Consistency of nearest neighbor classification under selective sampling

Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
9500 Gilman Drive #0404, La Jolla, CA 92093

Editor: Shie Mannor, Nathan Srebro, Bob Williamson

Abstract

This paper studies nearest neighbor classification in a model where unlabeled data points arrive in a stream, and the learner decides, for each one, whether to ask for its label. Are there generic ways to augment or modify any selective sampling strategy so as to ensure the consistency of the resulting nearest neighbor classifier?

1. Introduction

A binary classification problem is specified by an instance space $\mathcal{X}$, a label space $\mathcal{Y} = \{0, 1\}$, and a distribution $\mathbf{P}$ on $\mathcal{X} \times \mathcal{Y}$. For (X, Y) generated from $\mathbf{P}$, let μ denote the marginal distribution of X, and η the conditional expectation $\eta(x) = \mathbb{E}(Y|X = x)$. The error rate, or risk, of a rule $h : \mathcal{X} \to \mathcal{Y}$ is $\mathbf{P}(h(X) \neq Y)$. This is minimized by the rule $h^*(x) = 1(\eta(x) \geq 1/2)$, whose error rate is called the Bayes-optimal risk, R^*.

Now suppose that $\mathcal{X}$ is a metric space and that an infinite stream of examples, (X_1, Y_1), (X_2, Y_2), ..., is generated by independent draws from $\mathbf{P}$. For $n = 1, 2, \ldots$, there are various nearest neighbor classifiers T_n based on the first n observations. The 1-NN classifier assigns a point x the label of its nearest neighbor in $X_1, \ldots, X_n$. The k-NN classifier takes the majority label among x's k nearest neighbors; and the k_n-NN classifier does the same, for k_n growing with n. These schemes have all been shown to be *consistent*, or nearly so: as n grows, the expected risk of T_n goes to R^* for the k_n-NN classifier if $k_n = o(n)$, to $R^* + O(1/\sqrt{k})$ for k-NN, and to at most $2R^*$ for 1-NN (Fix and Hodges, 1951; Cover and Hart, 1967; Stone, 1977).

In this paper, we investigate how consistency is affected by *selective sampling*. Suppose the above setup is modified so that instances are free but labels have a unit cost. At each time step $n = 1, 2, \ldots$, a learning algorithm sees X_n and then decides on the spot whether or not to purchase Y_n. Nearest neighbor classifiers can be defined as before, except that now neighbors are chosen from the set of X_i whose labels are known.

The goal in selective sampling is to get a low-error classifier while buying as few labels as possible. There are plenty of sensible-sounding strategies for achieving this, for instance:

(S0) Given X_n, find its two nearest queried neighbors. Ask for Y_n if their labels differ.

But this, and many others like it, fail to preserve consistency. To see why, suppose that X is uniformly distributed in $[0, 1]$ and that Y is 0 except when $X \in [1/2 - \alpha, 1/2 + \alpha]$ (for small $\alpha > 0$), in which case it is 1. Rule (S0) is very likely to start with two 0-labels and then never ask for any other label. Thus it incurs an asymptotic risk of 2α whereas $R^* = 0$.

Taking a high-level view of the problem, the behavior of η inside any specific ball can be quite different from its behavior outside the ball (unless there are strong smoothness conditions). In order to guarantee consistency, it is therefore essential that *every ball in $\mathcal{X}$ of nonzero probability mass be queried infinitely often.* This is easy to achieve, for instance by querying each Y_n with probability $1/n$.

The first question we study is the following: in order to make a selective sampling strategy consistent, is it enough to simply ensure the following?

(R0) Every Y_n is queried with probability at least $1/n$.

The answer is yes, if $R^* = 0$ and the decision boundary isn't too strange (Theorem 2). Requirement (R0) is easily added to any sampling strategy, and incurs an overhead of just $\log n$ queries in n time steps.

On the other hand, if $R^* > 0$, consistency can fail dramatically under condition (R0). We construct a one-dimensional example in which R^* is arbitrarily small and yet the asymptotic risk is close to 1 (Theorem 3). We then propose a different sampling requirement that does yield consistency for any R^* (Theorem 13). It is simple and can be tacked on to any selective sampling strategy; however, it doubles the number of queries made by that strategy.

Finally, we consider rates of convergence. It has traditionally been somewhat tricky to give these for nearest neighbor schemes, unless either the data is one-dimensional or (μ, η) satisfy strong smoothness properties. We analyze a one-dimensional setting in which $R^* = 0$ but there are an unknown number of sign changes. We find that a scheme similar to (S0) works well, if augmented with additional sampling of type (R0). In fact, it attains error $\leq \epsilon$ after a number of queries proportional to just $\log(1/\epsilon)$ (Theorem 14).

Related work

After the pioneering work of Fix and Hodges (1951) on the k_n-NN rule, Cover and Hart (1967) showed consistency (within a factor of 2) of the 1-NN rule when $\mathcal{X}$ is a separable metric space and η is continuous; they also analyzed the k-NN case. Later, Stone (1977) showed consistency of the k_n-NN rule in Euclidean spaces, without any distributional assumptions; see also (Devroye, 1981). This was extended to strong consistency by Devroye et al. (1994). There has been some work that gives consistency results for nearest neighbor schemes under non-i.i.d. sampling (Kulkarni and Posner, 1995). However, these prohibit the querying decision for a point X_n from depending upon the results of earlier queries, which rules out selective sampling strategies.

On the active learning front, there have been many recent results focusing on learning parametric models like linear separators. Consistency has been found to be a basic hurdle because the learner's attempt to pick out informative examples leaves it with a labeled set that can be unrepresentative of the underlying distribution $\mathbf{P}$ (Dasgupta, 2011). There has also been some work on nonparametric settings (Castro and Nowak, 2008; Hanneke, 2011), studying the best rates achievable in some canonical cases. In this paper, we find that consistency is a problem for nearest neighbor methods as well, but that it permits a more lightweight and generic solution than has been obtained in the parametric setting.

2. Preliminaries

Let $(\mathcal{X}, d)$ be a metric space, and for any $x \in \mathcal{X}$ and $r \geq 0$, let $B(x, r)$ be the closed ball of radius r centered at x, that is, $\{z \in \mathcal{X} : d(x, z) \leq r\}$.

