Journal of Machine Learning Research 10 (2009) 827-856 Submitted 9/07; Revised 8/08; Published 4/09

Consistency and Localizability

Alon Zakai ALON.ZAKAI@MAIL.HUJI.AC.IL
Interdisciplinary Center for Neural Computation
Hebrew University of Jerusalem
Jerusalem, Israel 91904

Ya'acov Ritov YAACOV.RITOV@HUJI.AC.IL
Department of Statistics
Interdisciplinary Center for Neural Computation and Center for the Study of Rationality
Hebrew University of Jerusalem
Jerusalem, Israel 91905

Editor: Gabor Lugosi

Abstract

We show that all consistent learning methods—that is, that asymptotically achieve the lowest possible expected loss for any distribution on (X,Y)—are necessarily localizable, by which we mean that they do not significantly change their response at a particular point when we show them only the part of the training set that is close to that point. This is true in particular for methods that appear to be defined in a non-local manner, such as support vector machines in classification and least-squares estimators in regression. Aside from showing that consistency implies a specific form of localizability, we also show that consistency is logically equivalent to the combination of two properties: (1) a form of localizability, and (2) that the method's global mean (over the entire X distribution) correctly estimates the true mean. Consistency can therefore be seen as comprised of two aspects, one local and one global.

Keywords: consistency, local learning, regression, classification

1. Introduction

In a supervised learning problem we are given an i.i.d sample $S_n = \{(x_i,y_i)\}_{i=1..n}$ of size n from some distribution P; then a new pair (x,y) is drawn from the same P and our goal is to predict y when shown only x. Our prediction (also called estimate, or guess) of y is written $f(S_n,x)$, some function that depends on the training set and the point at which we estimate (note that it is slightly atypical to have both the training set and point as inputs to f, but this will be very convenient in our setting). We call f a *learning method*; in the context of regression we will also use the term *estimator*, and in the context of classification we will use the term *classifier*. In both of these settings, if f achieves the lowest possible expected loss as $n \to \infty$, for every distribution, then we call f *consistent* (we will formalize all of these definitions later on; for now we just sketch the general ideas). Consistent estimators are of obvious interest due to their capability to learn without knowing in advance anything about the actual distribution.

When we look at the learning methods known to be consistent, we can separate them into two general types. In the first of these we have methods that are defined in a local manner, for example, the k-nearest-neighbor (k-NN) classifier (Stone, 1977; Devroye et al., 1996). The k-NN classifier

guesses the class of a point x based on its nearest neighbors in the training set, thus, this classifier behaves in a 'local' way: only close-by points affect the estimate. More generally, by a locally-behaving method we mean one that, given a training set S_n and a point x at which to estimate the value of y, in some way treats the close-by part of the training set as the most important. A second type of consistent learning method is defined in a global manner, for example, support vector machines (see Vapnik, 1998; Steinwart, 2002, for a description and proof of consistency, respectively). It is not clear from the definition of support vector machines whether they behave locally or not: The separating hyperplane is determined based on the entire training set S_n, and furthermore does not depend on the specific point x at which we classify, perhaps leading us to expect that support vector machines do *not* behave locally. Thus, on an intuitive level we might think that some consistent methods behave locally and some do not.

This intuition also appears relevant when we consider regression: The k-NN regression estimator appears to behave locally, while on the other hand support vector regression (see, e.g., Smola and Schoelkopf, 1998), kernel ridge regression (Saunders et al., 1998), etc., seem not to have that property. Another example is that of orthogonal series estimation, that is, of using a weighted sum of fixed harmonic functions (Lugosi and Zeger, 1995); this method appears to not behave locally both because the harmonic functions are non-local and because the coefficients are determined in a way based on all of the data.

Despite the intuition that some consistent methods might not behave locally, we will see that in fact *all* of them necessarily behave in that manner. As mentioned before, we already know that some locally-behaving methods are consistent, since some are in fact defined in a local manner, for example, k-NN. What we will see is that all other consistent methods must *also* behave locally. In classification, this implies that, in particular, (properly regularized) support vector machines and boosting (Freund and Schapire, 1999; Zhang, 2004; Bartlett and Traskin, 2007) must behave locally, despite being defined in a way that appears global. In the area of regression, our results show that neural network estimators, orthogonal series estimators, etc., must behave locally if they are consistent, again, despite their being defined in a way that does not indicate such behavior.

In the rest of this introductory section we will present a summary of our approach and results as well as background regarding related work. While doing so we focus on regression problems since that setting allows for simpler and clearer definitions. For the same reasons we will also focus on regression in the main part of this paper; in a later section we will show how to apply our results to classification.

Our goal in regression is to estimate $f^*(x) = E(y|x)$, that is, the expected value of y conditioned on x, or the regression of y on x. Our hope is that $f(S_n, x)$ is close to $f^*(x)$. We say that f is consistent on a distribution P iff

$$L_n(f) \equiv E\,|f(S_n,x) - f^*(x)| \underset{n\to\infty}{\longrightarrow} 0$$

where the expected value is taken over training sets S_n and observations x both distributed according to P (which is suppressed in the notation). If a method is consistent on all P then we call it consistent (this is sometimes called *universal consistency*). Note that there are stronger notions of consistency, such as requiring that the loss converge to 0 with probability 1 (see, e.g., Gyorfi et al., 2002), but we will focus on the loss as just described. Note also that the L_n loss is 'global' in that we average over all x, which makes it all the more interesting to see whether methods that minimize it must end up behaving locally.

The notion of local behavior that we will consider will be called *localizability*, and it entails that the method returns a similar estimate for x when shown S_n in comparison to the estimate it would have returned when shown only the part of S_n that is close to x. In other words, if we define

$$S_n(x,r) = \{(x_i,y_i) \in S_n \ : \ ||x_i - x|| \leq r\}$$

then a localizable method has the property that

$$f(S_n,x) \approx f(S_n(x,r),x)$$

for some small $r > 0$ (the formal definitions of all of these concepts will be given in later sections). More specifically, for any sequence $R_n \searrow 0$ we can see $g(S_n,x) = f\left(S_n\left(x,R_n\right),x\right)$ as a 'localization' of f, since it applies f to the close-by part of the training set for the particular point at which we estimate. In other words, a localizable method is one that behaves similarly to a localization of itself. (Note the convenience of the $f(S_n,x)$ notation here, that is, of seeing f as a function of both S_n and x.)

Why is the concept of localizability of interest? The main motivation for us is that, as we will see later, consistency implies a form of localizability. That is, even an estimator defined in what seems to be a global manner, for example, by minimizing a global loss of the general form

$$\frac{1}{n}\sum_{i=1..n} l\big(f(x_i),y_i\big) + \lambda c(f)$$

where l is, for example, least-squares, and c is (optional) complexity penalization—then even such an estimator must be in some sense localizable, if it is consistent. Thus, our first motivation is to point out that not only locally-defined methods like k-NN behave locally, but also all other consistent methods as well.

Aside from this main motivation for investigating localizability, another reason is that it allows us to answer questions such as, "What might happen if we localize a support vector machine?" That is, we can apply a support vector machine (or some other useful method) to only the close-by part of the training set, perhaps motivated by the fact that training on this smaller set is more computationally efficient, at least if all we need is to generate estimates at a small number of points. If support vector machines are localizable, then we in fact know that such an approach can be consistent; and if they are not localizable, then we may end up with a non-consistent method with poor performance. Thus, localizability can have practical applications.

Note that one can consider other ways to define local behaviour than localizability. In one such approach, we can evaluate the behavior of a method when altering the far-off part of the training set, as opposed to removing it (which is what we do with localizability). Work along those lines (Zakai and Ritov, 2008) arrives at similar conclusions to the ones presented here. Comparing the two approaches, localizability has the advantage of relevance to practical applications, as mentioned in the previous paragraph.

Previous work related to localizability has been done in the context of learning methods that work by minimizing a loss function: We can 'localize' the loss function by re-weighting it so that close-by points are more influential; this has been investigated in the context of Empirical Risk Minimization (ERM; Vapnik, 1998) (Bottou and Vapnik, 1992; Vapnik and Bottou, 1993), as well as in the specific case of linear regression (see, e.g., Cleveland and Loader, 1995, and references therein); this approach has also lead to various applications (Atkeson et al., 1997). In this paper we

differ from these approaches in that we work in a more general context: Our approach is applicable to all learning methods, and not just those that are based on minimizing a loss function that can be re-weighted. Another difference is that we consider consistency in the sense of asymptotically arriving at the lowest possible loss achievable by any measurable function, and not in the sense of minimizing the loss within a set of finite VC dimension.

Another related work is that of Bengio et al. (2006), in which it was shown that kernel machines behave locally, in the sense of requiring a large number of examples in order to learn complex functions (because each local area must be learned separately). Our approach differs from this work in the way we define local behavior, and in that we are interested in all (consistent) learning methods, not just kernel machines. However, our conclusion is in agreement with theirs, that even methods that may appear to be global like support vector machines in fact behave locally.

We now sketch our main result, which is that consistency is logically equivalent to the combination of two properties (which will be given later, in Definitions 2 and 5): *Uniform Approximate Localizability* (UAL), which is a form of localizability, and *Weak Consistency in Mean* (WCM), which deals with the mean $E f(S_n,x)$ estimating the true mean $E f^*(x)$ reasonably well, where the expected values are taken over S_n and x. It will be easy to see that the UAL and WCM properties are 'independent' in the sense that neither implies the other, and therefore we can see consistency as comprised of two independent aspects, which might be presented as

$$\textbf{Consistency} \quad \Longleftrightarrow \quad \textbf{UAL} \oplus \textbf{WCM}.$$

This can be seen as describing consistency in terms of local (UAL) and global (WCM) aspects (WCM is 'global' in the sense of only comparing scalar values averaged over x).

Note that there are two issues here which might be surprising, the first of which has already been mentioned—that all consistent methods must behave locally. The second important issue is that WCM is sufficient, when combined with UAL, to imply consistency. That is, if our goal is to be consistent,

$$E|f(S_n,x) - f^*(x)| \longrightarrow 0 \tag{1}$$

then it is interesting that all that is needed in addition to behaving locally is a property close to

$$|E f(S_n,x) - E f^*(x)| \longrightarrow 0.$$

Note that the latter condition is very weak. For example, it is fulfilled by $g(S_n,x) = \frac{1}{n}\sum_i y_i$, the simple empirical mean of the y values in the sample S_n (ignoring the xs completely), since $\frac{1}{n}\sum_i y_i \longrightarrow Ey = E f^*(x)$ a.s., where Ey is the mean of y. On the other hand, g does not fulfill the stronger property (1) except on trivial distributions.

Our main result, which has just been described, will also be generalized to settings other than that of methods consistent on the set of all distributions. This will lead to peculiar consequences: Consider, for example, the following two sets of distributions:

$$\mathbb{P}_1 = \left\{P \,:\, y = f^*(x) + \varepsilon \,,\, E\varepsilon = 0 \,,\, \varepsilon \text{ is independent of } x\right\} \quad , \quad \mathbb{P}_2 = \left\{P \,:\, Ey = 0\right\}$$

(note that $\mathbb{P}_1$ is simply regression with additive noise). It turns out that a method consistent on $\mathbb{P}_1$ must behave locally on that set (just as with the set of all distributions), but the same is not true for $\mathbb{P}_2$, where a consistent method does not necessarily behave locally. The reason for this will be explained later.

