Appendix D Full Proofs for Near-Optimal Evasion

In this appendix, we give proofs for the theorems from Chapter 8. First, we show that the query complexity of K-STEP MULTILINESEARCH is $\mathcal{O}\left(L_\epsilon + \sqrt{L_\epsilon}|\mathbb{W}|\right)$ when $K = \lceil \sqrt{L_\epsilon} \rceil$. Second, we show three lower bounds for different cost functions. Each of the lower bound proofs follows a similar argument: We use classifiers based on the cost-ball and classifiers based on the convex hull of the queries to construct two alternative classifiers with different ϵ-*IMAC*s. This allows us to show results on the minimal number of queries required.

Proof of K-step MultiLineSearch Theorem

To analyze the worst case of K-STEP MULTILINESEARCH (Algorithm 8.3), we analyze the malicious classifier that seeks to maximize the number of queries. It is completely aware of the state of the adversary; i.e., the dimension of the space D, the adversary's goal L_ϵ, the cost function A, the bounds on the cost function C_t^+ and C_t^-, and so forth. In this proof, we refer to the querier as the *adversary*.

Proof of Theorem 8.5 At each iteration of Algorithm 8.3, the adversary chooses some direction, $\mathbf{e}$ not yet eliminated from $\mathbb{W}$. Every direction in $\mathbb{W}$ is feasible (i.e., could yield an ϵ-*IMAC*), and the malicious classifier, by definition, will make this choice as costly as possible. During the K steps of binary search along this direction, regardless of which direction $\mathbf{e}$ is selected or how the malicious classifier responds, the candidate multiplicative gap (see Section 8.1.3) along $\mathbf{e}$ will shrink by an exponent of 2^{-K}; i.e.,

$$\frac{B^-}{B^+} = \left(\frac{C^-}{C^+}\right)^{2^{-K}} \tag{D.1}$$

$$\log\left(G'_{t+1}\right) = \log\left(G_t\right) \cdot 2^{-K} \tag{D.2}$$

The primary decision for the malicious classifier occurs when the adversary begins querying other directions besides $\mathbf{e}$. At iteration t, the malicious classifier has two options:

Case 1 ($t \in \mathbb{C}_1$): Respond with "+" for all remaining directions. Here the bound candidates B^+ and B^- are verified, and thus the new gap is reduced by an exponent of 2^{-K}; however, no directions are eliminated from the search.

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press

Case 2 ($t \in \mathbb{C}_2$): Choose at least one direction to respond with "−". Here since only the value of C^- changes, the malicious classifier can choose to respond to the first K queries so that the gap decreases by a negligible amount (by always responding with "+" during the first K queries along $\mathbf{e}$, the gap only decreases by an exponent of $\left(1 - 2^{-K}\right)$). However, the malicious classifier must choose some number $E_t \geq 1$ of directions that will be eliminated.

By conservatively assuming the gap only decreases in case 1, the total number of queries is bounded for both cases independent of the order in which the malicious classifier applies them.

At the t^{th} iteration, the malicious classifier can either decide to be in case 1 ($t \in \mathbb{C}_1$) or case 2 ($t \in \mathbb{C}_2$). We assume that the gap only decreases in case 1. That is, we define $G_0 = C_0^- / C_0^+$ so that if $t \in \mathbb{C}_1$, then $G_t = G_{t-1}^{2^{-K}}$ whereas if $t \in \mathbb{C}_2$, then $G_t = G_{t-1}$. This assumption yields an upper bound on the algorithm's performance and decouples the analysis of the queries for $\mathbb{C}_1$ and $\mathbb{C}_2$. From it, we derive the following upper bound on the number of case 1 iterations that must occur before our algorithm terminates; simply stated, there must be a total of at least L_ϵ binary search steps made during the case 1 iterations and every case 1 iteration makes exactly K steps. More formally, each case 1 iteration reduces the gap by an exponent of 2^{-K} and our termination condition is $G_T \leq 1 + \epsilon$. Since our algorithm will terminate as soon as the gap $G_T \leq 1 + \epsilon$, iteration T must be a case 1 iteration and $G_{T-1} > 1 + \epsilon$ (otherwise the algorithm would have terminated earlier). From this the total number of iterations must satisfy

$$\log_2\left(G_{T-1}\right) > \log_2\left(1+\epsilon\right)$$

$$\underbrace{\log_2\left(G_0\right) \prod_{i \in C_1 \wedge i < T} 2^{-K}}_{\text{by Equation (D.2)}} > \log_2\left(1+\epsilon\right)$$

$$2^{-\sum_{i \in C_1 \wedge i < T} K} > \frac{\log_2\left(1+\epsilon\right)}{\log_2\left(G_0\right)}$$

