How Local Should a Learning Method Be?

Alon Zakai *
Interdisciplinary Center for Neural Computation
Hebrew University of Jerusalem
Jerusalem, Israel 91904
alon.zakai@mail.huji.ac.il

Ya'acov Ritov *
Department of Statistics
Hebrew University of Jerusalem
Jerusalem, Israel 91905
yaacov.ritov@huji.ac.il

Abstract

We consider the question of why modern machine learning methods like support vector machines outperform earlier nonparametric techniques like k-NN. Our approach investigates the *locality* of learning methods, i.e., the tendency to focus mainly on the close-by part of the training set when constructing a new guess at a particular location. We show that, on the one hand, we can expect *all* consistent learning methods to be local in some sense; hence if we consider consistency a desirable property then a degree of locality is unavoidable. On the other hand, we also claim that earlier methods like k-NN are local in a more strict manner which implies performance limitations. Thus, we argue that a degree of locality is necessary but that this should not be overdone. Support vector machines and related techniques strike a good balance in this matter, which we suggest may partially explain their good performance in practice.

1 Introduction

It is commonly seen in practice that modern methods in machine learning – such as kernel machines and more specifically support vector machines – outperform older techniques in nonparametric statistics such as k-NN [for a concrete example, see, e.g., Joa98]. The main approaches to explaining this phenomenon are margin-based bounds on the generalization error and that margin maximization in effect minimizes the VC dimension, again, arriving at a favorable bound on the generalization error [Vap98, STC00]. In this work we consider an alternative approach to investigating this matter, in hopes of showing the underlying issues in a different light.

We will focus on **local learning**, i.e., the property of a learning method that it uses mainly the close-by part of the training set to construct new guesses. That is, when an estimate is generated at a point x using a training set $S_n = \{(x_i, y_i)\}_{i=1..n}$ (i.e., we are trying to guess a corresponding value of y for x, using x and the training set), then a local method is one that is influenced mostly by the points (x_i, y_i) for which x_i is close to x. Many classical methods in nonparametric statistics are clearly of this sort, e.g., k-NN. This is often stated as a detriment of such methods, in particular since local learning is susceptible to the curse of dimensionality – in high-dimensional spaces, one needs a great many points in the training set in order for a sufficient amount to end up close-by to the point x currently being estimated. On the other hand, methods like support vector machines appear non-local in their definition – the separating hyperplane is determined by the entire training set, and furthermore does not depend on the particular point we intend to estimate at – and thus one might suspect that the superior performance of support vector machines and related techniques is connected to this matter.

However, whether this is the case is not immediately obvious. In fact, we might suspect that many kernel machines behave locally: consider that a typical kernel machine can be written as $\sum_i \alpha_i y_i k(x_i, x)$, where k is a kernel function, e.g., the RBF kernel $k_{\text{RBF}}(x, x') = \exp(-\gamma ||x - x'||^2)$ (we do not write the sign operation, which would appear here if our goal is classification and not regression, but the issue is the same in that case). This appears outwardly similar to weighted k-NN, whose general form is $\frac{\sum_i y_i s(x_i,x)}{\sum_i s(x_i,x)}$, where the sum is taken over the K nearest neighbors of x and s is a similarity measure; in fact, we can take $s = k_{\text{RBF}}$. Also similar in form is another classical statistical technique, kernel estimators, which can be written as $\frac{\sum_i y_i k(||x_i - x||)}{\sum_i k(||x_i - x||)}$ where $k : \mathbb{R} \to \mathbb{R}$ has compact support. It appears that the main difference between kernel machines and the earlier techniques lies in the coefficients; for kernel machines, α_i is determined in a manner based on the entire training set, and not just the local subset of it. Perhaps, then, this might lead to non-local behavior of some sort, and in conclusion it is not immediately obvious whether kernel machines behave locally or not. We are therefore in need of an analysis to give us an answer.

For convenience, we will from now on refer to kernel machines as '**modern methods**'; we mean mainly support vector machines and related techniques, specifically, ones that use both maximal-margin separation and the 'kernel trick' [Vap98], but to a lesser degree also boosting [FS99], which similarly appears to have good performance due to margin maximization [SFBL97]. By '**classical methods**' we will refer to older techniques studied in the statistical literature, the prime examples of which are, as mentioned in the previous paragraph, k-NN and kernel estimators; another example

*We would like to acknowledge support for this project from the Israel Science Foundation as well as partial support from NSF grant DMS-0605236.

is local regression [CL95]. Using this terminology, our goal is to explain, at least in part, the performance advantage of modern methods over classical ones.

As we have seen, we are in need of an analysis to tell us whether modern methods behave locally or not. One such analysis was carried out in [BDR06], with the conclusion that kernel machines do in fact behave locally in some sense. If so, then it might appear that being local cannot explain the performance advantage of modern methods over classical ones, since apparently both approaches have that property. We will argue against this notion, while at the same time agreeing with the results in [BDR06]. Specifically, we will first show that indeed both modern and classical methods behave locally in some sense, but that the underlying cause is the property of consistency, i.e., that the method is able to successfully learn given any distribution (we leave a formal definition to the next section). Importantly, however, classical methods are local in a far stricter manner, and we will show that such strict locality implies performance limitations. Thus, a degree of locality is necessary for consistency, but is detrimental if taken to excess. We hypothesize that modern methods in fact strike a good balance in this matter, which may help to explain their superior performance.

In more general terms, based on our results we will argue in the discussion (Section 7) that the challenge in devising useful learning methods is to combine a local aspect, which is necessary for consistency, with a global aspect, which is useful for improving performance; the prime example of such a performance-improving global aspect is of course maximal-margin separation. To see this point, consider first that k-NN is defined in a simple manner that immediately ensures it is local (from its very definition), which allows it to be consistent, as we will see, with little additional work. However, it is hard to incorporate into such a local technique a non-local regularization method like maximal-margin separation. On the other hand, if we start with a method using maximal-margin separation then it is not trivial to ensure that it behaves locally, which we will see is a precondition for consistency. In other words, we want our learning methods to (1) be local, so that they may be consistent, and (2) apply a global regularization method, since this improves performance in practice. Devising a method having both of these properties at the same time is not trivial to achieve, but support vector machines and related methods do manage to do so: on the one hand they utilize maximal-margin separation, while on the other via the 'kernel trick' they end up having sufficiently local behavior in order to be consistent (assuming we choose an appropriate kernel and so forth). We will discuss this argument more at length in the discussion at the end of this work.

Technically speaking, the analysis that we conduct in order to arrive at the conclusions just mentioned is based on definitions inspired by those in [ZR07]. The main difference from that work is that local behavior was defined there by comparing a method's response on the entire training set to the 'local training set', which contains only the close-by points. This approach has the advantage of having practical applications in that it can answer what might occur if we 'localize', say, a support vector machine (i.e., show it only close-by points, as done in k-NN). However, the comparison of a method's response on two training sets of different size (the local one is in all reasonable cases smaller) has the disadvantage that it is hard to talk about subtle degrees of locality, since the change in the size of the training set introduces a source of variability. Our goal in the present work is in fact to speak about such differences of degree. We therefore define locality differently, by considering changes to far-off points instead of removing them from the training set, which keeps the size of the training set fixed. Why this is helpful will become clear later on.

The structure of the rest of this work is as follows. In Section 2 we describe the formal setting of the problem and lay out notation. In Section 3 we define locality and other concepts and give an overview of our results. In Section 4 we present our results for consistency and its connection to weak locality. In Section 5 we turn to strict locality and its drawbacks. In Section 6 we deal with the application of our results to classification. Finally, in Section 7 we summarize and discuss our results.

2 Formal Setting

We now complete the formal description of the setting. We are given an i.i.d sample $S_n = \{(x_i, y_i)\}_{i=1..n}$ from some distribution P; then a new (independent) 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 measurable function that depends both on the training set and the point to be estimated (note that this notation, where the training set and new observation are of equal standing as inputs to f, is slightly atypical, but is very convenient in our setting, as will become clear later). We call f a *learning method* (sometimes *method* or *estimator*); note that it can produce guesses for any size training set and any x. One specific context is that of classification (also called pattern recognition) where $y \in \{-1, +1\}$; we call learning methods in this context classifiers and call y the class. While our results apply to classification, we will not focus on it in most of this work, since a regression-type setting is simpler to deal with. Later on, in Section 6, we will show how to apply our results to classification.