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be obtained by independent draws from distribution $\mathbf{P}$ on $\mathcal{X} \times \mathcal{Y}$. The learning algorithm sees the X_n in order, and for each, decides whether or not to query Y_n. Let Q_n denote the multiset of instances among $X_1, \ldots, X_n$ whose labels are queried; and let $\widehat{Y}_n(x)$ be the label of $x \in Q_n$. For any point $x \in \mathcal{X}$, define

$$\begin{aligned}\Gamma(x, S) &= \text{nearest neighbor of } x \text{ in } S\\ \Gamma_k(x, S) &= \text{multiset of } k \text{ nearest neighbors of } x \text{ in } S\end{aligned}$$

(breaking ties by preferring recent points, say). The 1-NN classifier, on input x, returns $\widehat{Y}_n(\Gamma(x, Q_n))$. The k-NN classifier returns the majority label amongst the $\widehat{Y}_n(z)$, for $z \in \Gamma_k(x, Q_n)$; ties are broken by, say, tossing a fair coin. And k_n-NN is like k-NN, except that k is a growing function of n.

We will denote a nearest neighbor classifier by $T_n : \mathcal{X} \to \mathcal{Y}$. This function, and its risk $R_n = \Pr_{(X,Y)}(T_n(X) \neq Y)$, are random variables depending on the training process. We say T_n is *consistent* if $\mathbb{E}R_n \to R^*$, taking expectation over the learning process up to time n, and we say it is *strongly consistent* if $R_n \to R^*$ almost surely.

3. Consistency in the realizable case

We start with the "realizable" case. Here we require $R^* = 0$, as well as a technical assumption about the boundary between the 0-label and 1-label regions. For $i \in \{0, 1\}$, let $\mathcal{X}_i$ consist of all $x \in \mathcal{X}$ for which there is a closed ball B centered at x with (i) $\mu(B) > 0$ and (ii) $\eta(z) = i$ for all $z \in B$. The specific realizability assumption is:

(A1) $\mu(\mathcal{X}_0 \cup \mathcal{X}_1) = 1$.

We will analyze selective sampling schemes which meet the following requirement.

(R1) There is a sequence of reals $(a_1, a_2, \ldots)$ such that the probability that Y_n is queried, conditional on X_n and all prior history, is at least a_n.

To make this formal, let $\mathcal{F}_n$ denote the σ-field of everything the learner has seen up to and including time n. If we define

$$Y_i' = \begin{cases} Y_i & \text{if queried} \\ ? & \text{if not queried} \end{cases}$$

we can write $\mathcal{F}_n = \sigma(X_1, Y_1', \ldots, X_n, Y_n')$. Our requirement for selective sampling is that

$$\Pr(Y_n \text{ is queried} \mid \mathcal{F}_{n-1}, X_n) \geq a_n.$$

Define $s_n = a_1 + \cdots + a_n$. We'll see that as long as $s_n \to \infty$, the k-NN estimator is consistent for any fixed k. Thus, for instance, setting $a_n = 1/n$ is good enough. More generally, for the k_n-NN estimator, a sufficient condition is $s_n/k_n \to \infty$. We will treat the three cases together, by considering a k_n-NN estimator in which the sequence (k_n) is allowed to be constant.

These results are based on the following consequence of the Borel-Cantelli lemma.

Lemma 1 *Pick any ball B with $\mu(B) > 0$. If k_n is a nondecreasing sequence of positive integers and $s_n/k_n \to \infty$, then there is almost surely some n_0 such that $|Q_n \cap B| \geq k_n$ for all $n \geq n_0$.*

Proof Let Z_t be the event that X_t lies in ball B *and* that Y_t is queried; so Z_t is $\mathcal{F}_t$-measurable. Define $\xi_t = \Pr(Z_t|\mathcal{F}_{t-1}) \geq \mu(B)a_t$. Since $\sum_{t=1}^n \xi_t \geq \mu(B)s_n \to \infty$, it follows by Levy's martingale version of the Borel-Cantelli lemma (Williams, 1991, Theorem 12.15) that

$$\frac{\sum_{t=1}^n \mathbf{1}(Z_t)}{\sum_{t=1}^n \xi_t} \to 1$$

almost surely. The numerator is $|Q_n \cap B|$ and the denominator is $\geq \mu(B)s_n$. Thus with probability one, there is some n_1 such that $|Q_n \cap B| \geq (1/2)\mu(B)s_n$ for any $n \geq n_1$.

By the condition on s_n, there is some n_2 such that $s_n/k_n \geq 2/\mu(B)$ whenever $n \geq n_2$. Thus, for $n \geq n_0 = \max(n_1, n_2)$, we have $|Q_n \cap B| \geq k_n$. ■

Theorem 2 *Let $(k_n : n = 1, 2, \ldots)$ be any nondecreasing sequence of positive integers. For a selective sampling scheme that meets requirement (R1), define $s_n = a_1 + \cdots + a_n$. If (A1) holds, and $s_n/k_n \to \infty$, then the resulting k_n-NN predictor is strongly consistent.*

Proof Let z denote a complete infinite instantiation $((x_1, y_1'), (x_2, y_2'), \ldots)$ of the training process. This $z \in \mathcal{Z} = (\mathcal{X} \times (\mathcal{Y} \cup \{?\}))^\infty$ is a draw from the distribution — call it γ — induced by (μ, η) and the learning algorithm. Any z specifies a sequence of k_n-NN classifiers $(T_{n,z} : n = 1, 2, \ldots)$.

For any z, the classifier $T_{n,z}$ has risk $R_{n,z} = \Pr_{(X,Y)}(T_{n,z}(X) \neq Y) = \mathbb{E}_X \text{err}_n(z, X)$, where for any $x \in \mathcal{X}$, we define

$$\text{err}_n(z, x) \;=\; \eta(x)1(T_{n,z}(x) \neq 1) + (1 - \eta(x))1(T_{n,z}(x) \neq 0).$$

We call a pair (z, x) "good" if this error goes to zero as n increases. More precisely, define $G = \{(z, x) : \limsup_n \text{err}_n(z, x) = 0\} \subset \mathcal{Z} \times \mathcal{X}$.

