[chapter]

A random approximate set model

with derivations of random approximate sets induced by set-theoretic operations with corresponding random binary classification measures.

Alexander Towell

atowell@siue.edu

Abstract

We define a *random approximate set* model and the probability space that follows. A random approximate set is a *probabilistic* set generated to *approximate* another set of objective interest. We derive several properties that follow from this definition, such as the expected *precision* in information retrieval. Finally, we demonstrate an application of approximate sets, approximate Encrypted Search with queries as a Boolean algebra, which generates random approximate result sets.

Contents

List of Tables

List of Figures

List of Algorithms

Nomenclature

The following nomenclature describes several symbols and notations in use.

$\mathtt{E}$	The absolute efficiency function.
ℓ	The bit length function.
$\mathtt{RE}$	The relative efficiency function.

Probability

$\mathrm{F}(])\mathrm{X}]$	The cumulative distribution function of random variable X.
$\mathrm{f}(])\mathrm{X}]$	The probability mass or density function of random variable X.

Random approximate sets

$\mathcal{A}^{\pm}$	A *random approximate set* of $\mathcal{A}$ where the false positive and negative rates are unspecified.
$\mathcal{A}_{-}^{\varepsilon}$	A *random approximate set* of $\mathcal{A}$ where the true positive rate is unspecified.
$\mathcal{A}_{-}^{\eta}$	A *random approximate set* of $\mathcal{A}$ where the false negative rate is unspecified.
$\mathcal{A}_{\omega}^{+}$	A *random approximate set* of $\mathcal{A}$ where the false positive rate is unspecified.
$\mathcal{A}_{\omega}^{\varepsilon}$	A *random approximate set* of $\mathcal{A}$ where the parameterizations are ω and ε.
ω	By convention, an *expected* false negative rate if nothing about it is known.
$\grave{\omega}$	By convention, an *observed* false negative rate if nothing about it is known.
ε	By convention, an *expected* false positive rate if nothing about it is known.
$\grave{\varepsilon}$	By convention, an *observed* false positive rate if nothing about it is known.
$\mathcal{A}^{+}$	A *negative* random approximate set of $\mathcal{A}$ where the true positive rate is unspecified.
$\mathcal{A}_{\omega}^{+}$	A *negative* random approximate set of $\mathcal{A}$ where the false negative rate is ω.
$\mathcal{A}^{-}$	A *positive* random approximate set of $\mathcal{A}$ where the false positive rate is unspecified.
$\mathcal{A}_{\varepsilon}^{-}$	A *positive* random approximate set of $\mathcal{A}$ where the false positive rate is ε.
η	By convention, an *expected* true negative rate if nothing about it is known.

$\grave{\eta}$ By convention, an *observed* true negative rate if nothing about it is known.

τ By convention, an *expected* true positive rate if nothing about it is known.

$\grave{\tau}$ By convention, an *observed* true positive rate if nothing about it is known.

Classical sets

$\mathcal{B}$ The binary set $\{0, 1\}$.

$\mathbb{N}$ The set of natural numbers $\{0, 1\}$.

$\mathbb{R}$ The set of reals $(-\infty, \infty)$.

$\mathcal{S}$ Sets are normally notated like this.

Set theory

$|\mathcal{A}|$ The cardinality of a set $\mathcal{A}$.

$\mathcal{P}(A)$ The power set of $\mathcal{A}$.

2^A The power set of $\mathcal{A}$.

$\mathbb{1}$ The indicator function, $\mathbb{1}_{\mathcal{A}} \colon \mathcal{X} \mapsto \{0, 1\}$ is defined by $\mathbb{1}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and otherwise equals 0.

$\mathcal{A}^{\pm} \cup \mathcal{B}^{\pm}$ The union of random approximate sets $\mathcal{A}^{\pm}$ and $\mathcal{B}^{\pm}$.

$\mathcal{A} \cup \mathcal{B}$ The union of $\mathcal{A}$ and $\mathcal{B}$.

$\mathcal{A} \times \mathcal{B}$ The Cartesian product of $\mathcal{A}$ and $\mathcal{B}$ defined as $\{ \langle a \, , b \rangle \mid a \in \mathcal{A} \wedge b \in \mathcal{B} \}$.

$\mathcal{B}^n$ The Cartesian product $\mathcal{B} \times \cdots \times \mathcal{B}$, n times.

Chapter 1

Introduction

An *approximate set* is a set that approximates another set of objective interest. It is *approximate* because with respect to the objective set, there are two types of errors, *false positives* and *false negatives*. The *Bloom filter* is a popular example of a *random approximate set*, but approximate sets arise in many different contexts and circumstances; often, such sets are generated explicitly, as in the Bloom filter, but they may also be *induced* by an uncertain process.

In chapter 2, we define the algebra of sets. In chapter 3, we provide a formal definition of the *random approximate set* model, in which the false positive and false negative rates are *expectations*. We describe the axioms of the random approximate set model such that, if satisfied, also satisfy the axioms of the algebra of sets. We further derive the probability distribution of random approximate sets entailed by the axioms. In section 5.2, we derive the probability distribution of random approximate sets that are generated from arbitrary set-theoretic operations on random approximate sets.

In chapter 8, we provide a treatment on the random approximate set model as an abstract data type and show how that, if the generative algorithm of an approximate set model is deterministic, the random approximate set model quantifies our ignorance or uncertainty.

In chapter 4, we derive several random variables that are fundamental to the approximate set model, such as the uncertain false positive rate.

In section 5.4, we derive several well-known binary classification performance measures of random approximate sets as a function of their error rates, such as *positive predictive value*.

Finally, in chapter 9, we consider Encrypted Search with secure indexes based on random approximate sets. To prove various properties of this model, such as expected precision, we only need to show that the *result sets* are approximate sets of the *objective* results and all the results immediately follow.

Chapter 2

Algebra of sets

A set is an unordered collection of distinct elements. A countable set is a *finite set* or a *countably infinite set*. A *finite set* has a finite number of elements. For example,

$$\mathcal{X} = \{1, 3, 5\}$$

is a finite set with three elements. A *countably infinite set* can be put in one-to-one correspondence with the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, 4, 5, \ldots\}\,. \tag{2.1}$$

The cardinality of a set $\mathcal{Y}$, denoted by $|\mathcal{Y}|$, is a measure of the number of elements in the set. The cardinality of a *finite set* is a natural number, e.g., $|\{1, 3, 5\}| = 3$.

Given two sets $\mathcal{A}$ and $\mathcal{B}$, a basic construction in set theory is the formation of an ordered pair $\langle a\,, b\rangle$, where $a \in \mathcal{A}$ and $b \in \mathcal{B}$, and whose main property is that $\langle a\,, b\rangle = \langle c\,, d\rangle$ *if and only if* $a = c$ and $b = d$.

Definition 2.1 (Cartesian product)**.** *The Cartesian product of* $\mathcal{A}$ *and* $\mathcal{B}$ *is given by* $\{\, \langle a\,, b\rangle \mid a \in \mathcal{A} \wedge b \in \mathcal{B}\,\}$ *and is denoted by the cross product* $\mathcal{A} \times \mathcal{B}$*.*

A *binary relation* over sets $\mathcal{A}$ and $\mathcal{B}$ is any subset of $\mathcal{A} \times \mathcal{B}$. A fundamental relation is the member-of relation, where $x \in \mathcal{A}$ denotes that an object x is a member of a set $\mathcal{A}$. A set $\mathcal{A}$ is a *subset* of a set $\mathcal{B}$ if every member of $\mathcal{A}$ is a member $\mathcal{B}$, denoted by $\mathcal{A} \subseteq \mathcal{B}$. The subset relation forms a *partial order*, i.e., if $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{B} \subseteq \mathcal{C}$ then $\mathcal{A} \subseteq \mathcal{C}$ and if $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{B} \subseteq \mathcal{A}$ then $\mathcal{A}$ and $\mathcal{B}$ are *equal*, denoted by $\mathcal{A} = \mathcal{B}$.

Two sets of particular importance are the empty set, denoted by $\varnothing$, which has no members, and the *universal set*, in which every element of interest is a member. The *power set* of a set $\mathcal{A}$, denoted by $2^{\mathcal{A}}$, is the set of sets that contains all of the possible subsets of $\mathcal{A}$, e.g., $2^{\{a,b\}} = \{\varnothing, \{a\}, \{b\}, \{a, b\}\}$.

In what follows, we consider functions with respect to these special sets. A predicate is a function that maps elements in its domain to true (denoted by 1) or false (denoted by 0). A predicate function of particular importance is the indicator function

$$\mathbb{1}_{\mathcal{A}} \colon \mathcal{X} \mapsto \{0, 1\} \tag{2.2}$$

defined as

$$\mathbb{1}_{\mathcal{A}}(x) \coloneqq \begin{cases} 0 & \text{if } x \notin \mathcal{A}\,, \\ 1 & \text{if } x \in \mathcal{A}\,. \end{cases} \tag{2.3}$$

The indicator function admits the construction of predicates for any relation, e.g., a binary predicate P for a binary relation $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{B}$ is defined as $\mathrm{P}(x_1, x_2) \coloneqq \mathbb{1}_{\mathcal{R}}(\langle x_1\,, x_2\rangle)$. Denoting the

universal set by $\mathcal{U}$, all the relations mentioned previously are *binary predicates*, such as $\in\colon \mathcal{U} \times 2^{\mathcal{U}} \mapsto \{0,1\}$ and $\subseteq\colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto \{0,1\}$.

Denoting the universal set by $\mathcal{U}$, in what follows some important functions on sets which map *pairs* of sets to a set, which are generally known as binary *operations*. The *union* operator, $\cup\colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$, is defined as

$$\mathcal{A} \cup \mathcal{B} \coloneqq \{\, x \in \mathcal{U} \mid x \in \mathcal{A} \vee x \in \mathcal{B} \,\} \tag{2.4}$$

where $\vee$ is the logical-connective *or*. The *intersection* operator, $\cap\colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$, is defined as

$$\mathcal{A} \cap \mathcal{B} \coloneqq \{\, x \in \mathcal{U} \mid x \in \mathcal{A} \wedge x \in \mathcal{B} \,\} \tag{2.5}$$

where $\wedge$ is the logical-connective *and*. If $\mathcal{A} \cap \mathcal{B} = \varnothing$, then we say $\mathcal{A}$ and $\mathcal{B}$ are *disjoint* sets.

The *relative complement* (set-difference) operator, $\setminus\colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$, is defined as

$$\mathcal{A} \setminus \mathcal{B} \coloneqq \{\, x \in \mathcal{U} \mid x \in \mathcal{A} \wedge x \notin \mathcal{B} \,\}\,. \tag{2.6}$$

The relative complement $\mathcal{U} \setminus \mathcal{A}$ is denoted by $\overline{\mathcal{A}}$ and is called the *complement* of $\mathcal{A}$.

2.1 Boolean algebras

An *algebra* denotes a mathematical structure in which a certain set of axioms hold. A *Boolean algebra* is given by the following definition.

Definition 2.2. *A Boolean algebra is a six-tuple* $(\mathcal{A}, \wedge, \vee, \neg, 0, 1)$ *where* $\mathcal{A}$ *is a set,* $\wedge$ *is the binary* meet *operation,* $\vee$ *is the binary* join *operation,* $\neg$ *is the unary* complement *operation,* 0 *is the* bottom *element, and* 1 *is the* top *element such that* $\forall a, b, c \in \mathcal{A}$ *the following axioms hold:*

1. *Associativity:* $a \vee (b \vee c) = (a \vee b) \vee c$ *and* $a \wedge (b \wedge c) = (a \wedge b) \wedge c$.
2. *Commutativity:* $a \vee b = b \vee a$ *and* $a \wedge b = b \wedge a$.
3. *Identity:* $a \vee 0 = a$ *and* $a \wedge 1 = a$.
4. *Distributivity:* $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ *and* $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$.
5. *Complementation:* $a \vee \neg a = 1$ *and* $a \wedge \neg a = 0$.

Every valid proposition in a Boolean algebra is derivable from the axioms in definition 2.2. A particularly useful result is *De Morgan's laws*,

$$a \vee b = \neg(\neg a \wedge \neg b) \tag{2.7}$$

and

$$a \wedge b = \neg(\neg a \vee \neg b)\,. \tag{2.8}$$

Postulate 2.1 (Algebra of sets)**.** *Given the universal set* $\mathcal{U}$ *and a set* $\Sigma \subseteq 2^{\mathcal{U}}$ *that is closed under unions, intersections, and complements,* $(\Sigma, \cup, \cap, \overline{\ }, \varnothing, \mathcal{U})$ *is a Boolean algebra.*

Trivially, $\Sigma = 2^{\mathcal{U}}$ forms a Boolean algebra, but later we demonstrate that implementations of the *random approximate set* model may form a Boolean algebra over some closed subset $\Sigma \subset 2^{\mathcal{U}}$.

The algebra of bit-wise operations on vectors of u bits is given by $(\{0,1\}^u, \wedge, \vee, \neg, \mathbf{0}, \mathbf{1})$ where $\wedge$ is bit-wise *and*, $\vee$ is bit-wise *or*, $\neg$ is bit-wise *negation*, $\mathbf{0}$ is vector of all zeros, and $\mathbf{1}$ is vector of all ones.

A bijection between the algebra of sets and the algebra of bit vectors is given by the following definition.

Definition 2.3. *Suppose there is some total order on $\mathcal{U}$, $u = |U|$, such that the j-th ranked element may be denoted by $x_{(j)}$. A bijection between the Boolean algebras $\left(2^{\mathcal{U}}, \cap, \cup, \bar{}, \varnothing, \mathcal{U}\right)$ and $(\{0,1\}^u, \wedge, \vee, \neg, \mathbf{0}, \mathbf{1})$ is given by mapping $\mathcal{X} \in 2^{\mathcal{U}}$ to $\mathbf{a} \in \{0,1\}^u$ where $a_j = \mathbb{1}_{\mathcal{X}}(x_{(j)})$. Additionally, $\vee \leftrightarrow \cup$, $\wedge \leftrightarrow \cap$, $\neg \leftrightarrow \bar{}$, $\mathbf{0} \leftrightarrow \varnothing$, and $\mathbf{1} \leftrightarrow \mathcal{U}$.*

This bijection allows us to use either representation interchangeably.

Chapter 3

Random approximate set model

The concept of a random approximate set depends upon the concept of an *approximate set*.

Given an objective set $\mathcal{S}$, any element that is a member of $\mathcal{S}$ is denoted a *positive* of $\mathcal{S}$ and any element that is *not* a member of $\mathcal{S}$ is denoted a *negative* of $\mathcal{S}$.

A set that is used as an *approximation* of $\mathcal{S}$ may be denoted by $\mathcal{S}^{\pm}$. If the *only* information we have about $\mathcal{S}$ is given by $\mathcal{S}^{\pm}$, then we may perform membership tests on $\mathcal{S}^{\pm}$ to *predict* the members (or non-members) of $\mathcal{S}$.

There are two ways a binary prediction can be false.

1. A *false positive* occurs if a negative of the objective set is predicted to be a positive. False positives are also known as *type I errors*. The complement of false positives are *true negatives*.

2. A *false negative* occurs if a positive of the objective set is predicted to be a negative. False negatives are also known as *type II errors*. The complement of false negatives are *true positives*.

