We investigate the problem of bounding the zero-one loss risk of the 1-NN binary classifier averaged with respect to the random draw of the training set. Under some assumptions on the data distribution $D$, we prove a bound of the form

$$E[\ell_D(A(S))] \leq 2\ell_D(f^*) + \varepsilon_m$$

(1)

where $A$ denotes the 1-NN algorithm, $S$ the training set of size $m$, $\ell_D(f^*)$ is the Bayes risk, and $\varepsilon_m$ is a quantity that vanishes for $m \to \infty$. Note that we are able to compare $E[\ell_D(A(S))]$ directly to the Bayes risk, showing that 1-NN is—in some sense—a powerful learning algorithm.

Recall that in binary classification we denote the joint distribution of $(X, Y)$ with the pair $(D_X, \eta)$, where $D_X$ is the marginal of $D$ on $X$ and $\eta(x) = P(Y = 1 \mid X = x)$. Fix $m$ and let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a training set of size $m$. we define the map $\pi(S, \cdot) : \mathbb{R}^d \to \{1, \ldots, m\}$ by

$$\pi(S, x) = \arg\min_{t=1,\ldots,m} \|x - x_t\|.$$

If there is more than one point $x_t$ achieving the minimum in the above expression, then $\pi(S, x)$ selects one of them using any deterministic tie-breaking rule; our analysis does not depend on the specific rule being used. The 1-NN classifier $h_S = A(S)$ is defined by $h_S(x) = y_{\pi(S,x)}$.

From now on, the training set $S$ is a sample $\{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ drawn i.i.d. from $D$. The expected risk is defined by

$$E[\ell_D(A(S))] = P(Y_{\pi(S,X)} \neq Y)$$

Where probabilities and expectations are understood with respect to the random draw of training set $S$ and of the example $(X, Y)$ with respect to which the risk of $A(S)$ is computed.

We now state a crucial lemma.

**Lemma 1.** The expected risk of the 1-NN classifier can be written as follows

$$E[\ell_D(h_S)] = E[\eta(X_{\pi(S,X)}) (1 - \eta(X))] + E\left[(1 - \eta(X_{\pi(S,X)})) \eta(X)\right]$$

To proceed with the analysis, we now need some assumptions on $D_X$ and $\eta$. First, all data points $X$ drawn from $D_X$ satisfy $\max_i |X_i| \leq 1$ with probability one. In other words, $X \in [-1,1]^d$ with probability 1. Let $X \equiv [-1,1]^d$ the subsets of data points with this property. Second we assume that $\eta$ is Lipschitz on $X$ with constant $c > 0$. We can thus write

$$\eta(x') \leq \eta(x) + c \|x - x'\|$$

(2)

$$1 - \eta(x') \leq 1 - \eta(x) + c \|x - x'\|$$

(3)
Figure 1: Bidimensional example of the construction used in the analysis of 1-NN. Left pane: \( \mathbf{X} \) (the red point) is in the same square \( C_i \) as its closest training point \( \mathbf{X}_{\pi(S,X)} \). Hence, \( \| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \| \) is bounded by the length of the diagonal of this square. Right pane: here there are no training points in the square where \( \mathbf{X} \) lies. Hence, \( \| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \| \) can only be bounded by the length of the entire data space (the large square).

Using (2) and (3), for all \( x, x' \in \mathcal{X} \) we have

\[
\eta(x)(1 - \eta(x')) + (1 - \eta(x))\eta(x') \\
\leq \eta(x)(1 - \eta(x)) + \eta(x)c \| x - x' \| + (1 - \eta(x))\eta(x) + (1 - \eta(x))c \| x - x' \| \\
= 2\eta(x)(1 - \eta(x)) + c \| x - x' \| \\
\leq 2 \min \{ \eta(x), 1 - \eta(x) \} + c \| x - x' \|
\]

where the last inequality holds because \( z(1 - z) \leq \min\{z, 1 - z\} \) for all \( z \in [0, 1] \). Therefore

\[
E[\ell_D(h_S)] \leq 2E \left[ \min \{ \eta(X), 1 - \eta(X) \} \right] + cE[\| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \|]
\]

Recalling that the Bayes risk for the zero-one loss is \( \ell_D(f^*) = E \left[ \min \{ \eta(X), 1 - \eta(X) \} \right] \) we have

\[
E[\ell_D(h_S)] \leq 2 \ell_D(f^*) + cE[\| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \|]
\]