Now, pick any $x \in \mathcal{X}_i$, for $i \in \{0, 1\}$. By definition, $\eta = i$ on some closed ball B around x of nonzero probability mass. By Lemma 1, with probability one (over z), there exists some n_0 such that for $n \geq n_0$, the k_n nearest queried neighbors $\Gamma_{k_n}(x, Q_n)$ lie within B, whereupon $T_n(x) = i = \eta(x)$ and hence $\text{err}_n(z, x) = 0$. In other words, for any $x \in \mathcal{X}_0 \cup \mathcal{X}_1$, we have $\gamma(\{z : (z, x) \in G\}) = 1$. Since $\mu(\mathcal{X}_0 \cup \mathcal{X}_1) = 1$, it follows that $(\gamma \times \mu)(G) = 1$, where $\gamma \times \mu$ is the product measure.

Strong consistency, namely the property that

$$\gamma(\{z : \lim_{n\to\infty} R_{n,z} = 0\}) \;=\; \gamma(\{z : \lim_{n\to\infty} \mathbb{E}_X \text{err}_n(z, X) = 0\}) \;=\; 1,$$

now follows by straightforward manipulations (Lemma 16). ■

4. An illustrative example for the general case

We'll now see that sampling requirement (R1), with $s_n = a_1 + \cdots + a_n \to \infty$, is no longer sufficient to guarantee consistency when the optimal risk is nonzero. In the counterexample we study, the optimal risk can be tuned to be arbitrarily close to 0, while the asymptotic risk of the 1-NN predictor gets close to 1.

4.1. Model and result

Let (X, Y) be a pair of random variables, where X is uniform distributed in $[0, 1]$ and $Y \in \{0, 1\}$ is independent of X, with $\mathbb{E}Y = \eta$ for some $0 < \eta < 1/2$. The Bayes-optimal prediction is zero everywhere, and has risk $R^* = \eta$.

Suppose the learner uses the following selective sampling strategy when it sees X_n: if X_n's nearest queried neighbor has label 0, then it queries Y_n, otherwise it queries Y_n with probability $1/n$. Recalling that $\widehat{Y}_t(\Gamma(x, Q_t))$ denotes the label of the nearest queried neighbor of x amongst $X_1, \ldots, X_t$, we can write:

(S1) Given X_n:

- If $\widehat{Y}_{n-1}(\Gamma(X_n, Q_{n-1})) = 0$ then query Y_n (call this a "type 0" query).
- If $\widehat{Y}_{n-1}(\Gamma(X_n, Q_{n-1})) = 1$ then query Y_n with probability $1/n$ ("type 1" query).

This strategy meets requirement (R1), but nevertheless produces a 1-NN classifier whose asymptotic expected risk is $1 - \eta$.

Theorem 3 *Let T_n denote the 1-NN classifier based upon the points queried up to time n under (S1). Pick any $0 < x < 1$. Then $\Pr(T_n(x) = 1) \to 1$ as $n \to \infty$.*

4.2. Analysis outline

Fix any point $0 < x < 1$. To simplify notation, we use $\widehat{Y}_n$ as a shorthand for the label of x's nearest queried neighbor, $\widehat{Y}_n(\Gamma(x, Q_n))$. We also use $\widehat{Y}_n^L$ to denote the (label of the) nearest queried neighbor to the left of x, and $\widehat{Y}_n^R$ the nearest queried neighbor to the right of x.

We will show that for large n, it is very likely that $\widehat{Y}_n^L = \widehat{Y}_n^R = 1$, and hence $\widehat{Y}_n = 1$. This results from two effects. First, if this condition holds at any time $t \geq n/\ln n$, it is likely to still hold at time n because the sampling rule dampens querying between 1-labels. Second, if the condition does not hold at any specific time t, then it has a constant probability of holding by the time two more points arrive that are close to x and on either side of it, since each of these points has a constant probability of being labeled 1.

The analysis is based on four events, each of which has probability $1 - o(1)$ of occurring. In what follows, let $w(n)$ be a decreasing function of n, and let $f(n)$ and $a(n)$ be functions that are slowly increasing. We will make these specific later. Moreover, let I_n denote the closed interval $[x - w(n), x + w(n)]$; we will consider n large enough that $I_n \subset [0, 1]$. We divide the first n time steps of the learning algorithm into two phases: denoting time by t,

Phase One: $t = 1, 2, \ldots, f(n)$.

Phase Two: $t = f(n) + 1, \ldots, n$.

Finally, we will call a specific time t an *arrival* if x's nearest neighbor changes at this time, that is, if $|x - X_t| < \min(|x - X_1|, \ldots, |x - X_{t-1}|)$; and we will let T_n be the time of the $a(n)^{th}$ arrival *after* phase one.

Here are the four events of interest.

(E_1) During phase one, at least one point is queried in each half of I_n (left/right of x).

(E_2) $\widehat{Y}_t^L = \widehat{Y}_t^R = 1$ for some $f(n) < t < T_n$.

(E_3) $T_n \leq n$.

(E_4) During phase two, no queries of type 1 are made in I_n.

Lemma 4 *If E_1, E_2, E_3, E_4 all occur, then $\widehat{Y}_n = 1$.*

Proof If E_1, E_2, E_3 occur, then there is a time $f(n) < t \leq n$ at which $\widehat{Y}_t^L = \widehat{Y}_t^R = 1$, and these correspond to data points in I_n, on either side of x. Call these points x_L and x_R. The first query to occur in the interval (x_L, x_R) after time t is necessarily a type-1 query; but E_4 tells us there is no such query. Thus $\widehat{Y}_n$ is the label of either x_L or x_R, and these are both 1. ■

4.3. Event E_1

For integers $n_1 \leq n_2$, let $H(n_1, n_2)$ denote the harmonic sum

$$H(n_1, n_2) \;=\; \frac{1}{n_1} + \frac{1}{n_1 + 1} + \cdots + \frac{1}{n_2} \;\approx\; \ln \frac{n_2}{n_1}. \tag{1}$$

Lemma 5 *If n is large enough that $I_n \subset [0,1]$, then* $\Pr(\overline{E_1}) \leq 2\exp(-w(n)H(1, f(n)))$.

