A. Proofs from Section 3

A.1. Proofs for Constant k

Proof. (Of Theorem 3.1) To show convergence in probability, we need to show that for all $\epsilon, \delta > 0$, there exists an $n(\epsilon, \delta)$ such that $\Pr(\rho(A_k(S_n, \cdot), x) \geq \epsilon) \leq \delta$ for $n \geq n_0(\epsilon, \delta)$.

The proof will again proceed in two stages. First, we show in Lemma A.1 that if the conditions in the statement of Theorem 3.1 hold, then there exists some $n(\epsilon, \delta)$ such that for $n \geq n(\epsilon, \delta)$, with probability at least $1 - \delta$, there exists two points x_+ and x_- in $B(x, \epsilon)$ such that (a) all k nearest neighbors of x_+ have label 1, (b) all k nearest neighbors of x_- have label 0, and (c) $x_+ \neq x_-$.

Next we show that if the event stated above happens, then $\rho(A_k(S_n, \cdot), x) \leq \epsilon$. This is because $A_k(S_n, x_+) = 1$ and $A_k(S_n, x_-) = 0$. No matter what $A_k(S_n, x)$ is, we can always find a point x' that lies in $\{x_+, x_-\} \subset B(x, \epsilon)$ such that the prediction at x' is different from $A_k(S_n, x)$. $\square$

Lemma A.1. *If the conditions in the statement of Theorem 3.1 hold, then there exists some $n(\epsilon, \delta)$ such that for $n \geq n(\epsilon, \delta)$, with probability at least $1 - \delta$, there are two points x_+ and x_- in $B(x, \epsilon)$ such that (a) all k nearest neighbors of x_+ have label 1, (b) all k nearest neighbors of x_- have label 0, and (c) $x_+ \neq x_-$.*

Proof. (Of Lemma A.1) The proof consists of two major components. First, for large enough n, with high probability there are many disjoint balls in the neighborhood of x such that each ball contains at least k points in S_n. Second, with high probability among these balls, there exists a ball such that the k neareast neighbors of its center all have label 1. Similarly, there exists a ball such that the k nearest neighbor of its center all have label 0.

Since μ is absolutely continuous with respect to Lebesgue measure in the neighborbood of x and η is continuous, then for any $m \in \mathbb{Z}+$, we can always find m balls $B(x_1, r_1), \cdots, B(x_m, r_m)$ such that (a) all m balls are disjoint, and (b) for all $i \in \{1, \cdots, m\}$, we have $x_i \in B(x, \epsilon)$, $\mu(B(x_i, r_i)) > 0$ and $\eta(x) \in (0, 1)$ for $x \in B(x_i, r_i)$. For simplicity, we use B_i to denote $B(x_i, r_i)$ and $c_i(n)$ to denote the number of points in $B_i \bigcap S_n$. Also, let $\mu_{\min} = \min_{i \in \{1, \cdots, m\}} \mu(B_i)$. Then by Hoeffding's inequality, for each ball B_i and for any $n > \frac{k+1}{\mu_{\min}}$,

$$\Pr[c_i(n) < k] \leq \exp(-2n\mu_{\min}^2/(k+1)^2),$$

where the randomness comes from drawing sample S_n. Then taking the union bound over all m balls, we have

$$\begin{aligned} &\Pr[\exists i \in \{1, \cdots, m\} \text{ such that } c_i(n) < k] \\ &\qquad \leq m \exp(-2n\mu_{\min}^2/(k+1)^2), \end{aligned} \tag{2}$$

which implies that when $n > \max\left(\frac{k+1}{\mu_{\min}}, \frac{[\log m - \log(\delta/3)](k+1)^2}{\mu_{\min}^2}\right)$, with probability at least $1 - \delta/3$, each of $B_1, \cdots, B_m$ contains at least k points in S_n.

An important consequence of the above result is that with probability at least $1 - \delta/3$, the set of k nearest neighbors of each center x_i of B_i is completely different from another center x_j's, so the labels of x_i's k nearest neighbors are independent of the labels of x_j's k nearest neighbors.