$$\sum_{i \in C_1 \wedge i < T} K > \underbrace{\log_2 \frac{\log_2\left(G_0\right)}{\log_2\left(1+\epsilon\right)}}_{= L_\epsilon \text{ by Equation (8.6)}}$$

$$(|C_1| - 1)K < L_\epsilon$$

where the factor $(|\mathbb{C}_1| - 1)$ comes as a result of excluding the last case 1 iteration, T. A similar derivation for $G_T \leq 1 + \epsilon$ yields $|\mathbb{C}_1| \cdot K \geq L_\epsilon$, and the only integer that satisfies both these conditions is

$$|\mathbb{C}_1| = \left\lceil \frac{L_\epsilon}{K} \right\rceil . \tag{D.3}$$

Now, at every case 1 iteration, the adversary makes exactly $K + |\mathbb{W}_t| - 1$ queries where $\mathbb{W}_t$ is the set of feasible directions remaining at the t^{th} iteration. While $\mathbb{W}_t$ is controlled by the malicious classifier, it is bounded by $|\mathbb{W}_t| \leq |\mathbb{W}|$. Using this and the relation from Equation (D.3), we bound the number of queries, Q_1, used in case 1 by

$$\begin{aligned}
Q_1 &= \sum_{t \in C_1} \left(K + |\mathbb{W}_t| - 1 \right) \\
&\leq \sum_{t \in C_1} \left(K + |\mathbb{W}| - 1 \right) \\
&= \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot \left(K + |\mathbb{W}| - 1 \right) \\
&\leq \left(\frac{L_\epsilon}{K} + 1 \right) \cdot K + \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot \left(|\mathbb{W}| - 1 \right) \\
&= L_\epsilon + K + \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot \left(|\mathbb{W}| - 1 \right).
\end{aligned}$$

For each case 2 iteration, the adversary makes exactly $K + E_t$ queries, and each eliminates $E_t \geq 1$ directions; hence, $|\mathbb{W}_{t+1}| = |\mathbb{W}_t| - E_t$. The malicious classifier will always make $E_t = 1$ in every case 2 instance since that maximally limits how much the adversary gains. Nevertheless, since case 2 requires the elimination of at least one direction, the following bound applies: $|\mathbb{C}_2| \leq |\mathbb{W}| - 1$. Moreover, regardless of the choice of E_t, $\sum_{t \in \mathbb{C}_2} E_t \leq |\mathbb{W}| - 1$ since each direction can be eliminated no more than once and at least one direction must remain. Thus,

$$\begin{aligned}
Q_2 &= \sum_{i \in \mathbb{C}_2} \left(K + E_t \right) \\
&\leq |\mathbb{C}_2| \cdot K + |\mathbb{W}| - 1 \\
&\leq \left(|\mathbb{W}| - 1 \right) \left(K + 1 \right).
\end{aligned}$$

The total number of queries used by Algorithm 8.3 is

$$\begin{aligned}
Q = Q_1 + Q_2 &\leq L_\epsilon + K + \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot \left(|\mathbb{W}| - 1 \right) + \left(|\mathbb{W}| - 1 \right) \left(K + 1 \right) \\
&= L_\epsilon + K + \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot |\mathbb{W}| - \left\lceil \frac{L_\epsilon}{K} \right\rceil + K \cdot |\mathbb{W}| - K + |\mathbb{W}| - 1 \\
&= L_\epsilon + \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot |\mathbb{W}| + K \cdot |\mathbb{W}| + |\mathbb{W}| - \left\lceil \frac{L_\epsilon}{K} \right\rceil - 1 \\
&\leq L_\epsilon + \left\lceil \frac{L_\epsilon}{K} \right\rceil \cdot |\mathbb{W}| + K \cdot |\mathbb{W}| + |\mathbb{W}| \\
&= L_\epsilon + \left(\left\lceil \frac{L_\epsilon}{K} \right\rceil + K + 1 \right) |\mathbb{W}|
\end{aligned}$$

Finally, choosing $K = \lceil \sqrt{L_\epsilon} \rceil$ minimizes this expression. Substituting this K into Q's bound and using the bound $L_\epsilon / \lceil \sqrt{L_\epsilon} \rceil \leq \sqrt{L_\epsilon}$, yield

$$Q \leq L_\epsilon + \left(2 \lceil \sqrt{L_\epsilon} \rceil + 1 \right) |\mathbb{W}|$$

so $Q = \mathcal{O}\left(L_\epsilon + \sqrt{L_\epsilon} |\mathbb{W}| \right)$. □

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press

Proof of Lower Bounds

Here we give proofs for the lower bound theorems from Section 8.2.1.2 using the same arguments for the multiplicative and additive cases. Recall that, for these lower bounds, D is the dimension of the space, $A : \Re^D \mapsto \Re_+$ is any positive convex function, and $0 < C_0^+ < C_0^-$ are initial upper and lower bounds on the MAC. We also have that $\hat{\mathcal{F}}^{\text{convex,"+"}} \subseteq \mathcal{F}^{\text{convex,"+"}}$ is the set of classifiers consistent with the constraints on the MAC; i.e., for $f \in \hat{\mathcal{F}}^{\text{convex,"+"}}$ the set $\mathcal{X}_f^+$is convex, $\mathbb{B}^{C_0^+}(A) \subseteq \mathcal{X}_f^+$ and $\mathbb{B}^{C_0^-}(A) \not\subseteq \mathcal{X}_f^+$. As in K-step MultiLineSearch, we again consider a malicious classifier.