Proof The probability that X_i lies in the left half of I_n (which has width $w(n)$) and is queried is at least $w(n)/i$. Thus the chance this never happens during phase one is at most

$$\prod_{i=1}^{f(n)} \left(1 - \frac{w(n)}{i}\right) \;\leq\; \exp\left(-w(n)H(1, f(n))\right).$$

and the same applies to the right half of I_n. ■

4.4. Event E_2

We will show that the condition $\widehat{Y}_t^L = \widehat{Y}_t^R = 1$ has a constant probability of being created between any two arrivals. More precisely, suppose that t is the time of an arrival, and that the desired condition does not hold right before t; in other words, one or both of $\widehat{Y}_{t-1}^L$, $\widehat{Y}_{t-1}^R$ are zero. Then there is a constant probability that both of these values will become one before the arrival subsequent to t.

Lemma 6 *Pick any arrival time t. Then*

$$\Pr\left(\exists\ t' \geq t \text{ before next arrival with } \widehat{Y}^L_{t'} = \widehat{Y}^R_{t'} = 1 \mid t \text{ arrival}, \widehat{Y}^L_{t-1}\widehat{Y}^R_{t-1} = 0\right) \ \geq\ \frac{1}{12}\eta^2.$$

We bound the probability that event E_2 fails by applying the previous lemma repeatedly.

Lemma 7 $\Pr(\overline{E_2}|E_1) \leq (1-\eta^2/12)^{a(n)-1}$.

Proof If event E_1 occurs, then by time $f(n)$, there are queried points to the left and right of x. We now apply Lemma 6 to each of the following $a(n)-1$ arrivals; in each case, there is at least an $\eta^2/12$ probability that the condition $\widehat{Y}^L_t = \widehat{Y}^R_t = 1$ will exist at some time t before the next arrival. ■

4.5. Event E_3

We need to analyze the likely number of arrivals between two times, say s and t. This is not complicated, because when $X_1, \ldots, X_t$ are ranked by distance from x, the result is a random permutation of $(1, 2, \ldots, t)$. So, let π denote such a random permutation, and define the indicator variable

$$Z_i \ = \ \begin{cases} 1 & \text{if } \pi(i) < \pi(1), \ldots, \pi(i-1) \\ 0 & \text{otherwise .} \end{cases}$$

Then $A(s,t) = Z_s + Z_{s+1} + \cdots + Z_t$ has exactly the same distribution as the number of arrivals in time steps $s, s+1, \ldots, t$. We now give a Chebyshev bound on $A(s,t)$.

Lemma 8 *For any $s \leq t$, and any $c > 0$,*

$$\Pr\left(|A(s,t) - H(s,t)| \geq c\sqrt{H(s,t)}\right) \ \leq\ \frac{1}{c^2}.$$

where $H(s,t)$ is the harmonic sum defined in (1).

Lemma 9 *If $a(n) \leq H(f(n)+1,n)/2$, then* $\Pr(\overline{E_3}) \leq 4/H(f(n)+1,n)$.

Proof Recalling that T_n is the time of the $a(n)^{th}$ arrival after time $f(n)$,

$$\Pr(T_n > n) \ = \ \Pr(A(f(n)+1,n) < a(n)) \ \leq\ \Pr(A(f(n)+1,n) < H(f(n)+1,n)/2),$$

and then we apply Lemma 8 with $c = (1/2)\sqrt{H(f(n)+1,n)}$. ■

4.6. Event E_4

Lemma 10 $\Pr(\overline{E_4}) \leq 2w(n)H(f(n)+1, n)$.

Proof The probability of a type-1 query in I_n at time i is at most $2w(n)/i$. Thus

$$\Pr(\overline{E_4}) \;\leq\; \sum_{i=f(n)+1}^{n} 2w(n) \cdot \frac{1}{i} \;=\; 2w(n)H(f(n)+1, n).$$

■

To finish the proof of Theorem 3, we ensure that $w(n)\log f(n)$, $a(n)$, and $\log(n/f(n))$ are increasing functions of n while $w(n)\log(n/f(n))$ is a decreasing function of n; for instance:

$$w(n) = \frac{1}{\sqrt{\ln n}}, \quad f(n) = \frac{n}{\ln n}, \quad a(n) = \sqrt{\ln\ln n}.$$

For n large enough that $[x - w(n), x + w(n)] \subset [0,1]$, we can then bound $\Pr(\widehat{Y}_n(\Gamma(x, Q_n)) \neq 1)$ by $\Pr(\overline{E_1} \vee \overline{E_2} \vee \overline{E_3} \vee \overline{E_4})$, which decreases to zero as $n \to \infty$.

5. Consistency in the general case

In the general case, when η takes values outside $\{0,1\}$, we have seen that requirement (R1) does not guarantee consistency. In the counterexample, we created a sampling rule satisfying this condition but favoring 1-labels, and saw that it caused a 1-NN classifier to eventually predict 1 everywhere, even if η were close to 0 everywhere.

To prevent the sampling strategy from being biased towards a particular label, we could stipulate, for instance, that the probability of querying the label of a new point x depends only on how homogeneous the labels of its k nearest queried neighbors are; that is, if the sum of these nearest labels is j, the probability should depend only on $\min(j, k-j)$. Something like this might work, but it creates statistical dependencies between the queried labels that are long-ranging and complex, making the analysis difficult.

Instead, we analyze a generic scheme which removes all complicated dependencies by dividing the queried labels into two groups, one of which is used for making selective sampling decisions, while the other is for use only by the final classifier. We call this latter multiset F ("future use"), with F_n denoting its state at the end of the n^{th} time step.

(R2) The querying strategy satisfies three rules: (1) Any queried point can be placed in F_n, and F_n contains only queried points. (2) The decision to query Y_n and/or place X_n in F_n depends on X_n and on all prior history except for the labels of points in F_{n-1}. (3) There is a sequence of reals $(a_1, a_2, \ldots)$ that the probability that X_n is placed in F_n is at least a_n, conditional upon X_n and prior history.