Our goal 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

$$L_{n,P}(f) \equiv E_{S_n, x\sim P} \left| f(S_n, x) - f^*(x) \right| \underset{n\to\infty}{\longrightarrow} 0$$

where the expected value is taken over all training sets S_n and observations x both distributed according to P. We will omit P from $L_{n,P}$ when the distribution is clear from the context, and we will generally further shorten our notation to write expressions of the form

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

when, again, the distribution is clear from the context. (Note that the choice of the absolute value in the L_n loss – i.e., the $\mathcal{L}_1$ norm – is only for convenience; our results hold in the more typical $\mathcal{L}_2$ norm as well.)

If a method is consistent on all P then we call it consistent (this is sometimes called *universal consistency*). Importantly for us, methods like support vector machines and boosting are consistent [Ste02, Zha04], or at least can be if the parameters are chosen accordingly. In fact such choices often turn out to lead to good performance in practice, and therefore we are interested in consistent versions of modern methods. We will return to this matter in the discussion.

The following notation will be used. Denote by μ_P (or just μ, if P is clear from the context) the marginal measure on x of a distribution P. We denote random variables by, for example, $(x, y) \sim P$ and $S_n \sim P$ where the latter indicates a random i.i.d sample of n elements from the distribution P. We will often abbreviate and 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.

As already implied, we write expected values in the form $E_{v \sim V} H(v)$ where v is a random variable distributed according to V. We denote probabilities by, e.g., $P_{v \sim V}(U(v))$, which is the probability of an event $U(v)$ taken over a random variable v. In both cases we will omit V when it is clear from the context.

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 the limiting of f^*'s domain to B). We denote $B_{x,r} = \{x' \in \mathbb{R}^d : ||x - x'|| \leq r\}$, the ball of radius r around x (using the Euclidean norm). Let $P_{x,r} = P_{B_{x,r}}$.

Finally, we make the following assumptions which are mainly for convenience. Our distributions P are on (X, Y) where $X \subset \mathbb{R}^d, Y \subset \mathbb{R}$ (i.e., we work in Euclidean spaces). 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, $|f(S_n, x)| \leq M_2$ for some $M_2 > 0$.[1] Let M be a constant fulfilling $M \geq M_1, M_2$.

3 Definitions and Overview

We will now define the main concepts that concern us, starting first with some convenient notation. For any training set $S_n = \{(x_i, y_i)\}$ and values $x \in X \ , \ r \geq 0 \ , \ \{\widetilde{y}_i\} \in Y^n$, let

$$S_n(x, r, \{\widetilde{y}_i\}) = \Big\{ \Big(x_i \ , \ 1\{||x - x_i|| \leq r\} y_i \ + \ 1\{||x - x_i|| > r\} \widetilde{y}_i \Big) \Big\}$$

That is, $S_n(x, r, \{\widetilde{y}_i\})$ does not change the locations x_i, and has the original y values y_i close-by to x (up to distance r), while replacing far-off labels with $\widetilde{y}_i$. We can now define one sense of locality: we call a method f **local** on a distribution P iff there exists a sequence $\{R_n\}$, $R_n \searrow 0$, for which

$$E_{\{x_i\},x} \sup_{\{y_i\},\{\widetilde{y}_i\}} |f(S_n, x) - f(S_n(x, R_n, \{\widetilde{y}_i\}), x)| \underset{n\to\infty}{\longrightarrow} 0$$

(Here $S_n = \{(x_i, y_i)\}$, following our usual notation, i.e., S_n is constructed from $\{x_i\}$ in the expectation and $\{y_i\}$ in the sup.)

This definition is fairly straightforward: a method is local if, asymptotically speaking, it returns very similar results when we change far-off labels. Thus, the method is influenced mainly by the close-by part of the training set, which is the intuition behind a local method. Note that, since $R_n \to 0$, in effect the method is influenced only by the local part of the training set in a strong sense. Note also that the definition speaks of locality on a single distribution P; as with consistency, if a method f is local on all P then we say that f is local (i.e., if P is not specified, we mean all P). We will use this convention with other definitions as well.

It turns out that there are more useful ways to define locality, for reasons which we will see later. One such definition of locality is weaker than that given before, and one is stronger. We start with the weaker:

Definition 1 *Call a method f **weakly local** on a distribution P iff, for every distribution $\widetilde{P}$ for which $\mu_P = \mu_{\widetilde{P}}$, there exists a sequence $\{R_n\}$, $R_n \searrow 0$ for which*

$$E_{\{x_i\},x} E_{\{y_i\} \sim P \mid \{x_i\} \ , \ \{\widetilde{y}_i\} \sim \widetilde{P} \mid \{x_i\}} |f(S_n, x) - f(S_n(x, R_n, \{\widetilde{y}_i\}), x)| \underset{n\to\infty}{\longrightarrow} 0$$

(Here $\{y_i\} \sim P \mid \{x_i\}$ means that each y_i has distribution P conditioned on x_i.)

Thus, a weakly local method is one which, if we replace far-off labels with labels from another distribution, is asymptotically not influenced by that change (note that we keep μ, the measure on x, fixed; we care only about changes to y values). This definition is weaker than the one given before in that instead of the supremum over all y, we sample alternate y values from a fixed distribution. However, since we require that this property occur for *all* distributions $\widetilde{P}$, we still have the essential behavior of being most influenced by the close-by part of the training set.

In one of our main results we will see that all consistent learnings methods are in fact very close to being weakly local (we will require a minor technical relaxation of the definition given above). Hence this is true for, e.g., support vector machines, assuming the kernel and parameters ensure consistency.

It is obvious that classical methods like k-NN and kernel estimators are weakly local, both because they are consistent [see DGKL94, GKP84, respectively], and by direct inspection, see Section 5. However, they seem to be local in a stronger sense than that appearing in weak locality. In fact they have the following stronger property:

Definition 2 *Call a method f **strictly local** on a distribution P iff there exists a sequence $\{R_n\}$, $R_n \searrow 0$, for which*

$$P_{\{x_i\},x} \Big(\forall \{y_i\}, \{\widetilde{y}_i\} \quad f(S_n, x) = f(S_n(x, R_n, \{\widetilde{y}_i\}), x) \Big) \underset{n\to\infty}{\longrightarrow} 1$$

Thus, a strictly local estimator is one for which we can replace far-off labels and this, with probability going to 1, will not affect our estimates at all (this is easily seen to be

[1] Note that this is a minor assumption since for essentially all modern and classical methods we have $\sup_x |f(S_n, x)| \leq C \cdot \max_i |y_i|$ for some $C > 0$, and the y_i values are already assumed to be bounded. Furthermore, we are concerned with consistent methods, i.e., that behave similarly to f^* in the limit, and f^* is bounded.

stronger than the original definition of locality due to the boundedness assumption on f^*). Note that we can consider stricter notions of locality, however, this definition is strict enough, since classical methods fulfill it.

We will see later in Section 5 that, unlike classical methods, many (if not all) modern methods are **not** strictly local, and that this has potentially important consequences, since strictly local methods have performance limitations.

In [ZR07] similar definitions appeared. In that work, local behavior was defined by comparing f's response to the response it would have given had far-off points been removed from the training set, whereas in the definitions given above we consider changes to their y values instead. As mentioned in the introduction, the reason for this is the need to consider varying degrees of locality. In our definitions, we can either change the y values to values sampled from a fixed distribution (weak locality) or consider all possible changes (locality, and, in a stronger sense, strict locality). We will see that these differences can in fact be of importance. A further reason for preferring our definitions over ones in which far-off examples are removed is that the latter approach changes the size of the training set, and in a data-dependent manner. This introduces a source of variability which then makes it hard to talk about concepts like strict locality, where we require that with high probability there be no change in the response; if n changes, this itself may cause an alteration (e.g., this occurs in the common case where a regularization constant is used whose value depends on n). Alternatively, we might have removed a fixed number of far-off observations depending on n (as in k-NN, in fact), but this causes other inconveniences in that the radius in which the remaining observations lie is now a random variable (which is, as before, a source of variability). Replacing far-off y values, as we have chosen to do, therefore seems the most productive choice.