Now let $\eta_{\min,+} = \min_{x \in B_1 \bigcup \cdots \bigcup B_m} \eta(x)$ and $\eta_{\min,-} = \min_{x \in B_1 \bigcup \cdots \bigcup B_m} (1 - \eta(x))$. Both $\eta_{\min,+}$ and $\eta_{\min,-}$ are greater than 0 by the construction requirements of $B_1, \cdots, B_m$. For any x_i,

$$\Pr[x_i\text{'s } k \text{ nearest neighbors all have label } 1] \geq \eta_{\min,+}^k$$

Then,

$$\begin{aligned} &\Pr[\exists i \in \{1, \cdots, m\} \\ &\quad \text{s.t. } x_i\text{'s } k \text{ nearest neighbor all have label } 1] \\ &\quad \geq 1 - (1 - \eta_{\min,+}^k)^m, \end{aligned} \tag{3}$$

which implies when $m \geq \frac{\log \delta/3}{\log(1 - \eta_{\min,+}^k)}$, with probability at least $1 - \delta/3$, there exists an x_i s.t. its k nearest neighbors all have label 1. This x_i is our x_+.

Similarly,

$$\begin{aligned} &\Pr[\exists i \in \{1, \cdots, m\} \\ &\quad \text{s.t. } x_i\text{'s } k \text{ nearest neighbor all have label } 0] \\ &\quad \geq 1 - (1 - \eta_{\min,-}^k)^m, \end{aligned} \tag{4}$$

and when $m \geq \frac{\log \delta/3}{\log(1 - \eta_{\min,-}^k)}$, with probability at least $1 - \delta/3$, there exists an x_i s.t. its k nearest neighbors all have label 0. This x_i is our x_-.

Combining the results above, we show that for

$$n > \max\left(\frac{k+1}{\mu_{\min}}, \frac{[\log m - \log(\delta/3)](k+1)^2}{\mu_{\min}^2}\right),$$

$$m \geq \max\left(\frac{\log \delta/3}{\log(1 - \eta_{\min,+}^k)}, \frac{\log \delta/3}{\log(1 - \eta_{\min,-}^k)}\right),$$

with probability at least $1 - \delta$, the statement in Lemma A.1 is satisfied. $\square$

A.2. Theorem and proof for k-nn robustness lower bound.

Theorem 3.1 shows that k-NN is inherently non-robust in the low k regime if $\eta(x) \in (0, 1)$. On the contrary, k-NN can be robust at x if $\eta(x) \in \{0,1\}$. We define the r-robust (p,Δ)-interior as follows:

$$\begin{aligned}\hat{\mathcal{X}}^+_{r,\Delta,p} &= \{x \in supp(\mu) | \forall x' \in B^o(x,r),\\ &\quad \forall x'' \in B(x', r_p(x')), \eta(x'') \geq 1/2 + \Delta\}\\ \hat{\mathcal{X}}^-_{r,\Delta,p} &= \{x \in supp(\mu) | \forall x' \in B^o(x,r),\\ &\quad \forall x'' \in B(x', r_p(x')), \eta(x'') \leq 1/2 - \Delta\}\end{aligned}$$

The definition is similar to the strict r-robust (p,Δ)-interior in Section 4, except replacing $<$ and $>$ with $\leq$ and $\geq$. Theorem A.2 show that k-NN is robust at radius r in the r-robust $(1/2, p)$-interior with high high probability. Corollary A.3 shows the finite sample rate of the robustness lowerbound.

Theorem A.2. *Let $x \in \mathcal{X} \cap supp(\mu)$ such that (a) μ is absolutely continuous with respect to the Lebesgue measure (b) $\eta(x) \in \{0,1\}$. Then, for fixed k, there exists an n_0 such that for $n \geq n_0$,*

$$\Pr[\rho(A_k(S_n,\cdot),x) \geq r] \geq 1-\delta$$

for all x in $\hat{\mathcal{X}}^+_{r,1/2,p} \bigcup \hat{\mathcal{X}}^-_{r,1/2,p}$ for all $p > 0, \delta > 0$.

In addition, with probability at least $1-\delta$, the astuteness of the k-NN classifier is at least:

$$\mathbb{E}(\mathbf{1}(X \in \hat{\mathcal{X}}^+_{r,1/2,p} \bigcup \hat{\mathcal{X}}^-_{r,1/2,p}))$$

Proof. The k-NN classifier $A_k(S_n,\cdot)$ is robust at radius r at x if for every $x' \in B^o(x,r)$, a) there are k training points in $B(x', r_p(x'))$, and b) more than $\lfloor k/2 \rfloor$ of them have the same label as $A_k(S_n,x)$. Without loss of generality, we look at a point $x \in \hat{\mathcal{X}}^+_{r,1/2,p}$. The second condition is satisfied since $\eta(x) = 1$ for all training points in $B(x', r_p(x'))$ by the definition of $\hat{\mathcal{X}}^+_{r,1/2,p}$.