Proof of Theorem 8.7 and 8.6 Suppose a query-based algorithm submits $N < D + 1$ membership queries $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \Re^D$ to the classifier. For the algorithm to be ϵ-optimal, these queries must constrain the family of all consistent classifiers, $\hat{\mathcal{F}}^{\text{convex,"+"}}$, to have a common point among their ϵ-$IMAC$ sets. Suppose that the responses to the queries are consistent with the classifier f defined as

$$f(\mathbf{x}) = \begin{cases} +1, & \text{if } A\left(\mathbf{x} - \mathbf{x}^A\right) < C_0^- \\ -1, & \text{otherwise} \end{cases}. \tag{D.4}$$

For this classifier, $\mathcal{X}_f^+$ is convex since A is a convex function, $\mathbb{B}^{C_0^+}(A) \subseteq \mathcal{X}_f^+$ since $C_0^+ < C_0^-$, and $\mathbb{B}^{C_0^-}(A) \not\subseteq \mathcal{X}_f^+$ since $\mathcal{X}_f^+$ is the open C_0^--ball, whereas $\mathbb{B}^{C_0^-}(A)$ is the closed C_0^--ball. Moreover, since $\mathcal{X}_f^+$ is the open C_0^--ball, $\nexists\, \mathbf{x} \in \mathcal{X}_f^-$ such that $A\left(\mathbf{x} - \mathbf{x}^A\right) < C_0^-$. Therefore, $MAC(f, A) = C_0^-$, and any ϵ-optimal points $\mathbf{x}' \in \epsilon$-$IMAC^{(*)}(f, A)$ must satisfy $C_0^- \leq A\left(\mathbf{x}' - \mathbf{x}^A\right) \leq (1 + \epsilon)C_0^-$. Similarly, any η-optimal points $\mathbf{x}' \in \eta$-$IMAC^{(+)}(f, A)$ must satisfy $C_0^- \leq A\left(\mathbf{x}' - \mathbf{x}^A\right) \leq C_0^- + \eta$.

Consider an alternative classifier g that responds identically to f for $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ but has a different convex positive set $\mathcal{X}_g^+$. Without loss of generality, suppose the first $M \leq N$ queries are positive and the remaining are negative. Let $\mathbb{G} = conv\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\right)$; that is, the convex hull of these M positive queries. Now let $\mathcal{X}_g^+$ be the convex hull of $\mathbb{G}$ and the C_0^+-ball of A: $\mathcal{X}_g^+ = conv\left(\mathbb{G} \cup \mathbb{B}^{C_0^+}(A)\right)$. Since $\mathbb{G}$ contains all positive queries and $C_0^+ < C_0^-$, the convex set $\mathcal{X}_g^+$ is consistent with the observed responses, $\mathbb{B}^{C_0^+}(A) \subseteq \mathcal{X}_g^+$ by definition, and $\mathbb{B}^{C_0^-}(A) \not\subseteq \mathcal{X}_g^+$ since the positive queries are all inside the open C_0^--sublevel set. Further, since $M \leq N < D + 1$, $\mathbb{G}$ is contained in a proper linear subspace of $\Re^D$ and hence the interior of $\mathbb{G}$ is empty; i.e., $int(\mathbb{G}) = \emptyset$. Thus, there is always some point from $\mathbb{B}^{C_0^+}(A)$ that is on the boundary of $\mathcal{X}_g^+$; i.e., $\mathbb{B}^{C_0^+}(A) \not\subseteq int(\mathbb{G})$ because $int(\mathbb{G}) = \emptyset$, hence, there must be at least one point from $\mathbb{B}^{C_0^+}(A)$ on the boundary of the convex hull of $\mathbb{B}^{C_0^+}(A)$ and $\mathbb{G}$. Hence, $MAC(g, A) = \inf_{\mathbf{x} \in \mathcal{X}_g^-}\left[A\left(\mathbf{x} - \mathbf{x}^A\right)\right] = C_0^+$. Since the accuracy $\epsilon < \frac{C_0^-}{C_0^+} - 1$, any $\mathbf{x} \in \epsilon$-$IMAC^{(*)}(g, A)$ must have