We now survey other related work. Research regarding locality was done in the context of learning methods that work by minimizing a loss function. Such loss functions can be 'localized' by re-weighting them so that close-by points are more influential; see [BV92], [VB93] for such an approach in the setting of Empirical Risk Minimization [ERM; Vap98] and [CL95] and references therein for the specific case of linear regression; see [AMS97] for a survey of applications in this area. The approach we follow differs from this one in that we focus on consistency in the sense of asymptotically arriving at the lowest possible loss achievable by any measurable function – i.e., in the nonparametric sense – and not in the sense of minimizing the loss within a set of finite VC dimension. The nonparametric sense is, we believe, the one most relevant to locality, and the best context in which to compare modern and classical methods.

We now briefly summarize our two main results. First, regarding the connection between consistency and weak locality, let us consider now a property weaker than consistency. Define the means of f and f^* by

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

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

the latter expression which is just the global mean of y, and define f, f^*'s Mean Absolute Deviations (MADs) by

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

(we prefer the MAD over the variance due to the choice of the $\mathcal{L}_1$ norm). We define

Definition 3 *Call a method f **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, $\forall P$,*

$$\left\{ \begin{array}{l} \limsup_{n\to\infty} |E_n(f) - E(f^*)| \\ \limsup_{n\to\infty} MAD_n(f) \end{array} \right\} \leq H\left(MAD(f^*)\right)$$

(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 – the mean and MAD, which are two scalar values – in a way that depends on the MAD, i.e., 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 that also 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 $c > 0$).

It is easy to see that WCM is weaker than consistency and implied by it. Assuming consistency,

$$\begin{aligned} |E_n(f) - E(f^*)| &= \left| E_{S_n,x}\big(f(S_n,x) - f^*(x)\big) \right| \\ &\leq \left| E_{S_n,x} \left| f(S_n,x) - f^*(x) \right| \right| \qquad (1) \\ &= |L_n(f)| \to 0 \end{aligned}$$

(note that here even $H(t) \equiv 0$ would have worked), and

$$\begin{aligned} \limsup_n \mathrm{MAD}_n(f) &= \limsup_n E_{S_n,x} \left| f(S_n,x) - E_n(f) \right| \\ &\leq \limsup_n \Big\{ E_{S_n,x} \left| f(S_n,x) - f^*(x) \right| \\ &\qquad + E_x \left| f^*(x) - E(f^*) \right| \\ &\qquad + \left| E(f^*) - E_n(f) \right| \qquad \Big\} \\ &= \mathrm{MAD}(f^*) \end{aligned}$$

using the consistency of f and (1); thus, $H(t) = t$ shows that the WCM property holds for all consistent methods.

We can now ask, what is missing in WCM that is present in consistency? Since WCM is a 'global' property (concerned only with two scalar values that are functions of the entire space), it seems apparent that what is missing in WCM is some 'local' aspect, i.e., of correctly learning in each small area separately. We will see that in fact a property very similar to weak locality can fill that role; we will call that definition Uniform Approximate Weak Locality (UAWL). We will then prove that consistency is logically equivalent to the combination of UAWL and WCM. From our definitions it will be easy to see that the UAWL and WCM properties are 'independent' in the sense that neither implies the other. Thus, we can see consistency as comprised of two independent properties, which might be presented as

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

Thus, our first conclusion is that a form of local behavior is fundamental to consistency; any consistent method must be in a sense local, no matter how it is defined. In fact, the difference between consistency and locality comes down to the additional requirement in consistency that we also are not far from estimating global properties of the distribution, as formalized by the WCM property.[2] This means that if we start with a method defined in an explicitly local manner, like k-NN, then we get 'for free' the property of UAWL. Then all we need to do to get consistency is to ensure the WCM property, which is relatively simple (we just need the scalar value representing our global mean to converge to the accurate one, and our MAD to not be too large). Since consistency is a desirable property, this explains some of the attractiveness of classical methods: achieving consistency with them is relatively simple.

Our second main result will show the drawbacks of this simplicity of classical methods, and will concern strict locality. To show the limitations of strict locality, we define the following property: call a method g **preferable** to another method f, over a set of distributions $\mathcal{P}$, iff, for every $P \in \mathcal{P}$,

$$L_n(g) < L_n(f)$$

for large enough n (possibly depending on P). That is, no matter what the true distribution is out of those in $\mathcal{P}$, g is eventually better than f. Our claim is then that, for every strictly local method f, we can always construct a non-strictly local g which is preferable to f. For convenience we will show this on a specific example, but argue that the result is a quite general one.

4 Weak Locality and Consistency

As hinted at before, it turns out that a slight complication of our definition of weak locality is necessary. To present the improved definition, we start with some preparatory notation. For any $q \geq 0$ and distribution P, let

$$\bar{f}^q(S_n, x) = E_{x' \sim P_{x,q}} f(S_n, x')$$

That is, $\bar{f}^q$ applies a 'smoothing' operation performed around the x being estimated (recall that $P_{x,q}$ is P conditioned on the ball of radius q around x). Note that if $q = 0$ then we interpret the expected value as a delta function and we get $\bar{f}^0 = f$. Note also that we require the actual unknown distribution P in the definition of $\bar{f}^q$, i.e., $\bar{f}^q$ cannot be directly implemented in practice – $\bar{f}^q$ is a construction for theoretical purposes.

We define the following set of sequences:

$$\mathcal{T} = \{\{T_n\} \ : \ T_n \searrow 0\}$$

and, for any sequence $T = \{T_n\} \in \mathcal{T}$, we define the set of its infinite subsequences and selection functions on them by

$$\mathcal{R}(T) = \{\{R_n\} \ : \ \{R_n\} \subseteq T \ , \ R_n \searrow 0\}$$

$$\mathcal{Q}(T) = \{Q : T \to T \ : \ Q(T_n) = o(T_n)\}$$

We now motivate these definitions. First, regarding $\mathcal{T}$: instead of allowing any possible value in $[0, \infty)$ for R_n and Q, we limit them to a countable set $\mathcal{T}$. The reason for this is that due to $[0, \infty)$ being an uncountable set it is not clear to the authors if additional conditions are not required to prove our results in that case. In any event, a countable set of possible values is of sufficient interest for any practical learning-theoretical purpose, since we end up using only a countable number of R_n, Q values (since $n \in \mathbb{N}$). Note that the set of possible values $\mathcal{T}$ can be chosen in whatever manner is desired, so long as this is done in advance.

$\mathcal{R}(T)$ contains **localizing sequences**, sequences of radii that determine how far off we alter the data shown when we perform $S_n(x, R_n, \{\widetilde{y}_i\})$. We require that $R_n \searrow 0$, as we are interested in learning methods that focus on the truly local part of the training set, i.e., having radius 0 asymptotically.

$\mathcal{Q}(T)$ contains functions of the possible values T that become negligibly small when T_n is small. We will use the values $Q(R_n)$ to determine radii on which to smooth, via $Q(R_n)$, which we might call the **smoothing radius**; note that since $Q(R_n) = o(R_n)$, we smooth on a radius much smaller than R_n, hence this is a fairly minor operation.

Finally, we define

$$\mathcal{R}^+(T) = \{\{R_n\} \ : \ \{R_n\} \subseteq T\}$$

$$\mathcal{Q}^+(T) = \{Q : T \to T\}$$

which are the same as before, but without the requirement of converging to 0. We now arrive at our main definition for this section, whose description is unavoidably technical:

Definition 4 *Call a learning method f* ***Uniformly Approximately Weakly Local (UAWL)*** *iff*

$$\begin{aligned}
&\forall P \ , \ \widetilde{P} \ , \ \mu_P = \mu_{\widetilde{P}} \\
&\forall T \in \mathcal{T} \\
&\exists Q \in \mathcal{Q}(T) \\
&\forall Q' \in \mathcal{Q}^+(T) \ , \ Q' \geq Q \\
&\exists \{R_n\} \in \mathcal{R}(T) \\
&\forall \{R'_n\} \in \mathcal{R}^+(T) \ , \ R'_n \geq R_n \\
&\quad E_{\{x_i\},x} E_{\{y_i\} \sim P \mid \{x_i\} \ , \ \{\widetilde{y}_i\} \sim \widetilde{P} \mid \{x_i\}} \\
&\qquad |f(S_n, x) - \bar{f}^{Q'(R'_n)}(S_n(x, R'_n, \{\widetilde{y}_i\}), x)| \underset{n \to \infty}{\longrightarrow} 0
\end{aligned}$$

(Here the expression $R'_n \geq R_n$ simply implies an inequality for the entire series, i.e., for all n. $Q' \geq Q$ implies $Q'(T_k) \geq Q(T_k)$ for all k.)

