Probabilistic Lipschitzness
A niceness assumption for deterministic labels

Ruth Urner
Cheriton School of Computer Science
Waterloo, ON, N2L 3G1
CANADA
rurner@uwaterloo.ca

Shai Ben-David
Cheriton School of Computer Science
Waterloo, ON, N2L 3G1
CANADA
shai@uwaterloo.ca

Abstract

We present Probabilistic Lipschitzness (PL), a notion of marginal label relatedness that is particularly useful for modeling niceness of distributions with deterministic labeling functions. We present convergence rates for Nearest Neighbor learning under PL. We further summarize reductions in labeled sample complexity for learning with unlabeled data (semi-supervised and active learning) under PL.

1 Introduction

The notion of learnability commonly employed in Machine Learning theory considers a worst case scenario: In order to meet the success criteria, an algorithm has to perform well with respect to all possible data generating distributions. However, applications often perform better than what is suggested by the lower bounds in that framework of analysis. An important challenge for theoretical machine learning research is therefore to develop models that provide a better explanation of our practical experience. Such models should employ assumptions about the learning tasks that are both realistic (in that we can expect real world tasks to comply with them), and accessible to mathematical analysis that yields more optimistic performance guarantees.

In this paper, we discuss Probabilistic Lipschitzness (PL), an assumption that aims to meet both these requirements. In a nutshell, PL is a measure of the coherence between the data's marginal distribution and its labeling rule. It quantifies the extent to which similar instances tend to have similar labels. Such a coherence is implicit in many machine learning algorithmic paradigms. We describe how, PL assumptions provably reduce the sample complexity of learning in two classification prediction regimes; nonparametric (Nearest Neighbor) learning of label-deterministic distributions, and utilizing unlabeled data (in semi-supervised and in active learning). We focus on label-deterministic distributions. We believe that the non-realizable deterministic labeling setting has not been fully addressed so far. In particular, we are not aware of any niceness assumptions for distributions with deterministic labellings that yield better than worst case bounds. However, our results for Nearest Neighbor learning under PL also hold under non-deterministic labeling rules.

The Tsybakov noise condition is probably the most prolific formal assumption of "data niceness" that yields improved sample complexity bounds (Mammen and Tsybakov [1999]). However, those results address learning with respect to a hypothesis class of bounded capacity, and rely on assuming that the Bayes optimal classifier belongs to that class, or at least can be well approximated by functions in the class (Tsybakov [2004],Boucheron et al. [2005]). When these assumptions are not met, even label-deterministic data distributions (that have zero noise) can result in slow convergence rates. In contrast, we consider non-parametric learning (nearest neighbor algorithms) and show how PL assumptions save samples in that setting.

The discrepancy between worst case lower bounds from learning theory and success in practice is particularly apparent for settings that exploit information from unlabeled data, such as semi-

supervised learning or active learning. For both these learning regimes, there are lower bounds on the label complexity in the presence of unlabeled data that match the sample complexity of learning without access to unlabeled data (Ben-David et al. [2008], Kääriäinen [2006], Raginsky and Rakhlin [2011]). These formally indicate that unlabeled data is not beneficial in worst-case scenarios. Despite these pessimistic prospects, many practical applications successfully boost learning performance by taking information from unlabeled data into account. Under PL assumptions, the label complexity of certain types of semi-supervised learning and active learning have label complexity that is provably below those worst-case lower bounds.

This work is a summary of previously published results on semi-supervised and active learning under PL (Urner et al. [2011, 2013]) as well as of the discussion in Chapter 3 of Urner [2013]. Probabilistic Lipschitzness was introduced by Urner et al. [2011]. A very similar notion (the *margin exponent*) has been proposed earlier by Steinwart and Scovel [2007]. There, the margin exponent is used to bound the approximation error of Gaussian kernels.

Notation We model a *learning task* as some distribution P over $\mathcal{X} \times \{0,1\}$ for some domain $\mathcal{X}$. We denote the marginal distribution of P over $\mathcal{X}$ by $P_{\mathcal{X}}$ and let $l : \mathcal{X} \to [0,1]$ denote the induced conditional label probability function, $l(x) = P(y = 1|x)$, and call it the *labeling function* of P. We say that the labeling function is *deterministic*, if $l(x) \in \{0,1\}$ for all $x \in \mathcal{X}$. For some *hypothesis* $h : \mathcal{X} \to \{0,1\}$ we define the *error* of h with respect to P as $\mathrm{Err}_P(h) = \Pr_{(x,y)\sim P}[y \neq h(x)]$.