$$A\left(\mathbf{x} - \mathbf{x}^A\right) \leq (1 + \epsilon)C_0^+ < \frac{C_0^-}{C_0^+} C_0^+ = C_0^-$$

whereas any $\mathbf{y} \in \epsilon\text{-}IMAC^{(*)}(f, A)$ must have $A\left(\mathbf{y} - \mathbf{x}^A\right) \geq C_0^-$. Thus, $\epsilon\text{-}IMAC^{(*)}(f, A) \cap \epsilon\text{-}IMAC^{(*)}(g, A) = \emptyset$, and we have constructed two convex-inducing classifiers f and g, which are both consistent with the query responses with no common $\epsilon\text{-}IMAC^{(*)}$. Similarly, since $\eta < C_0^- - C_0^+$, any $\mathbf{x} \in \eta\text{-}IMAC^{(+)}(g, A)$ must have

$$A\left(\mathbf{x} - \mathbf{x}^A\right) \leq \eta + C_0^+ < C_0^- - C_0^+ + C_0^+ = C_0^-$$

whereas any $\mathbf{y} \in \eta\text{-}IMAC^{(+)}(f, A)$ must have $A\left(\mathbf{y} - \mathbf{x}^A\right) \geq C_0^-$. Thus, $\eta\text{-}IMAC^{(+)}(f, A) \cap \eta\text{-}IMAC^{(+)}(g, A) = \emptyset$, and so the two convex-inducing classifiers f and g also have no common $\eta\text{-}IMAC^{(+)}$.

Suppose instead that a query-based algorithm submits $N < L_\epsilon^{(*)}$ membership queries (or $N < L_\eta^{(+)}$ for the additive case). Recall our definitions: C_0^- is the initial upper bound on the MAC, C_0^+ is the initial lower bound on the MAC, and $G_t^{(*)} = C_t^- / C_t^+$ is the gap between the upper bound and lower bound at iteration t ($G_t^{(+)} = C_t^- - C_t^+$ for the additive case). The malicious classifier f responds with

$$f\left(\mathbf{x}^{(t)}\right) = \begin{cases} +1, & \text{if } A\left(\mathbf{x}^{(t)} - \mathbf{x}^A\right) \leq \sqrt{C_{t-1}^- \cdot C_{t-1}^+} \\ -1, & \text{otherwise} \end{cases} \tag{D.5}$$

(for additive optimality, the condition for the first case is $A\left(\mathbf{x}^{(t)} - \mathbf{x}^A\right) \leq \frac{C_{t-1}^- + C_{t-1}^+}{2}$). When the classifier responds with "+", C_t^+ increases to no more than $\sqrt{C_{t-1}^- \cdot C_{t-1}^+}$ and so $G_t \geq \sqrt{G_{t-1}}$. Similarly when this classifier responds with "−", C_t^- decreases to no less than $\sqrt{C_{t-1}^- \cdot C_{t-1}^+}$ and so again $G_t \geq \sqrt{G_{t-1}}$. Thus, these responses ensure that at each iteration $G_t \geq \sqrt{G_{t-1}}$ (or in the additive case $G_t \geq \frac{G_{t-1}}{2}$) and since the algorithm cannot terminate until $G_N \leq 1 + \epsilon$, it must be the case that $N \geq L_\epsilon^{(*)}$ because of Equation (8.6) (or in the additive case, it must be the case that $N \geq L_\eta^{(+)}$ because of Equation (8.5)). Otherwise, there are still two convex-inducing classifiers with consistent query responses but with no common $\epsilon\text{-}IMAC$. The first classifier's positive set is the smallest cost-ball enclosing all positive queries, while the second classifier's positive set is the largest cost-ball enclosing all positive queries but no negatives. The MAC values for these classifiers differ by more than a factor of $(1 + \epsilon)$ if $N < L_\epsilon^{(*)}$ (or, for the additive case, by a difference of more than η if $N < L_\eta^{(+)}$), so they have no common $\epsilon\text{-}IMAC$. □

Proof of Theorem 8.12

For the proof of Theorem 8.12, we use the orthants (centered at $\mathbf{x}^A$)—i.e., an *orthant* is the D-dimensional generalization of a quadrant in 2-dimensions. There are 2^D orthants in a D-dimensional space. We represent each orthant by its *canonical representation*, which is a vector of D positive or negative ones; i.e., the orthant represented by

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press

$\mathbf{a} = (\pm 1, \pm 1, \ldots, \pm 1)$ contains the point $\mathbf{x}^A + \mathbf{a}$ and is the set of all points $\mathbf{x}$ satisfying

$$x_i \in \begin{cases} [0, +\infty], & \text{if } a_i = +1 \\ [-\infty, 0], & \text{if } a_i = -1 \end{cases}.$$

