Validation of $k$-Nearest Neighbor Classifiers Using Inclusion and Exclusion

Eric Bax, Lingjie Weng, and Xu Tian

Abstract

This paper presents a series of PAC error bounds for $k$-nearest neighbors classifiers, with $O(n^{-\frac{3}{2}})$ expected range in the difference between error bound and actual error rate, for each integer $r > 0$, where $n$ is the number of in-sample examples. The best previous expected bound range was $O(n^{-\frac{1}{2}})$. The result shows that $k$-nn classifiers, in spite of their famously fractured decision boundaries, come arbitrarily close to having Gaussian-style $O(n^{-\frac{3}{2}})$ expected differences between PAC (probably approximately correct) error bounds and actual expected out-of-sample error rates.

Index Terms

Learning systems, Machine learning, Nearest neighbor, Statistical learning, Supervised learning

I. INTRODUCTION

In machine learning, we begin with a set of labeled in-sample examples. We use those examples to develop a classifier, which maps from inputs to labels. The primary goal is to develop a classifier that performs well on out-of-sample data. This goal is called generalization. A secondary goal is to evaluate how well the classifier will perform on out-of-sample data. This is called validation. We do not want to sacrifice generalization for validation; we want to use all in-sample examples to develop the classifier and perform validation as well.

This paper focuses on validation of $k$-nearest neighbor ($k$-nn) classifiers. A $k$-nn classifier consists of the in-sample examples and a metric to determine distances between inputs. To label an input, a $k$-nn classifier first determines which $k$ in-sample examples have inputs closest to the input to be classified. Then the classifier labels the input with the label shared by a majority of those $k$ nearest neighbors. We assume that $k$ is odd. We also assume binary classification, meaning that there are only two possible labels.

The error bounds used to validate classifiers in this paper are probably approximately correct (PAC) bounds. PAC bounds consist of a range and a bound on the probability that the actual out-of-sample error rate is outside the range. An effective PAC bound has a small range and a small bound failure probability. PAC error bounds include bounds based on Vapnik-Chervonenkis (VC) dimension [1], bounds for concept learning by Valiant [2], compression-based bounds by Littlestone and Warmuth [3], Floyd and Warmuth [4], Blum and Langford [5], and Bax [6], and bounds based on worst likely assignments [7]. Langford [8] gives an overview and comparison of some types of PAC bounds for validation in machine learning.

The type of results in this paper are sometimes called conditional error bounds, because they are conditioned on a specific set of in-sample examples and hence a single classifier. There is also a history of research on the distributions of out-of-sample error rates over nearest neighbor classifiers based on random in-sample data sets, all with examples drawn i.i.d. using the same joint input-output distribution. Cover and Hart [9] prove that the leave-one-out error rate of a classifier is an unbiased estimate of the average error rate over classifiers based on one less example than in the in-sample set. Cover [10] shows that expected error rate converges to at most twice the optimal Bayes error rate as sample size increases. Psaltis, Snapp, and Venkatesh [11] analyze how input dimension affects the rate of this convergence. For more on nearest neighbors, see the books by Devroye, Győrfi, and Lugosi [12], Duda, Hart, and Stork [13], and Hastie, Tibshirani, and Friedman [14].

Prior research on validation of $k$-nn classifiers includes a method with an error bound range of $O(n^{-\frac{1}{2}})$, by Devroye, Győrfi, and Lugosi [12]. (We use “error bound range” to refer to the difference between an upper bound and the actual expected out-of-sample error rate for a classifier.) The idea is to use a holdout set of in-sample examples to bound the error rate of the classifier based on the remaining in-sample examples, called the holdout classifier, then bound the rate of disagreement between the held out classifier and the classifier based on all in-sample examples, called the full classifier. Let $m$ be the number of examples withheld, and assume they are drawn uniformly at random without replacement from the $n$ in-sample examples. The withheld examples can produce a bound on the holdout error rate with an $O(m^{-\frac{1}{2}})$ error bound range, using Hoeffding bounds [15] or any of the other sub-Gaussian bounds on sums of independent variables [16]. Now consider the rate of disagreement between the holdout classifier and the full classifier. The probability that at least one of the $k$ nearest neighbors to a randomly drawn out-of-sample example is in a randomly drawn size- $m$ holdout set is $O(mn^{-\frac{1}{2}})$, and this is a

Yahoo Labs, email: baxhome@yahoo.com
University of California, Irvine, email: weng1@uci.edu
University of California, Irvine, email: tianxu03@gmail.com
necessary condition for disagreement. To minimize the sum of the error bound range and the expected rate of the necessary condition for disagreement, select \( m = n^{\frac{3}{5}} \). Then

\[
O\left(m^{-\frac{1}{3}}\right) + O\left(\frac{m}{n}\right) = O\left(n^{-\frac{4}{15}}\right).
\]  

A more recent method by Bax [17] has an expected error bound range of \( O(n^{-\frac{2}{3}}) \). That method uses a pair of withheld data sets. The sets are used together to bound the error rate of the holdout classifier. Then each set is used to bound the rate of disagreement between holdout and full classifiers caused by withholding the examples in the other set. These bounds have ranges of \( O(m^{-\frac{1}{2}}) \). But we must also consider the rate of disagreement caused when both withheld data sets contain at least one of the \( k \) nearest neighbors to an out-of-sample example. This occurs with probability \( O((\frac{m}{n})^2) \). Selecting \( m = n^{\frac{2}{5}} \) minimizes the sum:

\[
O\left(m^{-\frac{1}{3}}\right) + O\left(\frac{m^2}{n^2}\right) = O\left(n^{-\frac{4}{15}}\right).
\]  

In this paper, we extend those error bounds by using more withheld data sets. We show that by using \( r > 0 \) withheld data sets, we can produce error bounds with \( O(n^{-\frac{2r}{(r+1)^2}}) \) expected bound range. The bounds use withheld data sets to bound rates of disagreement caused by withholding combinations of other withheld data sets, through a formula based on inclusion and exclusion.