Suppose we have an objective set $\mathcal{S}$ and an approximation $\mathcal{S}^{\pm}$. If we denote the set of false positives by $\mathcal{F}_p$, true positives by $\mathcal{T}_p$, false negatives by $\mathcal{F}_n$, and true negatives by $\mathcal{T}_n$, then the objective set $\mathcal{S}$ is equal to $\mathcal{F}_n \cup \mathcal{T}_p$ and the approximate set $\mathcal{S}^{\pm}$ is equal to $\mathcal{T}_p \cup \mathcal{F}_p$. See fig. 3.1 for an illustration.

If we only have access to the approximation $\mathcal{S}^{\pm}$, we cannot partition the universe into the sets $\mathcal{F}_p$, $\mathcal{T}_p$, $\mathcal{F}_n$, and $\mathcal{T}_n$ as demonstrated in fig. 3.1. However, we can quantify the degree of *uncertainty* about the elements that are predicted to be positive or negative.

The false positive and true negative *rates* are given by the following.

Definition 3.1. *The* false positive rate *is the proportion of predictions that are* false positives *as given by*

$$\grave{\varepsilon} = \frac{f_p}{f_p + t_n}\,, \tag{3.1}$$

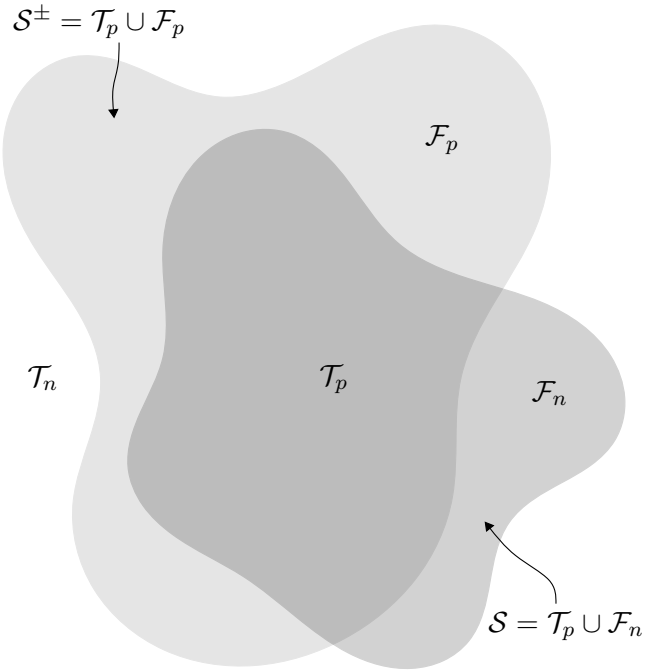
where f_p is the number of false positives *and t_n is the number of* true negatives*. In a complementary manner, the* true negative rate *is $\grave{\eta} = 1 - \grave{\varepsilon}$.*

The true positive and false negative *rates* are given by the following.

Definition 3.2. *The* true positive rate *is the proportion of predictions that are* true positives *as given by*

$$\grave{\tau} = \frac{t_p}{t_p + f_n}\,, \tag{3.2}$$

Figure 3.1: An approximate set $\mathcal{S}^{\pm}$ of an objective set $\mathcal{S}$



where f_n *is the number of* false negatives *and* t_p *is the number of* true positives. *In a complementary manner, the* false negative rate *is* $\grave{\omega} = 1 - \grave{\tau}$.

The *probabilities* of the four possible predictive outcomes are given by table 3.1.

	positive	**negative**
predict positive	$\grave{\tau} = 1 - \grave{\omega}$	$\grave{\varepsilon} = 1 - \grave{\eta}$
predict negative	$\grave{\omega} = 1 - \grave{\tau}$	$\grave{\eta} = 1 - \grave{\varepsilon}$

Table 3.1: The 2×2 contingency table of outcomes for approximate sets.

In the *random* approximate set model, we do not describe any particular approximation, but rather we describe the statistical properties of *algorithms* that *generate* approximations.

3.1 Axioms

The random approximate set is a *set* with an additional set of *probabilistic* axioms. As a function of the probabilistic nature of the random approximate set model, there are an additional set of functions that can defined on it.

Given an objective set $\mathcal{S}$, a *random* approximate set of $\mathcal{S}$ with an *expected* false positive rate ε and an *expected* true positive rate τ is denoted by

$$\mathcal{S}^{\pm}(\tau, \varepsilon)\,. \tag{3.3}$$

First, a random approximate set is a *set* and thus obeys all the axioms of sets. Additionally, there is at minimum a binary prediciate $\in : \mathcal{U} \times 2^{\mathcal{U}} \mapsto \{0, 1\}$ defined

Suppose Σ is a σ-algebra. A *random approximate set* model over the Boolean algebra $(\Sigma, \cup, \cap, \bar{}, \varnothing, \mathcal{U})$ defines a probability space given by the following two axioms.

Axiom 1. *The outcome of a membership test on any element in the negative set is an independent and identically distributed Bernoulli trial with a mean ε,*

$$\mathrm{P}\left[\mathbb{1}_{\mathcal{A}^{\pm}}(x;\tau,\varepsilon) \mid \neg\mathbb{1}_{\mathcal{A}}(x)\right] = \varepsilon\,. \tag{3.4}$$

Axiom 2. *The outcome of a membership test on any element in the negative set is an independent and identically distributed Bernoulli trial with a mean τ,*

$$\mathrm{P}\left[x \in \mathcal{S}^{\pm}(\varepsilon,\tau) \,\middle|\, x \in \mathcal{S}\right] = \tau\,. \tag{3.5}$$

By the above two axioms, the random approximate set $\varnothing^{\pm}$ has a false positive rate ε and the universal set $\mathcal{U}^{\pm}$ has a false negative rate ω where $\mathcal{U}$ denotes the universal set. To satisfy the identity and complementation axioms required by Boolean algebras, we make $\varnothing$ and $\mathcal{U}$ available in the random model as special cases.

Remark. *Alternatively, these axioms may be satisfied by making the empty set and the universal set* degenerate *cases, i.e.,* $\mathrm{P}\left[\varnothing^{\pm} = \varnothing\right] = 1$ *and* $\mathrm{P}\left[\mathcal{U}^{\pm} = \mathcal{U}\right] = 1$. △

The random approximate set model over a Boolean algebra $(\Sigma, \cup, \cap, \bar{}, \varnothing, \mathcal{U})$ is a Boolean algebra $(\Sigma', \cup, \cap, \bar{}, \varnothing, \mathcal{U})$, where $\Sigma' \subseteq \Sigma$.

If we *sample* from $\mathcal{A}^{\pm}(\varepsilon,\tau)$, some particular set will be realized. If we somehow *observe* the false positive and true positive rates, we may indicate which particular values the rates obtained, e.g., $\mathcal{A}^{\pm}(\varepsilon = \grave{\varepsilon}, \tau = \grave{\tau})$.

It may not be possible or practical to observe these rates, e.g., the objective set being approximated may not be precisely knowable from the information that is given. In **??** we derive the probability distributions for characteristics like the false positive rate. Thus, for instance, we may provide a *confidence interval* which contains the false positive rate with some probability α which is a function of parameters like the expected false positive rate ε.

A *positive approximate set* is a special case given by the following definition.

Definition 3.3. *A random approximate set $\mathcal{S}^{\pm}$ with a false negative rate equal to zero is a random* positive *approximate set denoted by $\mathcal{S}^{-}$.*

By this definition, any instance of $\mathcal{S}^{-}$ is a *superset* of $\mathcal{S}$.

As shown in section 5.2, positive approximate sets are closed under unions and intersections but not complements. We introduce the *negative* approximate set as a natural consequence.

Definition 3.4. *A random approximate set $\mathcal{S}^{\pm}$ with a false positive rate equal to zero is a random* negative *approximate set denoted by $\mathcal{S}^{+}$.*

By this definition, any instance of $\mathcal{S}^{+}$ is a *subset* of $\mathcal{S}$.

Negative approximate sets are *closed* under unions and intersections but not complements. The complement of a random positive (negative) approximate set is a random negative (positive) approximate set.

Every statistical property of the approximate set model is entailed by axioms 1 and 2. Furthermore, these assumptions generally hold in practice, e.g., the Bloom filter [**?**] and Perfect hash filter [**?**] are two separate implementations[1] of the random positive approximate set in which these assumptions hold.

[1] There may be a difference in that the algorithm may be deterministic; we address this point in **??**.

3.2 Probability space

Suppose the universal set is $\mathcal{U}$ and we have some process that generates approximations of some objective set $\mathcal{A}$ that is compatible with the axioms of the random approximate set model.

The process generates subsets of $\mathcal{U}$, or alternatively, the *sample space* is $\Sigma = 2^{\mathcal{U}}$. A primary objective in *probability modeling* is assigning *probabilities* to *events*. Suppose we have some *probability function* $\mathrm{P}\colon \Sigma \mapsto [0,1]$. The *probability* of some event $\mathcal{A} \in \Sigma$ is denoted by $\mathrm{P}[\mathcal{A}]$.

By definition 2.3, we use the Boolean algebras $\left(2^{\mathcal{U}}, \cap, \cup, \bar{}, \varnothing, \mathcal{U}\right)$ and $(\{0,1\}^u, \wedge, \vee, \neg, \mathbf{0}, \mathbf{1})$ interchangeably.

Consider an objective set $\mathcal{A}$ and a random approximate set $\mathcal{A}^{\pm}(\varepsilon, \tau)$ and suppose we are uncertain about which elements are their respective members. We model the uncertainty of the elements of $\mathcal{A}$ by the Boolean random vector $\mathbf{A} = \langle \mathrm{A}_1, \ldots, \mathrm{A}_u \rangle$ where $\mathrm{A_j} = \mathbb{1}_{\mathcal{A}}\left(x_{(j)}\right)$ for $j = 1, \ldots, u$. Similarly, we model the uncertainty of the elements of $\mathcal{A}^{\pm}$ by $\mathbf{A}^{\sigma} = \langle \mathrm{A}_1^{\sigma}, \ldots, \mathrm{A}_u^{\sigma} \rangle$.

The joint probability that $\mathbf{A}^{\sigma} = \mathbf{x}$ and $\mathbf{A} = \mathbf{y}$ is denoted by $\mathrm{P}[\mathbf{A}^{\sigma} = \mathbf{x}, \mathbf{A} = \mathbf{y}]$. By the axioms of probability, the joint probability may be rewritten as

$$\mathrm{P}[\mathbf{A}^{\sigma} = \mathbf{x}, \mathbf{A} = \mathbf{y}] = \mathrm{P}[\mathbf{A}^{\sigma} = \mathbf{x} \mid \mathbf{A} = \mathbf{y}]\, \mathrm{P}[\mathbf{A} = \mathbf{y}]\,. \tag{3.6}$$

By axioms 1 and 2, A_j^{σ} is only dependent on $\mathrm{A_j}$ for $j = 1, \ldots, u$ and thus by the axioms of probability

$$\mathrm{P}[\mathbf{A}^{\sigma} = \mathbf{x}, \mathbf{A} = \mathbf{y}] = \mathrm{P}[\mathbf{A} = \mathbf{y}] \prod_{j=1}^{u} \mathrm{P}\left[\mathrm{A}_j^{\sigma} = x_j \;\middle|\; \mathrm{A_j} = y_j\right]. \tag{3.7}$$

If it is given that $\mathbf{A} = \mathbf{y}$, i.e., the elements in the objective set are known, by the axioms of probability the conditional probability is

$$\mathrm{P}[\mathbf{A}^{\sigma} = \mathbf{x} \mid \mathbf{A} = \mathbf{y}] = \prod_{j=1}^{u} \mathrm{P}\left[\mathrm{A}_j^{\sigma} = x_j \;\middle|\; \mathrm{A_j} = y_j\right] \tag{3.8}$$

where $\varepsilon = \mathrm{P}\left[\mathrm{A}_j^{\sigma} \;\middle|\; \neg \mathrm{A_j}\right]$ and $\tau = \mathrm{P}\left[\mathrm{A}_j^{\sigma} \;\middle|\; \mathrm{A_j}\right]$.

These are the *elementary events* of the probability space. The *complete probability space* of the random approximate set model given an objective set $\mathcal{Y}$ is given by the triple

$$\left(\Omega = \{0,1\}^u, \mathcal{F} = 2^{\Omega}, \mathrm{P}[\,\cdot \mid \mathbf{y}]\right), \tag{3.9}$$

or in the other representation,

$$\left(\Omega = 2^{\mathcal{U}}, \mathcal{F} = 2^{\Omega}, \mathrm{P}[\,\cdot \mid \mathcal{Y}]\right). \tag{3.10}$$

The relative frequency of any event $\mathbf{x}$ in $\{0,1\}^u$ converges to $\mathrm{P}[\mathbf{X}^{\sigma} = \mathbf{x} \mid \mathbf{y}]$ as the number of times the random approximate set of $\mathbf{y}$ is generated goes to infinity.

Consider the following example.

Example 1 *Suppose we have an objective set $\{x_1\}$ and a universal set $\{x_1, x_2\}$ and consider an approximate set of $\{x_1\}$ parameterized by ε and τ. The Boolean vector $\langle 1, 0\rangle$ is an equivalent representation[2] of $\{x_1\}$. Let $\mathbf{A}^{\sigma}$ represent the approximation of $\mathbf{A}$. The distribution of $\mathbf{A}^{\sigma}$ conditioned*

[2]Boolean vectors $\{0,1\}^2$ over *bit-wise* operations with an identity $\langle 0, 0\rangle$ for bit-wise *or* and identity $\langle 1, 1\rangle$ for bit-wise *and*.

on $\mathbf{A} = \langle 1, 0 \rangle$ *is*

$$\mathrm{P}\big[\mathbf{A}^{\sigma} = \mathbf{x} \,\big|\, \mathbf{A} = \langle 1, 0 \rangle\big] = \begin{cases} (1-\tau)(1-\varepsilon) & \mathbf{x} = \langle 0, 0 \rangle \\ (1-\tau)\varepsilon & \mathbf{x} = \langle 0, 1 \rangle \\ \tau(1-\varepsilon) & \mathbf{x} = \langle 1, 0 \rangle \\ \tau\varepsilon & \mathbf{x} = \langle 1, 1 \rangle \end{cases} \tag{a}$$

Chapter 4

Probability distributions of parameters

The random approximate sets are *paramterized* by the *expected* rates of two types of error, false negative and false positive rates. In this section, we derive the distribution for these rates.

A random variable $\mathrm{W}\colon \Sigma \mapsto \mathsf{Y}$ is a function that maps outcomes in the σ-algebra to a measurable space Y. The probability that W realizes some measureable subset $\mathsf{Z} \subseteq \mathsf{Y}$ is given by $\mathrm{P}[\mathrm{W} \in \mathsf{S}] = \mathrm{P}\big[\{w \mid \mathrm{W}(w) \in \mathsf{Z}\}\big]$.

The number of false positives is a random variable given by the following theorem..

