Appendix A Background for Learning and Hyper-Geometry

This appendix contains background material we use throughout this book. In Section A.1, we introduce basic foundations and notation in mathematics and probability necessary for the machine learning concepts built upon in this book. Section A.2 summarizes technical properties of hyperspheres and spherical caps used in the proofs for near-optimal evasion.

A.1 Overview of General Background Topics

We use standard terms and symbols from several fields, as detailed below to avoid ambiguities. We expect that the reader is familiar with the topics in logic, set theory, linear algebra, mathematical optimization, and probability as reviewed in this section. We use $=$ to denote *equality* and $\triangleq$ to denote *defined as*.

Typesetting of Elements, Sets, and Spaces

The typeface Style of a character is used to differentiate between elements of a set, sets, and spaces as follows. Individual objects such as scalars are denoted with italic roman font (e.g., x), and multidimensional vectors are denoted with bold roman font (e.g., $\mathbf{x}$). As discussed below, sets are denoted using blackboard bold characters (e.g., $\mathbb{X}$). However, when referring to the *entire* set or universe that spans a particular kind of object (i.e., a space), we use calligraphic script such as in $\mathcal{X}$ to distinguish it from subsets $\mathbb{X}$ contained within this space.

Sequences and Indexes

In this book, we differentiate between two types of indexing of objects. The first type is used to refer to an element in a sequence of similar objects. This type of index occurs in the superscript following the referenced object and is enclosed within parentheses. For instance, $x^{(1)}$,$x^{(2)}$, …, $x^{(N)}$ are a sequence of objects with $x^{(t)}$ denoting the t^{th} object in the sequence. The second type of index refers to a component of a composite object (e.g., within a multidimensional object) and is indexed by the subscript following the object. For instance x_1,x_2, …, x_D are the components of the vector $\mathbf{x}$. Thus, $\mathbf{x}^{(t)}$ refers to the t^{th} vector in a sequence of vectors, $x_i^{(t)}$ refers to its i^{th} coordinate, and x_i^k is the k^{th} power of x_i.

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press

First-Order Logic

We next describe a formal syntax for expressing logical statements. The notation $a \wedge b$ denotes the logical *conjunction*, a and b; $a \vee b$ denotes the logical *disjunction*, a or b; $\neg a$ is the logical *negation*, not a; $a \Rightarrow b$ is the logical implication defined as $(\neg a) \vee b$; and $a \Leftrightarrow b$ is logical equivalence (i.e., *if and only if*) defined as $(a \Rightarrow b) \wedge (b \Rightarrow a)$. We use the symbols $\forall$ and $\exists$ for universal and existential quantification, respectively. When necessary, predicates can be formalized as functions such as $p(\,\cdot\,)$, which evaluates to true if and only if its argument exhibits the property represented by the predicate. The special *identity predicate* is defined as $\mathrm{I}[a] \Leftrightarrow a$. For convenience, we overload this notation for the indicator function, which instead evaluates to 1 if its argument is true and to 0 otherwise.

Sets

A set, or a collection of objects, is denoted using blackboard bold characters such as $\mathbb{X}$ as noted above and the empty set is given by $\emptyset$. To group a collection of objects as a set we use curly braces such as $\mathbb{X} = \{a, b, c\}$. To specify set membership we use $x \in \mathbb{X}$, and to explicitly enumerate the elements of a set we use the notation $\mathbb{X} = \{x_1, x_2, \ldots, x_N\}$ for a finite set and $\mathbb{X} = \{x_1, x_2, \ldots\}$ for a countably infinite sequence. To qualify the elements within a set, we use the notation $\mathbb{X} = \{x \mid A(x)\}$ to denote a set of objects that satisfy a logical condition represented here by the predicate $A(\,\cdot\,)$. We use $\mathbb{Y} \subseteq \mathbb{X}$ to denote that $\mathbb{Y}$ is a *subset* of $\mathbb{X}$; i.e., $\forall y\,(y \in \mathbb{Y} \Rightarrow y \in \mathbb{X})$. For finite sets, we use the notation $|\mathbb{X}|$ to denote the size of $\mathbb{X}$. We denote the *union* of two sets as $\mathbb{X} \cup \mathbb{Y} \triangleq \{a \mid (a \in \mathbb{X}) \vee (a \in \mathbb{Y})\}$, the *intersection* of two sets as $\mathbb{X} \cap \mathbb{Y} \triangleq \{a \mid (a \in \mathbb{X}) \wedge (a \in \mathbb{Y})\}$, and the *set difference* of two sets as $\mathbb{X} \setminus \mathbb{Y} \triangleq \{a \mid (a \in \mathbb{X}) \wedge (a \notin \mathbb{Y})\}$; i.e., the elements in $\mathbb{X}$ but not in $\mathbb{Y}$. We also use the predicate $\mathrm{I}_{\mathbb{X}}[\,\cdot\,]$ to denote the set indicator function for $\mathbb{X}$; i.e., $\mathrm{I}_{\mathbb{X}}[x] \triangleq \mathrm{I}[x \in \mathbb{X}]$ (again we overload this function to map onto $\{0, 1\}$ for convenience).