As the number of withheld data sets grows, the number of validations for combinations of withheld data sets also grows. So using larger values of \( r \) makes sense only for larger in-sample data set sizes \( n \). By increasing \( r \) slowly as \( n \) increases, we prove that \( k \)-nn classifiers can be validated with expected error bound range \( O(n^{-\frac{1}{2}}+\sqrt{\frac{\ln n}{n}}) \). This result shows that \( k \)-nn classifiers come arbitrarily close to having Gaussian-style \( O(n^{-\frac{1}{2}}) \) error bounds as \( n \to \infty \).

The paper is organized as follows. Section II outlines definitions and notation and reviews how to use concentration inequalities for validation of classifiers. Section III develops error bounds and theoretical results about them. Section IV discusses how to modify the error bounds for use on empirical data and presents test results. Then Section V discusses some potential directions for future work.

II. DEFINITIONS, NOTATION, EMPIRICAL MEANS, AND APPLYING CONCENTRATION INEQUALITIES

Let \( F \) be a set of \( n \) in-sample examples \((x, y)\) drawn i.i.d. from a joint input-output distribution \( D \). Inputs \( x \) are drawn from an arbitrary domain, and outputs \( y \) are drawn from \( \{0, 1\} \) (binary classification). Let \( g^* \) be the \( k \)-nn classifier based on all examples in \( F \).

Select \( r > 0 \) and \( m > 0 \) such that \( rm \leq n - k \). Let validation set \( V \) be a random size-\( rm \) subset of \( F \). Partition \( V \) at random into validation subsets \( V_1, \ldots, V_r \), each of size \( m \). For convenience, let \( R = \{1, \ldots, r\} \).

Let \( g_S \) be the \( k \)-nn classifier based on examples in

\[
(F - V) \cup (\cup_{j \in S} V_j).
\]  

In other words, \( g_S \) is based on all in-sample examples in \( F - V \) and in validation subsets indexed by \( S \). For example, \( g_R \equiv g^* \).

All probabilities are over out-of-sample examples \((x, y)\) drawn i.i.d. from \( D \), unless otherwise specified. For example,

\[
Pr\{g^*(x) \neq y\}
\]  

is the probability that \( g^* \) misclassifies an example drawn at random from \( D \). This is also called the out-of-sample error rate of \( g^* \), and it is the quantity we wish to bound. When a probability or expectation has a subscript, however, the probability or expectation is over a uniform distribution over examples in the set indicated by the subscript. For example,

\[
Pr_V\{g_R(x) \neq y\}
\]  

is the average error rate of a classifier based on \( F - V \) over the validation set \( V \). These averages are called empirical means. We will use them to bound out-of-sample error rates.

As mentioned in the introduction, our goal is to bound the error rate for a classifier based on a specific set of in-sample examples \( F \), not to bound the average error rate over random draws of in-sample data sets. So, unless otherwise specified, probabilities and expectations are conditioned on \( F \). The concentration results we use, such as the Hoeffding Inequality [15], rely on empirical means over i.i.d. examples drawn from \( D \) to bound means over \( D \). The concentration results require the examples used in each empirical mean to be independent of the function for which we bound the mean. For example, the Hoeffding Inequality implies that if \( f(x, y) \) is a function with range of length one, set \( A \) consists of examples drawn i.i.d. from \( D \), and \( A \cap B = \emptyset \), then for \( \delta > 0 \), with probability (over draws of \( B \)) at least \( 1 - \delta \),

\[
|E_A\{f(x, y)\} - E\{f(x, y)\}|B| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2|A|}}.
\]  

(6)
If $A \cap B \neq \emptyset$, this concentration inequality still holds if
\[ \forall (x, y) : f(x, y) \mid B = f(x, y) \mid B - A, \]
since the concentration inequality applies directly to $f(x, y) \mid B - A$, and $E \{ f(x, y) \mid B \} = E \{ f(x, y) \mid B - A \}$.

For example, consider the quantity we wish to bound:
\[ Pr \{ g^*(x) \neq y \mid F \} = E \{ I(g^*(x) \neq y) \mid F \}, \]
where $I()$ is the indicator function: one if the argument is true and zero otherwise. The examples in $V$ are not independent of the function $I(g^*(x) \neq y) \mid F$, because $g^*$ may use the validation examples as nearest neighbors for classification. However, $g_0$ uses only examples in $F - V$ for classification. So the examples in $V$ are independent of $g_0$:
\[ \forall (x, y) : I(g_0(x) \neq y) \mid F = I(g_0(x) \neq y) \mid F - V. \]

As a result, for $\delta > 0$, with probability (over draws of $F$) at least $1 - \delta$,
\[ |E_V \{ I(g_0(x) \neq y) \mid F \} - E \{ I(g_0(x) \neq y) \mid F \}| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2|V|}}, \]
or, equivalently (dropping conditional notation):
\[ |Pr_V \{ g_0(x) \neq y \} - Pr \{ g_0(x) \neq y \}| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2|V|}}. \]

This paper shows how to use inclusion and exclusion to bridge the gap from validation of $g_0$ to validation of $g^*$.

Let $h(x)$ be the distance from $x$ to its $k^{th}$ nearest neighbor in $F - V$. Let condition $c_S(x)$ be true if and only if $\forall i \in S$, $V_i$ has an example closer to $x$ than $h(x)$. Let $w \leq n - rm$ and let $W$ be a size-$w$ random subset of $F - V$. Let $c'_R(x)$ be the condition that each validation subset $V_i$ contains a nearer neighbor to $x$ than the $k^{th}$ nearest neighbor to $x$ in $(F - V) - W$. For $S \subseteq R$, let $V_S$ be the union of validation subsets indexed by $S$.