Theorem 4.1. *Given n negatives, the number of* false positives *in an approximate set with a false positive rate ε is a random variable denoted by FP_n with a distribution given by*

$$\mathrm{FP}_n \sim \mathrm{BIN}(n, \varepsilon) \ . \tag{4.1}$$

Proof. By axiom 1, the uncertain outcome that a negative element *tests* as positive is a Bernoulli trial with a mean ε. Since there are n such independent and identically distributed trials, the number of false positives is binomially distributed with a mean $n\varepsilon$. □

The false positive rate ε is an *expectation*. However, the false positive rate of a random approximate set $\mathcal{S}^{\pm}$ parameterized by ε is *uncertain*.

Theorem 4.2. *The false positive rate is the random variable, denoted by $\mathcal{E}_n$, defined as*

$$\mathcal{E}_n = \frac{\mathrm{FP}_n}{n} \,, \tag{4.2}$$

with an expectation ε, variance $\varepsilon(1-\varepsilon)/n$, and probability mass function

$$\mathrm{f}_{\mathcal{E}_n}(\grave{\varepsilon} \mid \varepsilon) = \mathrm{f}_{\mathrm{FP}_n}(\grave{\varepsilon} n \mid \varepsilon) \,. \tag{4.3}$$

over the support $\{\, \frac{j}{n} \in \mathbb{Q} \mid j \in \{0, \ldots, n\} \,\}$.

Proof. By definition 3.1, the false positive rate is given by the ratio of the number of false positives to the total number of negatives. By theorem 4.1, given that there are n negatives, the number of false positives is a random variable denoted by FP_n. Therefore, the false positive rate, as a function of FP_n, is the random variable $\frac{\mathrm{FP}_n}{n}$. The *expected* false positive rate is

$$\mathrm{E}\left[\frac{\mathrm{FP}_n}{n}\right] = \frac{1}{n}\,\mathrm{E}[\mathrm{FP}_n] = \varepsilon \tag{a}$$

and its variance is

$$\operatorname{var}\left[\frac{\mathrm{FP}_n}{n}\right] = \frac{1}{n^2}\operatorname{var}[\mathrm{FP}_n] = \frac{\varepsilon(1-\varepsilon)}{n}\,. \tag{b}$$

Finally, $\mathcal{E}_n = \mathrm{FP}_n/n$ is a *scaled* transformation of the binomial distribution. Thus, since $\mathrm{FP}_n = n\mathcal{E}_n$,

$$\mathrm{f}_{\mathcal{E}_n}(\hat{\varepsilon}_n \mid \varepsilon) = \mathrm{f}_{\mathrm{FP}_n}(n\hat{\varepsilon})\,. \tag{c}$$

□

The following corollary immediately follows.

Corollary 4.2.1. *Given n negatives, the number of* true negatives *in a random approximate set with a false positive rate ε is a random variable denoted by TN_n with a distribution given by*

$$\mathrm{TN}_n = n - \mathrm{FP}_n \sim \mathrm{BIN}(n, 1-\varepsilon)\,. \tag{4.4}$$

By definition, the true negative rate $\mathcal{N}_n = \mathrm{TN}_n/n = 1 - \mathcal{E}_n$.

By theorem 4.2, the more negatives there are, the lower the variance.

Corollary 4.2.2. *Given* countably infinite *negatives, a random approximate set with a false positive rate ε is* certain *to obtain ε.*

Proof. To say that the sequence $\mathcal{E}_1, \mathcal{E}_2, \ldots$ converges almost surely to ε means that

$$\mathrm{P}\left[\lim_{n\to\infty} \mathcal{E}_n = \varepsilon\right] = 1\,. \tag{a}$$

We know that the *expected* value for each of the random variables in this sequence is ε and the variance is $\varepsilon(1-\varepsilon)/n$. Immediately, we see that as n increases, the distribution of false positives must become more concentrated around ε. As $n \to \infty$, the variance goes to 0, i.e., the distribution becomes degenerate with all of the probability mass assigned to the mean. See **??** for a more rigorous proof. □

The fewer negatives, the greater the variance. The maximum possible variance, when $n = 1$ and $\varepsilon = 0.5$, is 0.25, may be used as the most *pessimistic* estimate given a situation where we have no information about the false positive rate ε and the cardinality of the universal set.

A degenerate case is given by letting $n = 0$, corresponding to a random approximate set of the universal set which has no negative elements that can be tested. Respectively, only random *negative* or *positive* approximate sets may be generated for the universal set or empty set.

The number of false negatives is given by the following theorem.

Theorem 4.3. *Given p positives, the number of* false negatives *in an approximate set $\mathcal{S}^{\pm}$ with a false negative rate ω is a random variable denoted by FN_p with a distribution given by*

$$\mathrm{FN}_p \sim \mathrm{BIN}(p, \omega)\,. \tag{4.5}$$

Proof. By axiom 1, the probability that a positive element *tests* as negative is ω. Thus, each test is a Bernoulli trial. Since there are $p = |\mathcal{S}|$ such independent and identically distributed trials with a probability of "success" ω, the number of false negatives is binomially distributed. □

The false negative rate ω is an *expectation*. However, the false false negative rate of an approximate set $\mathcal{S}^{\pm}$ parameterized by ω is *uncertain*.

Theorem 4.4. *The false negative rate realizes an uncertain value as given by*

$$\mathcal{W}_p = \frac{\mathrm{FN}_p}{p} \tag{4.6}$$

with a support $\{\, j/n \mid j = 0, \ldots, p \,\}$*, an expectation* ω*, and a variance* $\omega(1-\omega)/p$*.*

The proof follows the same logic as the proof for theorem 4.2, except we replace *negatives* with *positives*.

In section 5.2, we consider set-theoretic operations like *complements*. The *complement* operator applied to an approximate set of a set with countably infinite negatives is an approximate set of a set with countably infinite positives.

Corollary 4.4.1. *An approximate set of a set with countably infinite positives has a false negative rate that is* certain *to obtain* ω*.*

The proof follows the same logic as the proof for corollary 4.2.2, except we replace *negatives* with *positives*.

The number of true positives is given by the following corollary.

Corollary 4.4.2. *Given* p *positives, the number of* true positives *in an approximate set with a false negative rate* ω *is a random variable denoted by* TP_p *with a distribution given by*

$$\mathrm{TP}_p \sim \mathrm{BIN}(p, \tau)\,. \tag{4.7}$$

By definition, the true positive rate *is given by* $\mathcal{T}_p = 1 - \mathcal{W}_p$*.*

The proof follows the same logic as the proof for theorem 4.2.

Many other properties of random approximate sets follow from these distributions. For instance,

$$|\mathcal{A}^{\pm}(\tau, \varepsilon)| = \mathrm{TP}_p + \mathrm{FP}_n\,, \tag{4.8}$$

which has an expectation of $n\varepsilon + p\tau$ and variance of $n\varepsilon(1-\varepsilon) + p\tau(1-\tau)$, which is the generalization of the binomial distribution known as the *Poisson binomial distribution*.

If we do not know p, the cardinality $\mathcal{A}$, but have observed $\mathcal{A}^{\pm} = \mathcal{B}$, then $\mathcal{B}$ has a cardinality that tends to be centered around $u\varepsilon + p(\tau - \varepsilon)$ where u is the cardinality of the universal set. Solving for p yields a method of moments estimator

$$\widehat{p} = \frac{|\mathcal{B}| - u\varepsilon}{\tau - \varepsilon}\,. \tag{4.9}$$

If the universal set $\mathcal{U}$ is countably infinite, then this estimator is undefined.

4.1 Asymptotic limits

The false positive and false negative rates are a function of the cardinality of the objective and universal sets. The limiting distributions for the false positive and true positive rates are given by the following theorems.

Theorem 4.5. *By theorem 4.2, the uncertain false positive rate $\mathcal{E}_n$ converges in distribution to the normal distribution with a mean ε and a variance $\varepsilon(1-\varepsilon)/n$, written*

$$\mathcal{E}_n \xrightarrow{d} \mathcal{N}\big(\varepsilon, \varepsilon(1-\varepsilon)/n\big) \ . \tag{4.10}$$

Similarly, by **??**, *the uncertain true positive rate of an approximate set of p positives, denoted by $\mathcal{T}_p$, converges in distribution to the normal distribution with a mean τ and a variance $\tau(1-\tau)/p$, written*

$$\mathcal{T}_p \xrightarrow{d} \mathcal{N}\big(\tau, \tau(1-\tau)/p\big) \ . \tag{4.11}$$

Proof. By **??** in the proof of corollary 4.2.2, given n negatives, the false positive rate is

$$\mathcal{E}_n = \frac{\mathrm{X}_1}{n} + \cdots + \frac{\mathrm{X}_\mathrm{n}}{n} \,, \tag{a}$$

where $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}$ are n independent Bernoulli trials each with a mean ε and a variance $\varepsilon(1-\varepsilon)$. Therefore, by the central limit theorem, $\mathcal{E}_n$ converges in distribution to a normal distribution with a mean ε and a variance $\varepsilon(1-\varepsilon)/n$. The proof for the true positive rate follows the same logic. □

By eqs. (4.10) and (4.11),

$$\mathcal{N}_n \xrightarrow{d} \mathcal{N}\big(1-\varepsilon, \varepsilon(1-\varepsilon)/n\big) \text{ and } \mathcal{W}_n \xrightarrow{d} \mathcal{N}\big(1-\tau, \tau(1-\tau)/p\big) \ . \tag{4.12}$$

The random approximate set model is the *maximum entropy* probability distribution for the indicated false positive and true positive rates, e.g., any estimated α-confidence intervals are the largest intervals possible for the indicated α and therefore represent a worst-case uncertainty.

If we generate an approximate set, the uncertain false positive and true positive rates realize certain values, i.e., $\mathcal{E}_n = \grave{\varepsilon}$ and $\mathcal{T}_p = \grave{\tau}$. If the sample space is countably infinite, the distribution is degenerate, e.g., $\mathcal{E}_n = \varepsilon$ with probability 1. However, for finite sample spaces, the outcomes are uncertain. If these outcomes can be *observed*, e.g., it is not too costly to compute, the exact values $\grave{\varepsilon}$ and $\grave{\tau}$ may be recorded. If these outcomes cannot be observed, e.g., it is too costly to compute or the information to compute $\grave{\varepsilon}$ or $\grave{\tau}$ is not available, we may use the probabilistic model to inform us about the distribution of false positive rates.

Confidence intervals that contain the true false positive rate $\grave{\varepsilon}$ and the true true positive rate $\grave{\tau}$ are given by the following corollaries.

Theorem 4.6. *Given a random approximate set parameterized by ε and τ, asymptotic $\alpha \cdot 100\%$ confidence intervals for the false positive rate and true positive rate are respectively*

$$\varepsilon \pm \sqrt{\frac{\varepsilon(1-\varepsilon)}{n}} \Phi^{-1}(\alpha/2) \tag{4.13}$$

and

$$\tau \pm \sqrt{\frac{\tau(1-\tau)}{p}} \Phi^{-1}(\alpha/2) \,, \tag{4.14}$$

where $\Phi^{-1} \colon [0,1] \mapsto \mathbb{R}$ is the inverse cumulative distribution function of the standard normal.

As a worst-case (maximum uncertainty), we may let $n = p = 1$ in eqs. (4.13) and (4.14).

Chapter 5

Functions of random approximate sets

Given two sets $\mathcal{X}$ and $\mathcal{Y}$, the set of all possible functions from domain $\mathcal{X}$ to codomain $\mathcal{Y}$ is denoted by $\mathcal{X} \mapsto \mathcal{Y}$ (or $\mathcal{X}^{\mathcal{Y}}$ since there are a total of $|\mathcal{X}|^{|\mathcal{Y}|}$ functions in the set). The domain $\mathcal{X}$ may be an a Cartesian product of other sets, e.g., $\mathcal{X}_1 \times \mathcal{X}_2 \mapsto \mathcal{Y}$ denotes a set of binary functions.

A particular function in the set $\mathcal{X} \mapsto \mathcal{Y}$ may be given a label f and we declare that is a function in this set with the notation $\mathrm{f} \colon \mathcal{X} \mapsto \mathcal{Y}$.

If the domain of f is $\mathcal{X} \coloneqq 2^{\mathcal{U}}$, then we have functions that map sets. If we are interested in some objective value $\mathrm{f}(\mathcal{A}) \in \mathcal{Y}$, what happen if we replace $\mathcal{A}$ with $\mathcal{A}^{\pm}(\varepsilon, \tau)$? If f is not a *constant* function, $\mathrm{f}(\mathcal{A})$ denotes some probability distribution over the codomain $\mathcal{Y}$.

Example 2 *Suppose* $\mathrm{f} \colon 2^{\{0,1\}} \mapsto \{0,1\}$ *is defined as*

$$\mathrm{f}(\mathcal{A}) = \begin{cases} 1 & \mathcal{A} \in \{\{1\},\{0,1\}\} \ , \\ 0 & \textit{otherwise.} \end{cases} \tag{a}$$

The random approximate set $\{0\}^{\sigma}(\varepsilon = 0.25, \tau = 0.25)$ *has a probability distribution given by*

$$\mathrm{P}[\{0\}^{\sigma} = \mathcal{B}] = \begin{cases} 0.25 & \mathcal{B} \in \{\{1\},\{0,1\}\} \ , \\ 0.75 & \textit{otherwise.} \end{cases} \tag{b}$$

On inputs $\{1\}$ *and* $\{0,1\}$, f *maps to* 1 *and otherwise maps to* 0. *The random approximate set* $\{0\}^{\sigma}$ *maps to* $\{1\}$ *or* $\{0,1\}$ *with probability* 0.25. *Therefore,* $\mathrm{f}(\{0\}^{\sigma})$ *maps to* 1 *with probability* 0.25, *i.e.,* $\mathrm{f}(\{0\}^{\sigma}) \sim \mathrm{BER}\,(0.25)$.

We consider several classes of functions and the distributions induced by replacing the inputs with random approximate sets, e.g., operators like set-union or binary performance measures like positive predictive value.

In chapter 9, we consider a more complicated substitution that builds upon all the previous work where the function is Boolean search from set-theoretic queries and a collection of Boolean search indexes to documents satisfying the query. In this case, we consider replacing both the set-theoretic queries with approximate set-theoretic queries and replacing the Boolean search indexes with approximate Boolean indexes (approximate sets). In either case, the result is a function that maps to *random approximate result sets.*

5.1 Higher-order random approximate sets

Suppose we have an iterable set that is the output of some random approximation of some objective set of interest. We may wish to apply a more space-efficient data structure for random approximate sets, such as a Bloom filter [**?**]. In this case, the result is a *second-order* random approximate set; that is, a random approximate set *of* a random approximate set.

Theorem 5.1. *A random approximate set* $(\mathcal{A}^{\sigma})^{\sigma}(\tau', \varepsilon')$ *of a random approximate set* $\mathcal{A}^{\sigma}(\tau, \varepsilon)$ *is a random approximate set of* $\mathcal{A}$ *with a true positive rate* $\tau\tau' + \omega\varepsilon'$ *and false positive rate* $\varepsilon\tau' + \eta\varepsilon'$.

Proof. Suppose we have a random approximate set $\mathbf{A}^{\sigma}$ of some given set $\mathbf{B}^{\sigma}$.

$$\mathrm{P}[\mathrm{A}_2^{\sigma}, \mathrm{A}_1^{\sigma} \mid A_0] \tag{a}$$

□

Definition 5.1. *The* iterated *function* f^k *is defined as* k *compositions of* f *where* f^0 *denotes the* identity.