If we define

$$Y_i' \;=\; \begin{cases} Y_i & \text{if queried and not in } F \\ ! & \text{if queried and in } F \\ ? & \text{if not queried} \end{cases} \qquad .$$

then $\mathcal{G}_n = \sigma(X_1, Y_1', \ldots, X_n, Y_n')$ is all the data the learner uses to make decisions during training. What is hidden until prediction time is:

$$Y_i'' = \begin{cases} Y_i & \text{if queried and in } F \\ ! & \text{otherwise} \end{cases} .$$

The final estimator T_n, when given a query x, returns the majority label in $\widehat{Y}_n(\Gamma_{k_n}(x, F_n))$ (where k_n might be constant).

(R2) amounts to the following conditions:

- Decisions about querying Y_n and placing X_n in F_n are based only on $\mathcal{G}_{n-1}$ and X_n.
- $\Pr(X_n \text{ is put in } F_n \mid \mathcal{G}_{n-1}, X_n) \ \geq \ a_n$.

For instance, one could start from any querying strategy, then make sure to query each X_n with probability at least $2a_n$, and then place each queried point in F_n with probability 1/2.

We have adopted the metric space setting of Cover and Hart, and will use the same two assumptions.

(A2) The metric space $(\mathcal{X}, d)$ is separable (contains a countable dense subset).

The support of μ is defined as $\{x \in \mathcal{X} : \mu(B(x, r)) > 0 \text{ for all } r > 0\}$. A consequence of separability (Cover and Hart, 1967) is that this set has μ-mass 1.

(A3) For x chosen from μ, almost surely either $\mu(\{x\}) > 0$ or x is a continuity point of η.

As a result of (A3), there is a subset $\mathcal{X}' \subset \mathcal{X}$ with $\mu(\mathcal{X}') = 1$ that satisfies the following property: for any $x \in \mathcal{X}'$ and any $\epsilon > 0$, there is a closed ball B_x^ϵ centered at x with $\mu(B_x^\epsilon) > 0$, such that $|\eta(z) - \eta(x)| < \epsilon$ for all $z \in B_x^\epsilon$.

The asymptotic behavior of the selective sampling scheme for nearest neighbor rests upon the following fact, which is proved in exactly the same way as Lemma 1.

Lemma 11 *Pick any ball B with $\mu(B) > 0$. If k_n is a non-decreasing sequence of positive integers and $s_n/k_n \to \infty$ (where, as before, $s_n = a_1 + \cdots + a_n$), then there is almost surely some n_0 such that $|F_n \cap B| \geq k_n$ for all $n \geq n_0$.*

The remainder of the analysis cobbles together ideas from earlier nearest neighbor work, while clarifying that no unwanted dependencies are introduced by selective sampling.

As a consequence of Lemma 11, for large n, the nearest neighbors of a query point x will lie sufficiently close to it that their $\eta(\cdot)$ values will be close to $\eta(x)$, say within $\eta(x) \pm \epsilon$. What is the probability that the majority vote over these k nearest neighbors coincides with the label at x? This is a simple question about independent coin flips.

Lemma 12 *Pick any $0 < \epsilon, \eta < 1$. Let $Z, Z_1, \ldots, Z_k \in \{0, 1\}$ be the outcomes of independent coin flips with heads probabilities $\eta, \eta_1, \ldots, \eta_k$, respectively, where $|\eta_i - \eta| \leq \epsilon$. Let M_k be the majority vote over $Z_1, \ldots, Z_k$, breaking ties with a fair coin flip. Define $C(k, \epsilon, \eta)$ to be the supremum of $\Pr(M_k \neq Z)$ over all choices $\eta_1, \ldots, \eta_k \in [\eta - \epsilon, \eta + \epsilon]$. Then*

(a) $C(1, \epsilon, \eta) \leq 2 \min(\eta, 1 - \eta) + \epsilon$.

(b) If $k > 1$ and either $\eta = 1/2$ or $\epsilon \leq |1 - 2\eta|/4$, then $C(k, \epsilon, \eta) \leq \min(\eta, 1-\eta) + 2/\sqrt{k}$.

Theorem 13 *Let $(k_n : n = 1, 2, \ldots)$ be any nondecreasing sequence of positive integers. Suppose a selective sampling strategy meets requirement (R2). Define $s_n = a_1 + \cdots + a_n$. If assumptions (A2) and (A3) hold, and if $s_n/k_n \to \infty$, the asymptotic expected risk of the resulting k_n-NN classifier can be bounded thus:*

$$\limsup_n \mathbb{E}R_n \ \leq \ \begin{cases} 2R^* & \textit{if } (k_n) \equiv 1 \\ R^* + \frac{2}{\sqrt{k}} & \textit{if } (k_n) \equiv k \\ R^* & \textit{if } k_n \to \infty \end{cases}$$

Proof Pick any $\epsilon_o > 0$. For any $x \in \mathcal{X}$, define

$$\epsilon(x) \ = \ \begin{cases} \epsilon_o & \text{if } \eta(x) = 1/2 \\ \min(\epsilon_o, |1 - 2\eta(x)|/4) & \text{otherwise} \end{cases}$$

As in Theorem 2, denote an instantiation of the complete learning process by

$$z \ = \ ((x_1, y_1', y_1''), (x_2, y_2', y_2''), \ldots) \in \mathcal{Z} \ = \ (\mathcal{X} \times (\mathcal{Y} \cup \{?, !\}) \times (\mathcal{Y} \cup \{!\}))^\infty.$$

We can break this into the information used during learning, $z' = ((x_1, y_1'), (x_2, y_2'), \ldots)$, and the labels needed only for prediction, $z'' = (y_1'', y_2'', \ldots)$.

Each z defines an infinite sequence of future-use sets $F_n(z)$ and k_n-NN classifiers $T_{n,z}$. The expected risk at time n is

$$\begin{aligned} \mathbb{E}R_n &= \Pr_{Z,X,Y}(T_{n,Z}(X) \neq Y) \\ &\leq \Pr_{Z,X}(|F_n(Z) \cap B_X^{\epsilon(X)}| < k_n) + \Pr_{Z,X,Y}(T_{n,Z}(X) \neq Y, \, |F_n(Z) \cap B_X^{\epsilon(X)}| \geq k_n). \end{aligned}$$

Let's start with the first term. For $x \in \mathcal{X}'$, we know from Lemma 11 that almost surely over Z, $\lim_{n\to\infty} 1(|F_n(Z) \cap B_x^{\epsilon(x)}| < k_n) = 0$. Applying dominated convergence twice, we first get $\Pr_Z(|F_n(Z) \cap B_x^{\epsilon(x)}| < k_n) \to 0$, and then, since $\mu(\mathcal{X}') = 1$, we also have $\Pr_{Z,X}(|F_n(Z) \cap B_X^{\epsilon(X)}| < k_n) \to 0$.