The rest of this work is as follows. In Section 2 we present the formal setting and other preliminary matters. In Section 3 we present definitions of local properties (and specifically UAL) as well as some results concerning them. In Section 4 we define the global concepts that we need, specifically WCM. In Section 5 we present our main result, the equivalence of consistency to the combination of UAL and WCM. In Section 6 we extend our results to various sets of distributions. In Section 7 we use results from previous sections in order to derive consequences for classification. In Section 8 we summarize our results and discuss some directions for future work. Finally, proofs of our results appear in the Appendices.

2. Preliminaries

We now complete the description of the formal setting in which we work, as well as lay out notation useful later. Most of this section deals with the context of regression; details specific to classification will appear in Section 7.

We consider distributions P on (X,Y) where $X \subset \mathbb{R}^d, Y \subset \mathbb{R}$. We assume that X,Y are bounded, $\sup_{x\in X} ||x||, \sup_{y\in Y} |y| \leq M_1$ for some $M_1 > 0$ which is the same for all distributions. Thus, when we say 'all distributions' we mean all distributions bounded by the same value of M_1. We also assume that our learning methods return bounded responses, $\forall S,x \quad |f(S,x)| \leq M_2$.[1] Let M be a constant fulfilling $M \geq M_1, M_2$. Importantly, note that while these boundedness assumptions are non-trivial in the context of regression, they do not limit us when we consider classification, as we will see in Section 7.

Note that we wrote $f(S,x)$ instead of $f(S_n,x)$ in the previous paragraph. The reason is that n will always denote the size of the original training set, which in turn will always be written as S_n. Since we will also apply f to to other training sets (in particular, subsets of S_n), for clarity of notation we will therefore write S for a general (finite) set of pairs (x_i,y_i) and define learning methods via $f(S,x)$.

Formally speaking, a learning method $f(S,x)$ is defined as a sequence of measurable functions $\{f_k\}_{k\in\mathbb{N}}$, where each f_k is a function on training sets of size k, that is,

$$f_k : (X \times Y)^k \times X \longrightarrow Y.$$

For brevity, we will continue to write f instead of f_k since which f_k is used is determined by the size of the training set that we pass to f, that is, $f(S,x) = f_{|S|}(S,x)$. For example, we will often denote by $\widetilde{S}$ a subset of the original training set S_n. Then in an expression of the form $f(\widetilde{S},x)$ the actual function used is f_m, where $m = |\widetilde{S}|$, and in this example we expect to have $m \leq n$ (where, as mentioned before, n is the size of the original training set S_n).

For any distribution P on (X,Y), we write $f^*(x) = E(y|x)$, as already mentioned, and we denote the marginal distribution on X by μ. We write f_P^*, μ_P instead of f^*, μ to make explicit the dependence on P when necessary. Denote by $\mathrm{supp}_X(P) = \mathrm{supp}(\mu_P)$ the support of μ_P. The measure of sets $B \subseteq X$ will be written in the form $\mu(B)$.

1. Note that this is a minor assumption since for most methods we have $\sup_x |f(S,x)| \leq C \cdot \max_i |y_i|$ for some $C > 0$, and the y_i values are already assumed to be bounded. If this does not hold, then we might in any case want to consider enforcing boundedness based on the sample, that is, to truncate values larger in absolute value than $\max_i |y_i|$, and in doing so perhaps improve performance. Finally, recall that we are concerned with consistent methods, that is, that behave similarly to f^* in the limit, and f^* is bounded.

We denote random variables by, for example, $(x,y) \sim P$ and $S_n \sim P$ where in the latter case we intend a random i.i.d sample of n elements from the distribution P. We will often abbreviate and write $x,y \sim P$ instead of $(x,y) \sim P$; also, we will write $x \sim P$ where we mean $x \sim \mu_P$. To prevent confusion we always use x and y to indicate a pair (x,y) sampled from P.

We will write the mean and variance of random variables v as $E(v) = Ev$ and $\sigma^2(v)$, respectively. More generally, expected values will be denoted by $E_{v\sim V}H(v) = E_v H(v)$ where v is a random variable distributed according to V and $H(v)$ is some function of v; we may write $EH(v)$ when the random variables are clear from the context. The conditional expected value of $H(v)$ given w will be written in the form $E_{v|w}(H(v))$.

We will work mainly with the $\mathcal{L}_1$ loss, which we can now write formally as

$$L_{n,P}(f) \equiv E_{S_n,x\sim P}\left|f(S_n,x) - f^*_P(x)\right|,$$

or, more briefly,

$$L_n(f) \equiv E_{S_n,x}\left|f(S_n,x) - f^*(x)\right|.$$

Note that for purposes of consistency (i.e., $L_n(f) \to 0$) all $\mathcal{L}_p$ losses are equivalent, since

$$\forall 0 < p < q \qquad E|z|^p \le (E|z|^q)^{p/q} \le (2M)^{p(q-p)/q}(E|z|^p)^{p/q}$$

where $z = f(S_n,x) - f^*(x)$. Thus, if one $\mathcal{L}_p$ loss converges to 0, so do all the others. Hence our results apply to all $\mathcal{L}_p$ norms; we work mainly with the $\mathcal{L}_1$ norm for convenience.

For any set $B \subseteq X$, denote by P_B the conditioning of P on B, that is, the conditioning of μ_P on B (and leaving unchanged the behavior of y given x). Denote the ball of radius r around x by $B_{x,r} = \{x' \in \mathbb{R}^d \,:\, ||x - x'|| \le r\}$. Let $P_{x,r} = P_{B_{x,r}}$.

Finally, we mention two useful conventions. Note that we defined consistency on a single distribution, and then consistency in general as the property of being consistent on all distributions. More specifically, for any property **A** that can hold for a method f on particular distributions, we say that f has property **A** on a *set* of distributions $\mathbb{P}$ when f has property **A** on all $P \in \mathbb{P}$. We also say that f has property **A** (without specifying P or $\mathbb{P}$) when it has property **A** on the set of *all* distributions (with bounded support). This convention will be used for consistency as well as for UAL and other properties.

Similarly, when we start by defining a property **A** on the set of all distributions, we then use the convention that f has property **A** on a set $\mathbb{P}$ when we simply replace $\forall P$ with $\forall P \in \mathbb{P}$. This convention will be used with the WCM property.

3. Local Behavior

In this section we consider local properties of learning methods. We will present a series of definitions, leading up to a definition of Uniform Approximate Localizability (UAL).

We start with some introductory definitions. For any two learning methods f,g, we say that they are **mutually consistent** iff

$$D_n(f,g) \equiv E_{S_n,x}\left|f(S_n,x) - g(S_n,x)\right| \underset{n\to\infty}{\longrightarrow} 0.$$

That is, f and g are mutually consistent if they behave asymptotically similarly according to a distance metric D_n. The term 'mutual consistency' is used since $D_n(f,f^*) = L_n(f)$ (where we can

formally define $f^*(S,x) = f^*(x)$), that is, being mutually consistent with f^* is equivalent to being consistent. Thus, mutual consistency is a natural extension of consistency. Note that D_n obeys the triangle inequality,

$$\forall f,g,h \qquad D_n(f,g) \leq D_n(f,h) + D_n(h,g)$$

which will be convenient later.

We now define some useful notation for the topic of locality. For any training set S and $r \geq 0$, we call

$$S(x,r) = \{(x_i,y_i) \in S \,:\, ||x_i - x|| \leq r\}$$

a **local training set** for x within S, of radius r. For every learning method f and $r \geq 0$, let

$$f|_r(S,x) = f(S(x,r),x).$$

Note that, as mentioned previously, if $f = \{f_k\}_{k\in\mathbb{N}}$ then in the expression $f(S_n(x,r),x)$ (where S_n is, as always, the original training set of size n), we are passing $S_n(x,r),x$ to f_m, where $m = |S_n(x,r)|$, which will in general be smaller than $n = |S_n|$. Thus, formally speaking we might write

$$f|_r(S,x) = f_{|S(x,r)|}(S(x,r),x)$$

but for simplicity we will continue to drop the lower index on f. Note that $f|_r$ can also be formally defined as a series of functions $\{f|_{r,k}\}_{k\in\mathbb{N}}$, but again, for simplicity we avoid this.

In words, $f|_r$ is a learning method that results from forcing f to only work on local training sets of radius r around x, when estimating the value at x. For example, if f is a linear regression estimator then we can see $f|_r$ as performing local linear regression (Cleveland and Loader, 1995).

Continuing in our definitions, for any sequence $\{R_k\}_{k\in\mathbb{N}}, R_k \geq 0, R_k \to 0$, we call $f|_{\{R_k\}}$ a **local version**, or a **localization** of f; by this notation, we mean

$$f|_{\{R_k\}}(S,x) = f\big(S\big(x,R_{|S|}\big),x\big)$$

—that is, which R_k is used from the sequence $\{R_k\}$ depends on the size of the training set passed to $f|_{\{R_k\}}$. In particular, for the original training set S_n we have

$$f|_{\{R_k\}}(S_n,x) = f\big(S_n\big(x,R_n\big),x\big).$$

Note that, as a consequence, local versions are indeed 'local' in the limit since we have $R_n \to 0$.

We can now define one form of local behavior: Call a method f **localizable** on a distribution P iff there exists a local version $f|_{\{R_k\}}$ of f with which f is mutually consistent, that is,

$$D_n\left(f|_{\{R_k\}}, f\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Thus, a localizable method is one that is similar, in the sense of mutual consistency, to a local version of it, which implies that it gives similar results when seeing the entire training set versus only the local part of it; we can localize the method without changing the estimates significantly. Note once more that the requirement $R_n \to 0$ is what makes this definition truly define local behavior.

We will also need a notion of a method that, when localized, is consistent. Call a method f **locally consistent** on a distribution P iff there exists a local version $f|_{\{R_k\}}$ of f which is consistent,

$$L_n\left(f|_{\{R_k\}}\right) = D_n\left(f|_{\{R_k\}}, f^*\right) \underset{n\to\infty}{\longrightarrow} 0.$$

That is, there is a way to localize f so that it becomes consistent.

Are all consistent learning methods localizable and locally consistent? Consider the latter property: It appears as if any consistent method must be locally consistent, since a consistent method can successfully 'learn' given any underlying distribution, and when we localize such a method we are in effect applying that same useful behavior in every local area. Thus, it seems reasonable to expect a localization of a consistent method to be consistent as well, and if this were true then it might have useful practical applications, as mentioned in the introduction. Yet this intuition turns out to be false, and a similar failure occurs for localizability:

Proposition 1 *A learning method exists which is consistent but neither localizable nor locally consistent.*

The proof of Proposition 1 (appearing in Appendix A) is mainly technical, and consists in constructing a method $f = \{f_k\}$ which becomes less smooth as n rises, and asymptotically considerably less smooth than the true f^*. That is, the issue is that we have required merely that the functions f_k be measurable, and it turns out that without additional assumptions they can behave erratically in a manner that renders f not locally consistent. The specific example that we construct in the proof involves functions f_k that behave oddly on an area close to Ex but otherwise perform well. It is then possible to show that the 'problematic area' near Ex can be large enough so that all local versions of the method behave poorly, but small enough so that the original method is consistent.

Now, the counterexample constructed in the proof of Proposition 1 might rightfully be called a 'fringe case'. Yet, it suffices to show that not all consistent methods are localizable, contrary perhaps to intuition. There is therefore the question of what to do. One solution to this matter is to work with a property stronger than consistency, one that includes an additional smoothness requirement. The disadvantage of such an approach is that we cannot immediately derive consequences for the various methods known to be consistent, unless we also prove that they have the stronger property.

Instead, the approach that we will follow is to define more complex notions of local behavior which can be used to arrive at properties equivalent to consistency. Thus, one major goal of this paper is to arrive at suitable definitions for the topic of local behavior, that on the one hand capture the intuition correctly and on the other allow useful results to be proven. The simple definitions given before fail in the second matter; we will now give definitions that remedy that problem.

