d$. Assume that the labeling function l is deterministic and satisfies PL-Conditional with function ϕ_C. Then the sample complexity of Nearest Neighbor is upper bounded by:*

$$\Pr_S[\mathrm{er}_D(NN(S)) > \epsilon] \leq \min_{\lambda\in(0,1)}\left(\frac{1}{ne\epsilon}\left(\frac{\sqrt{d}}{\lambda}\right)^d + \frac{\phi_C(\lambda)}{\epsilon}\right). \tag{47}$$

Similarly, from Theorem 6, we obtain:

Theorem 16 *Let the domain be $\mathcal{X} \subset [0,1]^d$ and $S = \{(x_1, l(x_1), \ldots, (x_n, l(x_n)\}$ be the training set of n points sampled i.i.d. according to D. Assume that the labeling function l is deterministic and satisfies PL-Pairwise with function ϕ_P. Then for any $\epsilon > 0$:*

$$\Pr_S[\mathrm{er}_D(NN(S)) > \epsilon] \leq \min_{\lambda \in (0,1)} \left(\frac{1}{ne\epsilon} \left(\frac{\sqrt{d}}{\lambda} \right)^d + \frac{n\phi_P(\lambda)}{\epsilon} \right). \tag{48}$$

B.2.1. Proof of Theorem 7

Example 1. Let $\mathcal{X}_+$ be a set of points on $[-1,1]$ of form $\frac{1}{2^k}$ and let $\mathcal{X}_-$ be a set of points of form $-\frac{1}{2^k}$ for $k \geq 0$. Let $\mathcal{X}$ be a union of these two sets. All points in $\mathcal{X}_-$ are assigned label -1 and all points in X_+ are assigned label $+1$. Finally, assume the following marginal distribution:

$$p\left(\frac{1}{2^k}\right) = \frac{15}{32} \cdot \frac{1}{2^{4k}}$$
$$p\left(-\frac{1}{2^k}\right) = \frac{3}{8} \cdot \frac{1}{2^{2k}}$$

Fix $\lambda > 0$ such that $\frac{1}{2^k} < \lambda \leq \frac{1}{2^{k-1}}$ for some $k > 1$. Now we can compute the values of PL-Unary, PL-Pairwise and PL-Conditional for this distribution.

1. PL-Unary

$$\Pr_x \left(\Pr_y \left(l(x) \neq l(y) \wedge \mathrm{dist}(x,y) \leq \lambda \right) > 0 \right) = \sum_{n \geq k} p\left(-\frac{1}{2^n}\right) + \sum_{n \geq k} p\left(\frac{1}{2^n}\right) =$$
$$\sum_{n \geq k} \frac{3}{8} \cdot \frac{1}{2^{2n}} + \sum_{n \geq k} \frac{15}{32} \cdot \frac{1}{2^{4n}} = \frac{3}{8} \cdot \frac{1}{2^{2k}} \sum_{n \geq 0} \frac{1}{2^{2n}} + \frac{15}{32} \cdot \frac{1}{2^{4k}} \sum_{n \geq 0} \frac{1}{2^{4n}} = \frac{1}{2} \cdot \frac{1}{2^{2k}} + \frac{1}{2} \cdot \frac{1}{2^{4k}} \leq$$
$$0.5\lambda^2 + 0.5\lambda^4 \leq \lambda^2$$

2. PL-Pairwise

$$\Pr_{x,y} \left(l(x) \neq l(y) \wedge \mathrm{dist}(x,y) \leq \lambda \right) = \Pr_y \left(l(x) \neq l(y) \wedge \mathrm{dist}(x,y) \leq \lambda \middle| x \right) \Pr(x) =$$
$$\sum_{n \geq k} p\left(-\frac{1}{2^n}\right) \cdot \left(\sum_{m \geq k} p\left(\frac{1}{2^m}\right) \right) + \sum_{n \geq k} p\left(\frac{1}{2^n}\right) \cdot \left(\sum_{m \geq k} p\left(-\frac{1}{2^m}\right) \right) =$$
$$2 \left(\sum_{n \geq k} \frac{3}{8} \cdot \frac{1}{2^{2n}} \right) \cdot \left(\sum_{m \geq k} \frac{15}{32} \cdot \frac{1}{2^{4m}} \right) = \frac{2 \cdot 3 \cdot 15}{8 \cdot 32} \cdot \frac{1}{2^{2k}} \cdot \frac{1}{2^{4k}} \left(\sum_{n \geq 0} \frac{1}{2^{2n}} \right) \left(\sum_{m \geq 0} \frac{1}{2^{4m}} \right) \leq 0.5\lambda^6$$