2 Probabilistic Lipschitzness

Motivation Many common learning paradigms implicitly rely on the labeling function to comply with the geometry of the space, or, put more simply, they rely on close-by points being likely to have the same label (or conditional label probability). Obvious examples of such paradigms are Nearest Neighbor methods or algorithms that classify with halfspaces (or other geometrically defined classifiers). The success of such paradigms suggests that, for many label prediction tasks, there is a significant correlation between the geometry of the space, the marginal distribution over the data points and the labels. Under a suitable data representation, or feature choice, we expect that the closer two instances are, the less likely they are to have different labels. Probabilistic Lipschitzness is a measure that quantifies this correlation.

Definition We define Probabilistic Lipschitzness as a relaxation of standard Lipschitzness. While the standard Lipschitz condition can be readily applied to probabilistic labeling rules $l : X \to [0,1]$, it has strong implications in the case of deterministic labeling functions. A Lipschitz constant λ for a distribution with deterministic labeling function $l : X \to \{0,1\}$ forces a $1/\lambda$ gap between differently labeled points. Thus, it that the data lies in label homogeneous regions (clusters) that are separated by $1/\lambda$-margins, thus enforces the domain to be disconnected or the labeling function to be a constant. This is a rather strong assumption of label conform clusterability. PL weakens this assumptions by allowing the margins to "smoothen out". The relaxation from Lipschitzness to Probabilistic Lipschitzness is thus particularly relevant to the deterministic labeling regime. It allows to model the marginal-label relatedness without trivializing the setup.

Definition 1 (Probabilistic Lipschitzness)**.** Let $(\mathcal{X}, \mu)$ be some metric space and let $\phi : \mathbb{R} \to [0,1]$. We say that $f : \mathcal{X} \to \mathbb{R}$ is ϕ-Lipschitz with respect to a distribution $P_{\mathcal{X}}$ over $\mathcal{X}$ if, for all $\lambda > 0$:

$$\Pr_{x\sim P_{\mathcal{X}}}\left[\ \Pr_{y\sim P_{\mathcal{X}}}\ [\,|f(x)-f(y)|\ >\ 1/\lambda\,\mu(x,y))\,]\ >\ 0\right]\ \leq\ \phi(\lambda)$$

If, for some distribution $P = (P_{\mathcal{X}}, l)$, the labeling function l is ϕ-Lipschitz, then we also say that P satisfies the ϕ-Probabilistic Lipschitzness.

This definition generalizes the standard definition of Lipschitzness in the following way: If the labeling function l of a distribution is L-Lipschitz then it satisfies Probabilistic Lipschitzness with the function $\phi(\lambda) = 1$ if $\lambda \geq 1/L$ and $\phi(\lambda) = 0$ if $\lambda < 1/L$. If a distribution $P = (P_{\mathcal{X}}, l)$ with a deterministic labeling function l is ϕ-Lipschitz, then the weight of points x that have a positive mass of points of opposite label in an λ-ball around them, is bounded by $\phi(\lambda)$.

Examples of Probabilistic Lipschitzness

Linear Separators Let $P_{\mathcal{X}}$ be uniform over $\mathcal{X} = [0,1]^d$. If l is a linear separator, then $\phi(\lambda) = C \cdot \lambda$, for some constant C. This example also appears in Steinwart and Scovel [2007].

Ball Let $P_{\mathcal{X}}$ be the uniform distribution over $\mathcal{X} = [0,1]^d$. For some ball $B \subseteq \mathcal{X}$ we let l label points in B with label 1 and points outside B with label 0. Then the Probabilistic Lipschitzness is bounded by $\phi(\lambda) = C \cdot \lambda^d$ for some constant C.

Generalized Clusters To demonstrate how the marginal distribution influences the Probabilistic Lipschitzness, we consider distributions over $\mathcal{X} = [0,1]$ and let the labeling function l be 0 for $x \leq 1/2$ and 1 for $x > 1/2$. Now we let the density d of the distribution form clusters by setting $d(x) = c \cdot (x - 1/2)^\alpha$ for a suitable constant c (that ensures that d is a density function). Then we have $\phi(\lambda) = C \cdot \lambda^{\alpha+1}$ for some constant C.

For an example that satisfies an even stronger PL condition, we consider distributions over $\mathcal{X} = [0,1]$ and let the labeling function l be 0 for $x \leq 1/2$ and 1 for $x > 1/2$. Let the density d of the distribution be $d(1/2) = 0$ and $d(x) = c\mathrm{e}^{-1/|1/2-x|}$ for $x \neq 1/2$, for a suitable constant c (that ensures that d is a density function). Then we have $\phi(\lambda) = C \cdot \mathrm{e}^{-1/\lambda}$ for some constant C.