In order to formulate our improved definitions we first require some preparation. For any $r, q \geq 0$ and distribution P, let

$$\overline{f}|_r^q(S,x) = E_{x' \sim P_{x,q \cdot r}} f(S(x,r),x'). \tag{2}$$

Compared to $f|_r$, $\overline{f}|_r^q$ adds a smoothing operation performed around the x at which we estimate. Note that if $q = 0$ then we interpret the expected value as a delta function and we get $\overline{f}|_r^0 = f|_r$. Note also that we require the actual unknown distribution P in the definition of $\overline{f}|_r^q$, that is, $\overline{f}|_r^q$ cannot be directly implemented in practice—$\overline{f}|_r^q$ is a construction useful mainly for theoretical purposes. However, we can implement an approximate version of $\overline{f}|_r^q(S,x)$ by replacing the true expectation with the empirical one. We will return to this matter later.

We define the following set of sequences:

$$\mathcal{T} = \left\{ \{T_k\} \,:\, T_k \searrow 0 \text{ strictly} \right\}.$$

For any sequence $T = \{T_k\}$, let $\mathbf{T} = \{T_k \,:\, k \in \mathbb{N}\}$, that is, the set containing the elements in the sequence. For any such sequence $T = \{T_k\}$, we then define the set of its infinite subsequences and

selection functions on them by

$$\mathcal{R}(T) = \Big\{ R = \{R_k\} \quad : \quad \mathbf{R} \subseteq \mathbf{T}\ ,\ R_k \searrow 0 \Big\},$$
$$\mathcal{Q}(T) = \Big\{ Q : \mathbf{T} \to \mathbf{T} \quad : \quad Q(T_k) \searrow 0 \Big\}.$$

For any $T \in \mathcal{T}$ and $\{R_k\} \in \mathcal{R}(T)$, $Q \in \mathcal{Q}(T)$, we call $\overline{f}|_{\{R_k\}}^{\{Q(R_k)\}}$ a **smoothed local version** of f; by this notation, we mean to replace the q, r values in (2) in an appropriate manner, that is,

$$\overline{f}|_{\{R_k\}}^{\{Q(R_k)\}}(S,x) = E_{x' \sim P_{x,Q(R_{|S|})\cdot R_{|S|}}} f\big(S\big(x, R_{|S|}\big), x'\big).$$

Note that on the original training set S_n we get

$$\overline{f}|_{\{R_k\}}^{\{Q(R_k)\}}(S_n,x) = E_{x' \sim P_{x,Q(R_n)\cdot R_n}} f\big(S\big(x, R_n\big), x'\big).$$

We now elaborate on these definitions. T is the set of possible values that $Q(R_n), R_n$ can take; these values must approach 0 as our goal is to consider local behavior. $\mathcal{R}(T)$ contains sequences of radii of local training sets; we require that $R_n \searrow 0$, as we are interested in behavior on local training sets with radius descending to 0, that is, that become truly 'local' asymptotically. $\mathcal{Q}(T)$ contains functions that become small when T_k is small; the values $Q(R_n)$ determine radii on which to smooth, via $Q(R_n) \cdot R_n$. Since $Q(R_n) \cdot R_n = o(R_n)$, the smoothing is done on radii much smaller (asymptotically negligibly small) than the radii of the local training sets R_n, and therefore this is a minor operation. In conclusion, a smoothed local version is similar to a local version, but adds an averaging operation on small radii.

We now start with our main definitions. The idea behind them is not overly complex, but their description is necessarily somewhat technical.

Definition 2 *Call a learning method f* ***Uniformly Approximately Localizable (UAL)*** *iff*

$$\begin{aligned}
&\forall P \quad \forall T \in \mathcal{T} \\
&\exists Q \in \mathcal{Q}(T) \\
&\forall Q' \in \mathcal{Q}(T)\ ,\ Q' \geq Q \\
&\exists \{R_k\} \in \mathcal{R}(T) \\
&\forall \{R'_k\} \in \mathcal{R}(T)\ ,\ \{R'_k\} \geq \{R_k\} \\
&\qquad D_n\left(\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}}, f\right) \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

(Here the expression $\{R'_k\} \geq \{R_k\}$ simply implies an inequality for the entire series, that is, for all k. $Q' \geq Q$ implies $Q'(T_k) \geq Q(T_k)$ for all k.)

In essence, a UAL learning method is one for whom, for any choice of T, *all* large-enough choices of Q and $\{R_k\}$ are suitable in order to get similar behavior between f and a smoothed local version of f. That is, if we take $\{R_k\}$ and Q slowly enough to 0 then we get local behavior. Note that in any case taking $\{R_k\}$ to 0 very quickly is problematic since we may get empty local training sets, that is, $S_n(x, R_n) = \emptyset$.

Concerning our choice of name for this definition, 'uniformly' appears in the title 'uniformly approximately local' due to the requirement for all large-enough choices of $\{R_k\}, Q$ to be relevant, and 'approximately' appears because we allow smoothed local versions and not just local versions.

Both of these changes from the original definition of localizability are present in order to prevent odd counterexamples. A concrete counterexample was shown in Proposition 1 to make clear the need for smoothing; we do not present one in full for the uniformity requirement in order to save space.

Note that we only consider $\{R_k\}, Q$ taking values in some fixed T, and that while T is arbitrary it does need to be determined in advance. The issue is that if we instead allow all the values $[0,\infty)$ to appear in $\{R_k\}$ and Q then, due to this being an uncountable set, it is not clear to the authors if additional conditions are not required to prove our main results in that case. In any event, a countable set of possible values is of sufficient interest for any practical learning-theoretical purpose, and as already mentioned the actual set of possible values can be chosen in whatever manner is desired.

We also need a definition parallel to that of local consistency, as follows.

Definition 3 *Let **Uniform Approximate Local Consistency (UALC)** be the property defined exactly the same as UAL, except for replacing the last condition with*

$$L_n\left(\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}}\right) = D_n\left(\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}}, f^*\right) \underset{n\to\infty}{\longrightarrow} 0.$$

A UALC method is one whose smoothed local versions are consistent for any large-enough choice of $Q, \{R_k\}$.

We now mention some properties of UAL and UALC, noting first that they are independent, in that each can exist without the other. Consider the following two methods:

$$f_y(S,x) = \frac{1}{|S|}\sum_{i=1..|S|} y_i, \qquad\qquad f_0(S,x) = 0 \tag{3}$$

f_y (called thus because it considers only the y values) is UALC since a local version of it is simply a kernel estimator, using the 'window kernel' $k(x) = 1\{||x|| \le 1\}$, and thus consistent, for any $\{R_k\}$ fulfilling $nR_n^d \to \infty$ (Devroye and Wagner, 1980); $\{R_k\}$ acts as the bandwidth parameter of a kernel estimator. (Note that smoothing has no effect, as the guess does not depend on x.) f_y, however, clearly cannot be UAL (e.g., consider the simple example of x uniform on $[-1,1]$ and $y = \text{sign}(x)$); neither is it consistent. Turning to f_0, this method is clearly UAL but it is neither UALC nor consistent.

Our first main result is that consistency is equivalent to the combination of UAL and UALC:

Theorem 4 *A learning method is consistent iff it is both UAL and UALC.*

In light of the independence of UAL and UALC, mentioned before, we can summarize this as

$$\textbf{Consistency} \quad \Longleftrightarrow \quad \textbf{UAL} \oplus \textbf{UALC}.$$

The proof of $\Leftarrow$ is immediate: Pick any T. We derive some $Q, \tilde{Q}$ from the appropriate $\exists$ clauses of the definitions of UAL and UALC, respectively; let $Q' \in Q(T)$ fulfill $Q' \ge Q, \tilde{Q}$. Given Q', we can then derive some $\{R_k\}, \{\tilde{R}_k\}$ from the appropriate $\exists$ clauses of UAL and UALC; let $\{R'_k\} \in \mathcal{R}(T)$ fulfill $R'_k \ge R_k, \tilde{R}_k$. It is then clear that

$$D_n\left(\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}}, f^*\right), D_n\left(\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}}, f\right) \longrightarrow 0$$

due to UALC and UAL. By the triangle inequality we get

$$D_n(f,f^*) \le D_n\left(f,\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}}\right) + D_n\left(\overline{f}|_{\{R'_k\}}^{\{Q'(R'_k)\}},f^*\right) \longrightarrow 0 \tag{4}$$

that is, consistency. Thus, it is fairly immediate from the definitions of UAL and UALC that together they suffice for consistency. What is interesting is that they are equivalent to it. Instead of proving that fact directly, the remainder of the proof of Theorem 4 will be a corollary of the results in Section 5.

4. Global Properties

In this section we define global properties that will be useful in the next section, where we present our main results.

First we need some preliminary definitions. We define the means of f, f^* in the natural way,

$$E_n(f) \equiv E_{n,P}(f) \equiv E_{S_n,x} f(S_n,x),$$

$$E(f^*) \equiv E_P(f^*) \equiv E_x f^*(x) = E_x E(y|x) = Ey$$

the latter expression which is just the global mean of y. We also want to consider the Mean Absolute Deviation (MAD) of f and f^*,

$$\mathrm{MAD}_n(f) \equiv \mathrm{MAD}_{n,P}(f) \equiv E_{S_n,x}\left|f(S_n,x) - E_n(f)\right|,$$

$$\mathrm{MAD}(f^*) \equiv \mathrm{MAD}_P(f^*) \equiv E_x\left|f^*(x) - E(f^*)\right|.$$

We can now define the first version of the global property of interest to us: We say that f is **consistent in mean** iff

$$\forall P \qquad \lim_{n\to\infty} |E_n(f) - E(f^*)| \;=\; \lim_{n\to\infty} |\mathrm{MAD}_n(f) - \mathrm{MAD}(f^*)| \;=\; 0.$$

A consistent in mean learning method is required to correctly estimate $E(f^*)$ and $\mathrm{MAD}(f^*)$; that is, we require that the global behavior of f, averaged over x, be asymptotically equal to that of f^*. Note that we are only interested here in two scalar values which represent global averages of the behavior of f, f^*.

Consistency in mean is obviously a weaker property than requiring that, on average, f behave similarly to f^* on every x separately—that is, consistency—since

$$|E_n(f) - E(f^*)| = |E_{S_n,x} f(S_n,x) - E_x f^*(x)| \le E_{S_n,x}\left|f(S_n,x) - f^*(x)\right|. \tag{5}$$

By consistency the RHS converges to 0, and therefore so does the LHS. Consider now the MAD:

$$\begin{aligned}
\mathrm{MAD}_n(f) &= E_{S_n,x}|f(S_n,x) - E_n(f)| \\
&= E_{S_n,x}|f(S_n,x) - f^*(x) + f^*(x) - E(f^*) + E(f^*) - E_n(f)| \\
&\le E_{S_n,x}|f(S_n,x) - f^*(x)| + E_x|f^*(x) - E(f^*)| + |E(f^*) - E_n(f)| \\
&\le D_n(f,f^*) + \mathrm{MAD}(f^*) + |E(f^*) - E_n(f)|.
\end{aligned}$$

Of the last three expressions, the first converges to 0 by consistency, and the third by (5) (which was implied by consistency). Similarly,

$$\mathrm{MAD}_n(f) \geq \mathrm{MAD}(f^*) - D_n(f, f^*) - |E(f^*) - E_n(f)|$$

showing that $|\mathrm{MAD}_n(f) - \mathrm{MAD}(f^*)| \to 0$. Thus, unsurprisingly, consistency implies consistency in mean.