### III. Results

The main results concern the following PAC error bound:

**Theorem 1.** For any $\delta > 0$ and $\delta_W > 0$,
\[ Pr \{ |p^* - s_V| \geq \epsilon \} \leq \delta + \delta_W, \] where the probability is over draws of $F$,
\[ p^* \equiv Pr \{ g^*(x) \neq y \} \]
is the out-of-sample error rate we wish to bound,
\[ s_V \equiv \sum_{S \subseteq R} \sum_{T \subseteq R - S} (-1)^{|T|} Pr_V \{ c_{S \cup T}(x) \wedge g_S(x) \neq y \} \]
is a sum of empirical means based on terms of an inclusion and exclusion formula, and
\[ \epsilon \equiv r3^{-r-1} \sqrt{\frac{\ln \frac{2\epsilon}{\delta}}{2m} + 2r \left[ Pr_W \{ c'_R(x) \} + \sqrt{\frac{\ln \frac{2\epsilon}{\delta}}{2w}} \right]}. \]

For this bound we show:

**Theorem 2.** For constant $r$,
\[ E \{ \epsilon \} = O(n^{-\frac{1}{2} + \frac{r}{3r+1}}). \]

Then we show:

**Corollary 3.** For a method to increase $r$ as $n$ increases,
\[ E \{ \epsilon \} = O \left( n^{-\frac{1}{2} + \sqrt{\frac{r}{n}}} \right). \]

**Proof of Theorem 1.** Recall that $h(x)$ is the distance from $x$ to its $k^{th}$ nearest neighbor in $F - V$. Define condition $b_S(x)$ to be true if and only if $\forall i \in S$, $V_i$ has an example closer to $x$ than $h(x)$ and $\forall i \notin S$, $V_i$ has no example closer to $x$ than $h(x)$. Then, for each $x$, $b_S(x)$ holds for exactly one $S$, and
\[ b_S(x) \implies g_S(x) = g^*(x), \]
because the $k$ nearest neighbors to $x$ are in $F - V$ or validation subsets indexed by $S$. So

$$Pr \{ g^*(x) \neq y \} = \sum_{S \subseteq R} Pr \{ b_S(x) \land g_S(x) \neq y \}.$$  \hspace{1cm} (19)

Recall that condition $c_S(x)$ holds if and only if $\forall i \in S$, $V_i$ has an example closer to $x$ than $h(x)$. Note that $c_S$ is a looser condition than $b_S$, because $c_S$ does not require that $\forall i \notin S$, $V_i$ has no example closer to $x$ than $h(x)$. So, for each $x$, $c_S(x)$ may hold for multiple $S$. The following lemma states the out-of-sample error rate in terms of probabilities based on $c_S$.

**Lemma 4.**

$$Pr \{ g^*(x) \neq y \} = \sum_{S \subseteq T \subseteq R} \sum_{R - S} (-1)^{|T|} Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}.$$  \hspace{1cm} (20)

**Proof of Lemma.** Note that

$$b_S(x) = c_S(x) \land \neg \left[ \bigvee_{i \in R - S} c_{S \cup \{i\}}(x) \right],$$  \hspace{1cm} (21)

because $b_S$ requires that $\forall i \notin R - S$, $V_i$ has no examples closer to $x$ than $h(x)$. Similarly,

$$(b_S(x) \land g_S(x) \neq y)$$  \hspace{1cm} (22)

$$= (c_S(x) \land g_S(x) \neq y)$$  \hspace{1cm} (23)

$$\land \neg \left[ \bigvee_{i \in R - S} (c_{S \cup \{i\}}(x) \land g_S(x) \neq y) \right].$$  \hspace{1cm} (24)

So

$$Pr \{ b_S(x) \land g_S(x) \neq y \}$$  \hspace{1cm} (25)

$$= Pr \{ c_S(x) \land g_S(x) \neq y \}$$  \hspace{1cm} (26)

$$- Pr \left\{ \bigvee_{i \in R - S} (c_{S \cup \{i\}}(x) \land g_S(x) \neq y) \right\}.$$  \hspace{1cm} (27)

By inclusion and exclusion, the RHS is

$$= Pr \{ c_S(x) \land g_S(x) \neq y \}$$  \hspace{1cm} (28)

$$- \sum_{i \in R - S} Pr \left\{ c_{S \cup \{i\}}(x) \land g_S(x) \neq y \right\}$$  \hspace{1cm} (29)

$$+ \sum_{\{i,j\} \subseteq R - S} Pr \left\{ c_{S \cup \{i,j\}}(x) \land g_S(x) \neq y \right\}$$  \hspace{1cm} (30)

$$\pm \ldots$$  \hspace{1cm} (31)

$$= \sum_{T \subseteq R - S} (-1)^{|T|} Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}.$$  \hspace{1cm} (32)

Substitute this expression for each term in the RHS of Equation (19) to prove the lemma.  

Separate the RHS of Equation (20) into terms for which $S \cup T \subseteq R$:

$$t_V = \sum_{S \subseteq R} \sum_{T \subseteq R - S} (-1)^{|T|} Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}$$  \hspace{1cm} (33)

and terms for which $S \cup T = R$:

$$t_W = \sum_{S \subseteq R} (-1)^{|R - S|} Pr \{ c_R(x) \land g_S(x) \neq y \}.$$  \hspace{1cm} (34)

We will use empirical means over validation sets $V_i$ to bound $t_V$. Then we will bound $t_W$ using an empirical mean over $W$.  

Rewrite $t_V$ by gathering terms for each $i \in R$ that have $i \notin S \cup T$ and multiplying each term by $\frac{|V_i|}{|V_{R-(S \cup T)}|}$, so that the sum of these coefficients for each term is one.

$$t_V = \sum_{i=1}^{r} \left[ \sum_{s \subseteq R - \{i\}} \sum_{T \subseteq (R-\{i\})-S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} Pr\{c_{S \cup T}(x) \land g_S(x) \neq y\} \right].$$  

(35)

Convert the probability to the expectation of an indicator function, and use the linearity of expectation:

$$t_V = \sum_{i=1}^{r} E \left[ \sum_{s \subseteq R - \{i\}} \sum_{T \subseteq (R-\{i\})-S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} I(c_{S \cup T}(x) \land g_S(x) \neq y) \right].$$

(36)

Define

$$f_i(x, y) = \sum_{s \subseteq R - \{i\}} \sum_{T \subseteq (R-\{i\})-S} \frac{|V_i|}{|V_{R-(S \cup T)}|} (-1)^{|T|} I(c_{S \cup T}(x) \land g_S(x) \neq y).$$

(37)

Then

$$t_V = \sum_{i=1}^{r} E\{f_i(x, y)\}.$$  

(38)

Note that the examples in $V_i$ are independent of $f_i(x, y)$:

$$\forall (x, y) : f_i(x, y) | F = f_i(x, y) | F - V_i,$$

(39)

because $i \notin S \cup T$ implies

$$\forall (x, y) : I(c_{S \cup T}(x) \land g_S(x) \neq y) | F = I(c_{S \cup T}(x) \land g_S(x) \neq y) | F - V_i.$$  

(40)

So we can use the empirical mean $E_{V_i}\{f_i(x, y)\}$ to bound $E\{f_i(x, y)\}$ for each $i \in R$.