3 Learning under Probabilistic Lipschitzness

In this section, we consider the domain $\mathcal{X} = [0,1]^d$. We denote the set of all such distributions over $[0,1]^d \times \{0,1\}$ with deterministic labeling functions by $\mathcal{Q}^d_{\phi,\mathrm{det}}$. Given some PL-function ϕ and some ϵ, we let $\phi^{-1}(\epsilon)$ denote the smallest λ, such that $\phi(\lambda) \geq \epsilon$. In the following bounds, ϵ and δ denote the usual accuracy and confidence parameters.

Upper bound We start with bounding the sample complexity of Nearest Neighbor learning with respect to the class of distributions that have deterministic labeling functions and satisfy PL.

Theorem 2. *Let $\phi : \mathbb{R} \to [0,1]$. The sample complexity $m[\mathrm{NN}, \mathcal{Q}^d_{\phi,\mathrm{det}}]$ of the Nearest Neighbor algorithm with respect to the class $\mathcal{Q}^d_{\phi,\mathrm{det}}$, is bounded by*

$$m[\mathrm{NN}, \mathcal{Q}^d_{\phi,\mathrm{det}}](\epsilon, \delta) \leq \frac{2}{\epsilon\,\delta\,\mathrm{e}} \left(\frac{\sqrt{d}}{\phi^{-1}(\epsilon/2)} \right)^d .$$

Lower bound We now present a lower bounds on the sample complexity of any learning algorithm for distributions that satisfy Probabilistic Lipschitzness. It almost matches the upper bound in the theorem above.

Theorem 3. *Let $\phi : \mathbb{R} \to [0,1]$ with $\phi(1) \geq 1$. For every learning algorithm $\mathcal{A}$ and every $\epsilon > 0$ there exists a distribution $P \in \mathcal{Q}^d_{\phi,\mathrm{det}}$ such that, a sample size m with*

$$m < \frac{d}{32\epsilon} \left(\frac{1}{\phi^{-1}(8\epsilon)} \right)^{d-1}$$

implies that

$$\mathbb{E}_{S \sim P^m}[\mathrm{Err}_P(\mathcal{A}(S))] > \epsilon.$$

Faster Learning from easier distributions We now have a closer look at our bounds for distributions that satisfy the PL condition with functions $\phi(\lambda) = \lambda^n$ for some $n \in \mathbb{N}$. For these functions ϕ, the Probabilistic Lipschitzness is a stronger conditions (corresponds to a "nicer" data distribution) the larger n is. We have $\phi^{-1}(\epsilon) = \epsilon^{1/n}$. Thus, the lower bound on the sample complexity of learning in Theorem 3 becomes

$$\frac{d}{32\epsilon} \left(\frac{1}{(8\epsilon)^{1/n}} \right)^{d-1} = \Omega\left(\left(\frac{1}{\epsilon} \right)^{\frac{d-1+n}{n}} \right),$$

and the upper bound for Nearest Neighbor learning in Theorem 2 evaluates to

$$\frac{2}{\epsilon\,\delta} \left(\frac{2^{1/n}\sqrt{d}}{\epsilon^{1/n}} \right)^d = O\left(\left(\frac{1}{\epsilon} \right)^{\frac{d+n}{n}} \right).$$

Here, we use the Landau notation to illustrate the dependence on the accuracy parameter ϵ only. We see that the nicer the distribution (the larger n), the faster the Nearest Neighbor algorithm converges (and the weaker the lower bound becomes). Probabilistic Lipschitzness here thus serves as a parameter that quantifies the niceness of the data distribution in a way that allows to characterize the convergence behavior of the Nearest Neighbor algorithm. Our rate interpolates between a rate of $1/\epsilon^d$ for small $n = 1$ and a rate of $1/\epsilon$ for large n.

Learning with unlabeled data We now briefly summarize some results that were obtained for learning with unlabeled data (semi-supervised learning and active learning) under Probabilistic Lipschitzness. Intuitively, access to unlabeled data should facilitate the learning process when there is a correlation between the underlying marginal distribution and the labeling. This intuition is often referred to as the *cluster assumption*. PL can be viewed as one way to formalize and quantify the cluster assumption. The results summarized below illustrate that PL is a measure of niceness of a data generating distribution that is suitable for quantifying the benefits of unlabeled data for learning. The results below hold for distributions with Probabilistic Lipschitzness $\phi(\lambda) = \lambda^n$.