Now based on Lemma A.2, we give the required proof of Theorem 8.12:

Proof of Theorem 8.12 Suppose a query-based algorithm submits N membership queries $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \Re^D$ to the classifier. Again, for the algorithm to be ϵ-optimal, these queries must constrain all consistent classifiers in the family $\hat{\mathcal{F}}^{\text{convex,"+"}}$ to have a common point among their ϵ-*IMAC* sets. The responses described above are consistent with the classifier f defined as

$$f(\mathbf{x}) = \begin{cases} +1, & \text{if } A_p\left(\mathbf{x} - \mathbf{x}^A\right) < C_0^- \\ -1, & \text{otherwise} \end{cases}.$$

For this classifier, $\mathcal{X}_f^+$ is convex since A_p is a convex function for $p \geq 1$, $\mathbb{B}^{C_0^+}\left(A_p\right) \subseteq \mathcal{X}_f^+$ since $C_0^+ < C_0^-$, and $\mathbb{B}^{C_0^-}\left(A_p\right) \not\subseteq \mathcal{X}_f^+$ since $\mathcal{X}_f^+$ is the open C_0^--ball, whereas $\mathbb{B}^{C_0^-}\left(A_p\right)$ is the closed C_0^--ball. Moreover, since $\mathcal{X}_f^+$ is the open C_0^--ball, $\nexists\, \mathbf{x} \in \mathcal{X}_f^-$ such that $A_p\left(\mathbf{x} - \mathbf{x}^A\right) < C_0^-$; therefore $MAC\left(f, A_p\right) = C_0^-$, and any ϵ-optimal points $\mathbf{x}' \in \epsilon\text{-}IMAC^{(*)}\left(f, A_p\right)$ must satisfy $C_0^- \leq A_p\left(\mathbf{x}' - \mathbf{x}^A\right) \leq (1+\epsilon)C_0^-$.

Now consider an alternative classifier g that responds identically to f for $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ but has a different convex positive set $\mathcal{X}_g^+$. Without loss of generality suppose the first $M \leq N$ queries are positive and the remaining are negative. Consider a set that is the convex hull of the orthants of all M positive queries; that is,

$$\mathbb{G} = conv\left(orth\left(\mathbf{x}^{(1)}\right) \cap \mathcal{X}_f^+, orth\left(\mathbf{x}^{(2)}\right) \cap \mathcal{X}_f^+, \ldots, orth\left(\mathbf{x}^{(M)}\right) \cap \mathcal{X}_f^+\right) \tag{D.6}$$

where $orth\,(\mathbf{x})$ is some orthant that $\mathbf{x}$ lies within relative to the center, $\mathbf{x}^A$ (a data point may lie within more than one orthant, but it is only necessary to select one of the orthants that contains it to cover it). By intersecting each data point's orthant with the set $\mathcal{X}_f^+$ and taking the convex hull of these regions, $\mathbb{G}$ is convex, contains $\mathbf{x}^A$, and is a subset of $\mathcal{X}_f^+$ consistent with all the query responses of f; i.e., each of the M positive queries is in $\mathcal{X}_g^+$ and all the negative queries are in $\mathcal{X}_g^-$. Moreover, $\mathbb{G}$ is a superset of the convex hull of the M positive queries. Thus, the largest enclosed ℓ_p ball within $\mathbb{G}$ is an upper bound on $MAC\left(g, A_p\right)$, so we bound the size of this ℓ_p ball instead.

We now represent each orthant as a vertex in a D-dimensional hypercube graph—the Hamming distance between any pair of orthants is the number of different coordinates in their canonical representations, and two orthants are adjacent in the graph if and only if they have a Hamming distance of one. Using this notion of Hamming distance, we find a K-K-covering of $\mathbb{X}$ of the hypercube. We refer to the orthants used to construct $\mathbb{G}$ in Equation D.6 as *covering orthants* because they cover the M positive queries. The vertexes corresponding to these covering orthants form a covering of the hypercube. Suppose the M covering orthants are sufficient for a K covering but not $K-1$ covering; then there must be at least one vertex not in the covering that has at least a K Hamming distance to every vertex in the K-covering of $\mathbb{X}$. This vertex corresponds to an empty