It remains to check the first condition. Let B be a ball in $\mathbb{R}^d$ and $n(B)$ be the number of training points in B. Lemma 16 of (Chaudhuri and Dasgupta, 2010) suggests that with probability at least $1-\delta$, for all B in $\mathbb{R}^d$,

$$\begin{aligned}&\mu(B)\\ \geq &\frac{k}{n} + \frac{C_o}{n}\left(d\log n + \log\frac{1}{\delta} + \sqrt{k\left(d\log n + \log\frac{1}{\delta}\right)}\right)\end{aligned} \tag{5}$$

implies $n(B) \geq k$, where C_o is a constant term. Let $B = B(x', r_p(x'))$. By the definition of r_p, $\mu(B) \geq p > 0$. Then as $n \to \infty$, Inequality 5 will eventually be satisfied, which implies B contains at least k training points. The first condition is then met.

The astuteness result follows because $A_k(S_n,x) = y = 1$ in $\hat{\mathcal{X}}^+_{r,1/2,p}$ and $A_k(S_n,x) = y = 0$ in $\hat{\mathcal{X}}^-_{r,1/2,p}$ with probability 1. □

Corollary A.3. *For $n \geq \max(10^4, c^4_{d,k,\delta}/[(k+1)^2p^2])$ where*

$$c_{d,k,\delta} = 4(d+1) + \sqrt{16(d+1)^2 + 8(\ln(8/\delta) + k + 1)}$$

, with probability at least $1-2\delta$, $\rho(A_k(S_n,x)) \geq r$ for all x in $\hat{\mathcal{X}}^+_{r,1/2,p} \bigcup \hat{\mathcal{X}}^-_{r,1/2,p}$ and for all $p > 0, \delta > 0$.

In addition, with probability at least $1-2\delta$, the astuteness of the k-NN classifier is at least:

$$\mathbb{E}(\mathbf{1}(X \in \hat{\mathcal{X}}^+_{r,1/2,p} \bigcup \hat{\mathcal{X}}^-_{r,1/2,p}))$$

Proof. Without loss of generality, we look at a point $x \in \hat{\mathcal{X}}^+_{r,1/2,p}$. Let $B = B(x', r_p(x'))$, $J(B) = \mathbb{E}(Y \cdot \mathbf{1}(X \in B))$ and $\hat{J}(B)$ be the empirical estimation of $J(B)$. Notice that $\hat{J}(B)n$ is the number of training points in B, because $\eta(x) = 1$ for all $x \in B$ by the definition of r-robust $(1/2, p)$-interior. It remains to find a threshold n such that for all $n' > n$,

$$\hat{J}(B) \geq (k+1)/n' \tag{6}$$

By Lemma A.5, with probability $1-2\delta$,

$$\hat{J}(B) \geq p - 2\beta_n\sqrt{p} - 2\beta_n^2 \tag{7}$$

for all $B \in \mathbb{R}^d$. □

Therefore it suffices to find a threshold n that satisfies

$$p - 2\beta_n\sqrt{p} - 2\beta_n^2 \geq (k+1)/n, \tag{8}$$

where $\beta_n = \sqrt{(4/n)((d+1)\ln 2n + \ln(8/\delta))}$.

Solving this quadratic inequality yields

$$\beta_n \leq \frac{-\sqrt{p} + \sqrt{3p + (k+1)/n}}{2}, \tag{9}$$

which can be re-written as

$$(8/\sqrt{n})[(d+1)\ln(2n) + \ln(8/\delta) + (k+1)/8] \leq \sqrt{(k+1)p} \tag{10}$$

by substituting the expression for β_n. This inequality does not admit an analytic solution. Nevertheless, we observe that $n^{1/4} \geq \ln(2n)$ for all $n \geq 10^4$. Therefore it suffices to find an $n \geq 10^4$ such that

$$(8/\sqrt{n})[(d+1)n^{1/4} + \ln(8/\delta) + (k+1)/8] \leq \sqrt{(k+1)p}. \tag{11}$$

Let $m = n^{1/4}$. Inequality 11 can be re-written as

$$\sqrt{(k+1)p}m^2 - 8(d+1)m - (8\ln(8/\delta) + (k+1)) \geq 0. \tag{12}$$

Solving this quadratic inequality with respct to m gives

$$m \geq \frac{4(d+1) + \sqrt{16(d+1)^2 + 8(\ln(8/\delta) + k + 1)}}{\sqrt{(k+1)p}}. \tag{13}$$

Letting

$$c_{d,k,\delta} = 4(d+1) + \sqrt{16(d+1)^2 + 8(\ln(8/\delta) + k + 1)}$$

, we find a desired threshold

$$n = \max(10^4, m^4) \geq \max(10^4, c^4_{d,k,\delta}/[(k+1)^2p^2]). \tag{14}$$

The astuteness result follows in a similar way to Theorem A.2.