Urner et al. [2011] investigate *proper* semi-supervised learning, that is learning under the restriction that the output classifier needs to be a member of a pre-defined hypothesis class. That work establishes lower and upper bounds for proper learning under Probabilistic Lipschitzness. These show that access to unlabeled data reduces the labeled sample complexity of proper learning from $\frac{1}{\epsilon^2}$ to $\left(\frac{1}{\epsilon}\right)^{\frac{d+n}{n}}$ (this is a reduction whenever $n \geq d$).

Urner et al. [2013] show that enabling the learning algorithm to actively choose which points to query for labels reduces the labeled sample complexity even further. The reductions in labeled sample complexity achieved with the PLAL labeling procedure, presented in that work, are summarized in the table below. Note that the reductions increase with the parameter n of the PL condition.

$\phi(\lambda) = \lambda^n$	Passive lower bound	PLAL-active upper bound
Proper Learning of H	$\Omega\left(1/\epsilon^2\right)$	$O\left(\left(\frac{1}{\epsilon}\right)^{\frac{n+2d}{n+d}}\right)$
Unrestricted Learning of H	$\Omega\left(\frac{1}{\epsilon^{1.5}}\right)$	$O\left(\frac{1}{\epsilon}\right)$
Nearest Neighbor Learning	$\Omega\left(\left(\frac{1}{\epsilon}\right)^{1+\frac{d-1}{n}}\right)$	$O\left(\left(\frac{1}{\epsilon}\right)^{1+\frac{d^2}{n(n+d)}}\right)$

4 Discussion

Machine learning is often preceded by a process of feature selection or feature construction. Probabilistic Lipschitzness complies with an intuition of how such features should behave and can thus also provide a measure for the quality of a feature representation. It would be intriguing to investigate if PL, or a similar notion, could serve as an objective for developing a theoretical framework for feature learning.

We suggests to generalize the Probabilistic Lipschitzness condition as follows: Instead of just measuring the mass of (even only mildly) heterogeneous neighborhoods, we could consider the function that assigns every point the heterogeneity of a neighborhood around it. Interpreting this function as a new labeling function and imposing a low noise condition on this new labeling function (like the Tsybakov noise condition) provides a measure that is weaker than PL, but captures a similar property. It would be interesting to see if the results that are here derived under the assumption of PL generalize to this case.

References

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 33–44, 2008.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9(11):323–375, 2005.

Matti Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the Conference on Algorithmic Learning Theory (ALT)*, pages 63–77, 2006.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829, 1999.

Maxim Raginsky and Alexander Rakhlin. Lower bounds for passive and active learning. In *NIPS*, pages 1026–1034, 2011.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. 2014. To be published by Cambridge University Press.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines. *Annals of Statistics*, 35(2): 575–607, 2007.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

Ruth Urner. Learning with non-standard supervision. PhD Thesis, 2013. `http://uwspace.uwaterloo.ca/handle/10012/7925`.

Ruth Urner, Shai Ben-David, and Shai Shalev-Shwartz. Unlabeled data can speed up prediction time. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 641–648, 2011.

Ruth Urner, Sharon Wullf, and Shai Ben-David. Plal: Cluster-based active learning. In *Proceedings of the Conference on Learning Theory (COLT)*, 2013.

A Proofs

Proof of Theorem 3. The No-Free-Lunch theorem (Theorem 5.1 in Shalev-Shwartz and Ben-David [2014]) states that if a learner gets a sample of size less than half the size of the domain, then there is a distribution with a deterministic labeling function over this domain, such that the expected error of the learner for this sample size is at least $1/4$.

We construct a distribution on $[0,1]^d$ that satisfies the ϕ-Lipschitzness as follows: We set $P(\bar{0}) = 1 - 8\epsilon$ and distribute the remaining mass of 8ϵ uniformly on points of a grid G of sidelength $\lambda = \phi^{-1}(8\epsilon)$ "at the far side of the surface of" $[0,1]^d$, *i.e.* the points $x = (x_1, \dots x_d)$ where at least one of the x_i has value 1 and the others have values in $\{i\lambda \ : \ 1 \leq i \leq d\}$. Now P is ϕ-Lipschitz under any labeling of the grid points.