For the second term in the decomposition of $\mathbb{E}R_n$, we observe that the labels of points in F_n can be thought of as being exposed only at the time of prediction. Hence, for any x and any z',

$$\Pr_{Y,Z''}(T_{n,Z}(X) \neq Y, \, |F_n(Z) \cap B_X^{\epsilon(X)}| \geq k_n \mid X = x, Z' = z') \ \leq \ C(k_n, \epsilon(x), \eta(x))$$

for the quantity $C(\cdot)$ of Lemma 12. Thus

$$\begin{aligned} \Pr_{Z,X,Y}(T_{n,Z}(X) \neq Y, \, |F_n(Z) \cap B_X^{\epsilon(X)}| \geq k_n) \ &\leq \ \mathbb{E}_X C(k_n, \epsilon(X), \eta(X)) \\ &\leq \ \begin{cases} 2R^* + \epsilon_o & \text{if } (k_n) \equiv 1 \\ R^* + \frac{2}{\sqrt{k_n}} & \text{otherwise} \end{cases} \end{aligned}$$

The theorem follows by noting that this holds for any $\epsilon_o > 0$, and taking $n \to \infty$. ■

6. Rates of convergence in a simple setting

A simple model in which finite-sample rates of convergence can be obtained is when μ is continuous on $\mathcal{X} = [0,1]$, and η is piecewise $0-1$. More precisely, there exist values $0 = \theta_0 < \theta_1 < \cdots < \theta_k = 1$ and $\eta_1, \ldots, \eta_k \in \{0,1\}$ such that

$$\eta(x) = \eta_i \text{ for all } x \text{ in the interval } I_i = (\theta_{i-1}, \theta_i).$$

The locations of the θ_i, and the value of k, are unknown to the learning algorithm.

This setup has been addressed in previous active learning work (Balcan et al., 2010; Hanneke, 2012), though not with nearest-neighbor methods. We will study a procedure inspired by strategy (S0) in the introduction. Given X_n:

(S2) Find the nearest queried neighbors to the left and right of X_n. If either is absent, or if they have different labels, query Y_n ("type-I" query). Otherwise, query Y_n with probability $1/n^c$ ("type-II" query).

Here $0 < c \leq 1$ is some constant. Type-II queries are essential for consistency in this setting where there are an unknown number of sign changes. Since the rate at which they occur is easy to characterize, we will focus upon analyzing type-I queries.

The querying strategy cares only about the ordering of points in $\mathcal{X}$, and not the actual distances between them; hence, we may without loss of generality assume that μ is the uniform distribution on $[0,1]$. How would a supervised learner, that asked for every label, perform under these circumstances? After n examples, its estimate of each θ_i would be off by about $1/n$, and thus its error rate would be approximately k/n. This means it would need roughly k/ϵ labels before its error dropped below ϵ.

Now let's see what selective sampling does, when used for 1-NN classification. We will say that interval I_i has been *discovered* at time n if Q_n, the set of queries up to and including time n, contains a point in I_i. Let N_o be the random time at which all intervals are finally discovered, that is,

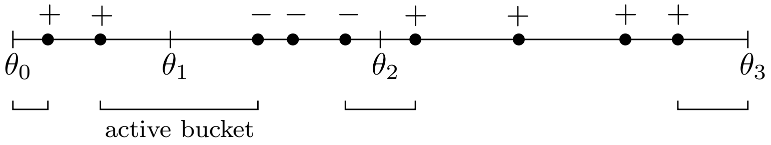
$$N_o = \min\{n : \text{every } I_i \text{ has been discovered at time } n\}.$$

We'll see that the learning proceeds in two phases: an initial phase up to time N_o, during which the rate of error reduction is similar to that of supervised learning, and a subsequent phase during which the error rate drops geometrically with each query.

Theorem 14 *There is an upper bound $0 \leq \Phi_n \leq 1$ on the error of the 1-NN classifier T_n under querying strategy (S2), such that for any $n > N_o$,*

$$\begin{aligned} \Pr(\text{type-I query at time } n) &= \Phi_{n-1} \\ \mathbb{E}(\Phi_n \mid \Phi_{n-1}, \text{type-I query at time } n) &\leq \left(1 - \frac{1}{4(k+1)}\right)\Phi_{n-1}. \end{aligned}$$

Proof Think of the queried points Q_{n-1} as partitioning $[0,1]$ into $|Q_{n-1}|+1$ *buckets*. We call a bucket *active* if it contains some θ_i, where $0 \leq i \leq k$. If $n > N_o$, no bucket contains more than one θ_i and thus there are exactly $k+1$ active buckets. Here is an example in which $k = 3$ and nine points have been queried so far:



If X_n falls into an inactive bucket B, its closest queried neighbors have the same label. If Y_n is queried, it will be a type-II query, and will split B into two inactive buckets.

If X_n falls into an active bucket B, it will be queried (type-I), and B will be split into two buckets, one active and the other inactive. A quick calculation shows that the expected length of the inactive portion is at least $|B|/4$, where $|B|$ is the length of B.

Now, define Φ_{n-1} to be the sum of $|B|$ over all active buckets at time $n-1$. This is exactly the probability that X_n falls in an active bucket, in other words, the probability of a type-I query at time n. Suppose this event occurs. Then, using our earlier calculation, the expected amount by which Φ shrinks is

$$\mathbb{E}(\Phi_{n-1} - \Phi_n \mid \Phi_{n-1}, X_n \text{ falls in an active bucket}) \;\geq\; \sum_{\text{active } B} \frac{|B|}{\Phi_{n-1}} \cdot \frac{|B|}{4} \;\geq\; \frac{1}{4(k+1)} \Phi_{n-1}$$

where the last step follows from the Cauchy-Schwarz inequality. ■

What this implies is that, after N_o, the error rate of T_n drops below ϵ after about $O(k/\epsilon)$ unlabeled points are seen, and during this time, $O(k \log(1/\epsilon))$ type-I queries are made.