That is, a UAWL method returns similar values even when we replace far-off data with different values of y; essentially the same idea as with weak locality, but allowing for minor smoothing, and requiring uniformity in Q, R_n. With a UAWL method, loosely speaking, for any large enough Q, R_n we get local behavior. Note that the notion of R_n being large enough is a natural one since taking R_n to 0 very quickly is problematic (doing so may lead to us getting few or no points in radius R_n, i.e., few or no points from the important distribution).

The reason for including smoothing in this definition is that, if all we assume is that learning methods are measurable

[2] Note that we need both the mean and the MAD to behave in an appropriate way, as appearing in the definition of the WCM property, because if only the mean is accurate then due to the variance we may estimate the global properties very poorly.

(and not smooth in some strong sense), then odd counterexamples exist to the connection between locality and consistency; see [ZR07] for details. By incorporating smoothing in our definition we remove the need to require it of the learning methods we consider, which lets us apply our results to any method known to be consistent. The reason for the second new aspect in this definition, that of allowing all large-enough Q', R'_n, is that this leads to an exact equivalence with consistency, as we will see in Theorem 5; furthermore, it would be odd for the locality of a method to depend much on the specific Q, R_n used for it. To make the matter concrete, note that the proof of Theorem 5 requires using the same Q, R_n over multiple distributions; without allowing all large-enough Q', R'_n there exist odd counterexamples in which each distribution has some appropriate Q, R_n but none exist that are appropriate for all of them simultaneously.

Our result for consistency is the following:

Theorem 5 *A learning method f is consistent iff f is both UAWL and WCM.*

We prove the $\Leftarrow$ direction, that UAWL and WCM imply consistency, in Appendix A. Note that it is clear from the proof that we can replace $\widetilde{P}$ in the definition of UAWL with all distributions having y constant, but we believe the definition given before is clearer.

For the $\Rightarrow$ direction, that consistency implies UAWL and WCM, it is immediately obvious that consistency implies WCM. Regarding UAWL, a proof of a slightly simpler claim (without uniformity in R_n, Q) appears in [ZR07]; using methods from other proofs in [ZR07], it is trivial to extend the proof to showing uniformity as well. For completeness we give a brief sketch of the proof appearing there: for fixed r, q instead of R_n, Q, we can use the consistency of f on the effective distributions seen (i.e., distributions that are altered to $\widetilde{P}$ far away from x) to see that the appropriate loss converges to 0, for every x separately. Since, again for every x, the overall loss converges to 0, this also occurs in the area with radius q, which is the one relevant to us. We then take R_n, Q to 0 slowly enough to complete the proof.

Theorem 5 can be summarized as follows:

$$\textbf{Consistency} \quad \Longleftrightarrow \quad \textbf{UAWL} \oplus \textbf{WCM}$$

Here we use the symbol $\oplus$ because each of the two properties UAWL and WCM can exist without the other: consider the following two methods,

$$f_y(S_n, x) = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad\qquad f_0(S_n, x) = 0$$

f_y (called thus because it considers only the y values) is WCM, since $E_n(f_y) \to E(f^*)$ and clearly f_y's MAD converges to 0. (In fact, f_y is WCM with $H \equiv 0$, i.e., in the strongest sense. That is, there are even 'weaker' methods that are WCM.) On the other hand, f_y is clearly not UAWL (consider, e.g., two distributions having $f^*(x) \equiv -1, f^*(x) \equiv +1$). On the other hand, f_0 is trivially UAWL, but not WCM.

5 Strict Locality

In this section we will deal with strict locality and its consequences.

It is immediately clear that kernel estimators are strictly local (use R_n equal to the bandwidth, and recall that k has compact support). For k-NN things are less obvious, but still fairly simple: k-NN is consistent if the number of neighbors k_n fulfills $k_n \to \infty, \frac{k_n}{n} \to 0$ [DGKL94]. From inspecting the proof of consistency it is clear that these conditions ensure that the k_n neighbors will fall in an area of radius going to zero, with probability going to 1. Thus (unsurprisingly) k-NN is strictly local: just like kernel estimators, it completely ignores far-off points, but it does so with very high probability instead of certainty (since there is always a chance, even though it becomes negligibly small, that we will need to look far for the k_n nearest neighbors).

We have seen that any consistent method must be in some sense local, specifically, UAWL. We can now ask, must a consistent method also be strictly local? It turns out that the answer is no. Consider, for example, kernel ridge regression [SGV98], which can be written in the kernel-induced space (via a transformation ϕ) as

$$L(w) = \frac{1}{n}\sum_i (w'\phi(x_i) - y_i)^2 + \lambda ||w||^2 \tag{2}$$

It is clear that under mild regularity conditions we will not get strict locality, since any change to the y_i values can cause a change to the resulting w, as is obvious from looking at the solution to (2); thus, kernel ridge regression is not strictly local. It appears clear that a similar phenomenon occurs for other types of kernel machines, as well as methods such as boosting (but we do not supply a formal proof), simply because there is always the possibility of influence by far-off points (as is also clear from these methods minimizing a global loss function which is an average of losses at individual points; any change to a point influences the overall loss, with potential consequences on the entire space). While the influence of far-off points wanes as n converges to infinity – which is necessary, as we have seen, in order for the method to be consistent – the far-off points are not simply ignored as with classical methods like k-NN. There is always the possibility of being influenced by the farther points, even if this is a rare occurrence.

We will now see that the property of potentially being influenced by far-off points can, in fact, be important. The reason is that strictly local methods have performance limitations. As is well known, to talk in a meaningful way about performance, we cannot make comparisons on the set of all distributions [see, e.g., DGL96]. We therefore consider limited sets of distributions, as is done in the minimax setting in statistics. We first begin with a brief reminder of the setting and how minimax losses can be achieved.

Assume for simplicity a Lipschitz set of functions $f^* \in \mathcal{L}(L)$ on $[0,1]^d$,

$$|f^*(x_1) - f^*(x_2)| \leq L \cdot ||x_1 - x_2||$$

and take x uniform on $X = [0,1]^d$; let $y = f^*(x) + \epsilon, \epsilon \sim N(0, \sigma^2)$. Consider a simple kernel estimator with radius r,

$$f(S_n, x) = \frac{\sum_i 1\{||x_i - x|| \leq r\} y_i}{\sum_i 1\{||x_i - x|| \leq r\}}$$

For every x_0, we receive on average on the order of nr^d points in radius r to estimate $f^*(x_0)$, so we can estimate

$Ef^*(x)$ in that area up to precision $\frac{\sigma}{\sqrt{nr^d}}$. It is also clear that $Ef^*(x)$ differs from $f^*(x_0)$ by up to Lr, giving us roughly $L_n(f) \leq \frac{\sigma}{\sqrt{nr^d}} + Lr$, an example of a bias-variance tradeoff (the bias is due to estimating $Ef^*(x)$ on $B_{x_0,r}$ and not $f^*(x_0)$ directly, and the variance is due to having only the order of nr^d points). From this simple analysis it can be concluded that a choice of $r = r_n = O(n^{-1/(d+2)})$ is appropriate, and that this will give us a loss of $O(n^{-1/(d+2)})$. This is in fact the minimax rate, i.e., the best-possible achievable rate, as shown in [Sto80, Sto82].

Importantly, notice how we must consider close-by points in order to arrive at the rate: if we look only at points at distance r or more, then $f^*(x_0)$ may differ by up to Lr and we would not be able to overcome this issue in a minimax sense. Furthermore, it is also obvious from the analysis that the close-by points are enough in order to achieve the rate, i.e., to be up to a constant factor of the actual minimax loss. This can be directly seen by the equality of the bias and variance factors when we minimize their sum.