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press

orthant that differs from all covered orthants in at least K coordinates of their canonical vertexes. Without loss of generality, suppose this uncovered orthant has the canonical vertex of all positive ones that is scaled to $C_0^-\mathbf{1}$. Now, consider the hyperplane with normal vector $\mathbf{w} = \mathbf{1}$ and displacement

$$d = \begin{cases} C_0^-(D-K)^{\frac{p-1}{p}} & \text{if } 1 < p < \infty \\ C_0^-(D-K) & \text{if } p = \infty \end{cases}$$

that specifies the discriminant function $s(\mathbf{x}) = \mathbf{x}^\top\mathbf{w} - d = \sum_{i=1}^{D} x_i - d$. For this hyperplane, the vertex $C_0^-\mathbf{1}$ yields

$$\begin{aligned} s\left(C_0^-\mathbf{1}\right) &= C_0^- D - d \\ &= C_0^- D - \left(C_0^- D - K\right)^{\frac{p-1}{p}} \\ &> C_0^- D - \left(C_0^- D - K\right) \\ &> 0. \end{aligned}$$

Also for any orthant $\mathbf{a}$ with Hamming distance at least K from this uncovered orthant, all points $\mathbf{x} \in orth(\mathbf{a}) \cap \mathcal{X}_f^+$ yield the following valuation of the function s, by definition of the orthant and $\mathcal{X}_f^+$:

$$\begin{aligned} s(\mathbf{x}) &= \sum_{i=1}^{D} x_i - d \\ &= \sum_{\{i \mid a_i=+1\}} \underbrace{x_i}_{\geq 0} + \sum_{\{i \mid a_i=-1\}} \underbrace{x_i}_{\leq 0} - d. \end{aligned}$$

Since all the terms in the second summation are nonpositive, the second sum is at most 0. Thus, maximizing the first summation upper bounds $s(\mathbf{x})$. The summation $\sum_{\{i \mid a_i=+1\}} x_i$ (with the constraint that $\|\mathbf{x}\|_p < C_0^-$, which is necessary for $\mathbf{x}$ to be in $\mathcal{X}_f^+$) has at most $D-K$ terms and is maximized by $x_i = C_0^-(D-K)^{-1/p}$ (or $x_i = C_0^-$ for $p = \infty$) for which the first summation is upper bounded by $C_0^-(D-K)^{\frac{p-1}{p}}$ or $C_0^-(D-K)$ for $p = \infty$; i.e., it is upper bounded by d and so $s(\mathbf{x}) \leq 0$. Thus, this hyperplane separates the scaled vertex $C_0^-\mathbf{1}$ from each set $orth(\mathbf{a}) \cap \mathcal{X}_f^+$ where $\mathbf{a}$ is the canonical representation of any orthant with a Hamming distance of at least K from the positive orthant represented by $\mathbf{1}$. This hyperplane also separates the scaled vertex from $\mathbb{G}$ by the properties of the convex hull. Since the displacement d defined above is greater than 0, by applying Lemma A.3, this separating hyperplane upper bounds the cost of the largest ℓ_p ball enclosed in $\mathbb{G}$ as

$$MAC\left(g, A_p\right) \leq C_0^-(D-K)^{\frac{p-1}{p}} \cdot \|\mathbf{1}\|_{\frac{p}{p-1}}^{-1} = C_0^-\left(\frac{D-K}{D}\right)^{\frac{p-1}{p}}$$

for $1 < p < \infty$ and

$$MAC\left(g, A_p\right) \leq C_0^-(D-K) \cdot \|\mathbf{1}\|_1^{-1} = C_0^- \frac{D-K}{D}$$

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press

for $p = \infty$. Based on this upper bound on the *MAC* of g and the *MAC* of f (i.e., C_0^-), if there is a common ϵ-*IMAC* between these classifiers, it must satisfy

$$(1+\epsilon) \geq \begin{cases} \left(\frac{D}{D-K}\right)^{\frac{p-1}{p}}, & \text{if } 1 < p < \infty \\ \frac{D}{D-K}, & \text{if } p = \infty \end{cases}.$$

Solving for the value of K required to achieve a desired accuracy of $1+\epsilon$ yields

$$K \leq \begin{cases} \frac{(1+\epsilon)^{\frac{p}{p-1}}-1}{(1+\epsilon)^{\frac{p}{p-1}}} D, & \text{if } 1 < p < \infty \\ \frac{\epsilon}{1+\epsilon} D, & \text{if } p = \infty \end{cases},$$

which bounds the size of the K-covering of $\mathbb{X}$ required to achieve the desired multiplicative accuracy ϵ.