Integers and Reals

Common sets include the set of all integers $\mathfrak{Z}$ and the set of all real numbers $\mathfrak{R}$. Special subsets of the integers are the natural numbers $\mathfrak{N} \triangleq \{z \in \mathfrak{Z} \mid z > 0\} = \{1, 2, \ldots\}$ and the whole numbers $\mathfrak{N}_0 \triangleq \{z \in \mathfrak{Z} \mid z \geq 0\} = \{0, 1, \ldots\}$. Similarly, special subsets of the reals are the positive reals $\mathfrak{R}_+ \triangleq \{r \in \mathfrak{R} \mid r > 0\}$ and the non-negative reals $\mathfrak{R}_{0+} \triangleq \{r \in \mathfrak{R} \mid r \geq 0\}$. *Intervals* are subsets of the reals spanning between two bounds; these are denoted by $(a, b) \triangleq \{r \in \mathfrak{R} \mid a < r < b\}$, $[a, b) \triangleq \{r \in \mathfrak{R} \mid a \leq r < b\}$, $(a, b] \triangleq \{r \in \mathfrak{R} \mid a < r \leq b\}$, and $[a, b] \triangleq \{r \in \mathfrak{R} \mid a \leq r \leq b\}$. For instance, $\mathfrak{R}_+ = (0, \infty)$ and $\mathfrak{R}_{0+} = [0, \infty)$.

Indexed Sets

To order the elements of a set, we use an index set as a mapping to each element. For a finite indexable set, we use the notation $\left\{x^{(i)}\right\}_{i=1}^{N}$ to denote the sequence of N objects, $x^{(i)}$, indexed by $\{1, \ldots, N\}$. More generally, a set indexed by some $\mathbb{I}$ is denoted $\left\{x^{(i)}\right\}_{i \in \mathbb{I}}$. An infinite sequence can be indexed by using infinite index sets such as $\mathfrak{N}$ or $\mathfrak{R}$ depending on the cardinality of the sequence.

Multidimensional Sets

Sets can also be coupled to describe multidimensional objects or *ordered tuples*. In refering to an object that is a tuple, we use a (lowercase) bold character such as $\mathbf{x}$ and use parenthetical brackets $(\cdot)$ to specify the contents of the tuple. The simplest tuple is an *ordered pair* $(x, y) \in \mathbb{X} \times \mathbb{Y}$, which is a pair of objects from two sets: $x \in \mathbb{X}$ and $y \in \mathbb{Y}$. The set of all such ordered pairs is called the *Cartesian product* of the sets $\mathbb{X}$ and $\mathbb{Y}$ denoted by $\mathbb{X} \times \mathbb{Y} \triangleq \{(x, y) \mid x \in \mathbb{X} \ \wedge y \ \in \mathbb{Y}\}$. This concept extends beyond ordered pairs to objects of any dimension. A *D-tuple* is an ordered list of D objects belonging to D sets: $(x_1, x_2, \ldots, x_D) \in \bigtimes_{i=1}^{D} \mathbb{X}_i$ where the generalized Cartesian product $\bigtimes_{i=1}^{D} \mathbb{X}_i \triangleq \mathbb{X}_1 \times \mathbb{X}_2 \times \ldots \times \mathbb{X}_D = \{(x_1, x_2, \ldots, x_D) \mid x_1 \in \mathbb{X}_1 \ \wedge x_2 \ \in \mathbb{X}_2 \ \wedge \ \ldots \ \wedge \ x_D \in \mathbb{X}_D\}$; i.e., the set of all such D-tuples. The *dimension* of this Cartesian product space and any member tuple is D, and the function $dim(\cdot)$ returns the dimension of a tuple. When each element of a D-tuple belongs to a common set $\mathbb{X}$, the generalized Cartesian product is denoted with exponential notation as $\mathbb{X}^D \triangleq \bigtimes_{i=1}^{D} \mathbb{X}$; e.g., the Euclidean space $\Re^D$ is the D-dimensional real-valued space.

Vectors

For our purposes, a vector is a special case of ordered D-tuples that we represent with a (usually lowercase) bold character such as $\mathbf{v}$; unlike general tuples, vector spaces are endowed with an addition operator and a scalar multiplication operator, which obey properties discussed here. Consider a D-dimensional vector $\mathbf{v} \in \mathbb{X}^D$ with elements in the set $\mathbb{X}$. The i^{th} element or *coordinate* of $\mathbf{v}$ is a scalar denoted by $v_i \in \mathbb{X}$ where $i \in \{1, 2, \ldots, D\}$. Special real-valued vectors include the all ones vector $\mathbf{1} = (1, 1, \cdots, 1)$, the all zeros vector $\mathbf{0} = (0, 0, \cdots, 0)$, and the coordinate or *basis* vector $\mathbf{e}^{(d)} \triangleq (0, \ldots, 1, \ldots, 0)$, which has a one only in its d^{th} coordinate and is zero elsewhere.