To apply the Hoeffding Inequality [15], we need to know the length of the range of $f_i(x, y)$. Each term in $f_i(x, y)$ has a range of length at most one, since $|V_i| \leq |V_{R-(S \cup T)}|$. There are as many terms as there ways to partition $R - \{i\}$ into three subsets: $S$, $T$, and $R - (S \cup T)$. So there are $3^{r-1}$ terms. So $f_i(x, y)$ has range length at most $3^{r-1}$.

Recall that $|V_i| = m$. Apply the Hoeffding Inequality with $\frac{2\ln \frac{2\delta}{\alpha}}{m}$ in place of $\delta$. Then

$$\forall i : Pr\left[ |E_{V_i}\{f_i(x, y)\} - E\{f_i(x, y)\}| \geq 3^{r-1} \sqrt{\frac{2\ln \frac{2\delta}{\alpha}}{m}} \right] \leq \frac{\delta}{r},$$

(41)

where the probability is over draws of $F$. Using the sum bound on the union of these probabilities:

$$Pr\left[ \left| \sum_{i=1}^{r} E_{V_i}\{f_i(x, y)\} - \sum_{i=1}^{r} E\{f_i(x, y)\} \right| \geq r3^{r-1} \sqrt{\frac{2\ln \frac{2\delta}{\alpha}}{m}} \right] \leq \delta.$$  

(42)

Note that

$$\sum_{i=1}^{r} E_{V_i}\{f_i(x, y)\} = s_V,$$

as defined in the statement of Theorem [1]. Substitute $s_V$ and Equation [38] into Inequality [42],

$$Pr\left[ |s_V - t_V| \geq r3^{r-1} \sqrt{\frac{2\ln \frac{2\delta}{\alpha}}{m}} \right] \leq \delta,$$

(48)
where the probability is over random draws of $F$. (We can get a similar result by applying the Hoeffding Inequality to each sum of terms that have the same set of validation data, $V_{R-(S\cup T)}$, then applying a union bound. See the appendix for details.)

Now consider $t_W$:

$$
\sum_{S \subseteq R} (-1)^{|R-S|} \Pr \{ c_R(x) \land g_S(x) \neq y \}. \\
(49)
$$

Note that

$$
t_W \in [-2^{r-1} \Pr \{ c_R(x) \}, 2^{r-1} \Pr \{ c_R(x) \}]. \\
(50)
$$

To estimate $\Pr \{ c_R(x) \}$, select a sample size $w$ and select sample $W$ by drawing $w$ examples from $F - V$ uniformly at random without replacement. Let $c_R'(x)$ be the condition that each validation subset $V_i$ contains a nearer neighbor to $x$ than the $k^{th}$ nearest neighbor to $x$ in $(F - V) - W$. Since $(F - V) - W \subseteq F - V$, $c_R(x)$ implies $c_R'(x)$. So

$$
\Pr \{ c_R'(x) \} \geq \Pr \{ c_R(x) \}. \\
(51)
$$

Hence

$$
t_W \in [-2^{r-1} \Pr \{ c_R'(x) \}, 2^{r-1} \Pr \{ c_R'(x) \}]. \\
(52)
$$

Use empirical mean $Pr_W \{ c_R'(x) \}$ to estimate $Pr \{ c_R'(x) \}$. Let

$$
\epsilon_W = \Pr \{ c_R'(x) \} - Pr_W \{ c_R'(x) \}. \\
(53)
$$

Again using the Hoeffding Inequality, for any $\delta_W > 0$,

$$
\Pr \left\{ |\epsilon_W| \geq \sqrt{\frac{\ln \frac{2}{w}}{2w}} \right\} \leq \delta_W. \\
(54)
$$

So

$$
\Pr \left\{ |t_W| \geq 2^r \left[ Pr_W \{ c_R'(x) \} + \sqrt{\frac{\ln \frac{2}{w}}{2w}} \right] \right\} \leq \delta_W. \\
(55)
$$

Combining this with the bound for $t_V$ from Inequality \[48\] the probability (over random draws of $F$) that the absolute value of the difference between the out-of-sample error rate of $g^*$ and the estimate of $t_V$ using empirical means exceeds

$$
\epsilon \equiv r2^{r-1} \sqrt{\frac{\ln \frac{2}{r}}{2m}} + 2^r \left[ Pr_W \{ c_R'(x) \} + \sqrt{\frac{\ln \frac{2}{w}}{2w}} \right] \\
(56)
$$

is at most $\delta + \delta_W$, which completes the proof of Theorem \[1\] ■

Here is a method to compute an upper bound, based on the steps in the proof of Theorem \[1\] (Though Theorem \[1\] concerns a two-sided bound, we present an upper bound here for simplicity and because we compute upper bounds in the test section later in this paper.) The following method returns a valid upper bound on the out-of-sample error rate of a $k$-nearest neighbor classifier with probability at least $1 - \delta - \delta_W$:

resultBound
1) inputs: data set $F$, $r$, $|V_1|, \ldots, |V_r|$, $\delta$, $w$, $\delta_W$
2) sum = 0.0.
3) // Bound $t_V$
4) Randomly partition: $F \rightarrow (F - V, V_1, \ldots, V_r)$.
5) for $i \in \{1, \ldots, r\}$:
   a) range = $|V_i| \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{1}{|V_{R-(S\cup T)}|}$.
   b) values = $\left( \forall (x, y) \in V_i : |V_i| \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} (-1)^{|T|} \frac{1}{|V_{R-(S\cup T)}|} I(c_{S\cup T}(x) \land g_S(x) \neq y) \right)$.
   c) sum = sum + hoeffdingBound(values, range, $\frac{1}{r}$).
6) // Bound $t_W$.
7) Randomly partition: $F - V \rightarrow (F - V - W, W)$.
8) range = 1.
9) values = $\left( \forall (x, y) \in W : c_R'(x) \right)$.
10) sum = sum + $2^{r-1}$ hoeffdingBound(values, range, $\delta_W$).
11) return sum.

hoeffdingBound
1) inputs: values, range, $\delta$
2) return mean(values) + range \( \sqrt{\frac{\ln \frac{1}{2(\text{values})}}{2^{\text{values}}}} \).