For the case $1 < p < \infty$, Lemma A.2 shows there must be

$$M \geq \exp\left\{\ln(2) \cdot D\left(1 - H\left(1 - (1+\epsilon)^{\frac{p}{1-p}}\right)\right)\right\}$$

vertexes of the hypercube in the K-covering of $\mathbb{X}$ to achieve any desired accuracy $0 < \epsilon < 2^{\frac{p-1}{p}} - 1$, for which

$$\delta = \frac{(1+\epsilon)^{\frac{p}{p-1}} - 1}{(1+\epsilon)^{\frac{p}{p-1}}} < \frac{1}{2}$$

to satisfy the condition required by the lemma. Thus, this theorem is applicable for any $\epsilon < 2^{\frac{p-1}{p}} - 1$. For example, for $p = 2$, the theorem is applicable for any $\epsilon < \sqrt{2} - 1$. Moreover, since $H(\delta) < 1$ for any $\delta < \frac{1}{2}$,

$$\alpha_{p,\epsilon} = \exp\left\{\ln(2)\left(1 - H\left(\frac{(1+\epsilon)^{\frac{p}{p-1}} - 1}{(1+\epsilon)^{\frac{p}{p-1}}}\right)\right)\right\} > 1$$

and

$$M \geq \alpha_{p,\epsilon}^D.$$

Similarly for $p = \infty$, applying Lemma A.2 requires $M \geq 2^{D\left(1-H\left(\frac{\epsilon}{1+\epsilon}\right)\right)}$ to achieve any desired accuracy $0 < \epsilon < 1$ (for which $\epsilon/(1+\epsilon) < 1/2$ as required by the lemma). Again, by the properties of entropy, the constant $\alpha_{\infty,\epsilon} = 2^{\left(1-H\left(\frac{\epsilon}{1+\epsilon}\right)\right)} > 1$ for any $0 < \epsilon < 1$ and $M \geq \alpha_{\infty,\epsilon}^D$. □

It is worth noting that the constants $\alpha_{p,\epsilon}$ and $\alpha_{\infty,\epsilon}$ required by Theorem 8.12 can be expressed in a more concise form by expanding the entropy function ($H(\delta) = -\delta \log_2(\delta) - (1-\delta)\log_2(1-\delta)$). For $1 < p < \infty$ the constant is given by

$$\alpha_{p,\epsilon} = 2 \cdot \left(1 - (1+\epsilon)^{\frac{p}{1-p}}\right) \cdot \exp\left(\ln\left(\frac{-1}{1-(1+\epsilon)^{\frac{p}{p-1}}}\right) \cdot (1+\epsilon)^{\frac{p}{1-p}}\right). \tag{D.7}$$

In this form, it is difficult to directly see that $\alpha_{p,\epsilon} > 1$ for $\epsilon < 2^{\frac{p-1}{p}} - 1$, but using the entropy form in the proof above shows that this is indeed the case. Similarly, for $p = \infty$

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press

the more concise form of the constant is given by

$$\alpha_{\infty,\epsilon} = \frac{2}{1+\epsilon}\exp\left(\ln(\epsilon)\cdot\left(\frac{\epsilon}{1+\epsilon}\right)\right). \tag{D.8}$$

Again, as shown in the proof above, $\alpha_{\infty,\epsilon} > 1$ for $\epsilon < 1$.

Proof of Theorem 8.13

For this proof, we build on previous results for K-covering of $\mathbb{X}$ hyperspheres. The proof is based on a covering number result from Wyner (1965) that first appeared in Shannon (1959). This result bounds the minimum number of spherical caps required to cover the surface of a hypersphere and is summarized in Appendix A.2.

Proof of Theorem 8.13 Suppose a query-based algorithm submits $N < D+1$ membership queries $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)} \in \Re^D$ to the classifier. For the algorithm to be ϵ-optimal, these queries must constrain all consistent classifiers in the family $\hat{\mathcal{F}}^{\text{convex,"+"}}$ to have a common point among their ϵ-*IMAC* sets. Suppose that all the responses are consistent with the classifier f defined as

$$f(\mathbf{x}) = \begin{cases} +1, & \text{if } A_2\left(\mathbf{x}-\mathbf{x}^A\right) < C_0^- \\ -1, & \text{otherwise} \end{cases}; \tag{D.9}$$

For this classifier, $\mathcal{X}_f^+$ is convex since A_2 is a convex function, $\mathbb{B}^{C_0^+}(A_2) \subseteq \mathcal{X}_f^+$ since $C_0^+ < C_0^-$, and $\mathbb{B}^{C_0^-}(A_2) \nsubseteq \mathcal{X}_f^+$ since $\mathcal{X}_f^+$ is the open C_0^--ball, whereas $\mathbb{B}^{C_0^-}(A_2)$ is the closed C_0^--ball. Moreover, since $\mathcal{X}_f^+$ is the open C_0^--ball, $\nexists\, \mathbf{x} \in \mathcal{X}_f^-$ such that $A_2\left(\mathbf{x}-\mathbf{x}^A\right) < C_0^-$ therefore $MAC(f, A_2) = C_0^-$, and any ϵ-optimal points $\mathbf{x}' \in \epsilon\text{-}IMAC^{(*)}(f, A_2)$ must satisfy $C_0^- \le A_2\left(\mathbf{x}'-\mathbf{x}^A\right) \le (1+\epsilon)C_0^-$.