The random approximate set $\mathcal{A}^{\sigma^k}$ denotes the k-th iteration of the random approximation[1] of $\mathcal{A}$ where the *zero-th* order random approximation is the degenerate case (identity) $\mathcal{A}^{\sigma^0} \mapsto \mathcal{A}$.

5.2 Compositions of random approximate sets

The set of n-arity operations on set $\mathcal{X}$ is given by $\mathcal{X}^n \mapsto \mathcal{X}$. In this section, we consider binary operations on $2^{\mathcal{U}}$, like $\cup\colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$, and the result of providing *random approximate sets* as input.

Even though we are generating random approximate sets from random approximate sets, we do not consider the compositions in this section to be higher-order approximations since we have zero control over the distribution of the parameters, like the false positive rate. Instead, such characteristics are a function of the random approximate sets being composed.

Given the Boolean algebra $\left(2^{\mathcal{U}}, \cap, \cup, \overline{\ }, \varnothing, \mathcal{U}\right)$ we derive the random approximate sets that result from the union or complement of random approximate sets. We use these results to derive the random approximate sets that result from arbitrary set-theoretic compositions of random approximate sets, e.g., $\mathcal{A}^{\pm} \smallsetminus \mathcal{B}^{\pm} = \overline{\overline{\mathcal{A}^{\pm}} \cup \mathcal{B}^{\pm}}$.

The random approximate sets that result from union operations on random approximate sets are given by the following theorems.

Theorem 5.2. *The union of two random approximate sets respectively with true negative rates* η_1 *and* η_2 *is a random approximate set with a true negative rate* $\eta_1\eta_2$.

Proof. Suppose we have two sets $\mathcal{A}$ and $\mathcal{B}$ with false positive rates ε_1 and ε_2. The false positive rate ε of $\mathcal{A}^{\pm} \cup \mathcal{B}^{\pm}$ is a probability conditioned on a negative for $\mathcal{A} \cup \mathcal{B}$ being a positive for $\mathcal{A}^{\pm} \cup \mathcal{B}^{\pm}$.

Switching to the Boolean vector representation, suppose we randomly select an element from the universe, denoted by x_j, such that $\neg \mathrm{A}_j \vee \neg \mathrm{B}_j$ is true.

The expected false positive rate of the union is defined by the probability

$$\varepsilon = \mathrm{P}[\mathrm{A}^{\sigma} \cup \mathrm{B}^{\sigma} \mid B_1 \cap B_2]\,. \tag{a}$$

[1] This is not necessarily a function, but the same logic applies.

By DeMorgan's law, the union of sets is the complement of the intersection of their complements. That is,

$$A_1 \cup A_2 \equiv \left(A_1' \cap A_2'\right)' \tag{b}$$

and thus

$$\varepsilon = \mathrm{P}\left[\left(A_1' \cap A_2'\right)' \,\middle|\, B_1 \cap B_2\right]. \tag{c}$$

Since either an event or the *complement* of the event is certain to occur, $\mathrm{P}[E] + \mathrm{P}[E'] = 1$, the above equation may be rewritten as

$$\varepsilon = 1 - \mathrm{P}\left[A_1' \cap A_2' \,\middle|\, B_1 \cap B_2\right]. \tag{d}$$

Since A_1' and A_2' are independent,

$$\varepsilon = 1 - \mathrm{P}\left[A_1' \,\middle|\, B_1 \cap B_2\right] \mathrm{P}\left[A_2' \,\middle|\, B_1 \cap B_2\right]. \tag{e}$$

Since A_1 is conditionally independent of B_2 and A_2 is conditionally independent of B_1, we may rewrite the above equation as

$$\varepsilon = 1 - \mathrm{P}\left[A_1' \,\middle|\, B_1\right] \mathrm{P}\left[A_2' \,\middle|\, B_2\right]. \tag{f}$$

A_j denotes $\mathrm{X} \in \mathcal{S}^j$, therefore A_j' denotes $\mathrm{X} \notin \mathcal{S}^j$. Substituting the definition of A_1', A_2', B_1, and B_2 into the above equation gives

$$\varepsilon = 1 - \mathrm{P}\left[\mathrm{X} \in \mathcal{A}^{\pm} \,\middle|\, \mathrm{X} \notin \mathcal{A}\right] \mathrm{P}\left[\mathrm{X} \notin \mathcal{B}^{\pm} \,\middle|\, \mathrm{X} \notin \mathcal{B}\right]. \tag{g}$$

By definition, $\mathrm{P}\left[\mathrm{X} \notin \mathcal{A}^{\pm} \,\middle|\, \mathrm{X} \notin \mathcal{A}\right]$ is the true negative rate η_1 and likewise for $\mathcal{B}^{\pm}$. Thus,

$$\varepsilon = 1 - \eta_1\eta_2\,. \tag{h}$$

□

The limiting probability distribution of the uncertain true negative rate of the union of $\mathcal{A}^{\pm}$ and $\mathcal{B}^{\pm}$ is thus

$$\mathcal{N}_n \sim \mathcal{N}\left(\eta_1\eta_2, \frac{\eta_1\eta_2(1-\eta_1\eta_2)}{n}\right) \tag{5.1}$$

where the number of negatives $n = u - |\mathcal{A} \cup \mathcal{B}| \leq u$, which is a value between 0 and u. Since this is a limiting distribution, presumably n is large, and as $n \to \infty$ the distribution converges in probability to $\tau_1\tau_2$.

Generally, the number of negatives n or positives p is not known, and so this serves a more analytic function, i.e., given around n negatives, what true negative rate η provides the desired level of confidence that the true negative rate will not realize a value less than some specified value?

Theorem 5.3. *The union of $\mathcal{A}^{\pm}(\omega_1, \eta_1)$ and $\mathcal{B}^{\pm}(\omega_2, \eta_2)$ is a random approximate set with an expected false negative rate*

$$\omega = \alpha_1\omega_1\eta_2 + \alpha_2\eta_1\omega_2 + (1 - \alpha_1 - \alpha_2)\omega_1\omega_2\,, \tag{5.2}$$

where

$$\begin{aligned} 0 \leq \alpha_1 &= \frac{|\mathcal{A} \smallsetminus \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}\,, \\ 0 \leq \alpha_2 &= \frac{|\mathcal{B} \smallsetminus \mathcal{A}|}{|\mathcal{A} \cup \mathcal{B}|}\,, \\ \alpha_1 + \alpha_2 &\leq 1\,. \end{aligned} \tag{5.3}$$

See appendix A.3 for a proof of theorem 5.3.

The complement of an approximate set is given by the following theorem.

Theorem 5.4. *The* complement *of a random approximate set with a false positive rate ε and false negative rate ω is an approximate set with a false positive rate ω and a false negative rate ε.*

Proof. The false positives in an approximate set are false negatives in its complement; likewise, the false negatives in an approximate set are the false positives in its complement set. □

Since many operations in $2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$ may be defined as a composition of unions and complements, theorems 5.2 to 5.4 may be used to derive many other random approximate sets, such as intersections and set-difference.

Remark. *Consider a sequence of positive (or negative) approximate sets $\mathcal{A}_i^-(\varepsilon_i)$ for $i = 1, \ldots, n$. Any* subsequence *contains strictly less information about $\mathcal{A}$. That is, positive (or negative) approximate sets are strictly* additive*, and at the limit $\cap_{i=1}^n \mathcal{A}^-(\varepsilon_i)$ (or $\cup_{i=1}^n \mathcal{A}^+(\omega_i)$) converges almost surely to $\mathcal{A}$ as $n \to \infty$.* △

5.3 Predicates: subset, equality

The *subset* predicate,

$$\subseteq\colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto \{0, 1\} \tag{5.4}$$

is defined as

$$\mathcal{A} \subseteq \mathcal{B} = \prod_{x \in \mathcal{A}} \mathbb{1}_{\mathcal{B}}(x)\,. \tag{5.5}$$

Suppose $\mathcal{A} \subseteq \mathcal{B}$. If we substitute $\mathcal{A}$ with $\mathcal{A}^{\pm}(\tau_1, \varepsilon_1)$ and $\mathcal{B}$ with $\mathcal{B}^{\pm}(\tau_2, \varepsilon_2)$, the result is a Boolean random variable defined as

$$\mathrm{Y} = \prod_{x \in \mathcal{A}^{\pm}} \mathbb{1}_{\mathcal{B}^{\pm}}(x)\,, \tag{5.6}$$

which is *Bernoulli* distributed, i.e.,

$$\mathrm{Y} \sim \mathrm{BER}(p) \tag{5.7}$$

where

$$p = (1 - \tau_1\omega_2)^{|\mathcal{A}|}\,(1 - \varepsilon_1\omega_2)^{|\mathcal{B}|-|\mathcal{A}|}\,(1 - \varepsilon_1\eta_2)^{|\mathcal{U}|-|\mathcal{B}|}\,. \tag{5.8}$$

By definition, the probability that the Bernoulli trial "succeeds" is p, thus

$$\mathrm{P}\left[\prod_{x \in \mathcal{A}^{\pm}} \mathbb{1}_{\mathcal{B}}(x) = 1\right] = p\,. \tag{5.9}$$

Recall in **??** we had shown that the power set of a random approximate set is a *constrained* random approximate set with a member-of relation on the subsets of the universal set, i.e., a probabilistic subset relation is induced. Here, we show that, more generally, the probabilistic subset relation is induced by the probabilistic member-of relation.

However, the induced probabilistic subset relation is not trivially defined and is a function of many other characteristics, i.e., cardinality of each of the involved sets.

The probabilistic *member-of* relation in the random approximate set model *induces* other probabilistic relations. For instance, we claim without proof that if $\mathcal{A} \subseteq \mathcal{B}$, then the subset relation $\mathcal{X}^{\pm}(\tau_1, \varepsilon_1) \subseteq \mathcal{Y}^{\pm}(\tau_2, \varepsilon_2)$ holds with a probability given by

$$(1 - \tau_1\omega_2)^{|\mathcal{A}|}\,(1 - \varepsilon_1\omega_2)^{|\mathcal{B}|-|\mathcal{A}|}\,(1 - \varepsilon_1\eta_2)^{|\mathcal{U}|-|\mathcal{B}|} \tag{5.10}$$

and, similarly, two independent observations of a random approximate set of $\mathcal{A}$ hold the equality relation with a probability given by

$$(\tau_1\tau_2 + \omega_1\omega_2 - \tau_1\tau_2\omega_1\omega_2)^{|\mathcal{A}|}(\varepsilon_1\varepsilon_2 + \eta_1\eta_2 - \varepsilon_1\varepsilon_2\eta_1\eta_2)^{|\mathcal{U}|-|\mathcal{A}|} \tag{5.11}$$

where $\omega_2 = 1 - \tau_2$ and $\eta_2 = 1 - \varepsilon_2$.

Relations like equality and subset have many structural properties, like transitivity, e.g., if $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{B} \subseteq \mathcal{C}$, then $\mathcal{A} \subseteq \mathcal{C}$. In the random approximate set model, such relationships *continue* to hold with some negligible probability when compared to the false positive and false negative rates of the member-of relations.

In chapter 9, we are *primarily* interested in two particular relations.

Theorem 5.5 (True subset rate)**.** *Given a non-random approximate set $\mathcal{A}$ that is a subset of $\mathcal{B}$, $\mathcal{A}$ is a subset of a random approximate set $\mathcal{B}^{\pm}(\tau, \cdot)$ with probability τ^k, $k = |\mathcal{A}|$.*

Proof. ? □

If $\varepsilon_1 > 0$ (and $\eta > 0$), as $|\mathcal{U}| \to \infty$ the conditional probability

$$\mathrm{P}\Big[\mathcal{A}^{\pm}(\tau_1, \varepsilon_1) \subseteq \mathcal{B}^{\pm}(\tau_2, \varepsilon_2) \;\Big|\; \mathcal{A} \subseteq \mathcal{B}\Big] \tag{5.12}$$

goes to 0.

The conditional probability

$$\mathrm{P}\Big[\mathcal{X}^{\pm}(\tau_1, \varepsilon_1) \subseteq \mathcal{Y}^{-}(\varepsilon_2) \;\Big|\; \mathcal{X} \subseteq \mathcal{Y}\Big] \tag{5.13}$$

is given by

$$\left(1 - \varepsilon_1(1 - \varepsilon_2)\right)^{|\mathcal{U}|-|\mathcal{Y}|} . \tag{5.14}$$

$$\mathrm{P}\Big[\mathcal{X}^{-}(\varepsilon_1) = \mathcal{Y}^{-}(\varepsilon_2) \;\Big|\; \mathcal{X} = \mathcal{Y}\Big] \tag{5.15}$$

is given by

$$\left[\left(1 - \varepsilon_1 + \varepsilon_2^2\right)\left(1 - \varepsilon_2 + \varepsilon_1^2\right)\right]^{|\mathcal{U}|-|\mathcal{X}|} . \tag{5.16}$$

If $\varepsilon = \varepsilon_1 = \varepsilon_2$, then the above simplifies to

$$\left(1 - \varepsilon + \varepsilon^2\right)^{2(|\mathcal{U}|-|\mathcal{X}|)} . \tag{5.17}$$

Suppose set $\mathcal{X} = \{x_{j_1}, \ldots, x_{j_k}\}$. The false subset rate is given by the probability

$$\varepsilon_k = \mathrm{P}\Big[\mathcal{X} \subseteq \mathcal{S}^{\pm} \;\Big|\; \mathcal{X} \not\subseteq \mathcal{S}\Big] , \tag{5.18}$$

which may be rewritten as

$$\varepsilon_k = \mathrm{P}\Big[\mathrm{B}_{\mathrm{j}_1} \cap \cdots \cap \mathrm{B}_{\mathrm{j}_k} \;\Big|\; \neg\left(\mathrm{A}_{\mathrm{j}_1} \cap \cdots \cap \mathrm{A}_{\mathrm{j}_k}\right)\Big] . \tag{5.19}$$

By **??**, $\mathrm{B}_1, \ldots, \mathrm{B}_{\mathrm{u}}$ are statistically independent. Making this simplification results in

$$\varepsilon_k = \prod_{p=1}^{k} \mathrm{P}\Big[\mathrm{B}_{\mathrm{j}_\mathrm{p}} \;\Big|\; \overline{\mathrm{A}_{\mathrm{j}_1} \cap \cdots \cap \mathrm{A}_{\mathrm{j}_k}}\Big] . \tag{5.20}$$

5.4 Binary classification measures

In the approximate set model, the distribution of random variables like the false positive, false negatives, true positives, and true negative rates are given respectively by parameters ε, ω, τ, and η. These parameters belong to a more general class of *binary performance measures*.

The above parameters are statements about the distribution of random approximate sets given corresponding objective sets of interest, e.g.,

$$\mathrm{P}\left[\mathbb{1}_{\mathcal{A}^{\pm}}(x;\tau,\varepsilon) \mid \mathbb{1}_{\mathcal{A}}(x)\right] = \tau\,. \tag{5.21}$$

The accuracy of *predictions* about objective sets given a corresponding approximate set is usually the more relevant performance measure. The *positive predictive value* is given by the following definition.

Definition 5.2. *The* positive predictive value *is a performance measure defined as*

$$\mathtt{ppv} = \frac{t_p}{t_p + f_p} \tag{5.22}$$

where t_p is the number of true positives *and f_p is the number of* false positives*.*

The positive predictive value of random approximate sets is a random variable given by the following theorem.