There are $|G| \geq d/(\lambda)^{d-1}$ such grid points. We show that with probability at least $1/2$, a sample of size at most m hits less than $|G|/2$ gridpoints. The expected number of such hits is bounded by $8\epsilon m$, formally

$$\mathbb{E}_{S\sim P^m}[|S \cap G|] \ = \ 8\epsilon m.$$

Now Markov's inequality yields

$$\Pr_{S\sim P^m}[|S \cap G| > |G|/2] \ \leq \ \frac{16\epsilon m}{|G|}.$$

Now $m < \frac{d}{32\epsilon}(\frac{1}{\phi^{-1}(8\epsilon)})^{d-1}$ and $|G| \geq \frac{d}{(\lambda)^{d-1}} = \frac{d}{(\phi^{-1}(8\epsilon))^{d-1}}$ implies

$$\Pr_{S\sim P^m}[|S \cap G| > |G|/2] \ < \ \frac{16\epsilon \cdot \frac{d}{32\epsilon}(\frac{1}{\phi^{-1}(8\epsilon)})^{d-1}}{\frac{d}{(\phi^{-1}(8\epsilon))^{d-1}}} = \frac{1}{2}.$$

The above mentioned no-free-lunch result implies that, there is a labeling for the points on G, such that the learner $\mathcal{A}$ has expected error at least $\frac{1}{4} \cdot 8\epsilon = 2\epsilon$ given that the sample hits at most half of the grid points. Since we have shown that this happens with probability at least $1/2$ for samples of size at most m, the learners' expected error over all samples of size at most m is at least ϵ. □

For the proof of the upper bound, we need the following technical lemma to bound the probability of points that do not have a close neighbor in the sample set S. This lemma and its proof can be found in Shalev-Shwartz and Ben-David [2014].

Lemma 4 (Lemma 19.2 in Shalev-Shwartz and Ben-David [2014])**.** *Let $C_1, C_2, \ldots, C_r$ be a sequence of subsets of some domain set $\mathcal{X}$ and let S be a set of points of size m, sampled i.i.d. according to some distribution $P_{\mathcal{X}}$ over $\mathcal{X}$. Then we have*

$$\mathbb{E}_{S\sim(P_{\mathcal{X}})^m}\left[\sum_{i:C_i\cap S=\emptyset} P_{\mathcal{X}}[C_i]\right] \leq \frac{r}{m\mathrm{e}}.$$

Proof of Theorem 2. Let $\lambda = \phi^{-1}(\epsilon)$. Note that we can cover $\mathcal{X} = [0,1]^d$ with $r = \left(\sqrt{d}/\lambda\right)^d$ boxes (axis-alligned rectangles) $C_1, C_2, \ldots C_r$ of side-length $\lambda/\sqrt{d}$ and diameter λ. For a domain point $x \in \mathcal{X} = [0,1]^d$ we denote the box (from the above cover) that contains x by $C(x)$.

We can bound the error of the Nearest Neighbor classifier $\mathrm{NN}(S)$ for a sample S as follows:

$$\begin{aligned}\mathrm{Err}_P(\mathrm{NN}(S)) &= \Pr_{x\sim P_{\mathcal{X}}}[\mathrm{NN}(S)(x) \neq l(x)] \\ &\leq \Pr_{x\sim P_{\mathcal{X}}}[S_{\mathcal{X}} \cap C(x) = \emptyset] + \Pr_{x\sim P_{\mathcal{X}}}[\,\exists\, y \in C(x) \,:\, l(y) \neq l(x)],\end{aligned}$$

where $S_{\mathcal{X}}$ denotes the projection of S to $\mathcal{X}$ ($S_{\mathcal{X}}$ contains the sample points in S without their labels). That is, a point will only be labeled erroneously by $\mathrm{NN}(S)$, if it falls into a box that was not hit by the sample S or if it falls into a box that contains points of the other label.

By the choice of λ, the probability that a (test-)point falls into a box that contains sample points of the opposite label is bounded by ϵ.

We now show that for $m \geq \frac{1}{\epsilon\delta}\left(\frac{\sqrt{d}}{\phi^{-1}(\epsilon)}\right)^d$, the total mass of boxes that are not hit by a sample of size m is bounded by ϵ as well (with probability at least $(1-\delta)$ over the sample). This implies that, for such samples, the error of the Nearest Neighbor classifier is bounded by 2ϵ.

With Markov's inequality, Lemma 4 implies that for any $\epsilon > 0$ and m we have

$$\Pr_{S\sim P^m}\left[\left[\sum_{i:C_i\cap S=\emptyset} P[C_i]\right] > \epsilon\right] \leq \frac{r}{\epsilon m\mathrm{e}} = \frac{\left(\sqrt{d}/\lambda\right)^d}{\epsilon m\mathrm{e}}.$$

Setting this to be smaller than δ and solving for m now shows that a sample of size

$$m > \frac{\left(\sqrt{d}/\lambda\right)^d}{\epsilon\delta\mathrm{e}}$$

suffices to guarantee that with probability at least $(1-\delta)$, the error of the output function $\mathrm{NN}(S)$ is at most 2ϵ. Substituting $\epsilon/2$ for ϵ yields the statement in the theorem. □