Thus, even a strictly local method like kernel estimators can achieve the minimax rate; in that sense, there is nothing to improve upon. In the example above the rate is $n^{-1/(d+2)}$, and kernel estimators can achieve it, but we have no assurance that they do so with a low constant factor; since such constant factors are hard to analyze, they are for the most part ignored in statistics. While this is reasonable in the sense that the rate is arguably the most important aspect in an asymptotic analysis, in actual practice – i.e., when working with some fixed finite n – the constant factor can be critical, since for fixed finite n we do not care about the asymptotic rate but only about the actual value of L_n. We will now make such a comparison of the actual values of L_n and claim that strictly local methods are limited in their ability to minimize it.

As defined previously, call a method g **preferable** to another method f, over a set of distributions $\mathcal{P}$, iff, for every $P \in \mathcal{P}$,

$$L_n(g) < L_n(f)$$

for large enough n (possibly depending on P). We will now see that in fact it is simple to construct a method preferable to any strictly local method, thus showing that strict locality brings with it performance limitations. The reason for the limitation is easy to see: by completely ignoring far-off points, there is no ability to adapt to rare occurrences in which those far-off points are in fact necessary for good performance. In statistical terms, while we have lower bias with the close-by points, we have lower variance with the farther-off ones due to their greater number. On average we prefer to balance these two out, as shown above, but in specific cases we can do better than such an average; consider, for example, the unlikely but possible case where the close-by points have bizarre values (e.g., their empirical variance is much larger than σ^2 in the example above); in such a case, based on the empirical sample we can tell that it would probably be better to focus on slightly farther off points. That is, while on average the close-by points are most relevant, there is a minority of cases in which they are in fact misleading, and in at least some of those cases we can tell when they occur, at least with high probability. We will now formalize this notion in a concrete result in a specific setting. While only one example, the underlying issue just mentioned should hold in a wide range of cases.

The following definition will make our result easier to state: call a method f **reasonable** iff, when all y_i in S_n have the same value, f returns that value. Note that practically every existing learning method has this property, including those of interest to us, and that in fact all consistent methods must have this property in an asymptotic sense in order to be consistent on distributions having a constant value of y. Then we claim the following:

Proposition 6 *Let $\mathcal{L}$ be the following set of distributions. Assume $X = [0,1]$ and that μ is uniform on X. Let $Y = \{-1,+1\}$, assume that all $f^*(x)$ are Lipschitz with constant $\leq L$, and that*

$$\mu\Big(f^*(x) \in \{-1,0,+1\}\Big) \;=\; 0 \qquad (3)$$

Assume that f is a strictly local method and that f is reasonable. Then there exists a reasonable method g for which, for every $P \in \mathcal{L}$, for large enough n we have

$$L_n(g) < L_n(f)$$

That is, g is preferable to f.

(Note that the assumption (3) is for convenience, and leaves us to deal with the most interesting cases.) The proof of the proposition appears in Appendix B.

Thus, any strictly local method can be improved upon due to its ignorance of far-off points. Given that support vector machines and other techniques used in machine learning are in fact local but not strictly local, there is the possibility (which we concede that we only argue towards, but do not prove) that this helps to explain their performance advantage over classical methods which are strictly local.

6 Classification

We will now show how our results apply to classification. First, we note that many theoretical analyses of classification methods such as support vector machines and boosting in fact work on the real-valued response of such methods, i.e., before the sign operation; see, e.g., [Zha04, BJM06]. In that sense these classification methods are treated similarly to regression estimators, and our results are of relevance to them. However, this connection is only an informal one, and therefore in this section we will show how it can be formalized.

In classification [see, e.g., DGL96] we deal with learning methods $c(S_n, x)$ which return values in $\{-1,+1\}$. The loss of interest is the 0-1 loss,

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

which is usually compared to the lowest possible loss (also known as the Bayesian loss), giving the excess loss, which is well-known to be equivalent to

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

where $\eta(x) = P(y = 1|x)$ and $c^*(x) = \text{sign}(f^*(x))$. This differs from the loss L_n studied in the main part of this work, but as shown in [ZR07], consistency-related results such as Theorem 5 can be adapted to classification, using a method that we now briefly summarize. The idea is to note that

$$\begin{aligned}\widetilde{L}_n(c) &\equiv E_{S_n,x}|c(S_n,x) - c^*(x)| \cdot |2\eta(x) - 1| \\ &= E_{S_n,x}|c(S_n,x) - c^*(x)| \cdot |f^*(x)| \\ &= E_{S_n,x}\Big|c(S_n,x) \cdot |f^*(x)| - f^*(x)\Big| \\ &\equiv E_{S_n,x}|f_c^*(S_n,x) - f^*(x)| \\ &= L_n(f_c^*)\end{aligned} \tag{4}$$

where we define $f_c^*(S_n,x) \equiv c(S_n,x) \cdot |f^*(x)|$. Now, a classifier c can be seen as estimating $\operatorname{sign}(f^*)$. For every such c we define a learning method f_c that estimates f^*, by

$$f_c(S_n,x) = c(S_n,x) f_{|\,|}(S_n,x)$$

where $f_{|\,|}$ is the absolute value of some pre-determined consistent method, i.e., a consistent estimator of $|f^*|$ (that is, c estimates the sign of f^* and $f_{|\,|}$ estimates the absolute value; together they estimate f^*). It is then straightforward to show that c is consistent (as a classifier) on a set of distributions precisely when f_c is consistent (as a regression-type estimator) on that same set, since f_c is asymptotically equivalent to f_c^*, and using $\widetilde{L}_n(c) = L_n(f_c^*)$ from (4).

Regarding our result for strict locality, Proposition 6, the proof can be modified to apply to classification as follows. First, note that already $Y = \{-1,+1\}$, and that if we replace f with a classifier c (i.e., a function into $\{-1,+1\}$) then g defined in the proof is also a classifier (in fact, the setting was chosen for its relevance to classification). Denote $d = g$ to avoid confusion; thus, our goal is to show that $\widetilde{L}_n(d) - \widetilde{L}_n(c) < 0$. Now, as shown in (4) we have $\widetilde{L}_n(c) = L_n(f_c^*)$, so our goal is to evaluate $L_n(f_d^*) - L_n(f_c^*)$. Note that, when event A occurs as defined in the proof, then instead of a response of 1 for f^* we now have a response of 1 for c, giving an overall response of $f_c^*(S_n,x) = |f^*(x)|$ (and vice versa for a response of -1), which leads to replacing $|1 - f^*(x)|$ with $\Big||f^*(x)| - f^*(x)\Big|$ and of $|-1-f^*(x)|$ with $\Big|-|f^*(x)| - f^*(x)\Big|$. In (5) we then get

$$\Big||f^*(x)| - f^*(x)\Big| - \Big|-|f^*(x)| - f^*(x)\Big| = -2f^*(x)$$

and $-2f^*(x)$ happens to be the exact same result as in the original proof. All the rest of the proof can remain as before, thus proving the claim in the context of classification.

7 Discussion

We have argued that (1) some degree of locality is unavoidable in learning, but that (2) if this is taken to an extreme then it brings with it performance limitations. We speculate that the superior performance of modern methods over classical ones may, in part, be due to the former striking a proper balance in this matter.

Regarding the unavoidability of local learning, this is a direct result of locality being implied by consistency. In fact, in consistency we require the ability to do well on *all* distributions, which includes distributions that only differ in very small localized ways. Thus, a consistent method must end up trusting only close-by points. The only way to avoid this issue is to dismiss consistency as a useful property. While in theory such an approach might make sense – say, if we know in advance that the true distribution belongs to some limited set – in practice many effective methods in machine learning are useful precisely because they make as few as possible assumptions on the distribution. In fact, this is the reason non-parametric methods are often more effective on real-world problems than parametric ones. Thus, generally speaking, consistency appears to be a property that we cannot easily discard. Since consistency implies a form of locality, locality is unavoidable as well.

As we have seen, the difference between consistency and the relevant form of locality, UAWL, turns out to be a fairly minor property, WCM. This means that if one of our goals is consistency then it makes sense to focus on achieving the UAWL property, since it is generally more difficult to ensure than WCM (ensuring WCM amounts to checking that two scalar values are within some reasonable bound). This may explain the historical appearance of and focus on classical methods like k-NN and kernel estimators: by defining them in an explicitly local manner, which is simple to do, the UAWL property is easily taken care of. Consequently, defining such local methods is convenient and proving their consistency relatively easy as well.