In order to bound the term containing the expected value of \( \| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \| \) we partition the data space \( \mathcal{X} \) in \( d \)-dimensional hypercubes with side \( \varepsilon > 0 \), see Figure 1 for a bidimensional example. Let \( C_1, \ldots, C_r \) the resulting hypercubes. We can now bound \( \| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \| \) using a case analysis. Assume first that \( \mathbf{X} \) belongs to a \( C_i \) in which there is at least a training point \( \mathbf{X}_1 \). Then \( \| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \| \) is at most the length of the diagonal of the hypercube, which is \( \varepsilon \sqrt{d} \), see the left pane in Figure 1. Now assume that \( \mathbf{X} \) belongs to a \( C_i \) in which there are no training points. Then \( \| \mathbf{X} - \mathbf{X}_{\pi(S,X)} \| \) is only bounded by the length of the diagonal of \( \mathcal{X} \), which is \( 2\sqrt{d} \), see the
right pane in Figure 1. Hence, we may write

$$
\mathbb{E}
\left[
\|X - X_{\pi(S,X)}\|
\right]
\leq
\varepsilon \sqrt{d} \sum_{i=1}^{r} \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{X \in C_i\}
+ 2\sqrt{d} \sum_{i=1}^{r} \mathbb{I}\{C_i \cap S = \emptyset\} \mathbb{I}\{X \in C_i\}
$$

$$
= \varepsilon \sqrt{d} \mathbb{E}
\left[
\sum_{i=1}^{r} \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{X \in C_i\}
\right]
+ 2\sqrt{d} \sum_{i=1}^{r} \mathbb{E}\left[\mathbb{I}\{C_i \cap S = \emptyset\} \mathbb{I}\{X \in C_i\}\right]
$$

where in the last step we used linearity of the expected value. Now observe that, for all \(S\) and \(X\),

$$
\sum_{i=1}^{r} \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{X \in C_i\} \in \{0,1\}
$$

because \(X \in C_i\) for only one \(i = 1, \ldots, d\). Therefore,

$$
\mathbb{E}
\left[
\sum_{i=1}^{r} \mathbb{I}\{C_i \cap S \neq \emptyset\} \mathbb{I}\{X \in C_i\}\right]
\leq 1
$$

To bound the remaining term, we use the independence between \(X\) and the training set \(S\),

$$
\mathbb{E}\left[\mathbb{I}\{C_i \cap S = \emptyset\} \mathbb{I}\{X \in C_i\}\right]
= \mathbb{E}\left[\mathbb{I}\{C_i \cap S = \emptyset\}\right] \mathbb{E}\left[\mathbb{I}\{X \in C_i\}\right]
= \mathbb{P}(C_i \cap S = \emptyset) \mathbb{P}(X \in C_i)
$$

Since \(S\) contains \(m\) data points independently drawn, for a generic data point \(X'\) we have that

$$
\mathbb{P}(C_i \cap S = \emptyset) = (1 - \mathbb{P}(X' \in C_i))^m \leq \exp(-m \mathbb{P}(X' \in C_i))
$$

where in the last step we used the inequality \((1 - p)^m \leq e^{-pm}\). Setting \(p_i = \mathbb{P}(X' \in C_i)\) we have

$$
\mathbb{E}
\left[
\left.\|X - X_{\pi(S,X)}\|\right|\right]
\leq \varepsilon \sqrt{d} + \left(2\sqrt{d}\right) \sum_{i=1}^{r} e^{-p_ip_i}
\leq \varepsilon \sqrt{d} + \left(2\sqrt{d}\right) \sum_{i=1}^{r} \max_{0 \leq p \leq 1} e^{-pm} p
\leq \varepsilon \sqrt{d} + \left(2\sqrt{d}\right) r \max_{0 \leq p \leq 1} e^{-pm} p .
$$

The concave function \(g(p) = e^{-pm} p\) is maximized for \(p = \frac{1}{m}\). Therefore,

$$
\mathbb{E}
\left[
\left.\|X - X_{\pi(S,X)}\|\right|\right]
\leq \varepsilon \sqrt{d} + \left(2\sqrt{d}\right) \frac{r}{em} = \sqrt{d} \left(\varepsilon + \frac{2}{em} \left(\frac{2}{\varepsilon}\right)^d\right)
$$

where we used the fact that the number \(r\) of hypercubes is equal to \((\frac{2}{\varepsilon})^d\). Putting everything together we find that

$$
\mathbb{E}[\ell_D(h_S)] \leq 2 \ell_D(f^*) + c \sqrt{d} \left(\varepsilon + \frac{2}{em} \left(\frac{2}{\varepsilon}\right)^d\right)
$$

Since this holds for all \(0 < \varepsilon < 1\), we can set \(\varepsilon = 2^{m^{-1/(d+1)}}\). This gives

$$
\varepsilon + \frac{2}{em} \left(\frac{2}{\varepsilon}\right)^d = 2m^{-1/(d+1)} + \frac{2^{d+1} - d - m^{d/(d+1)}}{em}
= 2m^{-1/(d+1)} \left(1 + \frac{1}{e}\right) \leq 4m^{-1/(d+1)} .\quad (4)
$$
Substituting this bound in (4), we finally obtain
\[ E[\ell_D(h_S)] t \leq 2 \ell_D(f^*) + c4m^{-1/(d+1)}\sqrt{d}. \]

Note that for \( m \to \infty \), \( \ell_D(f^*) \leq E[\ell_D(h_S)] \leq 2 \ell_D(f^*) \). Namely, the asymptotic risk of 1-NN lies between the Bayes risk and twice the Bayes risk.