A.3. Proofs for High k

A.3.1. Robustness of the Bayes Optimal Classifier

Proof. (Of Theorem 3.2) Suppose $x \in \mathcal{X}^+_{r,0,0}$. Then, $g(x) = 1$. Consider any $x' \in B^o(x, r)$; by definition, $\eta(x') > 1/2$, which implies that $g(x') = 1$ as well. Thus, $\rho(g, x) \geq r$. The other case ($x \in \mathcal{X}^-_{r,0,0}$) is symmetric.

Consider an $x \in \mathcal{X}^+_{r,0,0}$ (the other case is symmetric). We just showed that g has robustness radius $\geq r$ at x. Moreover, $p(y = 1 = g(x)|x) = \eta(x)$; therefore, g predicts the correct label at x with probability $\eta(x)$. The theorem follows by integrating over all x in $\mathcal{X}^+_{r,0,0} \cup \mathcal{X}^-_{r,0,0}$. □

A.3.2. Robustness of k-Nearest Neighbor

We begin by stating and proving a more technical version of Theorem 3.3.

Theorem A.4. *For any n and data dimension d, define:*

$$\begin{aligned} a_n &= \frac{C_0}{n}(d \log n + \log(1/\delta)) \\ b_n &= C_0\sqrt{\frac{d \log n + \log(1/\delta)}{n}} \\ \beta_n &= \sqrt{(4/n)((d+1)\ln 2n + \ln(8/\delta))} \end{aligned}$$

where C_0 is the constant in Theorem 15 of (Chaudhuri and Dasgupta, 2010). Now, pick k_n and Δ_n so that $\Delta_n \to 0$ and the following condition is satisfied:

$$\frac{k_n}{n} \geq \frac{2\beta_n + b_n + \sqrt{(2\beta_n + b_n)^2 + 2\Delta_n(2\beta_n^2 + a_n)}}{\Delta_n}$$

and set

$$\begin{aligned} p_n &= \frac{k_n}{n} + \frac{C_0}{n}\Big(d \log n + \log(1/\delta) \\ &\quad + \sqrt{k_n(d \log n + \log(1/\delta)}\Big) \end{aligned}$$

Then, with probability $\geq 1 - 3\delta$, k_n-NN has robustness radius r at all $x \in \mathcal{X}^+_{r,\Delta_n,p_n} \cup \mathcal{X}^-_{r,\Delta_n,p_n}$. In addition, with probability $\geq 1 - \delta$, the astuteness of k_n-NN is at least:

$$\mathbb{E}[\eta(X) \cdot \mathbf{1}(X \in \mathcal{X}^+_{r,\Delta_n,p_n})] + \mathbb{E}(1 - \eta(X)) \cdot \mathbf{1}(X \in \mathcal{X}^-_{r,\Delta_n,p_n})]$$

Before we prove Theorem A.4, we need some definitions and lemmas.

For any Euclidean ball B in $\mathbb{R}^d$, define $J(B) = \mathbb{E}[Y \cdot \mathbf{1}(X \in B)]$ and $\hat{J}(B)$ as the corresponding empirical quantity.

Lemma A.5. *With probability $\geq 1 - 2\delta$, for all balls B in $\mathbb{R}^d$, we have:*

$$|J(B) - \hat{J}(B)| \leq 2\beta_n^2 + 2\beta_n \min(\sqrt{J(B)}, \sqrt{\hat{J}(B)}),$$

where $\beta_n = \sqrt{(4/n)((d+1)\ln 2n + \ln(8/\delta))}$.

Proof. (Of Lemma A.5) Consider the two functions: $h^+_B(x, y) = \mathbf{1}(y = 1, x \in B)$ and $h^-_B(x, y) = \mathbf{1}(y = -1, x \in B)$. From Lemma A.6, both h^+_B and h^-_B are $0/1$ functions with VC dimension at most $d + 1$. Additionally, $J(B) = \mathbb{E}[h^+_B] - \mathbb{E}[h^-_B]$. Applying Theorem 15 of (Chaudhuri and Dasgupta, 2010), along with an union bound gives the lemma. □

Lemma A.6. *For an Euclidean ball B in $\mathbb{R}^d$, define the function $h^+_B : \mathbb{R}^d \times \{-1, +1\} \to \{0, 1\}$ as:*

$$h^+_B(x, y) = 1(y = 1, x \in B)$$

and let $\mathcal{H}_B = \{h^+_B\}$ be the class of all such functions. Then the VC-dimension of $\mathcal{H}_B$ is at most $d + 1$.