How long is the initial discovery period N_o? It depends on the lengths of the intervals I_i, and on the constant c in the sampling strategy: taking $c < 1$ reduces it substantially.

Lemma 15 *Let $\Delta = \min_i |I_i|$. Then for any n,*

$$\Pr(N_o \geq n) \;\leq\; \begin{cases} k \exp(-\Delta \ln n) & \text{if } c = 1 \\ k \exp(-\Delta n^{1-c}) & \text{if } 0 < c < 1 \end{cases}$$

Proof Pick any interval I_i. The probability that at time $m \leq n$, a point falls in this interval and is queried is at least Δ/m^c. Thus the probability that I_i is not yet discovered at time n is at most $\prod_{m=1}^n (1 - \Delta/m^c) \leq \exp(-\Delta \sum_{m=1}^n m^{-c})$. The lemma then follows by lower-bounding the summation in the two regimes of interest, and taking a union bound over all I_i. ■

Finally, how many queries are made before time N_o? A quick calculation shows that in the first m time steps, the expected number of type-I queries is $O(k \log(m/k))$, because of the shrinkage rate, while the expected number of type-II queries, because they are infrequent, is $O(\log m)$ if $c = 1$ and $O(m^{1-c})$ if $c < 1$.

Acknowledgments

The author is grateful for the support of the National Science Foundation, under grants IIS-0713540 and CNS-0932403, and the National Institutes of Health, under grant U54HL10846.

References

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. *Machine Learning*, 80(2-3):111–139, 2010.

R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

T. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.

L. Devroye. On the inequality of Cover and Hart in nearest neighbor discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(1):75–78, 1981.

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.

E. Fix and J. Hodges. Discriminatory analysis, nonparametric discrimination. *USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004, Report 4, Contract AD41(128)-31*, 1951.

S. Hanneke. Rates of convergence in active learning. *Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke. Activized learning: transforming passive to active with improved label complexity. *To appear, Journal of Machine Learning Research*, 2012.

S. Kulkarni and S. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

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

D. Williams. *Probability with Martingales.* Cambridge University Press, 1991.

Appendix A. Proof details

Lemma 6 *Pick any arrival time t. Then*

$$\Pr\left(\exists\, t' \geq t \text{ before next arrival with } \widehat{Y}^L_{t'} = \widehat{Y}^R_{t'} = 1 \mid t \text{ is an arrival}, \widehat{Y}^L_{t-1}\widehat{Y}^R_{t-1} = 0\right) \;\geq\; \frac{1}{12}\eta^2.$$

Proof Let's say $\widehat{Y}^L_{t-1}$ and $\widehat{Y}^R_{t-1}$ are the labels of points x_L and x_R, respectively (see Figure 1). Without loss of generality, $|x_L - x| \leq |x_R - x|$. Since X_t is an arrival, it lies in $[x_L, x_R]$ and is equally likely to be on either side of x. So $\Pr(X_t < x) = 1/2$, and we will henceforth condition upon this event.

There are three cases, depending on which of $\widehat{Y}^L_{t-1}$ and $\widehat{Y}^R_{t-1}$ are zero.

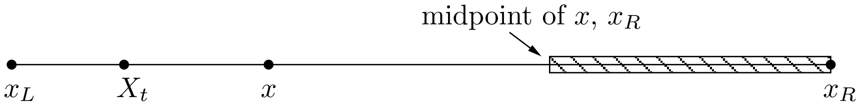


Figure 1: x_L and x_R are the nearest queried neighbors to the left and right of x at time $t-1$. The next point, X_t, is an arrival. In this figure, it is shown to the left of x, which occurs with probability 1/2.

Case 1: $\widehat{Y}^L_{t-1} = 0$ *and* $\widehat{Y}^R_{t-1} = 1$. In this case, X_t will be queried, and with probability η, will get label 1. Remembering that we need $X_t < x$, we have

$$\Pr(\widehat{Y}^L_t = \widehat{Y}^R_t = 1 \mid \text{case 1}) \;\geq\; \frac{1}{2}\eta.$$

Case 2: $\widehat{Y}^L_{t-1} = \widehat{Y}^R_{t-1} = 0$. The previous arrival (before X_t) was either x_L or something closer to x that did not get queried. Since X_t is uniformly distributed, we have

$$\Pr\left(|x - X_t| < \frac{1}{2}|x - x_L| \;\middle|\; X_t \text{ is an arrival}\right) \geq \frac{1}{2}.$$

Let E_o denote the event $|x - X_t| < |x - x_L|/2$. Say $X_{t'}$, $t' > t$, is the next point to fall in $[X_t, x_R]$.

$$\Pr(X_{t'} \text{ is in shaded region} \mid E_o) \;=\; \frac{|x - x_R|/2}{|x - x_R| + |x - X_t|} \;\geq\; \frac{1}{3}.$$

If this happens, $X_{t'}$ is further away from x than is X_t, and hence $X_{t'}$ is not an arrival. Also, both X_t and $X_{t'}$ will have their labels queried, and with probability η^2, both will be 1.

$$\begin{aligned} &\Pr(\exists\ t' \geq t \text{ before next arrival with } \widehat{Y}^L_{t'} = \widehat{Y}^R_{t'} = 1 \mid \text{case 2}) \\ &\quad\geq\; \Pr(X_t < x, E_o, X_{t'} \text{ in shaded region, both labels are 1}) \\ &\quad\geq\; \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot \eta^2 \;=\; \frac{1}{12}\eta^2. \end{aligned}$$

Case 3: $\widehat{Y}^L_{t-1} = 1$ *and* $\widehat{Y}^R_{t-1} = 0$. This is like case 2, except that the label of X_t will probably not be queried. Thus

$$\Pr(\exists\ t' \geq t \text{ before next arrival with } \widehat{Y}^L_{t'} = \widehat{Y}^R_{t'} = 1 \mid \text{case 3}) \;\geq\; \frac{1}{12}\eta^2.$$

■

Lemma 8 *For any $s \leq t$, and any $c > 0$,*

$$\Pr\left(|A(s,t) - H(s,t)| \geq c\sqrt{H(s,t)}\right) \;\leq\; \frac{1}{c^2}.$$

where $H(s,t)$ is the harmonic sum defined in (1).