Theorem 5.6. *Given n negatives, p positives, and a random approximate set with false positive and true positive rates ε and τ respectively, the* positive predictive value *is a random variable*

$$\mathrm{PPV} = \frac{\mathrm{TP}_p}{\mathrm{TP}_p + \mathrm{FP}_n} \tag{5.23}$$

with an expectation *given* approximately *by*

$$\mathtt{ppv}(\tau,\varepsilon,p,n) \approx \frac{\bar{t}_p}{\bar{t}_p + \overline{f}_p} + \frac{\bar{t}_p\sigma_{f_p}^2 - \overline{f}_p\sigma_{t_p}^2}{\left(\bar{t}_p + \overline{f}_p\right)^3}\,, \tag{5.24}$$

where $\bar{t}_p = p\tau$ is the expected true positive *frequency, $\overline{f}_p = n\varepsilon$ is the expected* false positive *frequency, $\sigma_{t_p}^2 = (1-\tau)\bar{t}_p$ is the variance of the* true positive *frequency, and $\sigma_{f_p}^2 = (1-\varepsilon)\bar{t}_p$ is the variance of* false positive *frequency.*

See appendix A.2 for a proof of theorem 5.6.

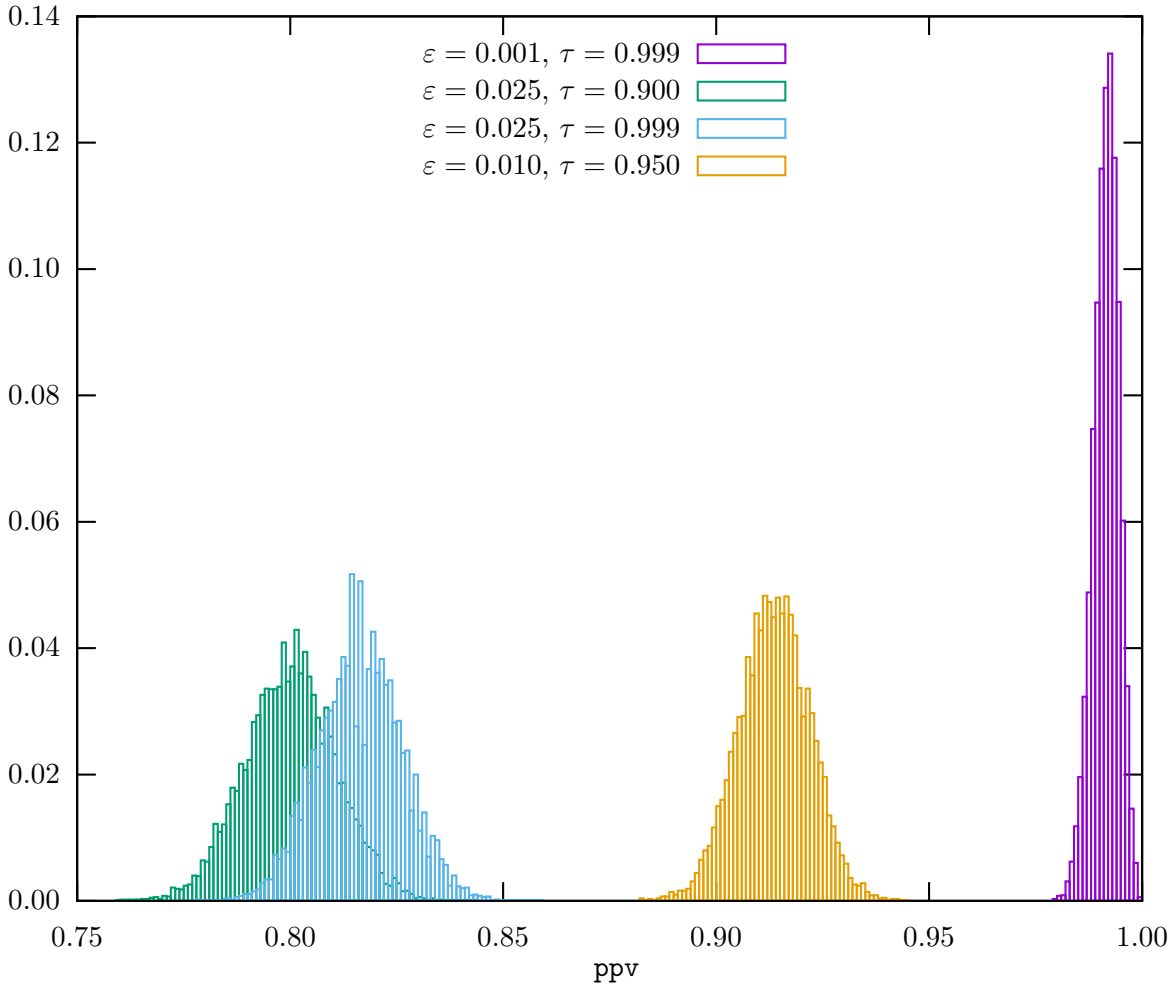
We make the following observations about eq. (5.24):

1. For sufficiently large approximate sets, $\mathtt{ppv} \approx \bar{t}_p/(\bar{t}_p + \overline{f}_p)$.
2. If $\varepsilon \neq 0$, as $n \to \infty$, $\mathtt{ppv} \to 0$.
3. As $\varepsilon \to 0$, $\mathtt{ppv} \to 1$.

Accuracy is given by the following definition.

Definition 5.3. *The* accuracy *is the proportion of true results (both* true positives *and* true negatives*) in the universe of positives and negatives, $(t_p + t_n)/(p+n)$, where t_p, t_n, p, and n are respectively the number of* true positives*,* true negatives*,* positives*, and* negatives*.*

Figure 5.1: Relative frequency of positive predicitive values for several different parameterizations of the false positive and true positive rates given $n = 900$ negatives and $p = 100$ positives.



The *expected* accuracy is given by the following theorem.

Theorem 5.7. *Given p positives and n negatives, a random approximate set with an* expected *false positive rate ε and an* expected *true positive rate τ is a random variable given by*

$$\mathrm{ACC}_{p+n} = \lambda \mathcal{T}_p + (1 - \lambda)\mathcal{N}_n \,. \tag{5.25}$$

has an expected *accuracy*

$$\mathtt{acc}(\tau, \varepsilon, n, p) = \lambda\tau + (1 - \lambda)\eta \tag{5.26}$$

with a variance

$$\frac{\lambda\omega\tau + (1 - \lambda)\varepsilon\eta}{p + n} \,, \tag{5.27}$$

where $\lambda = p/(p + n)$.

Proof. Suppose there the u elements in the universe can be partitioned into p positives and n negatives. An approximate set $\mathcal{S}^{\pm}$ with a false positive rate ε and false negative rate ω has an

uncertain accuracy

$$\mathrm{ACC}_{p+n} = \frac{\mathrm{TP}_p + \mathrm{TN}_n}{p+n}\,. \tag{a}$$

The expected accuracy is given by the expectation

$$\mathrm{E}[\mathrm{ACC}_{p+n}] = \mathrm{E}\left[\frac{\mathrm{TP}_p + \mathrm{TN}_n}{p+n}\right] \tag{b}$$

$$= \frac{p(1-\omega) + n(1-\varepsilon)}{p+n}\,. \tag{c}$$

Noting that $n/(p+n) = 1 - p/(p+n)$ and letting $\lambda = p/(p+n)$,

$$\mathrm{E}[\mathrm{ACC}_{p+n}] = \lambda(1-\omega) + (1-\lambda)(1-\varepsilon)\,. \tag{d}$$

The variance

$$\mathrm{var}[\mathrm{ACC}_{p+n}] = \mathrm{var}\left[\frac{\mathrm{TP}_p}{p+n}\right] + \mathrm{var}\left[\frac{\mathrm{TN}_n}{p+n}\right] \tag{e}$$

$$= \frac{1}{(p+n)^2}\,\mathrm{var}[\mathrm{TP}_p] + \frac{1}{(p+n)^2}\,\mathrm{var}[\mathrm{TN}_n] \tag{f}$$

$$= \frac{p\omega(1-\omega)}{(p+n)^2} + \frac{n\varepsilon(1-\varepsilon)}{(p+n)^2} \tag{g}$$

$$= \frac{\lambda\omega\tau + (1-\lambda)\varepsilon\eta}{p+n}\,. \tag{h}$$

□

Negative predictive value is given by the following definition.

Definition 5.4.

$$\mathtt{npv} = \frac{t_n}{t_n + f_n} \tag{5.28}$$

where t_n and f_n are respectively he number of true negatives *and* false negatives

The expected negative predictive value is given by the following theorem.

Theorem 5.8. *Given p positives, n negatives, and a random approximate set with false positive and true positive rates ε and τ respectively, the negative predictive value is a* random variable

$$\mathrm{NPV} = \frac{\mathrm{TN}_n}{\mathrm{TN}_n + \mathrm{FN}_p} \tag{5.29}$$

with an expectation *given approximately by*

$$\mathtt{npv}(\tau, \varepsilon, p, n) \approx \frac{\bar{t}_n}{\bar{t}_n + \overline{f}_n} + \frac{\bar{t}_n\sigma_{f_n}^2 - \overline{f}_n\sigma_{t_n}^2}{\left(\bar{t}_n + \overline{f}_n\right)^3}\,, \tag{5.30}$$

where $\bar{t}_n = n(1-\varepsilon)$ is the expected true negative *frequency, $\overline{f}_n = p(1-\tau)$ is the expected* false negative *frequency, $\sigma_{t_n}^2 = \varepsilon\bar{t}_n$ is the variance of the* true negative *frequency, and $\sigma_{f_n}^2 = \tau\overline{f}_n$ is the variance of the* false negative *frequency.*

Table 5.1: Various *expected* performance measures.

measure	**parameter**	**expected value**
true positive rate	$\texttt{tpr}(\tau)$	τ
false positive rate	$\texttt{fpr}(\varepsilon)$	ε
false negative rate	$\texttt{fnr}(\tau)$	$1-\tau$
true negative rate	$\texttt{tnr}(\varepsilon)$	$1-\varepsilon$
accuracy	$\texttt{acc}$	eq. (5.26)
positive predictive value	$\texttt{ppv}$	eq. (5.24)
negative predictive value	$\texttt{npv}$	eq. (5.30)
false discovery rate	$\texttt{fdr}$	$1-\texttt{ppv}$
false omission rate	$\texttt{for}$	$1-\texttt{npv}$

The proof for theorem 5.8 follows the same pattern as the proof for theorem 5.6.

Youden's J statistic is a measure of the performance of a binary test, defined as

$$J = \frac{t_p}{t_p + f_n} + \frac{t_n}{t_n + f_p} - 1\,, \tag{5.31}$$

with a range $[0, 1]$. In the case of the random approximate set model, J is a random variable

$$\mathrm{J} = \mathcal{T}_p - \mathcal{E}_n\,, \tag{5.32}$$

which has an *expectation*

$$\mathrm{E}[\mathrm{J}] = \tau - \varepsilon\,. \tag{5.33}$$

Table 5.1 may be used to reparameterize an approximate set.

Example 3 *Suppose we seek a* positive approximate set *with an expected accuracy* γ*. By table 5.1,*

$$\gamma = \texttt{acc}(\varepsilon, \omega = 0, \lambda) = 1 - \varepsilon(1-\lambda)\,. \tag{a}$$

Solving for ε *in terms of* γ *yields the result*

$$\varepsilon(\gamma, \lambda) = \frac{1-\gamma}{1-\lambda} \tag{b}$$

subject to $0 \leq \lambda \leq \gamma \leq 1$ *and* $\lambda < 1$*. Under this parameterization of the positive approximate set,* λ *must be known (or estimated). Note that if* $\lambda = 1$ *then* $\varepsilon(\gamma, \lambda = 1)$ *is undefined as expected, but as* λ *goes to* 1*,* $\varepsilon(\,\cdot\,; \lambda)$ *goes to* 1 *and* γ *goes to* 1*, which logically follows since if there are no negatives, there can be no false positives.*

5.5 Sampling distribution of arbitrary functions

TODO: add generative model as an algorithm for approximate sets? Add C++ implementation of the model? Do some simulations to see how rapidly it converges to the normal? TODO: feed in something like ppv function and see how well it matches the solution given in that one section. etc.

Suppose we have an objective function $\mathrm{f}\colon 2^{\mathcal{X}_1} \times \cdots \times 2^{\mathcal{X}_q} \mapsto \mathcal{Y}$, and we are interested in evaluating the loss when we replace one or more of the objective input sets with particular corresponding random approximate sets. The result of this substitution, as previously described, is a probability distribution over $\mathcal{Y}$.

The probability distribution of random approximate sets are precisely given; therefore, we may estimate the distribution of any function of random approximate sets by generating the random approximate sets and applying the function of interest.

Consider the m-by-q matrix where the (i, j)-th element is the random approximate set $\mathcal{A}^{i,j}(\tau_j, \varepsilon_j)$ such that they are all independely distributed and $\mathcal{A}^{i,j}$ for $i = 1, \ldots, m$ are also identically distributed. If we apply g to each row of the matrix,

$$\mathrm{Y_i} = \mathrm{g}\left(\mathcal{A}^{i,1}, \ldots, \mathcal{A}^{i,q}\right) \tag{5.34}$$

for $i = 1, \ldots, m$, we generate m i.i.d. random elements $\mathrm{Y}_1, \ldots, \mathrm{Y_m}$.

If $\mathcal{Y}$ is a measure space (discrete or continuous), consider the random variable

$$\overline{\mathrm{Y}}_\mathrm{m} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Y_i}\,. \tag{5.35}$$

If Y_1 has a well-defined mean and variance, then by the central limit theorem

$$\lim_{m \to \infty} \overline{\mathrm{Y}}_\mathrm{m} \tag{5.36}$$

converges in distribution to a normal with a mean $\mathrm{E}[\mathrm{Y}_1]$ and a variance $\mathrm{var}[\mathrm{Y}_1]/m$.

A general approach to estimating $\overline{\mathrm{Y}}_\mathrm{m}$ is given by generating a large sample of matrices and applying the function g to each to generate a large sample from Y_1.

We provide an implementation of the generative model and a tool set that permits one to analyze various properties of the distribution of the function of interest.

Chapter 6

Interval

We may not be certain about the false positive and true positive rate parameterizations. In this case, we have a jointly distributed random approximate set conditioned on random rates,

$$\mathcal{A}_{\mathcal{T}}^{\mathcal{E}} \times \mathcal{T} \times \mathcal{E}\,. \tag{6.1}$$

Note that $\mathcal{T}$ is not the same as $\mathcal{T}_p$ and the same for $\mathcal{E}$. $\mathcal{T}_p$ is the uncertain true positive rate that a realized random approximate set $\mathcal{A}_\tau^-$ while $\mathcal{T}$ is the uncertain *expected* true positive rate.

The simplest kind of uncertainty is given by a disjoint set of intervals, in which the true expected rate is uniformly distributed across the support.

Definition 6.1. *An interval is a convex set of real numbers. We denote by* $[x] = [\underline{x}, \bar{x}]$ *an interval with a lower-bound* $\underline{x}$ *and an upper-bound* $\bar{x}$*.*

A further simplification comes from mapping the disjoint set of intervals to the smallest interval that *spans* all of them.