Proof. (Of Lemma A.6) Let U be a set of $d + 2$ points in $\mathbb{R}^d$; as the VC dimension of balls in $\mathbb{R}^d$ is $d + 1$, U cannot be shattered by balls in $\mathbb{R}^d$. Let $U_L = \{(x, y)|x \in U\}$ be a labeling of U that cannot be achieved by any ball (with pluses inside and minuses outside); the corresponding $d+1$-dimensional points cannot be labeled accordingly by h^+_B. Since U is an arbitrary set of $d + 2$ points, this implies that any set of $d+2$ points in $\mathbb{R}^d \times \{-1, +1\}$ cannot be shattered by $\mathcal{H}_B$. The lemma follows. □

Lemma A.7. *Let* $\delta_p = \frac{C_0}{n}\Big(d \log n + \log(1/\delta) + \sqrt{k(d \log n + \log(1/\delta)}\Big)$. *Then, with probability $\geq 1 - \delta$, for all x, $\|x - X_{(k+1)}(x)\| \leq r_{k/n+\delta_p}(x)$, and $\mu(B(x, \|x - X_{(k+1)}(x)\|)) \geq \frac{k}{n} - \delta_p$.*

Proof. (Of Lemma A.7) Observe that by definition for any x, r_p is the smallest r such that $\mu(B(x, r_p(x)) \geq p$. The rest of the proof follows from Lemma 16 of (Chaudhuri and Dasgupta, 2010). □

Proof. (Of Theorem A.4)

From Lemma A.7, by uniform convergence of $\hat{\mu}$, with probability $\geq 1 - \delta$, for all x', $\|x' - X^{(k_n)}(x')\| \leq r_{p_n}(x')$ and $\mu(B(x, \|x - X^{(k_n)}(x)\|)) \geq \frac{k_n}{n} - \delta_p$. If $x' \in \mathcal{X}^+_{r,\Delta_n,p_n}$, this implies that for all $\tilde{x} \in B(x', X^{(k_n)}(x'))$, $\eta(\tilde{x}) \geq 1/2 + \Delta$. Therefore, for such an x', $J(B(x', X^{(k_n)}(x'))) \geq$

$(\frac{1}{2}+\Delta_n)\mu(B(x', X^{(k_n)}(x'))) \geq (\frac{1}{2}+\Delta_n)(k_n/n - \delta_p)$. Since for $B(x', X^{(k_n)}(x'))$, $\hat{\mu}(B(x', X^{(k_n)}(x'))) = \frac{k_n}{n}$, $\min(\hat{J}, J) \leq \frac{k}{n}$. Thus we can apply Lemma A.5 to conclude that

$$\hat{J}(B) > J(B) - 2\beta_n^2 - 2\beta_n\sqrt{k_n/n} > \frac{k_n}{2n},$$

which implies that $\hat{Y}(B) = \frac{1}{k_n}\sum_{i=1}^{k_n} Y^{(i)}(x) = \frac{n}{k_n}\hat{J}(B) > \frac{1}{2}$. The first part of the theorem follows.

For the second part, observe that for an $x \in \mathcal{X}^+_{r,\Delta_n,p_n}$, the label Y is equal to $+1$ with probability $\eta(x)$ and for an $x \in \mathcal{X}^-_{r,\Delta_n,p_n}$, the label Y is equal to -1 with probability $1-\eta(x)$. Combining this with the first part completes the proof. $\square$

B. Proofs from Section 4

We begin with a statement of Chernoff Bounds that we use in our calculations.

Theorem B.1. *(Mitzenmacher and Upfal, 2005) Let X_i be a $0/1$ random variable and let $X = \sum_{i=1}^m X_i$. Then,*

$$\Pr(|X - \mathbb{E}[X]| \geq \delta) \leq e^{-m\delta^2/2} + e^{-m\delta^2/3} \leq 2e^{-m\delta^2/3}$$

Lemma B.2. *Suppose we run Algorithm 1 with parameter r. Then, the points marked as red by the algorithm form an r-separated subset of the training set.*

Proof. Let $f(x_i)$ denote the output of Algorithm 2 on x_i. If $(x_i, 1)$ is a Red point, then $f(x_i) = 1 = f(x_j)$ for all $x_j \in B(x, r)$; therefore, $(x_j, -1)$ cannot be marked as Red by the algorithm as $f(x_j) \neq y_j$. The other case, where $(x_i, -1)$ is a Red point is similar. $\square$