Such definitions, however, make the resulting methods not only local in the necessary sense, but also *strictly* local. As we have seen, strict locality is not necessary for consistency and in fact implies some limitations on performance. Thus, being motivated by convenient definitions and proofs may lead to deficits in practice.

On the other hand, we can start with improving real-world performance. The primary method of doing so which we intend here is maximal-margin separation, which turns out to be very effective in practice, and has an appealing geometric intuition (keeping the classes as far apart as possible). This approach is clearly not a local one, since the maximal-margin hyperplane depends on the entire training set. Furthermore, in some sense it is reasonable to expect an effective regularization technique to in fact be non-local: if, as in soft-margin support vector machines, we consider the sum of deviations across the margin (i.e., of observations on the wrong side of it), then it would be hard to do so in a local manner. That is, if we expect to allow some total amount of deviations based on some rationale, it is hard to enforce this locally; if we do work locally, then we need to apply the same approach in every area, instead of being able to accept more deviations in some areas in return for smaller deviations elsewhere as well as a larger overall margin.

Thus, techniques like maximal-margin separation are effective and desirable, but non-local in their definition. This appears problematic if we also want the property of consistency, which as we have seen requires a degree of locality. Hence, in devising learning methods we come up against a difficulty: we want our learning methods to (1) be local, so that they may be consistent, but we also want to (2) apply some performance-improving technique like maximal-margin separation, which is non-local.

We can now try to explain the success of modern machine learning methods by their combining these two properties in an effective manner: by using the 'kernel trick' and choosing a universal kernel [Ste02] we can get sufficiently local behavior for consistency, while at the same time we are still applying the maximal-margin principle in a global manner, thus improving performance. It is this combined approach which may be missing from classical methods.[3]

A Proof of $\Leftarrow$ in Theorem 5

Denote $S_n(x,r,a) = S_n(x,r,\{a_i\})$ where $a_i = a$, i.e., $S_n(x,r,a)$ replaces the y values of all far-off points with a.

Fix some $T \in \mathcal{T}$ and some $r, q \in T$. For any $\alpha \in \mathbb{R}$, we have the trivial fact that

$$\begin{aligned} &|f(S_n,x) - f^*(x)| \leq \\ &\quad \left|f(S_n,x) - \bar{f}^q(S_n(x,r,\alpha),x)\right| \\ &\quad + \left|\bar{f}^q(S_n(x,r,\alpha),x) - f^*(x)\right| \end{aligned}$$

Let $A = \{\alpha_m\}$ be a countable set and let m_n be a sequence. Write

$$\begin{aligned} &|f(S_n,x) - f^*(x)| \\ &\leq \inf_{m \leq m_n} \Big(\left|f(S_n,x) - \bar{f}^q(S_n(x,r,\alpha_m),x)\right| \\ &\qquad + \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \Big) \\ &\leq \sup_{m \leq m_n} \left|f(S_n,x) - \bar{f}^q(S_n(x,r,\alpha_m),x)\right| \\ &\quad + \inf_{m \leq m_n} \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \end{aligned}$$

and thus

$$\begin{aligned} &E_{S_n,x}\,|f(S_n,x) - f^*(x)| \\ &\leq E_{S_n,x} \sup_{m \leq m_n} \left|f(S_n,x) - \bar{f}^q(S_n(x,r,\alpha_m),x)\right| \\ &\quad + E_{S_n,x} \inf_{m \leq m_n} \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \\ &\leq \sum_{m \leq m_n} E_{S_n,x} \left|f(S_n,x) - \bar{f}^q(S_n(x,r,\alpha_m),x)\right| \\ &\quad + E_{S_n,x} \inf_{m \leq m_n} \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \end{aligned}$$

By the UAWL property, for any $\alpha_m \in A$ we have

$$E_{S_n,x} \left|f(S_n,x) - \bar{f}^{Q(R_n)}(S_n(x,R_n,\alpha_m),x)\right| \to 0$$

for appropriate Q, R_n, since $S_n(x,R_n,\alpha)$ can be seen as sampled from a situation where $\widetilde{P}$ in the definition of UAWL has y constant and equal to α. This is then true in particular for $Q \equiv q, R_n \equiv r$, since by keeping these values fixed they necessarily eventually become appropriate in the sense of the definition of UAWL (i.e., as constants, they eventually become larger than the sequences from the definition of UAWL – both of which tend to 0 – that we compare them with in order to check if they are appropriate). It is therefore also clear that there exists a sequence $m_n \to \infty$ for which

$$\sum_{m \leq m_n} E_{S_n,x} \left|f(S_n,x) - \bar{f}^q(S_n(x,r,\alpha_m),x)\right| \to 0$$

(by taking $m_n \to \infty$ slowly enough, e.g., by keeping $m_n = k$ fixed and raising it to $k+1$ only when the sum of the first $k+1$ elements will, for all $n' \geq n$, be smaller than k^{-1}, which must eventually occur since the sum is of elements converging to 0). For this m_n we therefore have

$$\begin{aligned} &\limsup_{n \to \infty} E_{S_n,x}\,|f(S_n,x) - f^*(x)| \\ &\leq \limsup_{n \to \infty} E_{S_n,x} \inf_{m \leq m_n} \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \end{aligned}$$

We now pick $A = \{\alpha_m\}$ to be dense in $[-M, M]$ (recall that M is a bound on f^* and f), and turn to analyzing the expression on the last line. Fix some $x \in \operatorname{supp}(P)$, and consider the expression corresponding to x in the expected value. Then for large enough n we can find some $m(x) \in \{1, ..., m_n\}$ for which $|\alpha_{m(x)} - E_{P_{x,r}}(f^*)| < \epsilon$, for any $\epsilon > 0$ (due to A being dense). Then

$$\begin{aligned} &E_{S_n} \inf_{m \leq m_n} \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \\ &\quad\leq E_{S_n} \left|\bar{f}^q(S_n(x,r,\alpha_{m(x)}),x) - f^*(x)\right| \\ &\quad\leq E_{S_n} \left|E_{x' \sim P_{x,q}} \left[f(S_n(x,r,\alpha_{m(x)}),x') - f^*(x')\right]\right| \\ &\qquad + \left|E_{x' \sim P_{x,q}} f^*(x') - f^*(x)\right| \\ &\quad\leq E_{S_n} E_{x' \sim P_{x,q}} \left|f(S_n(x,r,\alpha_{m(x)}),x') - f^*(x')\right| \\ &\qquad + E_{x' \sim P_{x,q}} \left|f^*(x') - f^*(x)\right| \end{aligned}$$

The expression on the last line converges to 0 (for almost all x) when $q \to 0$, by the corollary to the following lemma:

Lemma 7 *[[Dev81]; Lemma 1.1] For any distribution P and measurable g, if $E_{x \sim P}|g(x)| < \infty$ then*

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

for almost all x.

Corollary 8 *For any distribution P and measurable g, if $E_{x \sim P}|g(x)| < \infty$ then*

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

for almost all x.

Thus, we arrive at

$$\begin{aligned} &\limsup_n E_{S_n} \inf_{m \leq m_n} \left|\bar{f}^q(S_n(x,r,\alpha_m),x) - f^*(x)\right| \\ &\leq \limsup_n E_{S_n} E_{x' \sim P_{x,q}} \left|f(S_n(x,r,\alpha_{m(x)}),x') - f^*(x')\right| \\ &\quad + \epsilon_1 \end{aligned}$$

where $\epsilon_1 > 0$ can be made arbitrarily small by picking q small enough.