3. PL-Conditional

$$\Pr_{x,y} \left(\mathrm{dist}(x,y) \leq \lambda \right) \geq \sum_{n \geq k} p\left(-\frac{1}{2^n}\right) \left(\sum_{m \geq k} p\left(-\frac{1}{2^m}\right) \right) = \frac{9}{64} \frac{1}{2^{2k}} \cdot \frac{1}{2^{2k}} \cdot \frac{16}{9} = \frac{1}{4} \cdot \frac{1}{2^{4k}} \geq 4\lambda^4$$

$$\Pr_{x,y}\left(\operatorname{dist}(x,y)\leq\lambda\right)\leq\sum_{n\geq k-1}p\left(-\frac{1}{2^n}\right)\left(\sum_{m\geq k-1}p\left(-\frac{1}{2^m}\right)+\sum_{m\geq k-1}p\left(\frac{1}{2^m}\right)\right)+$$
$$\sum_{n\geq k-1}p\left(\frac{1}{2^n}\right)\left(\sum_{m\geq k-1}p\left(-\frac{1}{2^m}\right)+\sum_{m\geq k-1}p\left(\frac{1}{2^m}\right)\right)=\left(\sum_{n\geq k-1}p\left(-\frac{1}{2^n}\right)\right)^2+$$
$$\left(\sum_{n\geq k-1}p\left(\frac{1}{2^n}\right)\right)^2+2\left(\sum_{n\geq k-1}p\left(-\frac{1}{2^n}\right)\right)\left(\sum_{n\geq k-1}p\left(\frac{1}{2^n}\right)\right)=\left(\frac{1}{2}\cdot\frac{1}{2^{4(k-1)}}\right)^2+$$
$$\left(\frac{1}{2}\cdot\frac{1}{2^{2(k-1)}}\right)^2+2\frac{1}{2}\cdot\frac{1}{2^{4(k-1)}}\cdot\frac{1}{2}\frac{1}{2^{2(k-1)}}=\frac{2^6}{2^{8k}}+\frac{4}{2^{4k}}+\frac{2^5}{2^{6k}}\leq$$
$$64\lambda^8+4\lambda^4+32\lambda^6\leq 100\lambda^4$$

$$\Pr_{x,y}\left(l(x)\neq l(y)\middle|\operatorname{dist}(x,y)\leq\lambda\right)=\frac{\Pr_{x,y}\left(l(x)\neq l(y)\wedge\operatorname{dist}(x,y)\leq\lambda\right)}{\Pr_{x,y}\left(\operatorname{dist}(x,y)\leq\lambda\right)}=\Theta(\lambda^2)$$

Consequently, bounds for sample complexity of the Nearest Neighbor classifier in terms of PL-Unary and PL-Conditional give:

$$O\left(\left(\frac{1}{\epsilon}\right)^{\frac{2+1}{2}}\right)=O\left(\left(\frac{1}{\epsilon}\right)^{\frac{3}{2}}\right)\tag{49}$$

while based on PL-Pairwise we obtain:

$$O\left(\left(\frac{1}{\epsilon}\right)^{\frac{6+1}{6-1}}\right)=O\left(\left(\frac{1}{\epsilon}\right)^{\frac{7}{5}}\right)\tag{50}$$

EXAMPLE 2. Let $\mathcal{X}$ be $[-1,1]$ and the marginal distribution have the following density function:

$$p(x)=|x|.\tag{51}$$

Fix $\frac{1}{2}\lambda>0$. Then:

1. PL-Unary

$$\Pr_x\left(\Pr_y\left(l(x)\neq l(y)\wedge\operatorname{dist}(x,y)\leq\lambda\right)>0\right)=\int_{-\lambda}^{\lambda}|x|dx=\lambda^2$$

2. PL-Pairwise

$$\Pr_{x,y}\left(l(x)\neq l(y)\wedge\operatorname{dist}(x,y)\leq\lambda\right)=2\int_0^{\lambda}x\int_{x-\lambda}^{0}(-y)dydx=\frac{5}{12}\lambda^4$$

3. PL-Conditional

$$\Pr_{x,y}\left(\operatorname{dist}(x,y)\leq\lambda\right)\geq\int_0^{1-\lambda}x\int_x^{x+\lambda}ydydx=\frac{1}{2}\int_0^{1-\lambda}x(2\lambda x+\lambda^2)dx=$$
$$\frac{\lambda(1-\lambda)^3}{3}+\frac{\lambda^2(1-\lambda)^2}{4}\geq\frac{\lambda}{24}$$

$$\Pr_{x,y}\left(l(x) \neq l(y) \middle| \operatorname{dist}(x,y) \leq \lambda\right) = \frac{\Pr_{x,y}\left(l(x) \neq l(y) \wedge \operatorname{dist}(x,y) \leq \lambda\right)}{\Pr_{x,y}\left(\operatorname{dist}(x,y) \leq \lambda\right)} \leq \frac{5 \cdot 24 \cdot \lambda^4}{12\lambda} = 10\lambda^3$$