A vector space, $\mathcal{X}$, is a set of vectors that can be added to one another or multiplied by a scalar to yield a new element within the space; i.e., the space is closed under vector addition and scalar multiplication operations that obey associativity, commutativity, and distributivity and have an identity (vector and scalar, respectively) as well as additive inverses. For example, the Euclidean space $\Re^n$ is a vector space for any $n \in \mathfrak{N}$ under the usual vector addition and real multiplication. A convex set $\mathbb{C} \subseteq \mathcal{X}$ is a subset of a vector space with real scalars, with the property that $\forall \alpha \in [0, 1], \quad x, y \in \mathbb{C} \ \Rightarrow \ (1 - \alpha)x + \alpha y \in \mathbb{C}$; i.e., $\mathbb{C}$ is closed under convex combinations. A vector space $\mathcal{X}$ is a *normed vector space* if it is endowed with a norm function $\|\cdot\| : \mathcal{X} \to \Re$ such that for all vectors $x, y \in \mathcal{X}$, *i*) there is a zero element 0 that satisfies $\|x\| = 0 \ \Leftrightarrow \ x = 0$, *ii*) for any scalar α, $\|\alpha x\| = |\alpha| \, \|x\|$, and *iii*) the triangle inequality holds: $\|x + y\| \leq \|x\| + \|y\|$. A common family of norms are the ℓ_p norms defined as

$$\|\mathbf{x}\|_p \triangleq \sqrt[p]{\textstyle\sum_{i=1}^{D} |x_i|^p} \tag{A.1}$$

for $p \in \Re_+$. An extension of this family includes the ℓ_∞ norm, which is defined as $\|\mathbf{x}\|_\infty \triangleq \max\,[|x_i|]$.

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press

Matrices

Usually denoted with an uppercase bold character such as $\mathbf{A}$, a *matrix* is a multi-dimensional object with two indices, which represent a row and column. The $(i, j)^{\text{th}}$ element of $\mathbf{A}$ is denoted by $A_{i,j} \in \mathbb{X}$ where $i \in \{1, 2, \ldots, M\}$ and $j \in \{1, 2, \ldots, N\}$. The full matrix can then be expressed element-wise using the bracket notation:

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,N} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ A_{M,1} & A_{M,2} & \cdots & A_{M,N} \end{bmatrix}.$$

As suggested by this notation, a matrix's first index specifies its *row* and the second specifies its *column*. Each row and column are themselves vectors and are denoted by $\mathbf{A}_{i,\bullet}$ and $\mathbf{A}_{\bullet,j}$ respectively. We also use the bracket notation $[\,\cdot\,]_{i,j}$ to refer to the $(i, j)^{\text{th}}$ element of a matrix-valued expression; e.g., $[\mathbf{A} + \mathbf{B}]_{i,j}$ is the $(i, j)^{\text{th}}$ element of the matrix $\mathbf{A} + \mathbf{B}$. Special matrices include the identity matrix $\mathbf{I}$, with 1's along its diagonal and 0's elsewhere, and the zero matrix $\mathbf{0}$ with zero in every element. The transpose of an $M \times N$-dimensional matrix is an $N \times M$-dimensional matrix denoted as $\mathbf{A}^\top$ and defined as $\left[\mathbf{A}^\top\right]_{i,j} = A_{j,i}$.

Matrix Multiplication

Here we consider vectors and matrices whose elements belong to a scalar field $\mathcal{X}$ endowed with multiplication and addition (e.g., $\mathfrak{Z}$, $\mathfrak{R}$). For the purpose of matrix multiplication, we represent an N-dimensional vector as the equivalent $N \times 1$ matrix for notational convenience. The *inner product* between two vectors $\mathbf{v}$ and $\mathbf{w}$, with $\mathit{dim}\,(\mathbf{v}) = \mathit{dim}\,(\mathbf{w}) = N$, is the scalar denoted by $\mathbf{v}^\top\mathbf{w} = \sum_{i=1}^{N} v_i \cdot w_i$. The *outer product* between M-dimensional vector $\mathbf{v}$ and N-dimensional vector $\mathbf{w}$ is an $M \times N$-dimensional matrix denoted by $\mathbf{v}\mathbf{w}^\top$ with elements $\left[\mathbf{v}\mathbf{w}^\top\right]_{i,j} = v_i \cdot w_j$. The product between an $M \times N$-dimensional matrix $\mathbf{A}$ and an N-dimensional vector $\mathbf{w}$ is denoted $\mathbf{A}\mathbf{w}$ and defined as the M-dimensional vector of inner products between the i^{th} row $\mathbf{A}_{i,\bullet}$ and the vector $\mathbf{w}$; i.e., $[\mathbf{A}\mathbf{w}]_i = \mathbf{A}_{i,\bullet}^\top\mathbf{w}$. It follows that $\mathbf{v}^\top\mathbf{A}\mathbf{w}$ is a scalar defined as $\mathbf{v}^\top\mathbf{A}\mathbf{w} = \sum_{i,j} v_i \cdot A_{i,j} \cdot w_j$. The *matrix product* between an $M \times N$-dimensional matrix $\mathbf{A}$ and an $N \times K$-dimensional matrix $\mathbf{B}$ is an $M \times K$-dimensional matrix denoted by $\mathbf{AB}$ whose $(i, j)^{\text{th}}$ element is the inner product between the i^{th} row of $\mathbf{A}$ and the j^{th} column of $\mathbf{B}$; i.e., $[\mathbf{AB}]_{i,j} = \mathbf{A}_{i,\bullet}^\top\mathbf{B}_{\bullet,j}$.