Note that we can see $S_n(x,r,\alpha)$ as sampled from the distribution $P_{x,r,\alpha}$, by which we mean a distribution having the same μ as P, equal to P on $B_{x,r}$, and having constant y

[3] Note that an additional advantage of kernel machines is that we can easily make them non-consistent, by choosing an appropriate kernel, i.e., a non-universal one.

equal to α elsewhere. Then

$$\begin{aligned}
&E_{S_n} E_{x' \sim P_{x,q}} \left| f(S_n(x,r,\alpha_{m(x)}), x') - f^*(x') \right| \\
&= \frac{1}{\mu\left(B_{x,q}\right)} E_{S_n, x' \sim P} \\
&\qquad \left| f(S_n(x,r,\alpha_{m(x)}), x') - f^*(x') \right| 1\left\{x' \in B_{x,q}\right\} \\
&\leq \frac{1}{\mu\left(B_{x,q}\right)} E_{S_n, x' \sim P} \left| f(S_n(x,r,\alpha_{m(x)}), x') - f^*(x') \right| \\
&= \frac{1}{\mu\left(B_{x,q}\right)} E_{S_n, x' \sim P_{x,r,\alpha_{m(x)}}} \left| f(S_n, x') - f^*(x') \right| \\
&= \frac{1}{\mu\left(B_{x,q}\right)} E_{S_n, x' \sim P_{x,r,\alpha_{m(x)}}} \\
&\qquad \big| f(S_n, x') - E_n(f) + E_n(f) - E(f^*) + \\
&\qquad E(f^*) - f^*(x') \big| \\
&\leq \frac{1}{\mu\left(B_{x,q}\right)} \Bigg[\mathrm{MAD}_{n, P_{x,r,\alpha_{m(x)}}}(f) + \\
&\qquad \left| E_{n, P_{x,r,\alpha_{m(x)}}}(f) - E_{P_{x,r,\alpha_{m(x)}}}(f^*) \right| + \\
&\qquad \mathrm{MAD}_{P_{x,r,\alpha_{m(x)}}}(f^*) \Bigg]
\end{aligned}$$

where the expected values $E_n(f), E(f^*)$ on the equation before last are w.r.t $P_{x,r,\alpha_{m(x)}}$ (the omission is for clarity).

Using the WCM property, we can therefore bound

$$\begin{aligned}
&\limsup_n E_{S_n} \inf_{m \leq m_n} \left| \bar{f}^q(S_n(x,r,\alpha_m), x) - f^*(x) \right| \\
&\leq \frac{1}{\mu\left(B_{x,q}\right)} \Bigg[\mathrm{MAD}_{P_{x,r,\alpha_{m(x)}}}(f^*) + \\
&\qquad 2H\left(\mathrm{MAD}_{P_{x,r,\alpha_{m(x)}}}(f^*) \right) \Bigg] + \epsilon_1
\end{aligned}$$

We now turn to consider the MAD of $P_{x,r,,\alpha_{m(x)}}$. Notice first that

$$\mathrm{MAD}(P_{x,r,\alpha_{m(x)}}) \leq \mathrm{MAD}(P_{x,r}) + \epsilon$$

because $|\alpha_{m(x)} - E_{P_{x,r}}(f^*)| < \epsilon$. Consider now the effect of changing r. First, by Lemma 7 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,

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

using Corollary 7 for the last equality, and thus

$$\limsup_{r \to 0} \mathrm{MAD}(P_{x,r,\alpha_{m(x)}}) \leq \epsilon$$

We can pick q to make ϵ_1 arbitrarily small, and then r to make $\mathrm{MAD}(P_{x,r,\alpha_{m(x)}})$ arbitrarily small as well (note that we thus counter the $\frac{1}{\mu(B_{x,q})}$ factor), and therefore, for almost every x,

$$\lim_{n \to \infty} E_{S_n} |f(S_n, x) - f^*(x)| = 0$$

where we also use the continuity of H at 0. This in turn implies, along with the dominated convergence theorem, that

$$\lim_{n \to \infty} L_n(f) = 0$$

thus proving that f is consistent.

B Proof of Proposition 6

Denote $S_n \cap B = \{(x_i, y_i) \in S_n \ : \ x_i \in B\}$. Fix some $P \in \mathcal{L}$ as in the statement of the proposition. Let R_n be the radii from the definition of strict locality for f. Note that, by the definition of strict locality, we can replace R_n with any $R'_n \geq R_n, R'_n \searrow 0$ and strict locality will still hold. WLOG we can therefore assume that $nR_n \to \infty$.

Define $R(S_n, x, r)$ as the property that

$$\forall \{y_i\}, \{\widetilde{y_i}\} \qquad f(S_n, x) = f(S_n(x, r, \{\widetilde{y_i}\}), x)$$

That is, R is the property that strict locality in fact occurs; by the definition of strict locality we know that the probability of $R(S_n, x, R_n)$ rises to 1.

We define the following additional properties. Denote by $A_X = A_X(x, r)$ the property that $x \in (\sqrt{r}, 1 - \sqrt{r})$ (which makes sense for $r \leq 1/4$, and is indeed the case concerning us as the values replacing r will tend to 0). Denote by $A_0 = A_0(S_n, x, r)$ the property that

$$\begin{aligned}
&|S_n \cap B_{x,r}| \in [nr, 3nr] \\
&\left| S_n \cap \left(B_{x,\sqrt{r}} \setminus B_{x,r} \right) \right| \in [n\sqrt{r}, 3n\sqrt{r}]
\end{aligned}$$

and that $R(S_n, x, r)$ holds. Note that $A_0(S_n, x, R_n)$ occurs with probability going to 1, due to the marginal distribution μ being uniform on $[0, 1]$ (and using Bernstein's Inequality), i.e., A_0 implies that the number of observations in the regions $B_{x,R_n}, B_{x,\sqrt{R_n}}$ are in the ranges of values we would expect them to be, up to a constant. The additional requirement that $R(S_n, x, R_n)$ holds does not change the probability of $A_0(S_n, x, R_n)$ going to 1, since the probability of $R(S_n, x, R_n)$ goes to 1.

Define also $A_+(S_n, x, r)$ as the property where $A_X(x, r)$, $A_0(S_n, x, r)$ hold, and in addition we have

$$\begin{aligned}
(x_i, y_i) \in S_n \cap B_{x,r} \quad &\longrightarrow \quad y_i = -1 \\
(x_i, y_i) \in S_n \cap \left(B_{x,\sqrt{r}} \setminus B_{x,r} \right) \quad &\longrightarrow \quad y_i = +1
\end{aligned}$$

i.e., the majority of points in $B_{x,\sqrt{r}}$ have label $+1$, while the minority in the smaller enclosed region $B_{x,r}$ have label -1, and strict locality occurs. Hence if f were applied to S_n, x, its response would be -1 (due to f being reasonable), despite the numerous slightly farther-off points with label $+1$. Likewise define $A_-(S_n, x, r)$ as the same property with reversed signs. Finally, let $A(S_n, x, r)$ be the property that either $A_+(S_n, x, r)$ or $A_-(S_n, x, r)$ holds. Note that for small R_n we expect that the probability of $A(S_n, x, R_n)$ be very small, i.e., it is an odd occurrence.

We define a new method g as follows:

$$g(S_n, x) = \begin{cases} f(S_n, x) & \neg A(S_n, x, R_n) \\ -f(S_n, x) & A(S_n, x, R_n) \end{cases}$$

(i.e., we return one value if the property $A(S_n, x, R_n)$ holds, and another otherwise). That is, on 'normal' training sets g is the same as f; however, on odd training sets with property A, g guesses the opposite of f: it trusts the large number of points within radius $(R_n, \sqrt{R_n})$ over the smaller number in radius $(0, R_n)$; g also behaves the same as f for x close to the boundaries $0, 1$ and only changes f's behavior when g takes into account the points in radius R_n and ignores the rest. Note that g is strictly local, like f, albeit with larger radius. This suffices to prove the proposition and thus make the claim that strict locality has performance limitations, since it shows that we would always want to raise R_n to improve performance. In fact we can continue to raise R_n while the close-by points comprise an 'odd' training set in a sense similar to that mentioned above, which will lead to a non-strictly local method (since we may end up with large R_n, even $O(1)$, albeit with small probability).