It turns out that a weaker property than consistency in mean is sufficient for our purposes:

Definition 5 *We say that f is **Weakly Consistent in Mean (WCM)** iff there exists a function $H : \mathbb{R} \to \mathbb{R}$, $H(0) = 0$, $\lim_{t\to 0} H(t) = 0$, for which*

$$\forall P \qquad \limsup_{n\to\infty} \left|E_n(f) - E(f^*)\right| , \limsup_{n\to\infty} MAD_n(f) \;\leq\; H\big(MAD(f^*)\big).$$

(Note that the same H is used for all P.) A WCM learning method is required only to do 'reasonably' well in estimating the global properties of the distribution, in a way that depends on the MAD, that is, on the difficulty: We only require that performance be good when the learning task is overall quite easy, in the sense of $f^*(x)$ being almost constant. Note that when $H(\mathrm{MAD}(f^*)) \geq 2M$ we require nothing of f for such f^* (since $|f|, |f^*| \leq M$), and also that for small $\mathrm{MAD}(f^*)$ we may allow the MAD of f to be significantly larger than that of f^* (consider, for example, $H(t) = c \cdot (\sqrt{t} + t)$ for large $c > 0$). Note also that we take the lim sups, that is, we do not even require that the limits exist (except in the trivial case where $\mathrm{MAD}(f^*) = 0$).

To see the justification for the adjective 'weak', note first that consistency in mean immediately implies WCM, using $H(t) = t$. Second, recall the example hinted at in the introduction: $f_y(S,x) = \frac{1}{n}\sum_i y_i$ is clearly not consistent, since it ignores the xs, nor is it consistent in mean, since

$$\mathrm{MAD}_n(f_y) = E_{S_n,x}|f_y(S_n,x) - E_n(f_y)| = E_{y_1,\ldots,y_n}\left|\frac{1}{n}\sum_i y_i - Ey\right| \longrightarrow 0.$$

However, f_y is WCM, since $E_n(f_y) \to E(f^*)$ and, as just mentioned, f_y's MAD converges to 0. (In fact, f_y is WCM with $H \equiv 0$, that is, in the strongest sense. That is, there are even 'weaker' methods that are WCM.)

5. Main Result

In this section we present our main result, the logical equivalence of consistency to the combination of the UAL and WCM properties. This will be arrived at during the course of proving the final step of Theorem 4.

The logical relationships between the concepts of consistency, UAL, UALC and WCM are shown in graph form in Figure 1. Note that we already saw that consistency implies WCM (since consistency implies consistency in mean, which implies WCM), and that UAL combined with UALC implies consistency (shown immediately after the statement of Theorem 4). Instead of proving directly that consistency implies UAL and UALC, we will prove that WCM implies UALC, and hence that consistency implies UALC. Consistency combined with UALC, in turn, implies UAL, since

$$D_n\left(f, \overline{f}|_{\{R_k\}}^{\{Q(R_k)\}}\right) \leq D_n(f, f^*) + D_n\left(\overline{f}|_{\{R_k\}}^{\{Q(R_k)\}}, f^*\right) \to 0$$

similarly to (4). Thus, in order to complete the picture sketched in Figure 1, it remains to show that

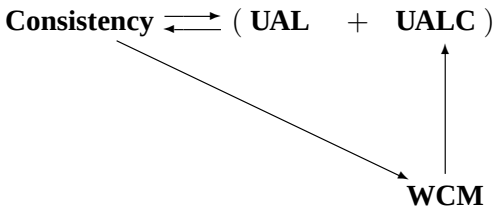

Figure 1: Logical relations between consistency, Uniform Approximate Localizability (UAL), Uniform Approximate Local Consistency (UALC) and Weak Consistency in Mean (WCM). That is, consistency is equivalent to the combination of UAL and UALC; consistency implies WCM; and WCM implies UALC.

Lemma 6 *A learning method that is WCM is UALC.*

The proof of Lemma 6 is the most complex of our results; we now sketch it very briefly (the full proof appears in Appendix B). The initial idea is to consider $S_n(x,r)$ as $|S_n(x,r)|$ points derived from $P_{x,r}$. The expected loss of a smoothed local version can then be seen to be approximately equal to a mean of expected losses over $P_{x,r}$ for various x. The expected losses over $P_{x,r}$ can, in turn, be bounded by the MADs of $f^*_{P_{x,r}}$ and f as well as the difference between their means; we bound the latter two using the WCM property. Finally, we construct in a recursive manner Q and $\{R_k\}$ that fulfill the requirements of UALC, using some results from measure theory regarding the asymptotic behavior of $f^*_{P_{x,r}}$.

Based on our results thus far, as summarized in Figure 1, it is obvious that we can also conclude the following:

Corollary 7 *A learning method is consistent iff it is both UAL and WCM.*

This can be stated as

$$\textbf{Consistency} \iff \textbf{UAL} \oplus \textbf{WCM}$$

since just like UAL and UALC, UAL and WCM are clearly independent in that neither implies the other, in fact, the same two examples seen in (3) apply here: f_y has already been mentioned to be WCM, while clearly it is not UAL, whereas f_0 is UAL but not WCM.

6. Localizable Sets of Distributions

Thus far we have been concerned with consistency in the sense of the set of all (appropriately bounded) distributions; this is a very general case. However, many specific types of learning problems consider more limited sets of distributions. Our goal in this section is to apply our results to such problems. In addition, this section will provide some of the tools used in Section 7 to derive results for classification.

This section relies on the following definition:

Definition 8 *We call a set of distributions $\mathbb{P}$* ***localizable*** *iff*

$$P \in \mathbb{P} \qquad \Longrightarrow \qquad \forall r \geq 0, x \in \text{supp}_X(P) \quad P_{x,r} \in \mathbb{P}.$$

That is, a localizable set of distributions contains all conditionings of its distributions onto small balls. Simple inspection of the proof of Lemma 6 reveals that it holds true on localizable sets of distributions, and not just on the set of all distributions, simply because we apply the WCM property of f only on distributions $P_{x,r}$. Consequently, it is easy to see that our results—specifically, Theorem 4 and Corollary 7—apply to localizable sets in general, and not just to the set of all distributions bounded by some constant $M > 0$ (which is just one type of example of a localizable set). Note that the requirement that a set of distributions be localizable is in a sense the minimal requirement we would expect, since if f is consistent on a distribution P but not on some $P_{x,r}$ then there is no reason to expect f to be UALC. We will say more about this matter in Section 8.

Are all standard learning problems defined (perhaps implicitly) on localizable sets of distributions? The answer is no; for example, if we assume that $Ey = 0$, which implies $E_x f^*(x) = 0$, then this is clearly not necessarily preserved when we consider some $P_{x,r}$. However, the converse is true for many standard setups in statistics and machine learning. Specific examples include the following, to all of which our results apply (assuming the boundedness assumption is upheld):

1. $y = f^*(x) + \varepsilon$, ε is independent of x , $E\varepsilon = 0$

   This is the standard regression model with additive noise (and random design) appearing in statistics. It is clearly a setup that implicitly works on a localizable set since $P_{x,r}$ retains the property that $y = f^*(x) + \varepsilon$.

2. As a subcase of the previous example, we can assume that $f^* \in \mathcal{F}$, where $\mathcal{F}$ is the set of all continuous functions, or alternatively some 'smoothness class', for example, Lipschitz-continuous functions, etc. Note, however, that if the density of the marginal distribution on X is assumed to be bounded then this is no longer a localizable set.

3. $P\big(y = f^*(x)\big) = 1$, $\forall S, x \;\; f(S,x), f^*(x) \in \{-1,+1\}$

   This is a noiseless classification problem, or set-estimation problem with non-overlapping sets, since

   $$P\big(f(S_n,x) \neq y\big) = E_{S_n,(x,y)} 1\big\{f(S_n,x) \neq y\big\} = \frac{1}{2} E_{S_n,x} \big|f(S_n,x) - f^*(x)\big| = \frac{1}{2} L_n(f).$$

   That is, the 0-1 loss used in classification is equal to (half) the $\mathcal{L}_1$ loss in this case. Note, however, that we assume $f(S,x) \in \{-1,+1\}$, and smoothed local versions do not have this property; for them the equality between the 0-1 and $\mathcal{L}_1$ losses is not valid. If we are willing to use the $\mathcal{L}_1$ norm for classification, however, then our results apply here.

   In the next section we will see a way to derive results for the 0-1 loss, as well as allow noisy distributions, that is, the standard classification setup. This will require somewhat different definitions than those used for regression.

4. $x = \phi(z)$ for some random variable z, where $\phi : R^d \to R^D$ is smooth and $D > d$.

   We conclude with this final somewhat more complex example in order to show how localizable sets can be present even in settings where we might not expect them. In the setting described here, the original data (z,y) lies on some low-dimensional space, but we observe $(x,y) = (\phi(z),y)$, which lies on a low-dimensional manifold inside a high-dimensional space. This sort of setting is considered in the manifold-learning field in unsupervised learning (see,

e.g., Roweis and Saul, 2000; Belkin and Niyogi, 2003); recently such methods have been applied to supervised learning (see, e.g., Kouropteva et al., 2003; Li et al., 2005). Note also the relevance to the kernel trick, in which a kernel implicitly defines such a transformation ϕ.

The assumption of $x = \phi(z)$ (and that ϕ is smooth) is a nontrivial requirement; however, it is easy to see that if we consider the set of all ϕ and z then this set is a localizable set of distributions.

7. Classification

As mentioned in the previous section, noiseless classification in the $\mathcal{L}_1$ norm can be dealt with using the results we have seen thus far, but this is quite limiting. Therefore in this section we consider the standard case of classification, using the natural 0-1 loss, and allowing for the possibility of noise.

In classification (also known as pattern recognition; Devroye et al., 1996) we consider only distributions for whom $Y = \{-1,+1\}$; when we say 'all distributions' in this section we mean only distributions of this sort. Note that such distributions form a localizable set, and thus we might expect our results to apply to them. Note also that $f^*(x) = E(y|x) = 2\eta(x) - 1$ where $\eta(x) = P(y = 1|x)$, the conditional probability. We call a learning method c a **classifier** iff $\forall S,x \;\; c(S,x) \in \{-1,+1\}$. We will use c,d, etc., to denote classifiers, to differentiate them from general learning methods, which we generally denote f,g.

In classification the natural loss is the 0-1,

$$R_{0-1}(c) = P\left(c(S_n,x) \neq y\right) = E_{S_n,(x,y)} 1\{c(S_n,x) \neq y\}$$

and the minimal (Bayesian) loss is

$$R^*_{0-1} = \inf_h E_{x,y} 1\{h(x) \neq y\}$$

where the infimum is taken over all measurable h. Denote the optimal (Bayesian) classification rule by

$$c^*(x) = \text{sign}(f^*(x)) = \text{sign}(2\eta(x) - 1)$$

which clearly minimizes R_{0-1}. We are interested in the relative loss $\Delta R_{0-1}(c) = R_{0-1}(c) - R^*_{0-1}$. This is well-known to be equal to (half the value of)

$$\widetilde{L}_n(c) \equiv E_{S_n,x}|c(S_n,x) - c^*(x)| \cdot |2\eta(x) - 1| = E_{S_n,x}|c(S_n,x) - c^*(x)| \cdot |f^*(x)|.$$

That is, we have the usual $\mathcal{L}_1$ loss but it is weighted according to the distance of $\eta(x)$ from $1/2$. Another difference is that we compare c to $c^* = \text{sign}(f^*)$ and not to f^*. These differences between $\widetilde{L}_n$ and the loss L_n we considered in the main part of this work prevent an immediate application of our results. We will therefore present alternative definitions that will allow us to get around this problem.