We also use the *Hadamard (element-wise) product* of vectors and matrices that we denote with the $\odot$ operator. The Hadamard product of vectors $\mathbf{v}$ and $\mathbf{w}$, with $\mathit{dim}\,(\mathbf{v}) = \mathit{dim}\,(\mathbf{w})$, is a vector defined as $[\mathbf{v} \odot \mathbf{w}]_i \triangleq v_i \cdot w_i$. Similarly, the Hadamard product of matrices $\mathbf{A}$ and $\mathbf{B}$, with $\mathit{dim}\,(\mathbf{A}) = \mathit{dim}\,(\mathbf{B})$, is the matrix $[\mathbf{A} \odot \mathbf{B}]_{i,j} \triangleq A_{i,j} \cdot B_{i,j}$.

Functions

We denote a function using regular italic font; e.g., the function g. However, for common named functions (such as logarithm and sine) we use the non-italicized Roman typeface (e.g., log and sin). A function is a mapping from its *domain* $\mathbb{X}$ to its *co-domain* or *range*

$\mathbb{Y}$; $g : \mathbb{X} \to \mathbb{Y}$. To apply g to x, we use the usual notation $g(x)$; $x \in \mathbb{X}$ is the argument and $g(x) \in \mathbb{Y}$ is the value of g at x. We also use this notation to refer to parameterized objects, but in this case, we will name the object according to its type. For instance, $\mathbb{B}^C(g) \triangleq \{x \mid g(x) < C\}$ is a set parameterized by the function g called the C-ball of g, and so we call attention to the fact that this object is a set by using the set notation $\mathbb{B}$ in naming it.

A convex function is any real-valued function $g : \mathbb{X} \to \Re$ whose domain $\mathbb{X}$ is a convex set in a vector space such that, for any $x^{(1)}, x^{(2)} \in \mathbb{X}$ and any $\alpha \in [0, 1]$, the function satisfies the inequality

$$f\left(\alpha x^{(1)} + (1-\alpha)x^{(2)}\right) \leq \alpha f\left(x^{(1)}\right) + (1-\alpha) f\left(x^{(2)}\right).$$

Families of Functions

A family of functions is a set of functions, for which we extend the previous concept of multidimensional sets. Functions can be defined as tuples of (possibly) infinite length—instead of indexing the tuple with natural numbers, it is indexed by the domain of the function; e.g., the reals. To represent the set of all such functions, we use the generalized Cartesian product over an index set $\mathbb{I}$ as $\bigtimes_{i\in\mathbb{I}} \mathbb{X}$ where $\mathbb{X}$ is the co-domain of the family of functions. For instance, the set of all real-valued functions is $\mathcal{G} = \bigtimes_{x\in\Re} \Re$; i.e., every function $g \in \mathcal{G}$ is a mapping from the reals to the reals: $g : \Re \to \Re$. We also consider special subsets such as the set of all continuous real-valued functions $\mathcal{G}^{(\text{continuous})} = \{g \in \mathcal{G} \mid \textit{continuous}(g)\}$ or the set of all convex functions $\mathcal{G}^{(\text{convex})} = \{g \in \mathcal{G} \mid \forall\, t \in [0, 1] \;\; g(tx + (1-t)y) \leq tg(x) + (1-t)g(y)\}$. Particularly, we use the family of all classifiers (as defined in Section 2.2.2) in a D-dimensional space in Chapter 8. This family is the set of functions mapping $\Re^D$ to the set $\{"-", "+"\}$ and is denoted by $\mathcal{F} \triangleq \bigtimes_{x\in\Re^D} \{"-", "+"\}$.

Optimization

Learning theory makes heavy use of mathematical optimization. Optimization typically is cast as finding a *best* object x from a space $\mathcal{X}$ in terms of finding a minimizer of an objective function $f : \mathcal{X} \to \Re$:

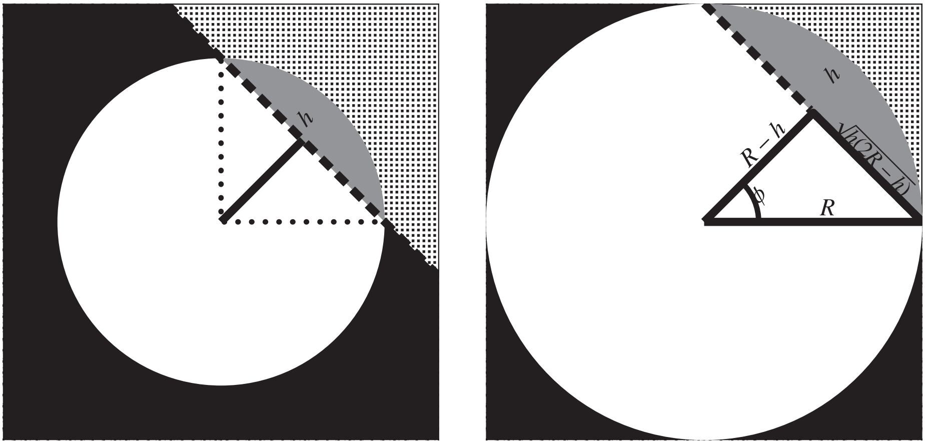
$$x^\star \in \underset{x\in\mathcal{X}}{\operatorname{argmin}}\,[f(x)]$$