Consequently, bounds for sample complexity of the Nearest Neighbor classifier in terms of PL-Unary and PL-Conditional give:

$$O\left(\left(\frac{1}{\epsilon}\right)^{\frac{2+1}{2}}\right) = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{3}{2}}\right) \tag{52}$$

and:

$$O\left(\left(\frac{1}{\epsilon}\right)^{\frac{3+1}{3}}\right) = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{4}{3}}\right) \tag{53}$$

while based on PL-Pairwise we obtain:

$$O\left(\left(\frac{1}{\epsilon}\right)^{\frac{4+1}{4-1}}\right) = O\left(\left(\frac{1}{\epsilon}\right)^{\frac{5}{3}}\right) \tag{54}$$

Appendix C. Active learning

C.1. Proof of Theorem 9

We begin by repeating arguments from the proof of Theorem 7 in (Kpotufe et al., 2015).

Fix one task t and any level l. The number of labels requested up to level l (including those requested at level l) is bounded by $n \cdot \epsilon \cdot 2^{\kappa l}(2\kappa l)$. Now we bound the number of labels yet to request at later levels. For any x let $\mathcal{C}^l(x)$ denote the cluster at level l that it belongs to. Define:

$$\mathcal{X}_l = \left\{x \in \mathcal{X} : \left|\eta_{C^l(x)} - \frac{1}{2}\right| \geq \tau_\epsilon\right\} \tag{55}$$

Consider any $x \in \mathcal{X}_l$. If $\mu_n(\mathcal{C}^l(x)) < \epsilon$ by construction x is labeled at level l. If instead $\mu_n(\mathcal{C}^l(x)) \geq \epsilon$ by combining Lemma 16 and Corollary 14 in (Kpotufe et al., 2015) one can show that with high probability $\hat{\eta}_S \leq \epsilon/3$ and thus all of $\mathcal{C}^l(x)$ is labeled by the procedure. Combining it all together, Kpotufe et al. (2015) have shown that with probability at least $1 - 7\delta$ for every level l, for every $x \in \mathcal{X}_l$ cluster $\mathcal{C}^l(x)$ is labeled by the procedure at level l. Thus, the number of points left to label is at most $|X \setminus \mathcal{X}_l|$. In the original proof of Kpotufe et al. (2015) this quantity is upper-bounded using the notion of clusterability of labels. We, instead will use PL-Pairwise.

Define $\bar{\mathcal{C}}^l$ to be the set of all clusters at level l for which:

$$\left|\eta_C - \frac{1}{2}\right| < \tau_\epsilon. \tag{56}$$

We need to bound $\mu(\mathcal{X} \setminus \mathcal{X}_l) = \mu(\cup_{C \in \bar{\mathcal{C}}^l} C) = \sum_{C \in \bar{\mathcal{C}}^l} \mu(C)$. First, note that for every $C \in \bar{\mathcal{C}}^l$:

$$\Pr[l_t(x) \neq l_t(x') | x, x' \in C] = 2\eta(C)(1 - \eta(C)) \geq \frac{1}{2} - 2\tau_\epsilon^2. \tag{57}$$

With this we obtain the following sequence of derivations with $r = 2^{-l}$:

$$\Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') \wedge \|x - x'\| \leq r] \geq \sum_{C \in \bar{\mathcal{C}}^l} \Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') \wedge x, x' \in C] =$$

$$\sum_{C \in \bar{\mathcal{C}}^l} \Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') | x, x' \in C] \Pr_{x \sim D_t}[x \in C] \Pr_{x' \sim D_t}[x' \in C] \geq \sum_{C \in \bar{\mathcal{C}}^l} \left(\frac{1}{2} - 2\tau_\epsilon^2\right) \mu^2(C) \geq$$

$$\left(\frac{1}{2} - 2\tau_\epsilon^2\right) \frac{\left(\sum_{C \in \bar{\mathcal{C}}^l} \mu(C)\right)^2}{|\bar{\mathcal{C}}^l|} \geq \left(\frac{1}{2} - 2\tau_\epsilon^2\right) \frac{\left(\sum_{C \in \bar{\mathcal{C}}^l} \mu(C)\right)^2}{2^{\kappa l}}.$$