We will now prove that g has the property described in the proposition, i.e., that it is preferable to f. Consider some fixed $x \in (0, 1)$, then the corresponding element for x from the loss $L_n(g) = E_{S_n,x}|g(S_n, x) - f^*(x)|$ obeys

$$\begin{aligned} &E_{S_n}|g(S_n, x) - f^*(x)| = \\ &\quad E_{S_n} 1\{A(S_n, x, R_n)\}|g(S_n, x) - f^*(x)| \\ &\quad + E_{S_n} 1\{\neg A(S_n, x, R_n)\}|g(S_n, x) - f^*(x)| \end{aligned}$$

The last expression is equal to

$$E_{S_n} 1\{\neg A(S_n, x, R_n)\}|f(S_n, x) - f^*(x)|$$

so when comparing $L_n(f)$ to $L_n(g)$ it cancels out. We are left with evaluating

$$l_x(g) \equiv E_{S_n} 1\{A(S_n, x, R_n)\}|g(S_n, x) - f^*(x)|$$

which we compare to

$$l_x(f) \equiv E_{S_n} 1\{A(S_n, x, R_n)\}|f(S_n, x) - f^*(x)|$$

As mentioned before, when $A_+(S_n, x, r)$ holds then f returns -1, because f considers only the points in radius r, all of whom have label -1, and because f is reasonable. Consequently in this case g returns $+1$, and vice versa for A_-. To consider the difference $L_n(g) - L_n(f)$, which we want to prove is negative, we can then write

$$\begin{aligned} &l_x(g) - l_x(f) \qquad (5) \\ &= E_{S_n} 1\{A(S_n, x, R_n)\} \\ &\qquad (|g(S_n, x) - f^*(x)| - |f(S_n, x) - f^*(x)|) \\ &= E_{S_n} 1\{A_+(S_n, x, R_n)\} \\ &\qquad (|1 - f^*(x)| - |-1 - f^*(x)|) \\ &\quad + E_{S_n} 1\{A_-(S_n, x, R_n)\} \\ &\qquad (|-1 - f^*(x)| - |1 - f^*(x)|) \\ &= E_{S_n}|1 - f^*(x)| \\ &\qquad (1\{A_+(S_n, x, R_n)\} - 1\{A_-(S_n, x, R_n)\}) \\ &\quad - E_{S_n}|-1 - f^*(x)| \\ &\qquad (1\{A_+(S_n, x, R_n)\} - 1\{A_-(S_n, x, R_n)\}) \\ &= E_{S_n}\left[1\{A_+(S_n, x, R_n)\} - 1\{A_-(S_n, x, R_n)\}\right] \\ &\qquad (|1 - f^*(x)| - |-1 - f^*(x)|) \\ &= -2f^*(x)\Big[P_{S_n}\big(A_+(S_n, x, R_n)\big) \\ &\qquad\qquad - P_{S_n}\big(A_-(S_n, x, R_n)\big)\Big] \end{aligned}$$

Our goal is to show that the expected value over x of this last expression is negative. For convenience we will write $A(S_n, x, R_n) \equiv A$, $A_+(S_n, x, R_n) \equiv A_+$ and likewise for A_-. Note that, for any x fulfilling $1\{A_X\}$, we have that the probability of A_0 converges to 1 as mentioned before. Thus, we are left to consider the sign of

$$-E_x 1\{A_X\} f^*(x)\Big[P_{S_n}(A_+|A_0) - P_{S_n}(A_-|A_0)\Big]$$

Denote

$$F_n(K, k) = P_{S_n}(A_+|A_0, K, k) - P_{S_n}(A_-|A_0, K, k)$$

where K is the number of observations in radius $(R_n, \sqrt{R_n})$ and k is the number in $(0, R_n)$, both around x; hence the relevant set of values for K is $[n\sqrt{R_n}, 3n\sqrt{R_n}]$, and for k is $[nR_n, 3nR_n]$. Note that F_n depends on x, but we omit it for clarity for reasons which will soon be obvious.

Let $p_n(K, k)$ be the probability of the values K, k for any x fulfilling $1\{A_X\}$. Then

$$\begin{aligned} &-E_x 1\{A_X\} f^*(x)\Big[P_{S_n}(A_+|A_0) - P_{S_n}(A_-|A_0)\Big] \\ &\quad = -\sum_{K,k} p_n(K, k) E_x 1\{A_X\} f^*(x) F_n(K, k) \end{aligned}$$

where the sum is over the set of relevant values for K, k as mentioned before.

We will now show that large enough n we have, for all relevant K, k, that $E_x 1\{A_X\} f^*(x) F_n(K, k) > 0$; note that this is enough to finish the proof.

Consider some fixed K, k and some fixed x fulfilling $1\{A_X\}$. Assume WLOG that $0 < f^*(x) < 1$ (due to the symmetry in the problem, the other case arrives at the same result). Using the Lipschitz property of f^*, and since $P(y = 1|x) = \frac{1}{2}(1 + f^*(x))$, we can bound the conditional probabilities on A_0, K, k (and assuming x fulfills $1\{A_X\}$) in the following manner (note that the conditional probabilities only depend on the behavior of y_i values):

$$\begin{aligned} &P_{S_n}(A_+|A_0, K, k) \geq \\ &\qquad \frac{\left(1 + f^*(x) - \sqrt{R_n}L\right)^K (1 - f^*(x) - R_nL)^k}{2^{K+k}} \\ &P_{S_n}(A_-|A_0, K, k) \leq \\ &\qquad \frac{(1 + f^*(x) + R_nL)^k \left(1 - f^*(x) + \sqrt{R_n}L\right)^K}{2^{K+k}} \end{aligned}$$

Note that these bounds depend only on $f^*(x)$ and not x itself. Note also that in particular

$$F_n(K, k) \geq -\frac{(2 + R_nL)^k \left(1 + \sqrt{R_n}L\right)^K}{2^{K+k}} \qquad (6)$$

Now, consider $2^{K+k} E_x 1\{A_X\} f^*(x) F_n(K, k)$. We claim that according to the bounds above, for every x fulfilling $1\{A_X\}$ we have

$$\inf_{K,k} 2^{K+k} F_n(K, k) \to \infty \qquad (7)$$

where the infimum is taken over all relevant K, k. To see this, recall the assumption that $0 < f^*(x) < 1$, and consider the behavior of the bound for $2^{K+k}P_{S_n}(A_+|A_0,K,k)$: by taking the logarithm we get

$$K\log(1+f^*(x)-\sqrt{R_n}L)+k\log(1-f^*(x)-R_nL)$$

which clearly converges to infinity, even when taking the infimum over K,k, since $R_n \to 0$ and all relevant K converge to infinity faster than all k (recall the ranges of values of K,k, and that $nR_n \to \infty$, so they all converge to infinity). Similarly we can see that $2^{K+k}P_{S_n}(A_-|A_0,K,k)$ converges to 0, thus showing (7).

In a similar manner we can see that, for every x fulfilling $1\{A_X\}$, for large enough n we have

$$\inf_{K,k} 2^{K+k}F_n(K,k) > \sup_{K,k}(2+R_nL)^k\left(1+\sqrt{R_n}L\right)^K \tag{8}$$

Note that the RHS is related to the lower bound of $F_n(K,k)$ as shown in (6).

Taken together, the facts just stated imply that the measure of points x fulfilling both (7) and (8) converges to 1 (formally, using the dominated convergence theorem on the identifier function on that set). Due to (6), it is clear that the values of the other points cannot overcome them from causing the overall integral to be positive, and we conclude that $E_x 1\{A_X\}f^*(x)F_n(K,k) > 0$ for large enough n in a manner that does not depend upon K,k (since we have used the sup, inf over relevant K,k values), proving the result.

References

[AMS97] C. G. Atkeson, A. W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

[BDR06] 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.

[BJM06] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BV92] L. Bottou and V. N. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.

[CL95] 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`.

[Dev81] L. Devroye. On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics*, 9:1310–1319, 1981.

[DGKL94] L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385, 1994.

[DGL96] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, NY, 1996.

[FS99] Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Soc. for Artif. Intel.*, 14(5):771–780, 1999.

[GKP84] W. Greblicki, A. Krzyzak, and M. Pawlak. Distribution-free pointwise consistency of kernel regression estimate. *Annals of Statistics*, 12(4):1570–1575, 1984.

[Joa98] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, volume 1398, pages 137–142. Springer Verlag, Heidelberg, DE, 1998.

[SFBL97] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. 14th International Conference on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997.

[SGV98] 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.

[STC00] J. Shawe-Taylor and N. Cristianini. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.

[Ste02] I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.

[Sto80] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8:1348–1360, 1980.

[Sto82] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

[Vap98] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, NY, 1998.

[VB93] V. N. Vapnik and L. Bottou. Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6):893–909, 1993.

[Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–134, 2004.

[ZR07] A. Zakai and Y. Ritov. Local behavior of consistent learning methods, 2007. Submitted for Publication.