(For a lower bound, change the plus signs to minus signs in line 10 of resultBound and line 2 of hoeffdingBound.)

Proof of Theorem 2: Our goal is to bound the expected value of \( \epsilon \), the range for the two-sided bound, from Equation 56. To do that, we prove the following lemma about the expected value of \( \Pr_W \{ c'_R(x) \} \).

Lemma 5.

\[
E \{ \Pr_W \{ c'_R(x) \} \} \leq \left( \frac{(k + r - 1)m}{n - w} \right)^r e^r.
\]

(57)

where the expectation is over drawing \( n \) examples i.i.d. from \( D \) to form \( F \), drawing a random size-\( w \) subset of \( F \) to form \( W \), drawing a random size-\( rm \) subset of \( F - W \) to form \( V \), and randomly partitioning \( V \) into \( r \) size-\( m \) subsets \( V_1, \ldots, V_r \).

Proof of Lemma 5: Define \( c''_R(x) \) to be the condition that the \( k + r - 1 \) nearest neighbors to \( x \) in \( F - W \) include at least \( r \) examples from \( V \). Condition \( c''_R(x) \) is a necessary condition for \( c'_R(x) \), because otherwise the \( k^{th} \) nearest neighbor to \( x \) in \( F - W \) is closer to \( x \) than the nearest neighbor from at least one of \( V_1, \ldots, V_r \). So

\[
E \{ \Pr_W \{ c''_R(x) \} \} \geq E \{ \Pr_W \{ c'_R(x) \} \},
\]

(58)

where both expectations are over the process outlined in the theorem statement. Note that

\[
E \{ \Pr_W \{ c''_R(x) \} \} = E_W \{ E_{(x,y) \in W} \{ E_V \{ I(c''_R(x)) \} \} \}.
\]

(60)

In other words, for \( \Pr_W \{ c''_R(x) \} \), the expectation over the process from the theorem statement is the same as the expectation over drawing \( W \) i.i.d. from \( D \), drawing \( (x, y) \) at random from \( W \), drawing \( F - W \) i.i.d. from \( D \), and drawing \( V \) at random from \( F - W \). But

\[
E_V \{ I(c''_R(x)) \}
\]

(61)

is the same for all \( W \), \( (x, y) \in W \), and \( F - W \). It is the probability that a random subset \( V \) of \( F - W \) contains \( r \) or more examples from a specified subset of \( k + r - 1 \) examples (the nearest neighbors to \( x \) in \( F - W \)). So it is the tail of a hypergeometric distribution:

\[
\sum_{i=r}^{k+r-1} \frac{(k+r-1)!}{i!(n-w-r)!} \left( \frac{n-w}{n-w-r} \right)^r
\]

(62)

Using a hypergeometric tail bound from Chvátal [18], this is

\[
\leq \left( \frac{(k + r - 1)m}{n - w} \right)^r \left( 1 + \frac{1}{m-1} \right)^{(m-1)r}
\]

(63)

\[
\leq \left( \frac{(k + r - 1)m}{n - w} \right)^r \left( \left( 1 + \frac{1}{m-1} \right)^{m-1} \right)^r
\]

(64)

\[
\leq \left( \frac{(k + r - 1)m}{n - w} \right)^r e^r.
\]

(65)

Lemma 5 implies

\[
E \{ \epsilon \} \leq r3^{r-1} \sqrt{\ln \frac{2r}{2m}} + 2^r \left( \frac{(k + r - 1)m}{n - w} \right)^r e^r + \sqrt{\ln \frac{2}{2w}}.
\]

(66)

Let \( w = m \). Let

\[
m = \frac{(n - m)^{\frac{1}{r} + \frac{1}{2}}}{k + r - 1}.
\]

(67)

(In practice, use the nearest integer to the solution for \( m \).) Then

\[
E \{ \epsilon \} \leq (n - m)^{- \frac{r}{ \alpha_1 + \alpha_2}} \left[ r3^{r-1} \sqrt{\frac{1}{2}(k + r - 1) \ln \frac{2r}{2m}} + 2^r \left( e^r + \sqrt{\frac{1}{2}(k + r - 1) \ln \frac{2}{2w}} \right) \right].
\]

(68)

Treating \( r \) as a constant,

\[
E \{ \epsilon \} = O(n^{- \frac{r}{ \alpha_1 + \alpha_2}}),
\]

(69)
which completes the proof of Theorem 2.

de Proof of Corollary 3. Suppose we do not hold $r$ constant, and instead increase $r$ slowly as $n$ increases. For example, let $r = C \sqrt{\ln n}$. Then

$$E \{ \epsilon \} = O \left( n^{- \frac{C \sqrt{\ln n}}{2C \sqrt{\ln n} + 1}} e^{2C \sqrt{\ln n}} \right).$$

(70)

Expand the fraction in the exponent on $n$ and convert the exponent on $e$ to an exponent on $n$:

$$E \{ \epsilon \} = O \left( n^{- \frac{1}{2} + \frac{1}{4C \sqrt{\ln n} + 1}} \right).$$

(71)

Let $C = \frac{1}{\sqrt{8}}$. Then

$$E \{ \epsilon \} = O \left( n^{- \frac{1}{2} + \sqrt{\frac{1}{n}} \right).$$

(72)

IV. TESTS

This section presents test results to show that using $r > 2$ can improve error bounds even for medium-sized data sets. We start with some modifications to make the results from the previous section produce stronger bounds. Then we present test results.