where argmin $[\,\cdot\,]$ is a mapping from the space of all objects $\mathcal{X}$ to a subset $\mathbb{X}' \subseteq \mathcal{X}$, which is the set of all objects in $\mathcal{X}$ that minimize f (or equivalently maximize $-f$). Optimizations can additionally be restricted to obey a set of *constraints*. When specifying an optimization with constraints, we use the following notation:

$$\begin{array}{ll} \operatorname{argmin}_{x\in\mathcal{X}} & [f(x)] \\ \text{s.t.} & C(x) \end{array}$$

where f is the function being optimized and C represents the constraints that need to be satisfied. Often there will be several constraints C_i that must be satisfied in the optimization.

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press



(a) A Spherical Cap on a Circle

(b) An Angular Cap on a Circle

Figure A.1 This figure shows various depictions of spherical caps. **(a)** A depiction of a spherical cap of height h, which is created by a halfspace that passes through the sphere. The gray region represents the area of the cap. **(b)** The geometry of the spherical cap: The intersecting halfspace forms a right triangle with the centroid of the hypersphere. The length of the side of this triangle adjacent to the centroid is $R - h$, its hypotenuse has length R, and the side opposite the centroid has length $\sqrt{h(2R - h)}$. The half-angle ϕ, given by $\sin(\phi) = \frac{\sqrt{h(2R-h)}}{R}$, of the right circular cone is used to parameterize the cap.

Probability and Statistics

We denote a probability distribution over the space $\mathcal{X}$ by $P_{\mathcal{X}}$. It is a function that is defined on the subsets in a σ-field of $\mathcal{X}$ (i.e., a set of subsets $\mathbb{A}^{(i)} \subseteq \mathcal{X}$ that is closed under complements and countable unions) and satisfies (i) $P_{\mathcal{X}}\left(\mathbb{A}^{(i)}\right) \geq 0$ for all subsets $\mathbb{A}^{(i)}$, (ii) $P_{\mathcal{X}}(\mathcal{X}) = 1$, and (iii) for pairwise disjoint subsets $\mathbb{A}^{(1)}, \mathbb{A}^{(2)}, \ldots$, it yields $P_{\mathcal{X}}\left(\bigcup_i \mathbb{A}^{(i)}\right) = \sum_i P_{\mathcal{X}}\left(\mathbb{A}^{(i)}\right)$. For a more thorough treatment, we refer the interested reader to Billingsley (1995). A random variable drawn from distribution $P_{\mathcal{X}}$ is denoted by $X \sim P_{\mathcal{X}}$—notice that we do not use a special notation for the random variable, but we make it clear in the text that they are random. The expected value of a random variable is denoted by $\mathrm{E}_{X\sim P_{\mathcal{X}}}[X] = \int x \, dP_{\mathcal{X}}(x)$ or simply by $\mathrm{E}[X]$ when the distribution of the random variables is known from the context. The family of all probability distributions on $\mathcal{X}$ is denoted by $\mathcal{P}_{\mathcal{X}}$; as above, this is the family of all functions that assign probability to elements of the σ-field of $\mathcal{X}$.

A.2 Covering Hyperspheres

Here we summarize the properties of hyperspheres and spherical caps and a covering number result provided by Wyner (1965) and Shannon (1959). This covering result will be used to bound the number of queries required by any evasion algorithm for ℓ_2 costs in Appendix D.2.

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press

A D-dimensional *hypersphere* is simply the set of all points with ℓ_2 distance less than or equal to its radius R from its centroid (in Chapter 8, $\mathbf{x}^A$); i.e.,; the ball $\mathbb{B}^R(A_2)$. Any D-dimensional hypersphere of radius R, $\mathbb{S}^R$, has volume

$$vol\left(\mathbb{S}^R\right) = \frac{\pi^{\frac{D}{2}}}{\Gamma\left(1+\frac{D}{2}\right)} \cdot R^D \tag{A.2}$$

and surface area

$$surf\left(\mathbb{S}^R\right) = \frac{D\cdot\pi^{\frac{D}{2}}}{\Gamma\left(1+\frac{D}{2}\right)} \cdot R^{D-1}.$$

A D-dimensional *spherical cap* is the outward region formed by the intersection of a halfspace and a hypersphere as depicted in Figure A.1(a). The cap has a height of h, which represents the maximum length between the plane and the spherical arc. A cap of height h on a D-dimensional hypersphere of radius R will be denoted by $\mathbb{C}_h^R$ and has a volume

$$vol\left(\mathbb{C}_h^R\right) = \frac{\pi^{\frac{D-1}{2}}R^D}{\Gamma\left(\frac{D+1}{2}\right)}\int_0^{\arccos\left(\frac{R-h}{R}\right)} \sin^D(t)\ dt$$

and surface area

$$surf\left(\mathbb{C}_h^R\right) = \frac{(D-1)\cdot\pi^{\frac{D-1}{2}}R^{D-1}}{\Gamma\left(\frac{D+1}{2}\right)}\int_0^{\arccos\left(\frac{R-h}{R}\right)} \sin^{D-2}(t)\ dt.$$