Definition 6.2. *Given a disjoint interval set* $\mathcal{X}$*,* $\texttt{span}(\mathcal{X})$ *maps to an interval with lower and upper bounds that are the lower and upper bounds of* $\mathcal{X}$*.*

A confidence interval, for instance, may be specified in this notation. Here, however, we consider an algebra for interval arithmetic and put it to use quantifying our ignorance about the distribution of parameters after, for instance, a union operation.

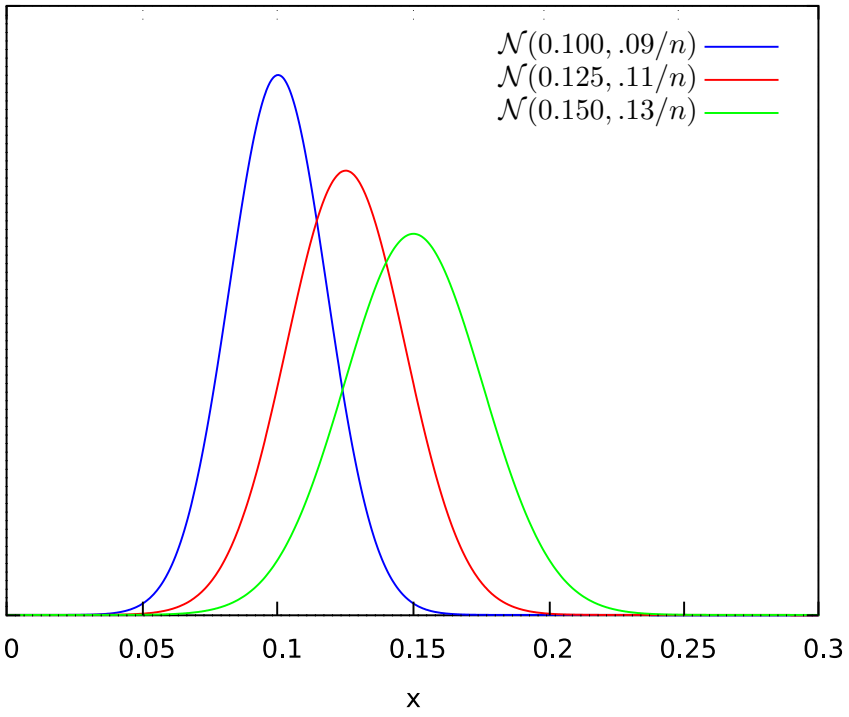
The performance measures summarized by table 5.1 depend upon the false positive rate ε, false negative rate ω, and proportion of positives λ being *known*. Any parameters that are not known with certainty may be replaced in the above table by intervals that (are assumed to) contain the expected value. As a consequence, the performance measure will also be an interval.

Maximum uncertainty is when the parameter value is in the interval $[0, 1]$, e.g., $[\lambda] = [0, 1]$, and *minimum* uncertainty is when the parameter is some value in the degenerate interval $[x, x]$, e.g., $[\varepsilon] = [.2, .2]$. The more certain–the smaller the width of the intervals–the more certain the performance measure.

When using interval arithmetic, the *dependency problem* can lead to overly pessimistic bounds. In our case, the formulae are simple enough to ensure dependencies are satisfied. We show the results of an uncertain proportion of positives $[\lambda]$ for the *accuracy* measure in the following example.

Example 4 *Suppose we wish to determine the expected accuracy given that the proportion of positives is known to be some value in the interval* $[\lambda]$*. Then, the expected accuracy is some value*

Figure 6.1: Relative frequency of positive predicitive values for several different parameterizations of the false positive and true positive rates given $n = 900$ negatives and $p = 100$ positives.



in the interval

$$\texttt{acc}([\varepsilon],[\omega];[\lambda]) = \Big[\mathrm{f}(\overline{\varepsilon},\overline{\omega})(1-\overline{\omega}) + (1-\mathrm{f}(\overline{\varepsilon},\overline{\omega}))\,(1-\overline{\varepsilon}),\\ \mathrm{f}(\underline{\omega},\underline{\varepsilon})(1-\underline{\omega}) + (1-\mathrm{f}(\underline{\omega},\underline{\varepsilon}))\,(1-\underline{\varepsilon})\Big], \tag{a}$$

where $\mathrm{f}(x,y) = \overline{\lambda}[x < y] + \underline{\lambda}[y \leq x]$. *If we have complete ignorance about* λ *then* $[\lambda] = [0,1]$. *As a special case, if we have complete ignorance about lambda and* $\omega = 0$ *(positive approximate set), then* $\texttt{acc}([\varepsilon], 0; [0,1]) = [1-\overline{\varepsilon}, 1]$.

However, the expected rates may not be known, e.g., the values of α_1 and α_2 in eq. (5.2) may not be known. Alternatively, we may not be interested in the *expected* value, but the smallest set of values such that with probability $1-\alpha$ the true rate realizes some value in the set, which is typically an *interval*, i.e., a confidence interval.

Intervals represent an uncertainty and they manifest themselves in two independent ways. The common notion of the *confidence interval* is a product of the probabilistic model, i.e., the realized true positive rate $\hat{\tau}$, which is normally centered around the expected true positive rate τ as discussed in section 4.1. We may use *interval arithmetic* and replace point values interval values, point values being a degenerate case. Basic interval arithmetic is presented in [**?**].

A set sampled from $\mathcal{A}^{\pm}(\varepsilon,\tau)$ is an approximate set such that the $(1-\alpha)\%$ asymptotic confidence

interval for the false negative and false positive rates given respectively by

$$[\omega] =? \tag{6.2}$$

and

$$[\varepsilon] =? \,. \tag{6.3}$$

By **??**, $\mathcal{A}^{\pm} \cup \mathcal{B}^{\pm}$, the observation $\mathcal{A}^{\pm}(\varepsilon, \tau) = \mathcal{A}$ is an approximate set with a false negative rate

$$\grave{\omega} \in [\omega_1](1 - [\varepsilon_2]) \cup [\omega_2](1 - [\varepsilon_1]) \cup [\omega_1][\omega_2] \tag{6.4}$$

and a false positive rate

$$\grave{\varepsilon} \in 1 - (1 - [\varepsilon_1])(1 - [\omega_2]) \,. \tag{6.5}$$

Equation (6.6) represents a disjoint set of intervals. However, we are only interested in the best and worst case of the false negative rate. Thus, we map the disjoint set to a *minimum width* interval that contains every point in the disjoint set.

Definition 6.3. *Given a set $\mathcal{X}$,* $\texttt{span}(\mathcal{X})$ *maps to an interval with lower and upper bounds that are the lower and upper bounds of $\mathcal{X}$.*

Theorem 6.1. *The* union *of two approximate sets with uncertain false negative rates $[\omega_1]$ and $[\omega_2]$ and uncertain false positive rates $[\varepsilon_1]$ and $[\varepsilon_2]$ is an approximate set with an uncertain false negative rate*

$$\begin{aligned} [\omega] &= \texttt{span}([\omega_1](1 - [\varepsilon_2]) \cup [\omega_2](1 - [\varepsilon_1]) \cup [\omega_1][\omega_2]) \\ &= \Big[\min\{\underline{\omega}_1(1 - \overline{\varepsilon}_2), \underline{\omega}_2(1 - \overline{\varepsilon}_1), \underline{\omega}_1\underline{\omega}_2\} \,, \\ &\qquad \max\{\overline{\omega}_1(1 - \underline{\varepsilon}_2), \overline{\omega}_2(1 - \underline{\varepsilon}_1), \overline{\omega}_1\overline{\omega}_2\}\Big] \end{aligned} \tag{6.6}$$

and an uncertain true negative rate

$$\begin{aligned} [\eta] &= [\eta_1][\eta_2] \\ &= [\underline{\eta}_1\underline{\eta}_2, \overline{\eta}_1\overline{\eta}_2] \\ &= 1 - (1 - [\varepsilon_1])(1 - [\varepsilon_2]) \\ &= [\underline{\varepsilon}_1 + \underline{\varepsilon}_2 - \underline{\varepsilon}_1\underline{\varepsilon}_2 \,, \overline{\varepsilon}_1 + \overline{\varepsilon}_2 - \overline{\varepsilon}_1\overline{\varepsilon}_2] \,. \end{aligned} \tag{6.7}$$

Proof. By **??**, the false positive rate of $\mathcal{A}^{\pm} \cup \mathcal{B}^{\pm}$ is

$$[\varepsilon] = [\varepsilon_1] + [\varepsilon_2] - [\varepsilon_1]\,[\varepsilon_2] \,. \tag{a}$$

and the false negative rate is

$$\begin{aligned} [\grave{\omega}] = \alpha_1\,[\grave{\omega}_1]\,(1 - [\varepsilon_2]) + \alpha_2\,[\grave{\omega}_2]\,(1 - [\varepsilon_1]) \\ + (1 - \alpha_1 - \alpha_2)\,(1 - [\varepsilon_1] + [\grave{\omega}_2]\,[\varepsilon_1]) \,, \end{aligned} \tag{5.2 revisited}$$

where $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 \leq 1$. Thus, to maximize (minimize) this equation, we simply need to put all of the *weight* into the largest (smallest) term. □

Theorem 6.2. *The* complement *of an approximate set with a false negative rate $[\omega]$ and false positive rate $[\varepsilon]$ is an approximate set with a false negative rate $[\varepsilon]$ and false positive rate $[\omega]$.*

Table 6.1: The smallest intervals that contain the false positive and false negative rates of the approximate sets that result from the corresponding set-theoretic operations on approximate sets $\mathcal{A}^{\pm}([\omega_1],[\varepsilon_1])$ and $\mathcal{B}^{\pm}([\omega_2],[\varepsilon_2])$.

op	**param**	**interval**
$\mathcal{A}^{\pm} \cup \mathcal{B}^{\pm}$	$[\varepsilon]$	$1-(1-[\varepsilon_1])(1-[\varepsilon_2]$
	$[\omega]$	$\mathtt{span}([\omega_1](1-[\varepsilon_2]) \cup [\omega_2](1-[\varepsilon_1]) \cup [\omega_1][\omega_2])$
$\mathcal{A}^{\pm} \cap \mathcal{B}^{\pm}$	$[\varepsilon]$	$\mathtt{span}([\omega_1](1-[\omega_2]) \cup [\varepsilon_2](1-[\omega_1]) \cup [\omega_1][\varepsilon_2])$
	$[\omega]$	$1-(1-[\omega_1])(1-[\omega_2]$
$\mathcal{A}^{\pm} \smallsetminus \mathcal{B}^{\pm}$	$[\varepsilon]$	$\mathtt{span}([\omega_1](1-[\varepsilon_2]) \cup [\omega_2](1-[\omega_1]) \cup [\omega_1][\omega_2])$
	$[\omega]$	$[\underline{\omega}_1 + \underline{\varepsilon}_2(1-\overline{\omega}_1), \overline{\omega}_1 + \overline{\varepsilon}_2(1-\underline{\omega}_1)]$
$\overline{\mathcal{A}^{\pm}}$	$[\varepsilon]$	$[\omega_1]$
	$[\omega]$	$[\varepsilon_1]$

Table 6.2: The tightest intervals that contain the false positive and false negative rates of the positive or negative approximate sets that result from the corresponding set-theoretic operations.

(a) $\mathcal{A}^{-}([\varepsilon_1])$ and $\mathcal{A}^{-}([\varepsilon_2])$.

op	**param**	**interval**
$\mathcal{A}^{-} \cup \mathcal{B}^{-}$	$[\varepsilon]$	$1-\left(1-[\varepsilon_1]\right)\left(1-[\varepsilon_2]\right])$
$\mathcal{A}^{-} \cap \mathcal{B}^{-}$	$[\varepsilon]$	$[\underline{\varepsilon}_1\underline{\varepsilon}_2, \max(\overline{\varepsilon}_1, \overline{\varepsilon}_2)]$
$\mathcal{A}^{-} \smallsetminus \mathcal{B}^{-}$	$[\varepsilon]$	$[0, \overline{\varepsilon}_1(1-\overline{\varepsilon}_2]$
	$[\omega]$	$[\varepsilon_2]$
$\overline{\mathcal{A}^{-}}$	$[\omega]$	$[\varepsilon_1]$

(b) $\mathcal{A}^{+}([\omega_1])$ and $\mathcal{B}^{+}([\omega_2])$.

op	**param**	**interval**
$\mathcal{A}^{+} \cup \mathcal{B}^{+}$	$[\omega]$	$[\underline{\omega}_1\underline{\omega}_2, \max(\overline{\omega}_1, \overline{\omega}_2)]$
$\mathcal{A}^{+} \cap \mathcal{B}^{+}$	$[\omega]$	$1-(1-[\omega_1])(1-[\omega_2])$
$\mathcal{A}^{+} \smallsetminus \mathcal{B}^{+}$	$[\varepsilon]$	$[0, \overline{\omega}_2(1-\underline{\omega}_1)]$
	$[\omega]$	$[\omega_1]$
$\overline{\mathcal{A}^{+}}$	$[\varepsilon]$	$[\omega_1]$

Since any set-theoretic composition is reducible to a combination of unions and complements, we may use theorems 6.1 and 6.2 to compute the bounds for any set-theoretic composition of approximate sets. See **??** for a summary of a several well-known operations.

Example 5 *Suppose we have three sets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ and consider the random approximate set*

$$\mathcal{D}^{\pm}(\varepsilon,\tau) = \left(\mathcal{A}^{-}(\varepsilon) \cap \mathcal{B}^{-}(\varepsilon)\right) \smallsetminus \mathcal{C}^{-}(\varepsilon)\,. \tag{a}$$

The intersection of $\mathcal{A}^{-}$ and $\mathcal{B}^{-}$ is an approximate set

$$\mathcal{A}^{-} \cap \mathcal{B}^{-}([\varepsilon]) = \overline{\overline{\mathcal{A}^{\pm}} \cup \overline{\mathcal{B}^{\pm}}}\,. \tag{b}$$

Chapter 7

Random approximate sets with invariants

Sometimes, a set must satisfy certain invariants. We have already encountered two invariants in the form of *positive* and *negative* random approximate sets in which if $x \in \mathcal{A}$ then $x \in \mathcal{A}^-$ and if $x \notin \mathcal{A}$, then $x \notin \mathcal{A}^+$. These invariants rely upon the usual partition of the universal set into *positives* and *negatives*, but now we wish to complicate this somewhat. For instance, if a power set has a member $\{a, b\}$ then it *necessarily* has members $\{a\}$, $\{b\}$, and $\varnothing$. In what follows we consider other kinds of invariants.

We restrict our attention to invariants of a set-theoretic quality. Relational invariants, like *symmetry* or *functional* on *binary relations* are a separate topic. See [**?**].

7.1 Random approximate Cartesian products

For any Cartesian product $\mathcal{A} \times \mathcal{B}$, if $\langle a_1 , b_1 \rangle$ and $\langle a_2 , b_2 \rangle$ are positives, then $\langle a_1 , b_2 \rangle$ must also be a positive. However, the *random approximate set* of $\mathcal{A} \times \mathcal{B}$ may generate a set that violates this invariant. For example, in the random approximate set $\langle a_1 , b_1 \rangle$ and $\langle a_2 , b_2 \rangle$ will be true positives with some probability and $\langle a_1 , b_2 \rangle$ will be a false negative with some probability.