The random variables corresponding to validation examples in $t_v$ tend to have absolute values that are small compared to their ranges: $c_{S \cup T}(x)$ becomes more unlikely as $|S \cup T|$ grows, and $g_S(x) \neq y$ is rare for accurate classifiers, making $I(c_{S \cup T}(x) \land g_S(x) \neq y)$ zero in many cases. So to make the bounds stronger, we use empirical Bernstein bounds in place of Hoeffding bounds for sums of those random variables. Empirical Bernstein bounds were first developed by Audibert [19], [20], [21] and are based on Bernstein bounds [22]. We use the version of empirical Bernstein bounds by Maurer and Pontil [23]. (For a variety of similar bounds, refer to Boucheron, Lugosi, and Massart [16].) Empirical Bernstein bounds are stronger than the standard Hoeffding bounds when the random variables have small standard deviations compared to their ranges. In effect, empirical Bernstein bounds bound the variance of the random variable, then rely on a small variance to produce a strong bound on the mean. (Hoeffding [15] includes a strong bound for low-variance random variables, but the standard version of Hoeffding bounds is based on a worst-case assumption about the variance.)

As $r$ increases, the random variables in $Pr_W \{ c'_R(x) \}$ become increasingly likely to be zeroes. Since these random variables have value either zero or one, we use directly computed binomial tail bounds, as described by Langford [8] and Hoel [24] (page 208). When $Pr \{ c'_R(x) \}$ is near zero, these bounds take advantage of the low variance in $Pr_W \{ c'_R(x) \}$ terms.

To further improve the bounds, we truncate the inclusion and exclusion formulas:

$$Pr \{ b_S(x) \land g_S(x) \neq y \} = \sum_{T \subseteq R - S} (-1)^{|T|} Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}. \quad (73)$$

To truncate, select an even $u \geq 0$, and

$$Pr \{ b_S(x) \land g_S(x) \neq y \} \leq \sum_{T \subseteq R - S, |T| \leq u} (-1)^{|T|} Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}. \quad (74)$$

(For more on truncation of inclusion and exclusion, refer to Linial and Nisan [25].) For a depth parameter $0 \leq d \leq r$, let $u(S) = \max(2\lceil \frac{|S| - |S|}{2} \rceil, 0)$. Then a truncated version of Theorem 4 is

$$Pr \{ g^*(x) \neq y \} \leq \sum_{S \subseteq R} \sum_{T \subseteq R - S, |T| \leq u(S)} (-1)^{|T|} Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}. \quad (75)$$

When we truncate to depth $d$, we estimate only the terms in this truncated formula, ignoring the others. This decreases the range of the random variables that correspond to validation examples, and it decreases the number of terms with $|S \cup T| = |R|$ that we bound based on $Pr_W \{ c'_R(x) \}$. In fact, for $d < r$, there is only one such term: $Pr \{ c_R(x) \land g_R(x) \neq y \}$. So, in $t_w$, the bound on $Pr \{ c'_R(x) \}$ is multiplied by one instead of a coefficient that is exponential in $r$.

Here is pseudocode for a one-sided (upper) bound, incorporating empirical Bernstein bounds, directly computed binomial tail bounds, and truncated inclusion and exclusion:

testBound

1) inputs: data set $F$, $r$, $|V_1|, \ldots, |V_r|$, $\delta$, $w$, $\delta_W$, $d$
2) sum = 0.0.
3) // Bound $t_v$:
4) Randomly partition: $F \rightarrow (F - V, V_1, \ldots, V_r)$.
5) for $i \in \{1, \ldots, r\}$:
   a) range = $|V_i| \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S, |T| \leq \max(2\lceil \frac{|S| - |S|}{2} \rceil, 0)} \frac{1}{|W_{R - \{S \cup T\}}|}$. 


near an example, due to condition tend to be small compared to the remaining terms, because the truncated terms require more validation sets to have neighbors near

\[ V \cap (R - \{i\}) \] reduces the range by removing terms from the double sum and also by reducing the coefficient on

\[ \Pr \left( x \in W \right) \] reduces the impact of this range on the bound. Using empirical Bernstein (vs. Hoeffding) bounds reduces the impact of this range on the bound. Using truncated inclusion

whether the number of negative components in \( x \) increases exponentially with \( r \). The tradeoff is that truncation introduces a bias into the bound. However, the truncated terms

were about 15% for \( k \) and about 12% for \( k = 3 \), making the standard deviations of the estimates of the means over the 100 trials about 0.1%. (So, in the figures, the differences between plotted points are statistically significant.)

we randomly partitioned the examples into \( F - V - W, V_1, \ldots, V_r, W \) with \( m = |V_t| = cn \) (rounded to the nearest integer), and \( |W| = |V_t| \). Then we computed an upper bound on expected out-of-sample error rate using each truncation depth \( 0 \leq d < r \). We recorded differences between bounds and estimated out-of-sample error rates on the 100,000 out-of-sample examples. Out-of-sample error rates were about 15% for \( k = 3 \) and about 12% for \( k = 7 \). Each result is the average of 100 tests, and the standard deviations for differences between bounds and estimated out-of-sample error rates are about 1%, making the standard deviations of the estimates of the means over the 100 trials about 0.1%. (So, in the figures, the differences between plotted points are statistically significant.)