Lemma B.3. *Let $x \in \mathcal{X}$ such that Algorithm 1 finds a Red x_i within $B^o(x, \tau)$. Then, Algorithm 1 has robustness radius at least $r - 2\tau$ at x.*

Proof. For all $x' \in B(x, \tau)$, we have:

$$\|x' - x_i\| \leq \|x - x_i\| + \|x - x'\| < 2\tau$$

Since x_i is a Red point, from Lemma B.2, any x_j in training set output by Algorithm 1 with $y_j \neq y_i$ must have the property that $\|x_i - x_j\| > r$. Therefore,

$$\|x' - x_j\| \geq \|x_i - x_j\| - \|x' - x_i\| > r - 2\tau$$

Therefore, Algorithm 1 will assign x' the label y_i. The lemma follows. $\square$

Lemma B.4. *Let B be a ball such that: (a) for all $x \in B$, $\eta(x) > \frac{1}{2} + \Delta$ and (b) $\mu(B) \geq \frac{2C_0}{n}(d\log n + \log(1/\delta))$. Then, with probability $\geq 1-\delta$, all such balls have at least one x_i such that $x_i \in |B \cap X_n|$ and $y_i = 1$.*

Proof. Observe that $J(B) \geq \frac{C_0}{n}(d\log n + \log(1/\delta))$. Applying Theorem 16 of (Chaudhuri and Dasgupta, 2010), this implies that $\hat{J}(B) > 0$, which gives the theorem. $\square$

Lemma B.5. *Fix Δ and δ, and let $k_n = \frac{3\log(2n/\delta)}{\Delta^2}$. Additionally, let*

$$p_n = \frac{k_n}{n} + \frac{C_0}{n}(d\log n + \log(1/\delta) + \sqrt{k_n(d\log n + \log(1/\delta)}),$$

where C_0 is the constant in Theorem 15 of (Chaudhuri and Dasgupta, 2010). Define:

$$\begin{aligned} S_{RED} &= \{(x_i, y_i) \in S_n | x_i \in \mathcal{X}^+_{r,\Delta,p_n} \cup \mathcal{X}^-_{r,\Delta,p}, \\ & \quad y_i = \frac{1}{2}sgn\left(\eta(x_i) - \frac{1}{2}\right) + \frac{1}{2}\} \end{aligned}$$

Then, with probability $\geq 1-\delta$, all $(x_i, y_i) \in S_{RED}$ are marked as Red by Algorithm 1 run with parameters r, Δ and δ.

Proof. Consider a $(x_i, y_i) \in S_{RED}$ such that $x_i \in X_n \cap \mathcal{X}^+_{r,\Delta,p_n}$, and consider any $(x_j, y_j) \in S_n$ such that $x_j \in B(x_i, r)$. From Lemma A.7, for all such x_j, $\|x_j - X^{(k_n)}(x_j)\| \leq r_{p_n}(x_j)$; this means that all k_n-nearest neighbors x'' of such an x_j have $\eta(x'') > \frac{1}{2} + \Delta$.

Therefore, $\mathbb{E}[\sum_{l=1}^{k_n} Y^{(l)}(x_j)] \geq k_n(1/2 + \Delta)$; by Theorem B.1, this means that for a specific x_j, $\Pr(\sum_{l=1}^{k_n} Y^{(l)}(x_j) < 1/2) \leq 2e^{-k_n\Delta^2/3}$, which is $\leq \delta/n$ from our choice of k_n. By an union bound over all such x_j, with probability $\geq 1-\delta$, we see that Algorithm 2 reports the label $g(x_i)$ on all such x_i, which is the same as y_i by the definition of interiors; x_i therefore gets marked as Red. $\square$

Finally, we are ready to prove the main theorem of this section, which is a slightly more technical form of Theorem 4.2.

Theorem B.6. *Fix a Δ_n, and pick k_n and p_n as in Lemma B.5. Suppose we run Algorithm 1 with parameters r, Δ_n and δ. Consider the set:*

$$\begin{aligned} X_R &= \Bigg\{ x \Big| x \in \mathcal{X}^+_{r+\tau,\Delta_n,p_n} \cup \mathcal{X}^-_{r+\tau,\Delta_n,p_n}, \\ & \quad \mu(B(x,\tau)) \geq \frac{2C_0}{n}(d\log n + \log(1/\delta)) \Bigg\}, \end{aligned}$$

where C_0 is the constant in Theorem 15 of (Chaudhuri and Dasgupta, 2010). Then, with probability $\geq 1 - 2\delta$ over the