However, the Cartesian product of random approximate sets is a Cartesian product as given by the following theorem.

Theorem 7.1. *The Cartesian product* $\mathcal{A}^{\pm}(\tau_1, \varepsilon_1) \times \mathcal{B}^{\pm}(\tau_2, \varepsilon_2)$ *is a random approximate set* $(\mathcal{A} \times \mathcal{B})^{\sigma}(\tau, \varepsilon)$ *that satisfies the constraints of Cartesian products where*

$$\begin{aligned} \tau &= ?\,, \\ \varepsilon &= ?\,. \end{aligned} \tag{7.1}$$

Proof. Prove true positive and false positive rates. Then, show how it obeys the constraints of Cartesian products. □

7.2 Random approximate power sets

The iterated power set $\mathcal{P}^k$ over the Boolean algebra $\left(2^{\mathcal{U}}, \cap, \cup, \bar{}, \varnothing, \mathcal{U}\right)$ generates the Boolean algebra $\left(\mathcal{P}^k(\mathcal{U}), \cap, \cup, \bar{}, \varnothing, \mathcal{P}^{k-1}(\mathcal{U})\right)$.

Theorem 7.2. *Given the Boolean algebra* $\left(2^{\mathcal{U}}, \cap, \cup, \bar{}, \varnothing, \mathcal{U}\right)$, $\mathcal{P}^k\left(\mathcal{A}^{\pm}(\tau, \varepsilon)\right)$ *is a random approximate power set of* $\mathcal{P}^k(\mathcal{A})$ *with a random true positive rate*

$$\mathcal{T}_{2^{\mathcal{A}^{\pm}}} = 2^{\mathrm{TP}_p - p} \tag{7.2}$$

with an approximate expectation

$$\tau_{2^{\mathcal{A}^{\pm}}} = \frac{\left(2^{p\tau} + cp\tau(1-\tau)\right)}{2^p} \tag{7.3}$$

and a random false positive rate

$$\mathcal{E}_{2^{\mathcal{A}^{\pm}}} = \mathcal{T}_{2^{\mathcal{A}^{\pm}}} \frac{2^{\mathrm{FP}_n} - 1}{2^n - 1} \tag{7.4}$$

with an approximate expectation

$$\varepsilon_{2^{\mathcal{A}^{\pm}}} = \tau_{2^{\mathcal{A}^{\pm}}} \frac{2^{n\varepsilon} + cn\varepsilon(1-\varepsilon) - 1}{2^n - 1} \tag{7.5}$$

where $c = \frac{1}{2}\ln^2 2$.

Proof. The joint distribution of random true positives and false positives, TP_p and FP_n respectively, is given by

$$\mathrm{f}_{\mathrm{TP}_p, \mathrm{FP}_n}(t_p, f_p) = \binom{n}{f_p}\binom{p}{t_p}\varepsilon^{f_p}(1-\varepsilon)^{n-f_p}\tau^{t_p}(1-\tau)^{p-t_p} \tag{a}$$

False positives occur in the random approximate power set if the random approximate set we are applying the power set to has any false positives. Given t_p true positives and f_p false negatives, the total number of false positives is given by the summation

$$\sum_{i=0}^{t_p}\sum_{j=1}^{f_p}\binom{t_p}{i}\binom{f_p}{j}, \tag{b}$$

since we must choose at least one false positive (out of the f_p false positives) and we can pick any number of true positives. When we sum over all possibilities, that generates the total number of false positives generated by the power set given t_p true positives and f_p false positives. Notice that each combinations is indexed by only of the summations, so we may rewrite this as

$$\left(\sum_{i=0}^{t_p}\binom{t_p}{i}\right)\left(\sum_{j=1}^{f_p}\binom{f_p}{j}\right). \tag{c}$$

The left summation is just 2^{t_p} and the right summation is just $2^{f_p} - 1$.

So, when there are t_p true positives and f_p false positives in the approximate set, the powerset of the approximation contains $2^{t_p}(2^{f_p} - 1)$ false positives, i.e., the powerset has a random number of false positives given by

$$\mathrm{FP}_{2^{\mathcal{A}^{\pm}}} = 2^{\mathrm{TP}_p}\left(2^{\mathrm{FP}_n} - 1\right). \tag{d}$$

Note that this distribution is no longer binomially distributed. It is a *constrained* random approximate set, since a powerset has a certain structure.

$$\frac{\mathrm{E}\left[2^{\mathrm{TP}_p}\right]\left(2^{\mathrm{E}[\mathrm{FP}_n]} - 1\right)}{2^p(2^n - 1)}. \tag{e}$$

□

We see that letting $\varepsilon = 0$ yields a random approximate power set with a false positive rate 0 and letting $\varepsilon = 1$ yields a random approximate power set with, approximately, a false positive rate $2^{-p(1-\tau)}$.

With a slight shift in perspective, the *power set* of an approximate set $\mathcal{X}^{\pm}$ is equivalently an *approximate subset relation* for $\mathcal{X}$, i.e.,

$$\mathrm{P}\Big[\mathcal{A} \subseteq \mathcal{X}^{\pm} \;\Big|\; \mathcal{A} \subseteq \mathcal{X}\Big] = \varepsilon\,. \tag{7.6}$$

More generally, instead of random *member-of* unary relations, we may define any random approximate relation. The *powerset* on random approximate sets is one way of constructing random approximate subset relations, but they may also be constructed *directly*. The *member-of* predicate may then be defined as

$$a \in \mathcal{X}^{\pm} \coloneqq \{a\} \subseteq \mathcal{X}^{\pm}\,. \tag{7.7}$$

7.3 Random approximate disjoint unions

Suppose we have a family of sets $\{\mathcal{A}_i : i \in \mathcal{I}\}$ indexed by $\mathcal{I}$. The disjoint union of sets $\mathcal{A}_i$ and $\mathcal{A}_j$ is given by

$$\mathcal{A}_i + \mathcal{A}_j = \{\,(x,i) \mid x \in \mathcal{A}_i\,\} \cup \{\,(y,j) \mid y \in \mathcal{A}_j\,\}\,. \tag{7.8}$$

The indexes are auxilliary; they are only used to keep track of which set an element in a disjoint union of sets is from. In a computer system, the indexes may come in the form of a *type* such that if you take the disjoint union of sets $\mathcal{A}$ amd $\mathcal{B}$ parameterized by types X and Y, the result is a set of type $X + Y$ that contains all the elements from $\mathcal{A}$ and $\mathcal{B}$.

Theorem 7.3. *Suppose we have a indexed family of sets with $\mathcal{A}_i$ indexed by i and $\mathcal{A}_j$ indexed by j. The disjoint union of random approximate sets $\mathcal{A}_i^{\pm}(\omega_i, \varepsilon_i)$ and $\mathcal{A}_j^{\pm}(\omega_j, \varepsilon_j)$ is a random approximate set with the invariant that it has a false positive rate ε_l and a false negative rate ω_l on elements indexed by $l \in \{i, j\}$. If the element is not indexed by i or j, then the false positive and negative rates are both* 0.

Proof. Suppose we select an element x assigned index i. Since an element assigned index i is only a candidate in $\mathcal{A}_i^{\pm}$, we only test x for membership in $\mathcal{A}_i^{\pm}$. If x is a *negative*, it will test as a false positive at a rate of ε_i and if x is a *positive*, it will test as a false negative at a rate of ω_i. The same logic applies to an element x assigned index j.

The outcome of a membership test on elements assigned an index $k \notin \{i, j\}$ is given by the following logic. Since, by construction, any such element is in the *negative* set, the false negative rate is undefined, i.e., the false negative rate is the ratio of the number of elements that tested negative to the number of elements in the positive set. However, we can say that the false negative rate is trivially 0. Similarly, the false positive rate must be 0 since, by construction, an element indexed by k is only a candidate in some set indexed by k, but no such such participates in the disjoint union, and therefore tests negative. □

The *plus* symbol for disjoint union is suggestive. The *cardinality* $|\mathcal{A}_1 + \mathcal{A}_2|$ is the *sum* $|\mathcal{A}_1| + |\mathcal{A}_2|$, similarly the cardinality of $|\mathcal{A}^1(\tau_1, \varepsilon_1) + \mathcal{A}^2(\tau_2, \varepsilon_2)|$ is

$$\varepsilon_1 u_1 + \varepsilon_2 u_2 + (\tau_1 - \varepsilon_1)|\mathcal{A}_1| + (\tau_2 - \varepsilon_2)|\mathcal{A}_2|\,, \tag{7.9}$$

where u_1 and u_2 are respectively the cardinalities of the universe of elements with the indexes 1 and 2.

Chapter 8

Abstract data type of the random approximate set

A *data type* is a set and the elements of the set are called the *values* of the data type. We impose a *structure* on sets (data types) by defining morphisms between them, such as operations like *intersection* or relations like *subset*. Morphisms are also types. Any data type needs one or more *value constructors*, functions that map to values of the type.

The random approximate set is an abstract data type that models a *set* with an additional set of *probabilistic* axioms described in **??**.

Suppose T is a data type that overloads the *member-of* predicate $\in\colon \mathcal{U} \times T \mapsto \{0, 1\}$ and has a *value constructor* f that is a *conditional probability distribution* over values of T given elements of type $2^{\mathcal{U}}$.

Data type T *models* the abstract data type of the random approximate set over elements in $\mathcal{U}$ with a false positive rate ε and true positive rate τ if axioms 1 and 2 are satisfied, i.e.,

$$\mathrm{P}\big[x \in \mathrm{f}(\mathcal{S}) \,\big|\, x \notin \mathcal{S}\big] = \varepsilon \tag{8.1}$$

and

$$\mathrm{P}\big[x \in \mathrm{f}(\mathcal{S}) \,\big|\, x \in \mathcal{S}\big] = \tau\,. \tag{8.2}$$

An instance of T also models a classic set by its membership predicate, i.e., two sets are *equal* if and only if they have the same members. We denote that an instance of T models a set $\mathcal{A}$ by $T(\mathcal{A})$.

Normally, two different data types that model an abstract data type are *exchangable* over a set of *regular functions* without changing the result. However, random approximate sets are *probabilistic* so this strict definition of exchangability does not capture the intended meaning. The random approximate set model is a *frequentistic probability* model where an event's probability is defined as the *limit* of its relative frequency in a large number of trials. Thus, we relax the definition of exchangability and conclude that two data types that model random approximate sets (or any other probabilistic abstract data type) should produce the same *limit* of the relative frequency of results in a large number of *independent runs*.

An important distinction must be made with respect to *independent runs*. The most straightforward meaning is, given any set $\mathcal{A} \in 2^{\mathcal{U}}$, at the limit, repeated applications of $\mathrm{f}(\mathcal{A})$ generates a frequency distribution equal to $\mathcal{A}^{\pm}(\tau, \varepsilon)$. However, we also wish to allow for *deterministic* value constructors.[1]

[1] Deterministic algorithms that generate random approximate sets are common, but frequently have an auxiliary seed which indexes a particular approximation in a family.

8.1 Deterministic value constructors

Value constructors can come in many forms. For example, in chapter 9 we show that queries in Boolean search in which search indexes are replaced with random approximate sets generate random approximate *result sets*. In this case, the value constructor is a *function* (deterministic) from queries to approximate result sets.

Suppose we have a deterministic value constructor that constructs objects of type T which model random approximate sets parameterized by τ and ε. The value constructor, being deterministic, generates the same instance of type T given the same input. That is to say, the value constructor is a *total function*, $\mathrm{f} \colon 2^{\mathcal{U}} \mapsto T$.

Recall that T also models the abstract data type of the set, thus there is a unique bijection between T and $2^{\mathcal{U}}$, i.e., every instance of T models a specific subset of $\mathcal{U}$. Thus, we may view the value constructor as a function $\mathrm{f} \colon 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$ with an *image*

$$\texttt{image}(\mathrm{f}) = \{\, \mathrm{f}(\mathcal{A}) \mid \mathcal{A} \in 2^{\mathcal{U}} \,\} \subseteq 2^{\mathcal{U}} \,. \tag{8.3}$$

Since the value constructor f may map multiple input sets to the same output set and some sets in the codomain may not be mapped to by any set in the domain, f is (possibly) a non-surjective, non-injective function.

Definition 8.1. *A σ-algebra is closed under countable unions, intersections, and complements.*

The image of f is not necessarily a σ-algebra. However, the subsets of $\mathcal{U}$ that may be constructed by countable complements, unions, and intersections for elements of the image along with the empty set $\varnothing$ and the universal set $\mathcal{U}$ is by definition a σ-algebra and is denoted by $\sigma(\mathrm{f})$.

Since $\sigma(\mathrm{f})$ is a set of sets closed under unions, intersections, and complements, it is a Boolean algebra defined by the six-tuple $\left(\sigma(\mathrm{f}), \cup, \cap, \bar{}, \varnothing, \mathcal{U}\right)$, e.g., set-theoretic operations over the above Boolean algebra are of the form

$$\sigma(\mathrm{f}) \times \sigma(\mathrm{f}) \mapsto \sigma(\mathrm{f}) \,. \tag{8.4}$$

Suppose we have two Boolean algebras, $\left(\sigma(\mathrm{f}), \cup, \cap, \bar{}, \varnothing, \mathcal{U}\right)$ and $\left(\sigma(\mathrm{g}), \cup, \cap, \bar{}, \varnothing, \mathcal{U}\right)$, where f and g are value constructors for approximate sets over $2^{\mathcal{U}}$. Set-theoretic operations over both Boolean algebras is the Boolean algebra $\left(\Sigma(\mathrm{f}, \mathrm{g}), \cup, \cap, \bar{}, \varnothing, \mathcal{U}\right)$ where $\Sigma(\mathrm{f}, \mathrm{g}) = \sigma\big(\sigma(\mathrm{f}) \cup \sigma(\mathrm{g})$. Note that $\sigma(\mathrm{f}), \sigma(\mathrm{g}) \subseteq \Sigma(\mathrm{f}, \mathrm{g})$, so $\Sigma(\mathrm{f}_1, \cdots, \mathrm{f}_{\mathrm{n}})$ converges to $2^{\mathcal{U}}$ as $n \to \infty$, where $\mathrm{f}_1, \ldots, \mathrm{f}_{\mathrm{n}}$ are different mappings.

Remark. *It is often trivial to implement a family of deterministic value constructors $2^{\mathcal{U}} \mapsto T = \{\mathrm{f}_1, \ldots, \mathrm{f}_{\mathrm{n}}\}$ with distinct σ-algebras where T models random approximate sets over $2^{\mathcal{U}}$. Additionally, assuming each time an approximate set is constructed, a "random" value constructor from $2^{\mathcal{U}} \mapsto T$ is invoked, then repeated invocations on some set $\mathcal{A} \in 2^{\mathcal{U}}$ generates a frequency distribution of sets that converges to $\mathcal{A}^{\pm}$ as $n \to \infty$, e.g., "randomly" seeding a Bloom filter's hash function.* △

How do we reconcile a deterministic value constructor $\mathrm{f} \colon 2^{\mathcal{U}} \mapsto T$ with the *probabilistic model*? In this context, the notion of *probability* quantifies our *ignorance*:

1. Given a set $\mathcal{S}$, we do not have complete *a priori* knowledge about the set the value constructor maps to. The approximate set model only provides *a priori* knowledge about the probability distribution $\mathcal{S}^{\pm}$. We acquire *a posterior* knowledge[2] by observing $\mathrm{f}(\mathcal{S})$.