Figures 1 and 2 show results for \( k = 3 \) and \( k = 7 \), respectively. For each \( r \) value, the figures show the curve for the \( d \) value that yields the tightest bound. For \( k = 3 \), the smallest average difference between bounds and estimated out-of-sample error rates (over 100 trials) was 1.3%, achieved with \( r = 3, d = 2 \), and \( m = 6.25\% \) of \( n \). So 3 \cdot 6.25\%, or about 20\%, of the in-sample examples were used for validation (or 25\% if we count examples in \( W \) as well as \( V \)). For \( k = 7 \), the smallest average difference between bounds and estimated out-of-sample error rates is 10.5\%, achieved with \( r = 3, d = 2 \), and \( m = 3.75\% \) of \( n \). As \( k \) increases from 3 to 7, \( Pr \{ c'_{R}(x) \} \) increases; decreasing validation set sizes helps offset the increase. (In Figure 2) the tightest bound for \( r = 1 \) and \( d = 0 \) lies just beyond the left side of the figure: it is 14.5\%, achieved at \( m = 0.6\% \) of \( n \).) For \( k = 3 \) and \( n = 25,000 \) (not shown in figures), the minimum average gap between bound and test error was 6\%, also achieved with \( r = 3 \) and \( d = 2 \), but with \( m = 7.5\% \) of \( n \), indicating that as the number of in-sample examples shrinks, a larger fraction of them are needed for validation.

Choices of \( r \) and \( d \) mediate tradeoffs in bound tightness. From Equation 56 with \( d = r \), \( Pr_{W} \{ c'_{R}(x) \} \) tends to shrink exponentially in \( r \), since \( c'_{R}(x) \) requires all \( r \) validation sets to have neighbors near \( x \). However, the coefficients \( 3r^{3-1} \) and \( 2r \) increase exponentially with \( r \). (There is also an increase proportional to \( \sqrt{\ln r} \), because we bound over more validation sets.) But Equation 56 is an upper bound on the difference between the error bound and the out-of-sample error rate. In practice, we can replace \( 3r^{3-1} \) by the length of the range of each \( f_{i}(x, y) \):

\[
\sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{|V_t|}{|V_t - (S \cup T)|} = \sum_{S \subseteq R - \{i\}} \sum_{T \subseteq (R - \{i\}) - S} \frac{1}{|R - (S \cup T)|}.
\]

Using empirical Bernstein (vs. Hoeffding) bounds reduces the impact of this range on the bound. Using truncated inclusion and exclusion (\( d < r \)) reduces the range by removing terms from the double sum and also by reducing the coefficient on \( Pr_{W} \{ c'_{R}(x) \} \) from \( 2r \) to one. The tradeoff is that truncation introduces a bias into the bound. However, the truncated terms tend to be small compared to the remaining terms, because the truncated terms require more validation sets to have neighbors near an example, due to condition \( c_{S\cup T}(x) \).

V. DISCUSSION

One direction for future research is to improve the bound by averaging bounds over a set of random choices or over all possible choices of \( V_1, \ldots, V_r \). In practice, this should reduce bias. Using all choices may also allow us to compute \( Pr \{ c_{R}(x) \} \)
Fig. 1. Differences between upper bound on out-of-sample error rate and actual error rate over 100,000 out-of-sample examples, averaged over 100 tests. The tightest bound is achieved with $r = 3$, $d = 2$, and $m = 3125$ (which is 6.25% of $n$).

directly instead of estimating it, which would produce a stronger bound for terms with $|S \cup T| = R$. Whether it is possible to efficiently compute a bound based on using all choices is an open question. For a method to solve a similar problem for leave-one-out estimates, refer to Mullin and Sukthankar [26]. For some theory on averaging bounds to form a single bound, refer to Bax [27].

It may be possible to improve on the results in this paper by tightening some bounds used to derive the error bounds. In the random variable corresponding to a validation example, it may not be logically possible for all positive terms ($|T|$ even) to be nonzero while the negative terms ($|T|$ odd) are zero, or vice versa. So the range of these variables may be much less than the equations indicate, and perhaps not even exponential in $r$. Also, it may be possible to improve the bound in Lemma 5 by using the requirement of $c'_R(x)$ that all of $V_1, \ldots, V_r$ contribute nearer neighbors to $x$ than the $k^{th}$ nearest neighbor from $(F - V) - W$. 

Bound - Test Error ($k = 3$ and $n = 50,000$)
Finally, it would be interesting to apply the techniques from this paper to derive error bounds for network classifiers, where the data is a graph annotated with node and edge data, and the goal is to generalize from labels on some nodes to labels for unlabeled nodes, sometimes including nodes yet to be added to the graph. (See Sen, Namata, Bilgic, Getoor, Gallagher, and Eliassi-Rad [28] and Macskassy and Provost [29] for more background on collective classification.) An initial challenge is to adapt the methods in this paper to network settings where the classification rules are local – based only on neighbors or neighbors of neighbors in the graph – and where nodes are drawn i.i.d. In this setting, nodes are similar to examples, and neighborhoods in the graph have a role similar to near neighbor relationships in the \( k \)-nn setting. It will be more challenging to apply the techniques to settings where classification rules are not local or nodes are not drawn i.i.d. For some background on error bounds in such settings, refer to London, Huang, and Getoor [30] and Li, Sonmez, Cataltepe, and Bax [31], [32].
REFERENCES

[1] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[2] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[3] N. Littlestone and M. Warmuth. Relating data compression and learnability, 1986. Unpublished manuscript, University of California Santa Cruz.

[4] S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):1–36, 1995.

[5] A. Blum and J. Langford. PAC-MDL bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*, pages 344–357, 2003.

[6] E. Bax. Nearly uniform validation improves compression-based error bounds. *Journal of Machine Learning Research*, 9:1741–1755, 2008.

[7] E. Bax and A. Callejas. An error bound based on a worst likely assignment. *Journal of Machine Learning Research*, 9:581–613, 2008.

[8] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

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

[10] T. M. Cover. Rates of convergence for nearest neighbors procedures. In B. K. Kinariwala and F. F. Kuo, editors, *Proceedings of the Hawaii International Conference on System Sciences*, pages 413–415. University of Hawaii Press, 1968.

[11] D. Psaltis, R. Snapp, and S. Venkatesh. On the finite sample performance of the nearest neighbor classifier. *IEEE Transactions on Information Theory*, 40(3):264–280, 1994.

[12] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001.

[14] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, 2009.

[15] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 49:196–203, 1954.

[16] V. Vapnik. The nature of statistical learning theory. *Springer, 1995.*

[17] A. Blum and J. Langford. PAC-MDL bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*, pages 639–646, 2003.