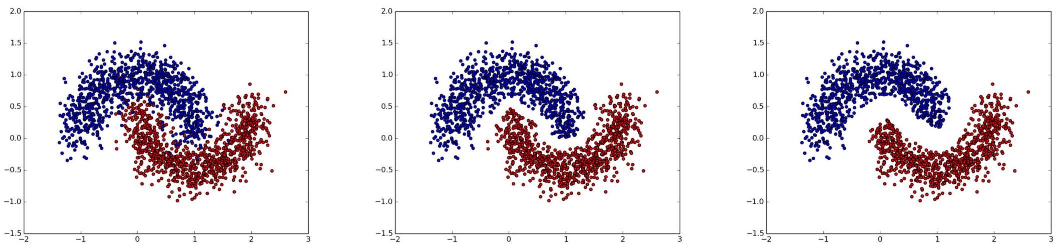

Figure 3. Visualization of the halfmoon dataset. 1) Training sample of size $n = 2000$, 2) subset selected by Robust_1NN with defense radius $r = 0.1$, 3) subset selected by Robust_1NN with defense radius $r = 0.2$.

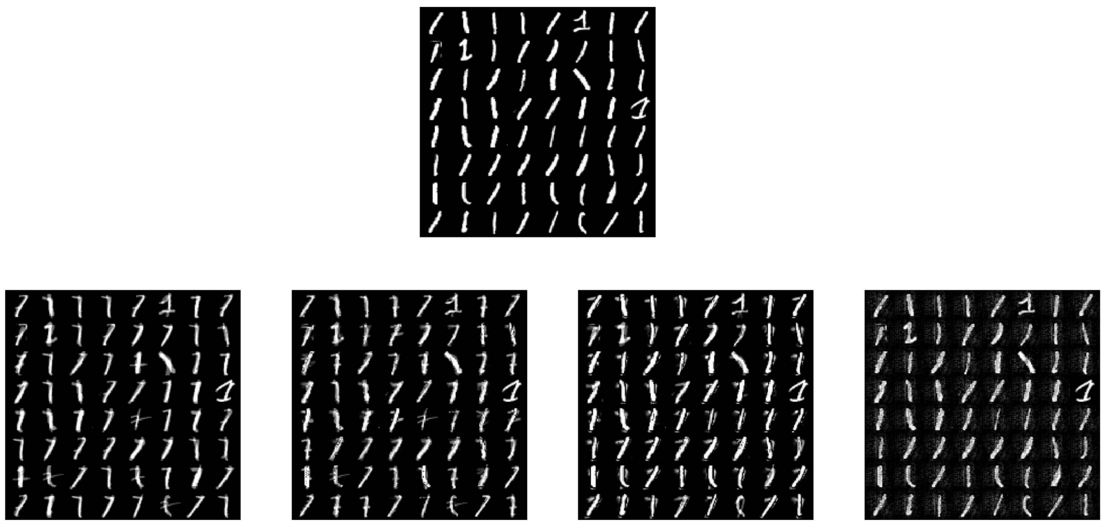

Figure 4. Adversarial examples of MNIST digit 1 images created by different attack methods. *Top row:* clean digit 1 test images. *Bottom row from left to right:* 1) direct attack, 2) white-box kernel attack, 3) black-box kernel attack, 4) black-box neural net substitute attack.

training set, Algorithm 1 has robustness radius $\geq r - 2\tau$ on X_R. Additionally, its astuteness at radius $r - 2\tau$ is at least $\mathbb{E}[\eta(X) \cdot \mathbf{1}(X \in \mathcal{X}^+_{r+\tau,\Delta_n,p_n})] + \mathbb{E}[(1-\eta(X)) \cdot \mathbf{1}(X \in \mathcal{X}^-_{r+\tau,\Delta_n,p_n})]$.

Proof. Due to the condition on $\mu(B(x,\tau))$, from Lemma B.4, with probability $\geq 1 - \delta$, all $x \in X_R$ have the property that there exists a (x_i, y_i) in S_n such that $y_i = g(x_i)$ and $x_i \in B(x,\tau)$. Without loss of generality, suppose that $x \in \mathcal{X}^+_{r+\tau,\Delta_n,p_n}$, so that $\eta(x) > 1/2 + \Delta_n$. Then, from the properties of r-robust interiors, this $x_i \in \mathcal{X}^+_{r,\Delta_n,p_n}$.