Alternatively, the cap can be parameterized in terms of the hypersphere's radius R and the half-angle ϕ about a central radius (through the peak of the cap) as in Figure A.1(b). A cap of half-angle ϕ forms the right triangle depicted in the figure, for which $R-h = R\cos(\phi)$ so that h can be expressed in terms of R and ϕ as $h = R*(1-\cos\phi)$. Substituting this expression for h into the above formulas yields the volume of the cap as

$$vol\left(\mathbb{C}_\phi^R\right) = \frac{\pi^{\frac{D-1}{2}}R^D}{\Gamma\left(\frac{D+1}{2}\right)}\int_0^{\phi} \sin^D(t)\ dt \tag{A.3}$$

and its surface area as

$$surf\left(\mathbb{C}_\phi^R\right) = \frac{(D-1)\cdot\pi^{\frac{D-1}{2}}R^{D-1}}{\Gamma\left(\frac{D+1}{2}\right)}\int_0^{\phi} \sin^{D-2}(t)\ dt.$$

Based on these formulas, we now bound the number of spherical caps of half-angle ϕ required to cover the sphere mirroring the result of Wyner (1965).

LEMMA A.1 ***(Result based on Wyner (1965)*** *Covering the surface of D-dimensional hypersphere of radius R, $\mathbb{S}^R$, requires at least*

$$\left(\frac{1}{\sin(\phi)}\right)^{D-2}$$

spherical caps of half-angle $\phi \in \left(0, \frac{\pi}{2}\right)$.

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press

Proof Suppose there are M caps that cover the hypersphere. The total surface area of the M caps must be at least the surface area of the hypersphere. Thus,

$$\begin{aligned} M &\geq \frac{surf\left(\mathbb{S}^R\right)}{surf\left(\mathbb{C}^R_\phi\right)} \\ &\geq \frac{\frac{D\cdot\pi^{\frac{D}{2}}}{\Gamma\left(1+\frac{D}{2}\right)}\cdot R^{D-1}}{\frac{(D-1)\cdot\pi^{\frac{D-1}{2}}R^{D-1}}{\Gamma\left(\frac{D+1}{2}\right)}\int_0^\phi \sin^{D-2}(t)\ dt} \\ &\geq \frac{D\sqrt{\pi}\Gamma\left(\frac{D+1}{2}\right)}{(D-1)\Gamma\left(1+\frac{D}{2}\right)}\left[\int_0^\phi \sin^{D-2}(t)\ dt\right]^{-1}, \end{aligned}$$

which is the result derived by Wyner (although applied as a bound on the packing number rather than the covering number). We continue by lower bounding the above integral. As demonstrated above, integrals of the form $\int_0^\phi \sin^D(t)\ dt$ arise in computing the volume or surface area of a spherical cap. To upper bound the volume of such a cap, note that *i*) the spherical cap is defined by a hypersphere and a hyperplane, *ii*) their intersection forms a $(D-1)$-dimensional hypersphere as the base of the cap, *iii*) the projection of the center of the first hypersphere onto the hyperplane is the center of the $(D-1)$-dimensional hyperspherical intersection, *iv*) the distance between these centers is $R-h$, and *v*) this projected point achieves the maximum height of the cap; i.e., continuing along the radial line achieves the remaining distance h—the height of the cap. We use these facts to upper bound the volume of the cap by enclosing the cap within a D-dimensional hypersphere. As seen in Figure A.1(b), the center of the $(D-1)$-dimensional hyperspherical intersection forms a right triangle with the original hypersphere's center and the edge of the intersecting spherical region (by symmetry, all such edge points are equivalent). That right triangle has one side of length $R-h$ and a hypotenuse of R. Hence, the other side has length $s=\sqrt{h(2R-h)}=R\sin(\phi)$. Moreover, $R\geq h$ implies $s\geq h$. Thus, a D-dimensional hypersphere of radius s encloses the cap, and its volume from Equation (A.2) bounds the volume of the cap as

$$vol\left(\mathbb{C}^R_\phi\right)\leq vol\left(\mathbb{S}^s\right)=\frac{\pi^{\frac{D}{2}}}{\Gamma\left(1+\frac{D}{2}\right)}\cdot\left(R\sin(\phi)\right)^D.$$

Applying this bound to the formula for the volume of the cap in Equation A.3 then yields the following bound on the integral:

$$\begin{aligned} \frac{\pi^{\frac{D-1}{2}}R^D}{\Gamma\left(\frac{D+1}{2}\right)}\int_0^\phi \sin^D(t)\ dt &\leq \frac{\pi^{\frac{D}{2}}}{\Gamma\left(1+\frac{D}{2}\right)}\cdot\left(R\sin(\phi)\right)^D \\ \int_0^\phi \sin^D(t)\ dt &\leq \frac{\sqrt{\pi}\Gamma\left(\frac{D+1}{2}\right)}{\Gamma\left(1+\frac{D}{2}\right)}\cdot\sin^D(\phi). \end{aligned}$$