[18] S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Combinatorica*, 10(4):349–365, 1990.

APPENDIX

In the main text, we do \( r \) separate validations, one for each validation subset \( V_i \), then we use a sum bound on a probability of a union to form a bound on \( |S_V - t_V| \). In this appendix, we show that, alternatively, we may do a separate validation for each combination of validation subsets and get a similar result. Recall (from Equality \( \text{(80)} \)) that

\[
t_V = \sum_{S \subseteq R} \sum_{T \subseteq R-S} (-1)^{|T|} \Pr \{ c_{S \cup T}(x) \land g_S(x) \neq y \}.
\]

(78)

Let \( A = S \cup T \), and rewrite \( t_V \) as a sum over \( A \):

\[
t_V = \sum_{A \subseteq R} \sum_{S \subseteq A} (-1)^{|A-S|} \Pr \{ c_A(x) \land g_S(x) \neq y \}.
\]

(79)

Define

\[
f_A(x, y) = \sum_{S \subseteq A} (-1)^{|A-S|} I(c_A(x) \land g_S(x) \neq y).
\]

(80)

Then

\[
t_V = \sum_{A \subseteq R} E \{ f_A(x, y) \}.
\]

(81)

Since

\[
\forall (x, y) : f_A(x, y) \mid F = f_A(x, y) \mid F - V_{R-A},
\]

(82)
we can apply the Hoeffding Inequality to each $f_A(x, y)$, using an empirical mean over $V_{R-A}$. So,

$$\forall A \subset R: Pr \left\{ \left| E_{V_{R-A}} \{f_A(x, y)\} - E \{f_A(x, y)\} \right| \geq 2^{|A|} \sqrt{\frac{\ln \frac{2}{\delta_A}}{2|V_{R-A}|}} \right\} \leq \delta_A,$$  \hspace{1cm} (83)

where $\delta_A > 0$ for all $A$, and the probability is over random draws of $F$. Using a sum bound for the probability of a union,

$$Pr \left\{ \left| \sum_{A \subset R} E_{V_{R-A}} \{f_A(x, y)\} - \sum_{A \subset R} E \{f_A(x, y)\} \right| \geq \sum_{A \subset R} 2^{|A|} \sqrt{\frac{\ln \frac{2}{\delta_A}}{2|V_{R-A}|}} \right\} \leq \sum_{A \subset R} \delta_A.$$

(84)

The first sum in the absolute value is $s_V$, and the second sum is $t_V$. Define

$$\epsilon_V = \sum_{A \subset R} 2^{|A|} \sqrt{\frac{\ln \frac{2}{\delta_A}}{2|V_{R-A}|}},$$

(85)

and let

$$\delta = \sum_{A \subset R} \delta_A.$$  \hspace{1cm} (86)

Then

$$Pr \{|s_V - t_V| \geq \epsilon_V\} \leq \delta.$$  \hspace{1cm} (87)

Since the $f_A(x, y)$ with different $|A|$ values have different ranges and different-sized validation sets $V_{R-A}$, it is not optimal to set all $\delta_A$ to the same value. To optimize, we could take the partial derivatives of $\epsilon_V$ with respect to each $\delta_A$, set those partial derivatives equal to each other, and solve (numerically) for the optimal $\delta_A$ values under the constraint that they sum to $\delta$. (To simplify this optimization, note that, by symmetry, it is optimal to set all $\delta_A$ with the same $|A|$ to the same value.)

For a straightforward result, let $\delta_j$ be the value of each $\delta_A$ having $|A| = j$. Then

$$\epsilon_V = \sum_{j=0}^{r-1} \binom{r}{j} 2^j \sqrt{\frac{\ln \frac{2}{\delta_j}}{\delta}}.$$  \hspace{1cm} (88)

and

$$\delta = \sum_{j=0}^{r-1} \binom{r}{j} \delta_j.$$  \hspace{1cm} (89)

Set

$$\delta_j = 2 \left( \frac{\delta}{2r \alpha(\delta)} \right)^{r-j},$$

(90)

where

$$\alpha(\delta) = \frac{\delta}{\ln(1 + \frac{2}{\delta})}.$$  \hspace{1cm} (91)

(Note that $\alpha(\delta)$ is close to one, since $z \approx \ln(1 + z)$ is a well-known approximation for small $z$.) Then

$$\epsilon_V = \sum_{j=0}^{r-1} \binom{r}{j} 2^j \sqrt{\frac{\ln \frac{2r \alpha(\delta)}{\delta}}{\delta}}.$$  \hspace{1cm} (92)

By binomial expansion,

$$\sum_{j=0}^{r} \binom{r}{j} 2^j = (1 + 2)^r = 3^r,$$

(93)

so

$$\sum_{j=0}^{r-1} \binom{r}{j} 2^j = 3^r - 2^r.$$  \hspace{1cm} (94)

So

$$\epsilon_V = (3^r - 2^r) \sqrt{\frac{\ln \frac{2r \alpha(\delta)}{\delta}}{\delta}}.$$  \hspace{1cm} (95)
To show that this bound is valid, we need to show that
\[
\sum_{j=0}^{r-1} \binom{r}{j} \delta_j \leq \delta. \tag{96}
\]

We can do this as follows:
\[
\sum_{j=0}^{r-1} \binom{r}{j} \delta_j \tag{97}
\]
\[
= \sum_{j=0}^{r-1} \binom{r}{j} 2 \left( \frac{\delta}{2r\alpha(\delta)} \right)^{r-j} \tag{98}
\]
\[
= 2 \left( -1 + \sum_{j=0}^{r} \binom{r}{j} \left( \frac{\delta}{2r\alpha(\delta)} \right)^{r-j} \right) \tag{99}
\]
\[
\leq 2 \left( -1 + e^{\frac{\delta}{2r\alpha(\delta)}} \right) \tag{100}
\]
\[
= 2 \left( -1 + e^{\ln(1+\frac{\delta}{2})} \right) \tag{102}
\]
\[
= \delta. \tag{103}
\]