Now consider an alternative classifier g that responds identically to f for $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}$ but has a different convex positive set $\mathcal{X}_g^+$. Without loss of generality suppose the first $M \le N$ queries are positive and the remaining are negative. Let $\mathbb{G} = conv\left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}\right)$; that is, the convex hull of the M positive queries. We assume $\mathbf{x}^A \in \mathbb{G}$, since otherwise, the malicious classifier can construct the set $\mathcal{X}_g^+$ as in the proof for Theorems 8.7 and 8.6 and achieve $MAC(f, A_2) = C_0^+$, thereby showing the desired result. Otherwise when $\mathbf{x}^A \in \mathbb{G}$, consider the points $\mathbf{z}^{(i)} = C_0^- \frac{\mathbf{x}^{(i)}}{A_2\left(\mathbf{x}^{(i)}-\mathbf{x}^A\right)}$; i.e., the projection of each of the positive queries onto the surface of the ℓ_2 ball $\mathbb{B}^{C_0^-}(A_2)$. Since each positive query lies along the line between $\mathbf{x}^A$ and its projection $\mathbf{z}^{(i)}$, by convexity and the fact that $\mathbf{x}^A \in \mathbb{G}$, the set $\mathbb{G}$ is a subset of $conv\left(\mathbf{z}^{(1)},\mathbf{z}^{(2)},\ldots,\mathbf{z}^{(M)}\right)$—we refer to this enlarged hull as $\hat{\mathbb{G}}$. These M projected points $\left\{\mathbf{z}^{(i)}\right\}_{i=1}^{M}$ must form a K-covering of $\mathbb{X}$ of the C_0^--hypersphere as the loci of caps of half-angle $\phi_\epsilon^\star = \arccos\left(\frac{1}{1+\epsilon}\right)$. If not, then there exists some point on the surface of this hypersphere that is at least an angle $\phi_\epsilon^\star$ from all $\mathbf{z}^{(i)}$ points, and the resulting $\phi_\epsilon^\star$-cap centered at this uncovered point is not in $\hat{\mathbb{G}}$ (since a cap is defined as the intersection of the hypersphere and a halfspace). Moreover, by definition of the $\phi_\epsilon^\star$-cap, it achieves a minimal ℓ_2 cost of $C_0^- \cos\phi_\epsilon^\star$. Thus, if the adversary fails to achieve a $\phi_\epsilon^\star$-K-covering of $\mathbb{X}$ of the C_0^--hypersphere, the alternative classifier g has $MAC\,(g, A_2) < C_0^- \cos\phi_\epsilon^\star = \frac{C_0^-}{1+\epsilon}$ and any $\mathbf{x} \in \epsilon\text{-}IMAC^{(*)}\,(g, A_2)$ must have

$$A_2\left(\mathbf{x} - \mathbf{x}^A\right) \leq (1+\epsilon)MAC < (1+\epsilon)\frac{C_0^-}{1+\epsilon} = C_0^-$$

whereas any $\mathbf{y} \in \epsilon\text{-}IMAC^{(*)}\,(f, A)$ must have $A\left(\mathbf{y} - \mathbf{x}^A\right) \geq C_0^-$. Thus, there are no common points in the $\epsilon\text{-}IMAC^{(*)}$ sets of these consistent classifiers (i.e., $\epsilon\text{-}IMAC^{(*)}\,(f, A) \cap \epsilon\text{-}IMAC^{(*)}\,(g, A) = \emptyset$), and so the adversary would have failed to ensure ϵ-multiplicative optimality. Thus, an $\phi_\epsilon^\star$-K-covering of $\mathbb{X}$ is necessary for ϵ-multiplicative optimality for ℓ_2 costs. Moreover, from our definition of $\phi_\epsilon^\star$, for any $\epsilon \in (0, \infty)$, $\phi^\star \in \left(0, \frac{\pi}{2}\right)$ and thus, Lemma A.1 is applicable for all ϵ. From Lemma A.1, to achieve an $\phi_\epsilon^\star$-K-covering of $\mathbb{X}$ requires at least

$$M \geq \left(\frac{1}{\sin\phi_\epsilon^\star}\right)^{D-2}$$

queries. Using the trigonometric identity $\sin\left(\arccos(x)\right) = \sqrt{1-x^2}$, and substituting for $\phi_\epsilon^\star$ yields the following bound on the number of queries required for a given multiplicative accuracy ϵ:

$$\begin{aligned} M &\geq \left(\frac{1}{\sin\left(\arccos\left(\frac{1}{1+\epsilon}\right)\right)}\right)^{D-2} \\ &\geq \left(\frac{(1+\epsilon)^2}{(1+\epsilon)^2 - 1}\right)^{\frac{D-2}{2}}. \end{aligned}$$

□

https://doi.org/10.1017/9781107338548.013 Published online by Cambridge University Press