First, we define mutual consistency in the context of classification, which we will call **mutual classification-consistency** (or, briefly, **mutual *C*-consistency**), using

$$\widetilde{D}_n(c,d) = E_{S_n,x}|c(S_n,x) - d(S_n,x)| \cdot |2\eta(x) - 1| \longrightarrow 0$$

—that is, we simply add the same weighting as in $\widetilde{L}_n$. This leads to defining **classification-consistency** (or, briefly, ***C*-consistency**) on a distribution P as

$$\widetilde{L}_n(c) = \widetilde{D}_n(c,c^*) = E_{S_n,x}|c(S_n,x) - c^*(x)| \cdot |2\eta(x) - 1| \longrightarrow 0$$

(denoting $c^*(S,x) = c^*(x)$). Note that, in a similar way as in regression, C-consistency is a specific case of mutual C-consistency.

A change needs to be made to smoothed local versions to ensure that they remain classifiers, since $\overline{c}|_r^q$ can return values not in $\{-1,+1\}$. Define, therefore,

$$\widetilde{c}|_r^q(S,x) = \operatorname{sign}\left(\overline{c}|_r^q(S,x)\right) = \operatorname{sign}\left(E_{x'\sim P_{x,qr}} c(S(x,r),x')\right).$$

When we talk of smoothed local versions in the context of classification, we intend $\widetilde{c}|_r^q$.

We can now define a version of UAL for the context of classification, in the following natural way. Let **C-UAL** be the same as UAL, but replace D_n with $\widetilde{D}_n$, f with c and $\overline{f}|_r^q$ with $\widetilde{c}|_r^q$.

Instead of proving results 'from scratch' for the definitions given in this section, we can build upon the previous ones, using the following general technique. Let $f_{\ker}$ be a consistent Nadaraya-Watson kernel estimator using the window kernel. Define

$$f_{|\cdot|}(S,x) = |f_{\ker}(S,x)|.$$

For any classifier c, define a learning method

$$f_c(S,x) = c(S,x)\, f_{|\cdot|}(S,x).$$

We can see $f_c(S,x)$ as an estimate of $f^*(x)$, using (in a plug-in manner) $c(S,x)$ to estimate $\operatorname{sign}(f^*(x))$ and $f_{|\cdot|}(S,x)$ to estimate $|f^*(x)|$.

We will now use our results on regression for estimators f_c in order to arrive at conclusions for classifiers c. Note that $|f_c(S,x)| \le 1$ (since $f_{\ker}$ uses the window kernel, it is bounded by the largest $|y_i|$ in the training set, which is 1). Given the additional fact that in classification we have $Y = \{-1,+1\}$, we can conclude that our boundedness assumption on learning methods is of no consequence to our treatment of classification, that is, to get completely general results for classification we need only rely on bounded regression with $M = 1$.

To arrive at results, we will need some minor facts:

Lemma 9 *The following hold true for every P, c, d:*

1. $D_n(f_c, f_d) \to 0 \quad \Longleftrightarrow \quad \widetilde{D}_n(c,d) \to 0.$
2. $D_n(f_c, f^*) \to 0 \quad \Longleftrightarrow \quad \widetilde{D}_n(c,c^*) \to 0.$
3. f_c *is UALC* $\quad \Rightarrow \quad D_n\left((\overline{f_c})|_{\{R_k\}}^{\{Q(R_k)\}}, f_{\widetilde{c}|_{\{R_k\}}^{\{Q(R_k)\}}}\right) \to 0$
   for large-enough $Q, \{R_k\}$ in the sense appearing in the definitions UAL,UALC.

The first part of the lemma indicates that the C-mutual consistency of any c,d are linked to the mutual consistency of f_c, f_d; the second does the same for C-consistency and consistency. The third part of the lemma shows that if f_c is UALC then smoothed local versions of f_c are asymptotically equivalent to $f_{c'}$, where c' is a smoothed local version in the sense of classification of c (in other words, we can either smooth f_c or first smooth c and then apply f; the result is similar). A proof of all parts of the lemma appears in Appendix E.

We can now present an example of deriving results for classification from those for regression. Assume that c is C-consistent on some localizable set $\mathbb{P}$. Then f_c is consistent on $\mathbb{P}$ by part 2 of

Lemma 9, hence by Theorem 4 (and what we have seen regarding localizable sets in Section 6) f_c is UAL on $\mathbb{P}$, that is, for every $P \in \mathbb{P}$ we have

$$D_n\left(f_c, (\bar{f_c})|_{\{R_k\}}^{\{Q(R_k)\}}\right) \to 0$$

for large-enough Q, R_k in the sense of the definition of UAL. f_c is also UALC on $\mathbb{P}$, and therefore for large enough Q, R_k we have

$$D_n\left((\bar{f_c})|_{\{R_k\}}^{\{Q(R_k)\}}, f_{\tilde{c}|_{\{R_k\}}^{\{Q(R_k)\}}}\right) \to 0$$

by part 3 of Lemma 9, and therefore we conclude by the triangle inequality that (again, for large enough Q, R_k)

$$D_n\left(f_c, f_{\tilde{c}|_{\{R_k\}}^{\{Q(R_k)\}}}\right) \to 0.$$

Hence, by part 1 of Lemma 9, c and $\tilde{c}|_{\{R_k\}}^{\{Q(R_k)\}}$ are mutually C-consistent for large-enough Q, R_k, that is, c is C-UAL on P. Since this was for every $P \in \mathbb{P}$, we conclude that c is C-UAL on $\mathbb{P}$. Since the entire argument was for an arbitrary localizable set $\mathbb{P}$, we conclude that

Theorem 10 *A C-consistent classifier on a localizable set $\mathbb{P}$ is C-UAL on $\mathbb{P}$.*

In particular, since the set of all distributions (having $Y = \{-1, +1\}$) is localizable, we get

Corollary 11 *A C-consistent classifier is C-UAL.*

In a similar way we can define in the context of classification concepts parallel to UALC and WCM, and prove corresponding results (we do not go into details to avoid repetition).

8. Concluding Remarks

Our analysis of consistency has led to the following result: Consistent learning methods must have two properties, first, that they behave locally (UAL); second, that their mean must not be far from estimating the true mean (WCM). Only a learning method having these two independent properties is consistent, and vice versa; their combination is logically equivalent to consistency.

To further elaborate on this result, note that UAL is clearly a local property, and WCM a global one. Thus, we can see consistency as comprised of two aspects, one local and one global. Note also that the global property, WCM, is a trivial consequence of consistency, and therefore what is worth noting about this result is that consistency implies local behavior. We can then ask, in an informal manner at least, why is this so?

As noted in the introduction, the loss L_n that we considered is a global one, in that we average over x, and hence it does not seem to directly imply local behavior. What does seem to be the crux of the matter is the requirement to perform well on *all* distributions, or more generally on a localizable set; note that the term 'localizable' here is a giveaway. Indeed, if we have a method that is consistent on a *non*-localizable set of distributions, it may not behave locally. As a simple example (but complex enough for the underlying issues to be evident), let us assume we work on distributions having $Ey = 0$ (which was mentioned in Section 6 as being non-localizable). We might consider the following approach on this problem: Let f be some general learning method known

to be useful on other data sets; specifically, assume that f is consistent. We can then calibrate its output so that it returns an empirical mean of 0, which makes sense since we know that $Ey = 0$. That is, let

$$g(S,x) = f(S,x) - \frac{1}{|S|} \sum_{i=1..|S|} f(S,x_i)$$

(alternatively, we might calibrate by multiplying by an appropriate scalar, etc.). Then, while g is consistent on the set of distributions with $Ey = 0$ (using the consistency of f on all distributions), g does not behave locally. This can be seen, for example, by considering g on a distribution with x uniform on $[-1,1]$ and $y = \text{sign}(x)$. This distribution fulfills $Ey = 0$, and g is consistent on it, but neither UAL nor UALC, since smoothed local versions of g in fact return values tending towards 0.

Thus, in summary, the fact that the set of all distributions is localizable is what causes consistency to imply local behavior. If we are concerned about that fact (e.g., because we suspect local behavior might lead to the curse of dimensionality; Bengio et al., 2006), then we must do away with consistency on the set of all distributions and instead talk about consistency on a more limited set, one which is not localizable. However, part of the reason why nonparametric methods often outperform parametric ones on real-world data is precisely because they make as few as possible assumptions about the unknown distribution. Consequently, we may find that local behavior is hard to avoid.

We now turn to directions for future work. One such direction is to apply our results towards proving the consistency of learning methods not yet known to be consistent. It appears that in many cases proving the WCM property should not be difficult; hence, what remains is to prove UAL. While not necessarily a simple property to show, it may in some cases be easier than proving consistency directly.

An additional area for possible future work is empirical investigation, as follows. Since a consistent method is necessarily UALC, we can consider using a smoothed local version of the original method, since if chosen appropriately it is also consistent.[2] Now, if for example we have a large training set and only a few points at which to estimate, then by training on the local parts of the original training set we may save time. Of course, there is no guarantee that this will be beneficial on a particular problem, since consistency is an asymptotic property. Future empirical work might therefore examine real-world data sets in detail to see how the performance of local versions of a method compare to the original.

Acknowledgments

We thank the anonymous reviewers of a previous version of this paper for their comments. We also acknowledge support for this project from the Israel Science Foundation as well as partial support from NSF grant DMS-0605236.

2. Note that we might not need smoothing in practice, since the smoothing radius becomes negligibly small, $Q(R_k)R_k = o(R_k)$. That is, local versions might behave similarly enough to smoothed local versions for practical purposes.

Appendix A. Proof of Proposition 1

For any sample S, denote the mean and standard deviation (of the xs) by

$$\hat{E}(S) = \frac{1}{|S|}\sum_{i=1..|S|} x_i \qquad , \qquad \hat{\sigma}(S) = \sqrt{\frac{1}{|S|}\sum_{i=1..|S|}\left(x_i - \hat{E}(S)\right)^2}.$$

Now, start with a consistent method g, which is *translation-invariant* in the sense that

$$\forall S, x, b \qquad g(S,x) = g(S+b, x+b)$$

where $S+b = \{(x_i+b, y_i) \ : \ (x_i,y_i) \in S\}$ (for example, we can take g to be a kernel estimator).

Define, for any $q > 0$,

$$f^q(S,x) = \begin{cases} g(S,x) & x \notin B_{\hat{E}(S),q} \text{ or } x \in \{x_i \ : \ (x_i,y_i) \in S\} \\ 0 & \text{otherwise} \end{cases}$$

and let

$$Q(S) = \frac{\hat{\sigma}(S)}{\log(|S|)}.$$

Finally, let

$$f(S,x) = f^{Q(S)}(S,x).$$

We will first show that f is consistent; then we will show that f is neither locally consistent nor localizable. A brief overview of the proof is as follows: As $n \to \infty$, we get that $Q(S_n) \approx \frac{\text{const}}{\log(n)}$. This means that f is forced to return 0 on an area with vanishing radius, and hence f behaves like g on an area with measure rising to 1, and is consistent. On the other hand, when we consider a local version, we get that $Q(S_n(x,R_n)) \approx O\left(\frac{R_n}{\log(m)}\right)$, where m is the size of $S_n(x,R_n)$. We also get that x, the point at which we are estimating, is at distance $\approx O\left(\frac{R_n}{\sqrt{m}}\right)$ from $\hat{E}(S_n(x,R_n))$, a distance which is asymptotically smaller than Q. Hence x will tend to be in the area on which f is forced to return 0, making f neither locally consistent nor localizable.