From Lemma B.5, with probability $\geq 1 - \delta$, this (x_i, y_i) is marked Red by Algorithm 1 run with parameters r, Δ_n and δ. The theorem now follows from an union bound and Lemma B.3. $\square$

C. Experiment Visualization and Validation

First, we show adversarial examples created by different attacks on the MNIST dataset in order to illustrate characteristics of each attack. Next, we show the subset of training points selected by Algorithm 1 on the halfmoon dataset. The visualization illustrates the intuition behind Algorithm 1 and also validates its implementation. Finally, we validate how effective the black-box subsitute classifiers emulate the target classifier.

C.1. Adversarial Examples Created by Different Attacks

Figure 4 shows adversarial examples created on MNIST digit 1 images with attack radius $r = 3$. First, we observe that the perturbations added by direct attack, white-box kernel attack and black-box kernel attack are clearly targeted: either a faint horizontal stroke or a shadow of digit 7 are added to the original image. The perturbation budget is used on "key" pixels that distinguish digit 1 and digit 7, therefore the attack is effective. On the contrary, black-box attacks with neural nets substitute adds perturbation to a large number of pixels. While such perturbation often fools a neural net classifier, it is not effective against nearest neighbors.

Table 1. An evaluation of the black-box substitute classifier. Each black-box substitute is evaluated by: 1) its accuracy on the its training set, 2) its accuracy on the test set, and 3) the percentage of predictions agreeing with the target classifier on the test set. A combination of high test accuracy and consistency with the original classifier indicates the black-box model emulates the target classifier well.

Abalone

	target f	% training accuracy	% test accuracy	% test f same as f
Kernel	StandardNN	100%	61.3%	72.6%
	RobustNN	100%	62.5%	90.9%
	ATNN	100%	61.4%	73.7%
	ATNN-All	100%	63.5%	73.5%
Neural Nets	StandardNN	69.1%	68.9%	68.6%
	RobustNN	87.2%	64.1%	86.9%
	ATNN	68.8%	68.4%	68.4%
	ATNN-All	66.5%	65.0%	66.6%

Halfmoon

	target f	% training accuracy	% test accuracy	% test same as f
Kernel	StandardNN	95.9%	95.6%	95.5%
	RobustNN	97.7%	94.9%	97.6%
	ATNN	96.4%	95.1%	96.0%
	ATNN-All	97.6%	96.8%	97.3%
Neural Nets	StandardNN	94.5%	94.0%	94.4%
	RobustNN	94.2%	90.5%	94.1%
	ATNN	95.3%	94.2%	95.2%
	ATNN-All	96.9%	96.2%	96.5%

MNIST 1v7

	target f	% training accuracy	% test accuracy	% test same as f
Kernel	StandardNN	100%	98.9%	99.3%
	RobustNN	100%	95.4%	97.6%
	ATNN	100%	98.9%	99.3%
	ATNN-All	100%	98.7%	99.3%
Neural Nets	StandardNN	99.9%	98.9%	99.1%
	RobustNN	99.8%	94.8%	98.7%
	ATNN	100%	98.8%	99.2%
	ATNN-All	99.7%	98.9%	99.3%

Consider a pixel that is dark in most digit 1 and digit 7 training images; adding brightness to this pixel increases the distance between the test image to training images from both classes, therefore may not change the nearest neighbor to the test image.

Figure 4 also illustrates the break-down attack radius of visual similarity. At $r = 3$, the true class of adversarial examples created by effective attacks becomes ambiguous even to humans. Our defense is successful as the Robust_1NN classifiers still have non-trivial classification accuracy at such attack radius. Meanwhile, we should not expect robustness against even larger attack radius since the adversarial examples at $r = 3$ are already close to the boundary of human perception.

C.2. Training Subset Selected by Robust_1NN

Figure 3 shows the training set selected by Robust_1NN on a halfmoon training set of size 2000. On the original training set, we see a noisy region between the two halfmoons where both red and blue points appear. Robust_1NN cleans training points in this region so as to create a gap between the red and blue halfmoons, and the gap width increases with defense radius r.

C.3. Performance of Black-box Attack Substitutes

We validate the black-box substitute training process by checking the substitute's accuracy on its training set, the clean test set and the percentage of predictions agreeing with the target classifier on the clean test set. The results are shown in Table 1. For the halfmoon and MNIST dataset, the substitute classifiers both achieve high accuracy on both the training and test sets, and are also consistent with the target classifier on the test set. The subsitutute classifiers do not emulate the target classifier on the Abalone dataset as close as on the other two datasets due to the high noise level in the Abalone dataset. Nonetheless, the substitute classifier still achieve test time accuracy comparable to the target classifier.