Proof We have $\mathbb{E}(Z_i) = 1/i$, so $\mathbb{E}(A(s,t)) = H(s,t)$. For $i \neq j$, $\mathbb{E}(Z_i Z_j) = 1/(ij)$. Thus

$$\begin{aligned} \mathbb{E}(A(s,t)^2) &= \mathbb{E}\Big(\big(\sum_{i=s}^{t} Z_i\big)^2 \Big) = \sum_i \mathbb{E}(Z_i^2) + \sum_{i \neq j} \mathbb{E}(Z_i Z_j) \\ &\leq H(s,t) + \sum_{i,j} \frac{1}{ij} = H(s,t) + H(s,t)^2 \end{aligned}$$

and the variance of $A(s,t)$ is $\mathbb{E}(A(s,t)^2) - (\mathbb{E}(A(s,t)))^2 \leq H(s,t)$. By Chebyshev's bound,

$$\Pr\left(|A(s,t) - \mathbb{E}A(s,t)| \geq c\sqrt{H(s,t)} \right) \leq \frac{\mathrm{var}(A(s,t))}{c^2 H(s,t)} \leq \frac{1}{c^2}.$$

■

Lemma 12 *Pick any $0 < \epsilon, \eta < 1$. Let $Z, Z_1, \ldots, Z_k \in \{0,1\}$ be the outcomes of independent coin flips with heads probabilities $\eta, \eta_1, \ldots, \eta_k$, respectively, where $|\eta_i - \eta| \leq \epsilon$. Let M_k be the majority vote over $Z_1, \ldots, Z_k$, breaking ties with a fair coin flip. Define $C(k, \epsilon, \eta)$ to be the supremum of $\Pr(M_k \neq Z)$ over all choices $\eta_1, \ldots, \eta_k \in [\eta - \epsilon, \eta + \epsilon]$. Then*

(a) $C(1, \epsilon, \eta) \leq 2\min(\eta, 1-\eta) + \epsilon$.

(b) If $k > 1$ and either $\eta = 1/2$ or $\epsilon \leq |1 - 2\eta|/4$, then $C(k, \epsilon, \eta) \leq \min(\eta, 1-\eta) + 2/\sqrt{k}$.

Proof Assume without loss of generality that $\eta \leq 1/2$, and let $p = \Pr(M_k = 1)$. Then

$$\Pr(M_k \neq Z) = \Pr(Z=1)\Pr(M_k = 0) + \Pr(Z=0)\Pr(M_k=1) = \eta + p(1-2\eta). \quad (2)$$

For (a), we have $p = \eta_1 \leq \eta + \epsilon$, and we're done.

For (b), if $\eta = 1/2$, then we are done immediately by (2).

Otherwise, the condition on ϵ implies $\eta_i \leq \eta + \epsilon < 1/2$. This means that $p \leq 1/2$. To make the argument carefully, we can first check by means of a coupling argument that p is maximized when all the η_i are set to their maximum value. So assume this; then, since any outcome $(Z_1, \ldots, Z_k)$ with majority vote 1 is at most as likely as the corresponding $(1 - Z_1, \ldots, 1 - Z_k)$ with majority vote 0, it follows that $p \leq 1/2$.

Now, consider the case when $|1 - 2\eta| \leq 4/\sqrt{k}$. Plugging directly into (2) above, we get

$$\Pr(M_k \neq Z) \leq \eta + p(1-2\eta) \leq \eta + \frac{1}{2} \cdot \frac{4}{\sqrt{k}} = \eta + \frac{2}{\sqrt{k}}.$$

On the other hand, if $|1 - 2\eta| > 4/\sqrt{k}$, write $S_k = Z_1 + \cdots + Z_k$, so that $\mathbb{E}S_k = \eta_1 + \cdots + \eta_k \leq \frac{k}{2} - \frac{k}{2}(\frac{1}{2} - \eta)$ and $\mathrm{var}(S_k) = \eta_1(1-\eta_1) + \cdots + \eta_k(1-\eta_k) \leq k/4$. Applying Chebyshev's bound, we have

$$p \leq \Pr\left(S_k \geq \frac{k}{2}\right) \leq \Pr\left(S_k \geq \mathbb{E}S_k + \frac{k}{2}\left(\frac{1}{2} - \eta\right)\right) \leq \frac{\mathrm{var}(S_k)}{\left(\frac{k}{2}\left(\frac{1}{2} - \eta\right)\right)^2} \leq \frac{1}{\left(\frac{1}{2} - \eta\right)\sqrt{k}}.$$

The result then follows from (2). ■

Lemma 16 *Let α, β be probability measures on spaces $\mathcal{A}, \mathcal{B}$, and let $\alpha \times \beta$ denote the product measure. Given a sequence of measurable functions $f_1, f_2, \ldots : \mathcal{A} \times \mathcal{B} \to [0, 1]$, define $G \subset \mathcal{A} \times \mathcal{B}$ by*

$$G = \{(a, b) : \limsup_n f_n(a, b) = 0\}.$$

If $(\alpha \times \beta)(G) = 1$ then $\alpha(\{a : \lim_{n\to\infty} \mathbb{E}_{B\sim\beta} f_n(a, B) = 0\}) = 1$.

Proof For any $a \in \mathcal{A}$, define $G_a = \{b : (a, b) \in G\} \subset \mathcal{B}$, and let $\mathcal{A}_0 \subset \mathcal{A}$ consist of all a such that $\beta(G_a) = 1$. We'll see that $\alpha(\mathcal{A}_0) = 1$. Since

$$\mathbb{E}_{A\sim\alpha} \beta(G_A) = \mathbb{E}_{A\sim\alpha} \mathbb{E}_{B\sim\beta} 1((A, B) \in G) = (\alpha \times \beta)(G) = 1,$$

it follows that $\beta(G_A) = 1$ (and hence $A \in \mathcal{A}_0$) almost surely when A is drawn from α.

Now, consider any $a \in \mathcal{A}_0$. Pick $B \sim \beta$ independently. Almost surely, $(a, B) \in G$ and $f_n(a, B) \to 0$. By dominated convergence, $\mathbb{E}_{B\sim\beta} f_n(a, B) \to 0$. ■