We now start with the formal proof, first showing that f is consistent. Since X is bounded, $\hat{\sigma}(S_n)$ is bounded, and hence $Q(S_n) = \frac{\hat{\sigma}(S_n)}{\log(n)} \to 0$. We first assume that there is not a point mass on Ex, which implies $\mu\left(B_{\hat{E}(S_n),Q(S_n)}\right) \to 0$ almost surely, which follows from the fact that, since $\hat{E}(S_n) \to Ex$ a.s. (by the LLN) and $Q(S_n) \to 0$, we must have $\limsup_n B_{\hat{E}(S_n),Q(S_n)} \subseteq \{Ex\}$ with probability 1. We can then use the consistency of g to see that

$$\begin{aligned}
&E_{S_n}E_x\left|f(S_n,x) - f^*(x)\right| \\
&\quad \le E_{S_n}E_x 1\{x \notin B_{\hat{E}(S_n),Q(S_n)} \text{ or } x \in \{x_i \ : \ (x_i,y_i) \in S_n\}\}\left|g(S_n,x) - f^*(x)\right| \\
&\qquad + E_{S_n} 2M\mu\left(B_{\hat{E}(S_n),Q(S_n)} \bigcap \{x_i \ : \ (x_i,y_i) \in S_n\}^C\right) \\
&\quad \le E_{S_n}\left[E_x\left|g(S_n,x) - f^*(x)\right| + 2M\mu\left(B_{\hat{E}(S_n),Q(S_n)}\right)\right] \\
&\quad \to 0
\end{aligned}$$

which proves that f is consistent. Consider now the case where there is a point mass on Ex. Recall that $\limsup_n B_{\hat{E}(S_n),Q(S_n)} \subseteq \{Ex\}$ a.s., and note that due to the point mass on Ex, we have $P\big(Ex \in \{x_i : (x_i,y_i) \in S_n\}\big) \to 1$. Because of these two facts, we get

$$E_{S_n}\mu\left(B_{\hat{E}(S_n),Q(S_n)} \bigcap \{x_i : (x_i,y_i) \in S_n\}^C\right) \to 0$$

hence once more f is consistent by the consistency of g, by a similar argument as before.

We will now show that f is not locally consistent. For this, it is sufficient to show a single distribution on which $L_n\big(f|_{\{R_k\}}\big)$ does not converge to 0, for any sequence $R_k \to 0$. Take $X = [0,1]$ (higher-dimensional cases can be proved similarly), let μ be uniform on X, and let $y = +1$ with probability $(1+f^*(x))/2$, and otherwise $y = -1$ (which makes sense if $f^*(x) \in [-1,+1]$). Define

$$f^*(x) = \begin{cases} +1/2 & x \in \left[0,\frac{1}{2}\right] \\ -1/2 & x \in \left(\frac{1}{2},1\right] \end{cases}.$$

Fix some $x \in (0,1)$. Note that μ has no point masses, so $P(x \in \{x_i : (x_i,y_i) \in S_n\}) = 0$, and the relevant condition in the definition of f^q is of no consequence. Now, for large enough n (that is, small enough R_n) we have $x \in [R_n, 1-R_n]$; we will now focus on that case.

Denote $\widetilde{S} = S_n(x,R_n)$ and $m = m(n,x,R_n) = |\widetilde{S}|$. Notice that m is the sum of i.i.d Bernoulli variables, and that

$$Em = 2nR_n \quad , \quad \sigma^2(m) \le 2nR_n \tag{6}$$

(recall that $Em, \sigma^2(m)$ denote the expected value and variance of the random variable m, respectively). Now, consider the case in which $nR_n \not\to \infty$. Then there is some bounded subsequence, $k_nR_{k_n} \le K$ for all n. Then (using Chebyshev) clearly m is less than $4k_nR_{k_n}$ with non-vanishing probability on this subsequence. Since $R_k \to 0$ then in order for f to be consistent we must, for large enough n, discriminate between the two possibilities $f^*(x) = +1/2, f^*(x) = -1/2$ in a way that does not depend upon x (due to the translation-invariance of f, which stems from the translation-invariance of g). But discriminating between the two cases with arbitrarily small error cannot be done with a bounded sample size, hence the loss cannot go to 0. Thus, we conclude that if $nR_n \not\to \infty$ then f is not locally consistent.

Consider now the other case, of $nR_n \to \infty$. We start by formulating bounds for m, $|x - \hat{E}(\widetilde{S})|$, and $\hat{\sigma}^2(\widetilde{S})$.

- m: Using Bernstein's Inequality and (6), we get

$$P(|m-2nR_n| \ge t) \le 2\exp\left(-\frac{1}{2}\frac{t^2}{2nR_n + t/3}\right).$$

Picking $t = t_n = nR_n$, we get

$$P(|m-2nR_n| \ge nR_n) \le 2\exp\left(-\frac{1}{6}nR_n\right). \tag{7}$$

Note that this bound converges to 0 since $nR_n \to \infty$.

- $|x-\hat{E}(\widetilde{S})|$: Clearly x is the mean of P_{x,R_n}, and also the mean of the individual observations in $\widetilde{S}$, as they are distributed i.i.d as P_{x,R_n}. Note that for every $x_i \in \widetilde{S}$ (i.e., taken from the distribution P_{x,R_n}) we have $\sigma^2(x_i) = \frac{R_n^2}{3}$. Then, by Hoeffding's Inequality,

$$P\left(\left|x-\hat{E}(\widetilde{S})\right| \geq t\right) \leq E_m\left\{2\exp\left(-\frac{mt^2}{2R_n^2}\right)\right\}.$$

Picking $t = t_n = R_n\sqrt{\frac{\log(nR_n)}{nR_n}}$, we get

$$\begin{aligned} P\left(\left|x-\hat{E}(\widetilde{S})\right| \geq R_n\sqrt{\frac{\log(nR_n)}{nR_n}}\right) &\leq E_m\left\{2\exp\left(-\frac{1}{2}\frac{m}{nR_n}\log(nR_n)\right)\right\} \\ &\leq 2\exp\left(-\frac{1}{2}\log(nR_n)\right) + 4\exp\left(-\frac{1}{6}nR_n\right) \\ &\to 0 \end{aligned} \tag{8}$$

using the bound for m above (7).

- A final bound that we will need relates to $\hat{\sigma}^2(\widetilde{S})$. Note that for any sample S and point x,

$$\hat{\sigma}^2(S) = \frac{1}{|S|}\sum_i (x_i-x)^2 - (x-\hat{E}(S))^2$$

Consider the first expression on the RHS. On our (sub)sample $\widetilde{S}$, we have for every $x_i \in \widetilde{S}$ that $E(x_i-x)^2 = \frac{R_n^2}{3}$, and note that $(x_i-x)^2 \leq R_n^2$. Then, by Hoeffding's Inequality,

$$P\left(\left|\frac{1}{m}\sum_{i=1}^m (x_i-x)^2 - \frac{R_n^2}{3}\right| \geq t\right) \leq E_m 2\exp\left(-\frac{mt^2}{2R_n^4}\right).$$

Picking $t = t_n = R_n^2\sqrt{\frac{\log(nR_n)}{nR_n}}$, we get

$$\begin{aligned} P\left(\left|\frac{1}{m}\sum_{i=1}^m (x_i-x)^2 - \frac{R_n^2}{3}\right| \geq R_n^2\sqrt{\frac{\log(nR_n)}{nR_n}}\right) \\ \leq E_m 2\exp\left(-\frac{1}{2}\frac{m}{nR_n}\log(nR_n)\right) \\ \leq 2\exp\left(-\frac{1}{2}\log(nR_n)\right) + 4\exp\left(-\frac{1}{6}nR_n\right) \\ \to 0 \end{aligned}$$

similarly as before. Combined with the bound from before for $|x-\hat{E}(\widetilde{S})|$, we get, for large enough n,

$$\begin{aligned}
&P\left(\hat{\sigma}^2(\widetilde{S}) \le \frac{1}{12}R_n^2\right) \\
&\quad = P\left(\hat{\sigma}^2(\widetilde{S}) \le \frac{1}{12}R_n^2\,,\ \left(x-\hat{E}(\widetilde{S})\right)^2 < R_n^2\frac{\log(nR_n)}{nR_n}\right) \\
&\qquad + P\left(\hat{\sigma}^2(\widetilde{S}) \le \frac{1}{12}R_n^2\,,\ \left(x-\hat{E}(\widetilde{S})\right)^2 \ge R_n^2\frac{\log(nR_n)}{nR_n}\right) \\
&\quad \le P\left(\frac{1}{m}\sum_{i=1}^{m}(x_i-x)^2 \le \frac{1}{6}R_n^2\right) + P\left(\left|x-\hat{E}(\widetilde{S})\right| \ge R_n\sqrt{\frac{\log(nR_n)}{nR_n}}\right) \\
&\quad \to 0.
\end{aligned}$$

We now have all the bounds we need. Using our results for $m=|\widetilde{S}|$ and $\hat{\sigma}^2(\widetilde{S})$ together, we get

$$\begin{aligned}
&\lim_{n\to\infty} P\left(\left|x-\hat{E}(\widetilde{S})\right| \ge Q(\widetilde{S})\right) \\
&\quad = \lim_{n\to\infty} P\left(\left|x-\hat{E}(\widetilde{S})\right| \ge \frac{\hat{\sigma}(\widetilde{S})}{\log(m)}\right) \\
&\quad = \lim_{n\to\infty} P\left(\left|x-\hat{E}(\widetilde{S})\right| \ge \frac{\hat{\sigma}(\widetilde{S})}{\log(m)}\,,\ |m-2nR_n| \le nR_n\,, \hat{\sigma}^2(\widetilde{S}) \ge \frac{1}{12}R_n^2\right) \\
&\quad \le \lim_{n\to\infty} P\left(\left|x-\hat{E}(\widetilde{S})\right| \ge \frac{R_n}{\sqrt{12}\log(3nR_n)}\right) \\
&\quad \to 0
\end{aligned}$$

where to reach the very last line we used what we know about $|x-\hat{E}(\widetilde{S})|$, as appearing in (8). Thus, for every $x \in \mathrm{supp}_X(P)$ we have $P\left(x \in B_{\hat{E}(\widetilde{S}),Q(\widetilde{S})}\right) \to 1$.

Finally, we can see that

$$\begin{aligned}
E_{S_n}\left|f|_{\{R_k\}}(S_n,x) - f^*(x)\right| &= E_{S_n}\left|f(\widetilde{S},x) - f^*(x)\right| \\
&= E_{S_n}1\{x \in B_{\hat{E}(\widetilde{S}),Q(\widetilde{S})}\}\left|f(\widetilde{S},x) - f^*(x)\right| \\
&\quad + E_{S_n}1\{x \notin B_{\hat{E}(\widetilde{S}),Q(\widetilde{S})}\}\left|f(\widetilde{S},x) - f^*(x)\right| \\
&\ge E_{S_n}1\{x \in B_{\hat{E}(\widetilde{S}),Q(\widetilde{S})}\}\,|f^*(x)| \\
&= \frac{1}{2}E_{S_n}1\{x \in B_{\hat{E}(\widetilde{S}),Q(\widetilde{S})}\} \\
&= \frac{1}{2}P\left(x \in B_{\hat{E}(\widetilde{S}),Q(\widetilde{S})}\right) \\
&\to \frac{1}{2}
\end{aligned}$$

where we used that $f^*(x) \in \{-1/2, +1/2\}$. Taking the expected value over x, the dominated convergence theorem gives us

$$\lim_{n\to\infty} L_n\left(f|_{\{R_k\}}\right) \geq \frac{1}{2}.$$

Thus, $f|_{\{R_k\}}$ is not consistent, that is, f is not locally consistent (note that this is even by a relatively large constant factor).