Therefore:

$$\sum_{C \in \bar{\mathcal{C}}^l} \mu(C) \leq \sqrt{\frac{2 \cdot 2^{\kappa l} \Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') \wedge \|x - x'\| \leq 2^{-l}]}{1 - 4\tau_\epsilon^2}}. \tag{58}$$

Thus, for every task t the expected number of label queries is bounded by:

$$\min_l \left(2^{\kappa l} \cdot 2\kappa l n \epsilon + 7\delta n + \sqrt{\frac{2 \cdot 2^{\kappa l} \Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') \wedge \|x - x'\| \leq 2^{-l}]}{1 - 4\tau_\epsilon^2}} \right). \tag{59}$$

Therefore, the expected number of queries, averaged across all T tasks is upper bounded by:

$$\min_l \left(2^{\kappa l} \cdot 2\kappa l n \epsilon + 7\delta n + \frac{1}{T} \sum_{t=1}^{T} \sqrt{\frac{2 \cdot 2^{\kappa l} \Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') \wedge \|x - x'\| \leq 2^{-l}]}{1 - 4\tau_\epsilon^2}} \right) \leq$$

$$\min_l \left(2^{\kappa l} \cdot 2\kappa l n \epsilon + 7\delta n + \sqrt{\frac{2 \cdot 2^{\kappa l} \frac{1}{T} \sum_{t=1}^{T} \Pr_{x,x' \sim D_t}[l_t(x) \neq l_t(x') \wedge \|x - x'\| \leq 2^{-l}]}{1 - 4\tau_\epsilon^2}} \right) \leq$$

$$\min_l \left(2^{\kappa l} \cdot 2\kappa l n \epsilon + 7\delta n + \sqrt{\frac{2 \cdot 2^{\kappa l} \Phi(2^{-l})}{1 - 4\tau_\epsilon^2}} \right),$$

which concludes the proof.

C.2. Proof of Theorem 10

EXAMPLE 1. Consider the marginal distribution with density function $p(x) = 0.5$ on $[-1, 0]$ and $p(x) = 2.5x^4$ on $(0, 1]$. Then:

PL-Unary:

$$\Pr_x \left(\Pr_y \left(l(x) \neq l(y) \wedge \text{dist}(x, y) \leq \lambda \right) > 0 \right) = \int_{-\lambda}^{0} 0.5 dx + \int_{0}^{\lambda} 2.5x^4 dx = 0.5\lambda + 0.5\lambda^5$$

PL-Pairwise:

$$\Pr_{x,y} \left(l(x) \neq l(y) \wedge \text{dist}(x, y) \leq \lambda \right) = \int_{-\lambda}^{0} 0.5 \int_{0}^{x+\lambda} 2.5y^4 dy dx + \int_{0}^{\lambda} 2.5x^4 \int_{x-\lambda}^{0} 0.5 dy dx =$$

$$\frac{5}{4} \int_{-\lambda}^{0} \frac{(x + \lambda)^5}{5} dx + \frac{5}{4} \int_{0}^{\lambda} x^4 (\lambda - x) dx = \frac{1}{4} \frac{\lambda^6}{6} + \frac{5}{4} \left(\frac{\lambda^6}{5} - \frac{\lambda^6}{6} \right) = \frac{\lambda^6}{12}$$

Suppose that as hierarchical partition one uses dyadic tree and so $\kappa = 1$. Then the guarantees provided by Theorem 9 result in label complexity of $\tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{10}{7}}\right)$. At the same time, based on PL-Unary, we obtain label complexity of $\tilde{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{3}{2}}\right)$.

Example 2. Consider example 2 from Theorem 7.

Algorithm 1 Labeling procedure from (Kpotufe et al., 2015)

Input parameters X, P, ϵ, δ
set the active cluster set $\mathcal{C}^l = P_0$
for $l = 0, 1, \ldots$ **do**
  $\delta_l = \delta / |\mathcal{C}^l| 2^{l+1}$
  $n_l(\epsilon) = 9' log(8/\delta_l)/\epsilon$
  for each $C \in \mathcal{C}^l$ **do**
    if $\mu_n(C) < \epsilon$ **then**
      request all labels for points in $C \cap X$ and skip to the next cluster in $\mathcal{C}^l$
    end if
    S =labeled sample from $C \cap X$ (with replacement) of size $n_l(\epsilon)$
    $\hat{\eta}_S$ - probability of label 1 in S
    if $\min\{\hat{\eta}_S, 1 - \hat{\eta}_S\} \leq \epsilon/3$ **then**
      label all $C \cap X$ with the majority label from S
    else
      add children of C to $\mathcal{C}^{l+1}$
    end if
  end for
  if all of X is labeled **then**
    return labeled sample X
  end if
end for