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press

Using this bound on the integral, the bound on the size of the covering from Wyner reduces to the following (weaker) bound:

$$M \geq \frac{D\sqrt{\pi}\Gamma\left(\frac{D+1}{2}\right)}{(D-1)\Gamma\left(1+\frac{D}{2}\right)}\left[\frac{\sqrt{\pi}\Gamma\left(\frac{D-1}{2}\right)}{\Gamma\left(\frac{D}{2}\right)}\cdot\sin^{D-2}(\phi)\right]^{-1}.$$

Finally, using properties of the gamma function, it can be shown that $\frac{\Gamma(\frac{D+1}{2})\Gamma(\frac{D}{2})}{\Gamma(1+\frac{D}{2})\Gamma(\frac{D-1}{2})} = \frac{D-1}{D}$ which simplifies the above expression to

$$M \geq \left(\frac{1}{\sin(\phi)}\right)^{D-2}.$$

□

It is worth noting that by further bounding the integral $\int_0^\phi \sin^D(t)\ dt$, the bound in Lemma A.1 is weaker than the original bound on the covering derived in Wyner (1965). However, the bound provided by the lemma is more useful for later results because it is expressed in a closed form (see the proof for Theorem 8.13 in Appendix D.2).

Of course, there are other tighter bounds on this power-of-sine integral. In Lemma A.1, this quantity is controlled using a bound on the volume of a spherical cap, but here we instead bound the integral directly. A naive bound can be accomplished by observing that all the terms in the integral are less than the final term, which yields

$$\int_0^\phi \sin^D(t)\ dt \leq \phi\cdot\sin^D(\phi),$$

but this bound is looser than the bound achieved in the lemma. However, by first performing a variable substitution, a tighter bound on the integral can be obtained. The variable substitution is given by letting $p=\sin^2(t)$, $t=\arcsin\left(\sqrt{p}\right)$, and $dt=\frac{dp}{2\sqrt{1-p}\sqrt{p}}$. This yields

$$\int_0^\phi \sin^D(t)\ dt = \frac{1}{2}\int_0^{\sin^2(\phi)} \frac{p^{\frac{D-1}{2}}}{\sqrt{1-p}}\,dp.$$

Within the integral, the denominator is monotonically decreasing in p since, for the interval of integration, $p\leq 1$. Thus it achieves its minimum value at the upper limit $p=\sin^2(\phi)$. Fixing the denominator at this value therefore results in the following upper bound on the integral:

$$\int_0^\phi \sin^D(t)\ dt \leq \frac{1}{2\cos(\phi)}\int_0^{\sin^2(\phi)} p^{\frac{D-1}{2}}\ dp = \frac{\sin^{D+1}(\phi)}{(D+1)\cos(\phi)}. \qquad \text{(A.4)}$$

This bound is not strictly tighter than the bound applied in Lemma A.1, but for large D and $\phi<\frac{\pi}{2}$, this result does achieve a tighter bound. We apply this bound for additional analysis in Section 8.3.1.4.

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press

A.3 Covering Hypercubes

Here we introduce results for covering D-dimensional *hypercube graphs*—a collection of 2^D nodes of the form $(\pm 1, \pm 1, \ldots, \pm 1)$ where each node has an edge to every other node that is Hamming distance 1 from it. The following lemma summarizes coverings of a hypersphere and is utilized in Appendix D for a general query complexity result for ℓ_p distances:

LEMMA A.2 *For any* $0 < \delta \leq \frac{1}{2}$ *and* $D \geq 1$*, to cover a D-dimensional hypercube graph so that every vertex has a Hamming distance of at most* $h = \lfloor \delta D \rfloor$ *to some vertex in the covering, the minimum number of vertixes in the covering is bounded by*

$$Q(D, h) \geq 2^{D(1-H(\delta))},$$

where $H(\delta) = -\delta \log_2(\delta) - (1-\delta)\log_2(1-\delta)$ *is the* entropy *of* δ*.*

Proof There are 2^D vertices in the D-dimensional hypercube graph. Each vertex in the covering is within a Hamming distance of at most h for exactly $\sum_{k=0}^{h}\binom{D}{k}$ vertexes. Thus, one needs at least $2^D / \left(\sum_{k=0}^{h}\binom{D}{k}\right)$ vertexes to cover the hypercube graph. Now we apply the following bound (cf. Flum & Grohe 2006, page 427)

$$\sum_{k=0}^{\lfloor \delta D \rfloor} \binom{D}{k} \leq 2^{H(\delta)D}$$

to the denominator,[1] which is valid for any $0 < \delta \leq \frac{1}{2}$. □

LEMMA A.3 *The minimum of the* ℓ_p *cost function* A_p *from the target* $\mathbf{x}^A$ *to the halfspace* $\mathbb{H}^{(\mathbf{w},\mathbf{b})} = \left\{\mathbf{x} \;\middle|\; \mathbf{x}^\top\mathbf{w} \geq \mathbf{b}^\top\mathbf{w}\right\}$ *can be expressed in terms of the equivalent hyperplane* $\mathbf{x}^\top\mathbf{w} \geq d$ *parameterized by a normal vector* $\mathbf{w}$ *and displacement* $d = \left(\mathbf{b} - \mathbf{x}^A\right)^\top \mathbf{w}$ *as*