Having shown that f is consistent but not locally consistent, we now show that in addition f cannot be localizable. This is immediate, since if f were localizable, then some sequence $R_k \to 0$ would exist for which $D_n\left(f, f|_{\{R_k\}}\right) \to 0$, and therefore

$$L_n\left(f|_{\{R_k\}}\right) = D_n\left(f|_{\{R_k\}}, f^*\right) \leq D_n\left(f|_{\{R_k\}}, f\right) + D_n(f, f^*) \to 0$$

by the localizability and consistency of f. But this result implies f is locally consistent, contradicting what we saw earlier.

Appendix B. Proof of Lemma 6

Define (as in the proof of Proposition 1) $\widetilde{S} = S_n(x,r)$, $m = m(n,x,r) = |S_n(x,r)|$, the size of the local training set. Note that we can see $\widetilde{S}$ as m points sampled from $P_{x,r}$. Then we get, for any $r, q > 0$,

$$\begin{aligned}
D_n\left(\overline{f}|_r^q, f^*\right) &= E_{S_n,x}\left|E_{x'\sim P_{x,qr}} f(S_n(x,r),x') - f^*(x)\right| \\
&= E_{x,m} E_{\widetilde{S}\sim P_{x,r}\,|\,m}\left|E_{x'\sim P_{x,qr}} f(\widetilde{S},x') - f^*(x)\right| \\
&\leq E_{x,m} E_{\widetilde{S}\sim P_{x,r}\,|\,m} E_{x'\sim P_{x,qr}}\left|f(\widetilde{S},x') - f^*(x)\right| \\
&\leq E_{x,m} E_{\widetilde{S}\sim P_{x,r}\,|\,m\,,\,x'\sim P_{x,qr}}\left|f(\widetilde{S},x') - f^*(x')\right| \\
&\quad + E_{x,x'\sim P_{x,qr}}\left|f^*(x') - f^*(x)\right|.
\end{aligned}$$

We now start to consider the limit behavior of these expressions when we replace $r > 0$ with $\{R_k\} \in \mathcal{R}(T)$ and $q > 0$ with $\{Q(R_k)\}, Q \in \mathcal{Q}(T)$, where $T \in \mathcal{T}$ is arbitrary. First, for any such $\{R_k\}, Q$ we have

$$\lim_{n\to\infty} E_{x,x'\sim P_{x,Q(R_n)R_n}}\left|f^*(x') - f^*(x)\right| = 0$$

by the corollary to the following lemma (and since $R_n, Q(R_n) \to 0$):

Lemma 12 *[Devroye, 1981; Lemma 1.1] For any distribution P and measurable g, if $E_{x\sim P}|g(x)| < \infty$ then*

$$\lim_{r\to 0} E_{x'\sim P_{x,r}} g(x') = g(x)$$

for almost all x.

Corollary 13 *For any distribution P and measurable g, if $E_{x\sim P}|g(x)| < \infty$ then*

$$\lim_{r\to 0} E_{x'\sim P_{x,r}}\left|g(x') - g(x)\right| = 0$$

for almost all x.

Write $\widetilde{S} = S_n(x, R_n)$ (a minor abuse of our notation, as we also write $\widetilde{S} = S_n(x,r)$, but r is always a placeholder for R_n in any case). Then

$$\limsup_{n\to\infty} D_n\left(\overline{f}|_{\{R_k\}}^{\{Q(R_k)\}}, f^*\right) \le \limsup_{n\to\infty} E_{x,m} E_{\widetilde{S}\sim P_{x,R_n} \mid m\,,\, x'\sim P_{x,Q(R_n)R_n}} \left| f(\widetilde{S},x') - f^*(x') \right|.$$

To simplify notation for this last expression, define, for any $x \in \mathrm{supp}_X(P), n \in \mathbb{N}, r, q \in T$,

$$C(x,n,r,q) = E_m E_{\widetilde{S}\sim P_{x,r} \mid m\,,\, x'\sim P_{x,qr}} \left| f(\widetilde{S},x') - f^*(x') \right|$$

and

$$C(n,r,q) = E_x C(x,n,r,q).$$

It is therefore our goal to show that $C(n, R_n, Q(R_n)) \to 0$ for appropriate $\{R_k\}, Q$. Towards that end, we consider the limit $\limsup_{n\to\infty} C(n,r,q)$, for fixed r, q. We have, for every $x \in \mathrm{supp}_X(P)$,

$$\begin{aligned}
&C(x,n,r,q) \\
&= E_m E_{\widetilde{S}\sim P_{x,r} \mid m\,,\, x'\sim P_{x,qr}} \left| f(\widetilde{S},x') - f^*(x') \right| \\
&= \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} E_m E_{\widetilde{S}\sim P_{x,r} \mid m\,,\, x'\sim P_{x,r}} \left| f(\widetilde{S},x') - f^*(x') \right| 1\left\{x' \in B_{x,qr}\right\} \\
&\le \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} E_m E_{\widetilde{S},x'\sim P_{x,r} \mid m} \left| f(\widetilde{S},x') - f^*(x') \right| \\
&= \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} E_m E_{\widetilde{S},x'\sim P_{x,r} \mid m} \left| f(\widetilde{S},x') - E_m(f) + E_m(f) - E(f^*) + E(f^*) - f^*(x') \right| \\
&\le \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} E_m \left[\mathrm{MAD}_{m,P_{x,r}}(f) + |E_{m,P_{x,r}}(f) - E_{P_{x,r}}(f^*)| + \mathrm{MAD}_{P_{x,r}}(f^*) \right]
\end{aligned}$$

where the expected values $E_m(f), E(f^*)$ on the line before last are w.r.t $P_{x,r}$, and the expected values and MADs on the last two lines are all conditional on m.

Now, clearly $m \to \infty$ almost surely (since x is in the support of μ, and $r > 0$ is fixed, so there is a positive probability for an observation to fall within $B_{x,r}$). Hence, by the WCM property of f on $P_{x,r}$ (a *fixed* distribution in the current view),

$$\limsup_{n\to\infty} C(x,n,r,q) \le \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} \left[\mathrm{MAD}_{P_{x,r}}(f^*) + 2H\left(\mathrm{MAD}_{P_{x,r}}(f^*)\right) \right].$$

Using this bound, we can then conclude that

$$\begin{aligned}
\limsup_{n\to\infty} C(n,r,q) &= \limsup_{n\to\infty} E_x C(x,n,r,q) \\
&\le E_x \limsup_{n\to\infty} C(x,n,r,q) \\
&\le E_x \min\left\{ 2M\,,\, \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} \left[\mathrm{MAD}_{P_{x,r}}(f^*) + 2H\left(\mathrm{MAD}_{P_{x,r}}(f^*)\right) \right] \right\}
\end{aligned} \tag{9}$$

using the Fatou-Lebesgue theorem for the first inequality, where we rely on the trivial fact that $C(x,n,r,q) \leq 2M$ based on our boundedness assumptions on f, f^*, and using those same boundedness assumptions for the second inequality as well. We define

$$C^*(r,q) = \limsup_{n\to\infty} C(n,r,q),$$
$$\bar{C}^*(r,q) = E_x \min\left\{2M\ ,\ \frac{\mu(B_{x,r})}{\mu(B_{x,qr})}\left[\mathrm{MAD}_{P_{x,r}(f^*)} + 2H\left(\mathrm{MAD}_{P_{x,r}}(f^*)\right)\right]\right\}.$$

Hence, based on (9) we have

$$C^*(r,q) \leq \bar{C}^*(r,q) \tag{10}$$

We will now need the following lemma:

Lemma 14 *[Devroye, 1981; see the proof of Lemma 2.2] For any measure μ, for almost every x,*

$$\lim_{r\to 0} \frac{r^d}{\mu(B_{x,r})} = c_x \quad , \quad |c_x| < \infty.$$

That is, the limit exists and is finite.

We now consider $\bar{C}^*$ for fixed q and varying r. By Lemma 14, we know that for almost every x,

$$\lim_{r\to 0} \frac{\mu(B_{x,r})}{\mu(B_{x,qr})} = \lim_{r\to 0} \frac{c_x r^d}{c_x q^d r^d} = \frac{1}{q^d} \tag{11}$$

and using Lemma 12 we can consider the MAD, as follows. First, by Lemma 12 we have, for almost every x,

$$\lim_{r\to 0} E_{x'\sim P_{x,r}} f^*(x') = f^*(x)$$

so, for almost every x,

$$\begin{aligned}
\lim_{r\to 0} \mathrm{MAD}_{P_{x,r}}(f^*) &= \lim_{r\to 0} E_{x'\sim P_{x,r}} |f^*(x') - E_{x''\sim P_{x,r}} f^*(x'')| \\
&\leq \lim_{r\to 0} E_{x'\sim P_{x,r}} |f^*(x') - f^*(x)| + |f^*(x) - E_{x''\sim P_{x,r}} f^*(x'')| \\
&= \lim_{r\to 0} E_{x'\sim P_{x,r}} |f^*(x') - f^*(x)| \\
&= 0
\end{aligned}$$

using Corollary 13 for the last equality. Combining this with (11), and using the properties of H, we get

$$\lim_{r\to 0} \bar{C}^*(r,q) = \lim_{r\to 0} E_x \min\left\{2M\ ,\ \frac{\mu(B_{x,r})}{\mu(B_{x,qr})}\left[\mathrm{MAD}_{P_{x,r}}(f^*) + 2H\left(\mathrm{MAD}_{P_{x,r}}(f^*)\right)\right]\right\} = 0$$

since we have convergence to 0 for almost every x in the expected value, and can apply the dominated convergence theorem (for which the bound of $2M$ is crucial). Consequently, due to (10) we have

$$\lim_{r\to 0} C^*(r,q) \leq \lim_{r\to 0} \bar{C}^*(r,q) = 0. \tag{12}$$

We now turn to defining $\{R_k\}$ and Q, using what we have seen thus far. Recall that $T = \{T_k\}$ is arbitrary and that $T_k \searrow 0$. Define in a recursive manner, for every $q \in \mathbf{T}$ (where recall that $\mathbf{T}$ is the set containing all the members in the sequence T),

$$K(T_0) = 0,$$
$$K(q) = \min\left\{k \in \mathbb{N} \quad : \quad \begin{array}{c} \forall q' \in \mathbf{T}, q' > q \\ k > K(q') \end{array} \text{ and } \begin{array}{c} \forall k' \geq k \, \forall q' \in \mathbf{T}, q' \geq q \\ C^*(T_{k'}, q') \leq q' \end{array}\right\}$$

(since $T_k \searrow 0$, we start by defining K for the largest value in T, which is T_0, and then proceed to lower ones). Note that the clause regarding C^* is fulfillable by (12) for individual q', and since we consider a finite number of such q', we can take k large enough for them all. Note also that we ensure that $K(T_k)$ strictly increases, for which we rely on the fact that T_k strictly descends.