$$\min_{\mathbf{x} \in \mathbb{H}^{(\mathbf{w},d)}} A_p\left(\mathbf{x} - \mathbf{x}^A\right) = \begin{cases} d \cdot \|\mathbf{w}\|_{\frac{p}{p-1}}^{-1}, & \textit{if } d > 0 \\ 0, & \textit{otherwise} \end{cases} \tag{A.5}$$

for all $1 < p < \infty$ *and for* $p = \infty$ *it is*

$$\min_{\mathbf{x} \in \mathbb{H}^{(\mathbf{w},d)}} A_\infty\left(\mathbf{x} - \mathbf{x}^A\right) = \begin{cases} d \cdot \|\mathbf{w}\|_1^{-1}, & \textit{if } d > 0 \\ 0, & \textit{otherwise} \end{cases}. \tag{A.6}$$

Proof For $1 < p < \infty$, minimizing A_p on the halfspace $\mathbb{H}^{(\mathbf{w},\mathbf{b})}$ is equivalent to finding a minimizer for

$$\min_{\mathbf{x}} \frac{1}{p} \sum_{i=1}^{D} |x_i|^p \quad \text{s.t.} \quad \mathbf{x}^\top \mathbf{w} \leq d.$$

Clearly, if $d \leq 0$ then the vector $\mathbf{0}$ (corresponding to $\mathbf{x}^A$ in the transformed space) trivially satisfies the constraint and minimizes the cost function with cost 0, which yields

[1] Gottlieb, Kontorovich, & Mossel (2011) present a tighter entropy bound on this sum of binomial coefficients, but it is unnecessary for our result.

the second case of Equation (A.5). For the case $d > 0$, we construct the Lagrangian

$$\mathcal{L}(\mathbf{x}, \lambda) \triangleq \frac{1}{p} \sum_{i=1}^{D} |x_i|^p - \lambda \left(\mathbf{x}^\top \mathbf{w} - d \right).$$

Differentiating this with respect to $\mathbf{x}$ and setting that partial derivative equal to zero yield

$$x_i^\star = \operatorname{sign}(w_i) \left(\lambda |w_i| \right)^{\frac{1}{p-1}}.$$

Plugging this back into the Lagrangian yields

$$\mathcal{L}(\mathbf{x}^\star, \lambda) = \frac{1-p}{p} \lambda^{\frac{p}{p-1}} \sum_{i=1}^{D} |w_i|^{\frac{p}{p-1}} + \lambda d,$$

which we differentiate with respect to λ and set the derivative equal to zero to yield

$$\lambda^\star = \left(\frac{d}{\sum_{i=1}^{D} |w_i|^{\frac{p}{p-1}}} \right)^{p-1}.$$

Plugging this solution into the formula for $\mathbf{x}^\star$ yields the solution

$$x_i^\star = \operatorname{sign}(w_i) \left(\frac{d}{\sum_{i=1}^{D} |w_i|^{\frac{p}{p-1}}} \right) |w_i|^{\frac{1}{p-1}}.$$

The ℓ_p cost of this optimal solution is given by

$$A_p \left(\mathbf{x}^\star - \mathbf{x}^A \right) = d \cdot \|\mathbf{w}\|_{\frac{p}{p-1}}^{-1},$$

which is the first case of Equation (A.5).

For $p = \infty$, once again if $d \leq 0$ then the vector $\mathbf{0}$ trivially satisfies the constraint and minimizes the cost function with cost 0, which yields the second case of Equation (A.6). For the case $d > 0$, we use the geometry of hypercubes (the equi-cost balls of a ℓ_∞ cost function) to derive the second case of Equation (A.6). Any optimal solution must occur at a point where the hyperplane given by $\mathbf{x}^\top \mathbf{w} = \mathbf{b}^\top \mathbf{w}$ is tangent to a hypercube about $\mathbf{x}^A$—this can either occur along a side (face) of the hypercube or at a corner. However, if the plane is tangent along a side (face), it is also tangent at a corner of the hypercube. Hence, there is always an optimal solution at some corner of the optimal cost hypercube.

The corner of the hypercube has the following property:

$$|x_1^\star| = |x_2^\star| = \ldots = |x_D^\star|;$$

that is, the magnitude of all coordinates of this optimal solution is the same value. Further, the sign of the optimal solution's i^{th} coordinate must agree with the sign of the hyperplane's i^{th} coordinate, w_i. These constraints, along with the hyperplane constraint, lead to the following formula for an optimal solution:

$$x_i = d \cdot \operatorname{sign}(w_i) \|\mathbf{w}\|_1^{-1}$$

for all i. The ℓ_∞ cost of this solution is simply $d \cdot \|\mathbf{w}\|_1^{-1}$. □

https://doi.org/10.1017/9781107338548.010 Published online by Cambridge University Press