We now define Q:

$$\begin{array}{rl} Q(T_0), ..., Q(T_{K(T_1)}) & = T_0, \\ Q(T_{K(T_1)+1}), ..., Q(T_{K(T_2)}) & = T_1, \\ \vdots & \\ Q(T_{K(T_k)+1}), ..., Q(T_{K(T_{k+1})}) & = T_k, \\ \vdots & \end{array} \tag{13}$$

Then according to these definitions, for any $Q' \in \mathcal{Q}(T), Q' \geq Q$, and any (large enough) $k' \in \mathbb{N}$,

$$Q'(T_{k'}) \geq Q(T_{k'}) = T_k \quad \text{for some } k \text{ fulfilling} \quad k' > K(T_k)$$

and hence, for all large enough k',

$$C^*(T_{k'}, Q'(T_{k'})) \leq Q'(T_{k'}). \tag{14}$$

We now work towards defining $R = \{R_k\}$, which just as with Q will be done in two stages. First, we define in a recursive manner, for every $r, q \in \mathbf{T}$,

$$N(T_0, T_0) = 0,$$
$$N(r,q) =$$
$$\min\left\{\tilde{n} \in \mathbb{N} \quad : \quad \begin{array}{c} \forall r', q' \in \mathbf{T}, r' > r, q' \geq q \\ \tilde{n} > N(r', q') \end{array} \text{ and } \begin{array}{c} \forall \tilde{n}' \geq \tilde{n} \, \forall r', q' \in \mathbf{T}, r' \geq r, q' \geq q \\ C(\tilde{n}', r', q') \leq 2C^*(r', q') \end{array}\right\}.$$

Note that the clause regarding C, C^* is fulfillable by the definition of C^* as the lim sup of C, and hence we can achieve it over a set of finite q, r as well. Note also that we ensure $N(r,q)$ strictly increases when r strictly descends (and q does not rise).

Define, for any $Q' \in \mathcal{Q}(T)$,

$$\begin{array}{rl} R_0, ..., R_{N(T_1, Q'(T_1))} & = T_0, \\ R_{N(T_1, Q'(T_1))+1}, ..., R_{N(T_2, Q'(T_2))} & = T_1, \\ \vdots & \\ R_{N(T_k, Q'(T_k))+1}, ..., R_{N(T_{k+1}, Q'(T_{k+1}))} & = T_k, \\ \vdots & \end{array}$$

Then, for any $R' \in \mathcal{R}(T), R' \geq R$, and if n is large enough,

$$R'_n \geq R_n = T_k \text{ for some } k \text{ fulfilling} \quad n > N(T_k, Q'(T_k)).$$

Note that $Q'(R'_n) \geq Q'(T_k)$. Thus, by the definition of $N(\cdot,\cdot)$,

$$C(n, R'_n, Q'(R'_n)) \leq 2C^*(R'_n, Q'(R'_n)).$$

If we now also assume that $Q' \geq Q$, where Q is as defined in (13), then by (14) we get

$$C(n, R'_n, Q'(R'_n)) \leq 2C^*(R'_n, Q'(R'_n)) \leq 2Q'(R'_n) \underset{n\to\infty}{\longrightarrow} 0$$

completing the proof.

Appendix C. Proof of Lemma 9

1. First, note that

$$\forall P \qquad E_{S_n,x}\left|f_{|\cdot|}(S_n,x) - |f^*(x)|\right| \leq E_{S_n,x}\left|f_{\ker}(S_n,x) - f^*(x)\right| \longrightarrow 0. \tag{15}$$

Now, notice that for any P, c, d,

$$\begin{aligned}
D_n(f_c, f_d) &= E_{S_n,x}\left|c(S_n,x) f_{|\cdot|}(S_n,x) - d(S_n,x) f_{|\cdot|}(S_n,x)\right| \\
&= E_{S_n,x}\left|c(S_n,x) - d(S_n,x)\right| f_{|\cdot|}(S_n,x) \\
&= E_{S_n,x}\left|c(S_n,x) - d(S_n,x)\right| |2\eta(x) - 1| \\
&\quad + E_{S_n,x}\left|c(S_n,x) - d(S_n,x)\right| \left(f_{|\cdot|}(S_n,x) - |2\eta(x)-1|\right) \\
&= \widetilde{D}_n(c,d) + E_{S_n,x}\left|c(S_n,x) - d(S_n,x)\right| \left(f_{|\cdot|}(S_n,x) - |2\eta(x)-1|\right).
\end{aligned}$$

The last expression converges to 0 by (15) since

$$E_{S_n,x}\left|c(S_n,x) - d(S_n,x)\right| \left(f_{|\cdot|}(S_n,x) - |2\eta(x)-1|\right) \leq 2E_{S_n,x}\left|f_{|\cdot|}(S_n,x) - |2\eta(x)-1|\right| \to 0$$

which leads directly to the result we wanted.

2. By part 1 we know that $D_n(f_c, f_{c^*}) \to 0 \iff \widetilde{D}_n(c, c^*) \to 0$, so it suffices to prove that $\forall P \quad D_n(f_{c^*}, f^*) \to 0$, which can be shown using (15):

$$D_n(f_{c^*}, f^*) = E_{S_n,x}\left|c^*(x) f_{|\cdot|}(S_n,x) - f^*(x)\right| = E_{S_n,x}\left|f_{|\cdot|}(S_n,x) - |f^*(x)|\right| \to 0.$$

3. We consider $D_n\left((\overline{f_c})|_r^q, f_{\widetilde{c}|_r^q}\right)$; later we will replace q, r with $\{Q(R_k)\}, \{R_k\}$. By the definitions,

$$\begin{aligned}
&D_n\left((\overline{f_c})|_r^q, f_{\widetilde{c}|_r^q}\right) \\
&= E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') f_{|\cdot|}(S_n(x,r),x') - \operatorname{sign}\left(E_{x'\sim P_{x,qr}} c(S_n(x,r),x')\right) f_{|\cdot|}(S_n,x)\right| \\
&\leq E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') - \operatorname{sign}\left(E_{x'\sim P_{x,qr}} c(S_n(x,r),x')\right)\right| |f^*(x)| \\
&\quad + E_{S_n,x}\left|\operatorname{sign}\left(E_{x'\sim P_{x,qr}} c(S_n(x,r),x')\right)\left(f_{|\cdot|}(S_n,x) - |f^*(x)|\right)\right| \\
&\quad + E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x')\left(f_{|\cdot|}(S_n(x,r),x') - |f^*(x)|\right)\right| \qquad (16) \\
&\leq E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') - c^*(x)\right| |f^*(x)| \\
&\quad + E_{S_n,x}\left|f_{|\cdot|}(S_n,x) - |f^*(x)|\right| \\
&\quad + E_{S_n,x} E_{x'\sim P_{x,qr}}\left|f_{|\cdot|}(S_n(x,r),x') - |f^*(x)|\right|
\end{aligned}$$

where we used the fact that $|a-\text{sign}(a)| \le |a-b|$ for any $b \in \{-1,+1\}$. We now consider the final three expressions separately, denoting them (1),(2),(3):

(1) :
$$\begin{aligned}
&E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') - c^*(x)\right| |f^*(x)| \\
&= E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') |f^*(x)| - f^*(x)\right| \\
&\le E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') \left(|f^*(x)| - f_{|\cdot|}(S_n(x,r),x')\right)\right| \\
&\quad + E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') f_{|\cdot|}(S_n(x,r),x') - f^*(x)\right| \\
&\le E_{S_n,x} E_{x'\sim P_{x,qr}} \left||f^*(x)| - f_{|\cdot|}(S_n(x,r),x')\right| \\
&\quad + E_{S_n,x}\left|E_{x'\sim P_{x,qr}} c(S_n(x,r),x') f_{|\cdot|}(S_n(x,r),x') - f^*(x)\right|.
\end{aligned}$$

Of the last two expressions, the first is equal to the last of the three expressions we arrived at in (16), so we can consider later on double the value of that expression instead of handling it here. Regarding the second, it is simply equal to $D_n\left((f_c)|_r^q, f^*\right)$, which we know to converge to 0 for large-enough $Q, \{R_k\}$ since f_c is UALC on P.

(2) : This converges to 0 by (15).

(3) : Since

$$E_{S_n,x} E_{x'\sim P_{x,qr}} \left|f_{|\cdot|}(S_n(x,r),x') - |f^*(x)|\right| \le E_{S_n,x} E_{x'\sim P_{x,qr}} \left|f_{\ker}(S_n(x,r),x') - f^*(x)\right|$$

we can use the same techniques as in the proof of Lemma 6:

$$\begin{aligned}
&E_{S_n,x} E_{x'\sim P_{x,qr}} \left|f_{\ker}(S_n(x,r),x') - f^*(x)\right| \\
&\le E_{S_n,x} E_{x'\sim P_{x,qr}} \left|f_{\ker}(S_n(x,r),x') - f^*(x')\right| \\
&\quad + E_x E_{x'\sim P_{x,qr}} \left|f^*(x') - f^*(x)\right| \\
&= E_x \frac{\mu(B_{x,r})}{\mu(B_{x,rq})} E_m E_{\widetilde{S}\sim P_{x,r}\,|\,m} E_{x'\sim P_{x,r}} \left|f_{\ker}(\widetilde{S},x') - f^*(x')\right| 1\{x'\in B_{x,rq}\} \\
&\quad + E_x E_{x'\sim P_{x,qr}} \left|f^*(x') - f^*(x)\right| \\
&\le E_x \frac{\mu(B_{x,r})}{\mu(B_{x,rq})} E_m E_{\widetilde{S},x'\sim P_{x,r}\,|\,m} \left|f_{\ker}(\widetilde{S},x') - f^*(x')\right| \\
&\quad + E_x E_{x'\sim P_{x,qr}} \left|f^*(x') - f^*(x)\right|
\end{aligned}$$

where, as in previous proofs, $m = |S_n(x,r)|$. The expression before last converges to 0 for every $x \in \text{supp}_X(P)$ and fixed $r,q>0$ by the consistency of $f_{\ker}$ on $P_{x,r}$ (using the fact that $m \to \infty$ a.s.), and therefore in a similar way as that performed in Lemma 6 we can see that the entire expression (with expected value over x) converges to 0 for slowly-enough descending $Q, \{R_k\}$. The final expression converges to 0 for any $qr \to 0$ by Corollary 13.

Thus, by replacing q, r with large-enough $Q, \{R_k\}$ we can see that the original expression converges to 0, as required.

References

C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

P. L. Bartlett and M. Traskin. Adaboost is consistent. In *Advances in Neural Information Processing Systems 19*, pages 105–112. MIT Press, Cambridge, MA, 2007.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Y. Bengio, O. Delalleau, and N. Le Roux. The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems 18*, pages 107–114. MIT Press, Cambridge, MA, 2006.

L. Bottou and V. N. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

W. Cleveland and C. Loader. Smoothing by local regression: Principles and methods. Technical report, AT&T Bell Laboratories, Murray Hill, NY., 1995. Available at `http://citeseer.ist.psu.edu/194800.html`.

L. Devroye. On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics*, 9:1310–1319, 1981.

L. Devroye and T. J. Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8(2):231–239, 1980.

L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, NY, 1996.

Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Soc. for Artif. Intel.*, 14(5): 771–780, 1999.

L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, Berlin, 2002.

O. Kouropteva, O. Okun, and M. Pietikine. Supervised locally linear embedding algorithm for pattern recognition. *Lecture Notes in Computer Science*, 2652:386–394, 2003.

H. Li, L. Teng, W. Chen, and I. Shen. Supervised learning on local tangent space. In L. Gyorfi, editor, *Lecture Notes in Computer Science*, pages 546–551. Springer-Verlag, 2005.

G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41(3):677–687, 1995.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

G. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proc. 15th International Conf. on Machine Learning*, pages 515–521. Morgan Kaufmann, San Francisco, CA, 1998.

A. J. Smola and B. Schoelkopf. A tutorial on support vector regression, 1998. Available at `http://citeseer.ist.psu.edu/smola03tutorial.html`.

I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.

C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, NY, 1998.

V. N. Vapnik and L. Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, 1993.

A. Zakai and Y. Ritov. How local should a learning method be? In *21st Annual Conference on Learning Theory (COLT)*, pages 205–216, 2008.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–134, 2004.