[2] A posteriori knowledge is dependent on experience.

2. Given $T(\mathcal{S})$, we do not have complete *a priori* knowledge about $\mathcal{S}$. According to the probabilistic model, the only *a priori* knowledge we have is given by the specified *expected* false positive and false negative rates.

   We may acquire *a posteriori* knowledge by evaluating $\mathrm{f}(\mathcal{A})$ for each $\mathcal{A} \in 2^{\mathcal{U}}$ and remembering the sets that map to $T(\mathcal{S})$.[3] However, since f is (possibly) non-injective, one or more sets may map to $T(\mathcal{S})$ and thus this process may not completely eliminate uncertainty. Additionally, the domain $2^{\mathcal{U}}$ has a cardinality $2^{|\mathcal{U}|}$ and thus exhaustive searches are impractical to compute even for relatively small domains.[4]

Suppose $\mathcal{U}$ is finite. The set of deterministic value constructors $2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$ has a cardinality $(2^u)^{(2^u)}$, and in a sense they are all compatible with the random approximate set model.

For instance, a Bloom filter (positive approximate set) may have a family of hash function that, for a particular binary coding of the elements of a given universal set, maps *every* element in the universal set to the same hash. Thus, for instance, no matter the objective set $\mathcal{X} \subseteq \mathcal{U}$, it will map to $\mathcal{U}$. The Bloom filter had a theoretically sound implementation, but only after empirical evidence was it discovered that it was not suitable. This is an extremely unlikely outcome in the case of large universal sets, but as the cardinality of the universal set decreases, the probability of such an outcome increases. Indeed, at $|U| = 2$, the probability of this outcome is ?.

Thus, *a priori* knowledge, e.g., a theoretically sound algorithm, is not in practice sufficient (although for large universal sets, the probability is negligible). The suitability of an algorithm can only be determined by acquring *a posterior* knowledge.

We could explore the space of functions in the family and only choose those which, on some sample of objective sets of interest, generates the desired expectations for the false positive and false negative rates with the desired variances. Most of them will if constructed in the right sort of way.

A family of functions that are compatible with the probabilistic model is given by observing a particular realization $\mathcal{X} = \mathcal{S}^{\pm}$ and outputting $\mathcal{X}$ on subsequent inputs of $\mathcal{S}$, i.e., caching the output of a *non-deterministic* process that conforms to the probabilistic model. This is essentially how well-known implementations like the Bloom filter work, where the pseudo-randomness comes from mechanical devices like hash functions that approximate random oracles.

The false positive rate of the approximate set corresponding to objective set $\mathcal{X}$ is given by

$$\grave{\varepsilon}(\mathcal{X}) = \frac{1}{n} \sum_{x \in \overline{\mathcal{X}}} \mathbb{1}_{\mathrm{f}(\mathcal{X})}(x)\,, \tag{8.5}$$

where $n = |\overline{\mathcal{X}}|$.

Let $\mathcal{U}_p$ denote the set of objective sets with cardinality p. The *mean* false positive rate,

$$\overline{\varepsilon} = \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X} \in \mathcal{U}_p} \grave{\varepsilon}(\mathcal{X})\,, \tag{8.6}$$

is an unbiased estimator of ε and the population variance

$$s_\varepsilon^2 = \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X} \in \mathcal{U}_p} \grave{\varepsilon}(\mathcal{X})\,, \tag{8.7}$$

is an unbiased estimator of $\mathrm{var}[\mathcal{E}_n] = \varepsilon(1-\varepsilon)/n$.

[3] If the approximate set is the result of the union, intersection, and complement of two or more approximate sets, then we must consider the closure.

[4] In the case of *countably infinite* domains, it is not even theoretically possible.

Proof. We imagine that the function f caches the output of a *non-deterministic* process that conforms to the probabilistic model. Thus, each time the function maps an objective set $\mathcal{X}$ of cardinality p to its approximation, the algorithm *observes* a realization of $\mathcal{E}_n = \grave{\varepsilon}$. Thus,

$$\overline{\varepsilon} = \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X}_\rangle \in \mathcal{U}_p} \grave{\varepsilon}(\mathcal{X}_\rangle) \tag{a}$$

$$= \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X}_\rangle \in \mathcal{U}_p} \mathrm{E}\Big[\mathcal{E}_n^{(i)}\Big] = \varepsilon\,. \tag{b}$$

□

8.2 Space complexity

If the finite cardinality of a universe is u and the set is *dense* (and the approximation is also dense, i.e., the false negative rate is relatively small), then

$$\mathcal{O}(u) \text{ bits} \tag{8.8}$$

are needed to code the set, which is independent of p, the false positive rate, and the false negative rate.

The lower-bound on the *expected* space complexity of a data structure that models the *random approximate set* where the elements are over a *countably infinite* universe is given by the following postulate.

Postulate 8.1. *The* information-theoretic lower-bound *of a data structure that implements the countably infinite* random approximate set *abstract data type has an* expected *bit length given by*

$$-\tau \log_2 \varepsilon \text{ bits/element}\,, \tag{8.9}$$

where $\varepsilon > 0$ is the false positive rate and τ is the true positive.

The *relative space efficiency* of a data structure X to a data structure Y is some value greater than 0 and is given by the ratio of the bit length of Y to the bit length of X,

$$\mathtt{RE}(X, Y) = \frac{\ell(Y)}{\ell(X)}\,, \tag{8.10}$$

where ℓ is the bit length function. If $\mathtt{RE}(X, Y) < 1$, X is less efficient than Y, if $\mathtt{RE}(X, Y) > 1$, X is more efficient than Y, and if $\mathtt{RE}(X, Y) = 1$, X and Y are equally efficient. The absolute space efficiency is given by the following definition.

Definition 8.2. *The absolute space efficiency of a data structure X, denoted by $\mathtt{E}(X)$, is some value between 0 and 1 and is given by the ratio of the bit length of the theoretical lower-bound to the bit length of X,*

$$\mathtt{E}(X) = \frac{\theta}{\ell(X)}\,, \tag{8.11}$$

where $\ell(X)$ denotes the bit length of X and θ denotes the bit length of the information-theoretic lower-bound. The closer $\mathtt{E}(X)$ is to 1, the more space-efficient the data structure. A data structure that obtains an efficiency of 1 is optimal.[5]

[5]Sometimes, a data structure may only obtain the information-theoretic lower-bound with respect to the limit of some parameter, in which case the data structure *asymptotically* obtains the lower-bound with respect to said parameter, the number of positives p being the most obvious parameter.

The *absolute* space efficiency of a data structure X implementing a random approximate set of an objective set with p elements with a false positive rate ε and true positive rate τ is given by

$$\mathtt{E}(X) = \frac{-p\tau \log_2 \varepsilon}{\ell(X)} \,. \tag{8.12}$$

A well-known implementation of countably infinite *positive approximate set* is the Bloom filter [**?**] which has an expected space complexity given by

$$-\frac{1}{\ln 2} \log_2 \varepsilon \text{ bits/element} \,, \tag{8.13}$$

thus the absolute efficiency of the Bloom filter is $\ln 2 \approx 0.69$. A practical implementation of the *positive random approximate set* is given by the *Perfect Hash Filter* [**?**], which compares favorably to the Bloom filter in may circumstances.[6]

In **??**, we claimed that the method of moments estimator for p of an objective set given a particular realization of a random approximation set is undefined for countably infinite universes. Suppose we have a data structure X that *models* random approximate sets with an *expected* space complexity proportional to p, i.e., $p \cdot \mathrm{b}(\tau, \varepsilon)$ bits, where b is the expected bits per *positive* element given a false positive rate ε and true positive rate τ. Then, given an object x of type X, an estimator of p is

$$\widehat{p} = \frac{\ell(x)}{\mathrm{b}(\tau, \varepsilon)} \,, \tag{8.14}$$

were ℓ is the bit length function. An expected *upper-bound* on the cardinality is obtained by plugging in the information-theoretic lower-bound $\mathrm{b}(\tau, \varepsilon) = -\tau \log_2 \varepsilon$ bits per element.

Space efficiency of *unions* and *differences*

As a way to implement *insertions* and *deletions*, we consider the space efficiency of the set-theoretic operations of unions and differences of approximate sets.

Let $\mathcal{S}_1 = \{x_{j_1}, \ldots, x_{j_m}\}$ and suppose we wish to insert the elements $x_{k_1}, \ldots, x_{k_p}$ into $\mathcal{S}_1$. If X_1 is a mutable object, then an *insertion* operator may be applied on X_1 for each x_{k_i}, $i = 1, \ldots, p$.

Alternatively, if X_1 is immutable, then we may construct another object, X_2, that implements the set $\mathcal{S}_2 = \{x_{k_1}, \ldots, x_{k_p}\}$, and then apply the union function,

$$X_1 \cup X_2 \,. \tag{8.15}$$

If we replace X_1 and X_2 by objects that implement *positive approximate sets* of $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively, then by **??**, the false positive rate of the resulting approximate set is $\grave{\varepsilon}_1 + \grave{\varepsilon}_2 - \grave{\varepsilon}_1\grave{\varepsilon}_2$.

The space efficiency of this positive approximate set is given by the following theorem.

Theorem 8.1. *Given two countably infinite positive approximate sets $\mathcal{S}_1^-$ and $\mathcal{S}_2^-$ respectively with false positive rates $\grave{\varepsilon}_1$ and $\grave{\varepsilon}_2$, the approximate set $\mathcal{S}_1^- \cup \mathcal{S}_2^-$, which has an induced false positive rate $\grave{\varepsilon}_1 + \grave{\varepsilon}_2 - \grave{\varepsilon}_1\grave{\varepsilon}_2$, has an expected upper-bound on its absolute efficiency given by*

$$\mathtt{E}(\grave{\varepsilon}_1, \grave{\varepsilon}_2 | \alpha_1, \alpha_2) = \frac{\log_2(\grave{\varepsilon}_1 + \grave{\varepsilon}_2 - \grave{\varepsilon}_1\grave{\varepsilon}_2)}{\alpha_1 \log_2 \grave{\varepsilon}_1 + \alpha_2 \log_2 \grave{\varepsilon}_2} \,, \tag{8.16}$$

[6] The *Singular Hash Set* [**?**] is an example of a data structure that obtains optimality using *brute-force* search, so it is not practical for even relatively small objective sets. However, its primary purpose is analytic tractability.

where

$$\begin{aligned} 0 < \alpha_1 &= \frac{|\mathcal{S}_1|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \leq 1\,, \\ 0 < \alpha_2 &= \frac{|\mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \leq 1\,, \\ 1 &\leq \alpha_1 + \alpha_2\,. \end{aligned} \tag{8.17}$$

As $\grave{\varepsilon}_j \to 1$ *or* $\grave{\varepsilon}_j \to 0$ *for* $j = 1, 2$*, or* $(\grave{\varepsilon}_1, \grave{\varepsilon}_2) \to (1, 1)$*, the absolute efficiency goes to* 0. [7]

Proof. The proof comes in two parts. First, we prove eq. (8.16), and then we prove the bounds on α_1 and α_2 given by eq. (8.17).

Let X and Y denote optimally space-efficient data structures that respectively implement positive approximate sets $\mathcal{S}_1^-$ and $\mathcal{S}_2^-$ with false positive rates $\grave{\varepsilon}_1$ and $\grave{\varepsilon}_2$. By **??**, their union has an induced false positive rate given by

$$\grave{\varepsilon}_1 + \grave{\varepsilon}_2 + \grave{\varepsilon}_1\grave{\varepsilon}_2\,. \tag{a}$$

The information-theoretic lower-bound of the approximate set of $\mathcal{S}_1 \cup \mathcal{S}_2$ with the above false positive rate is given by

$$-|\mathcal{S}_1 \cup \mathcal{S}_2| \log_2(\grave{\varepsilon}_1 + \grave{\varepsilon}_2 + \grave{\varepsilon}_1\grave{\varepsilon}_2) \text{ bits}\,. \tag{b}$$

Since we assume we only have X and Y and it is not possible to enumerate the elements in either, we must implement their union by storing and separately querying both X and Y. Since X and Y are optimal, $\ell(X) = -|\mathcal{S}_1| \log_2 \grave{\varepsilon}_1$ and $\ell(Y) = -|\mathcal{S}_2| \log_2 \grave{\varepsilon}_2$. Making these substitutions yields an absolute efficiency ???

$$\mathtt{E} = \frac{|\mathcal{S}_1 \cup \mathcal{S}_2| \log_2(\grave{\varepsilon}_1 + \grave{\varepsilon}_2 + \grave{\varepsilon}_1\grave{\varepsilon}_2)}{|\mathcal{S}_1| \log_2 \grave{\varepsilon}_1 + |\mathcal{S}_2| \log_2 \grave{\varepsilon}_2}\,. \tag{c}$$

Letting

$$\alpha_1 = \frac{|\mathcal{S}_1|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \text{ and } \alpha_2 = \frac{|\mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|}\,, \tag{d}$$

we may rewrite eq. (c) as

$$\frac{\log_2(\grave{\varepsilon}_1 + \grave{\varepsilon}_2 - \grave{\varepsilon}_1\grave{\varepsilon}_2)}{\alpha_1 \log_2 \grave{\varepsilon}_1 + \alpha_2 \log_2 \grave{\varepsilon}_2}\,. \tag{8.16 revisited}$$

In the second part of the proof, we prove the bounds on α_1 and α_2 as given by eq. (8.17). Both α_1 and α_2 must be non-negative since each is the ratio of two positive numbers (cardinalities). If $|\mathcal{S}_1| \ll |\mathcal{S}_2|$, then $\alpha_1 \approx 0$. If $\mathcal{S}_1 \supset \mathcal{S}_2$, then $\alpha_1 = 1$. A similar argument holds for α_2. Finally,

$$\alpha_1 + \alpha_2 = \frac{|\mathcal{S}_1| + |\mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \tag{e}$$

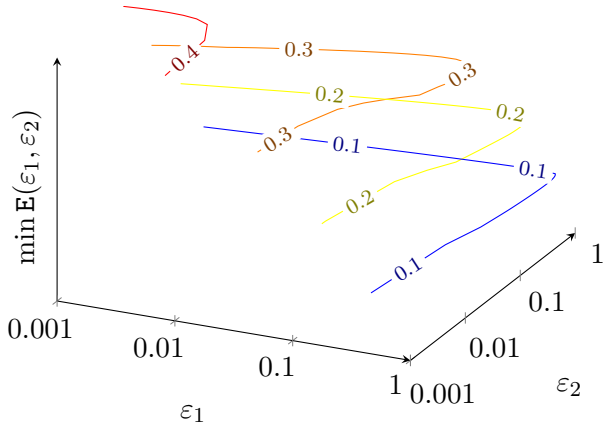
has a minimum value by assuming that $\mathcal{S}_1$ and $\mathcal{S}_2$ are disjoint sets (i.e., their intersection is the empty set), in which case

$$\alpha_1 + \alpha_2 = \frac{|\mathcal{S}_1| + |\mathcal{S}_2|}{|\mathcal{S}_1| + |\mathcal{S}_2|} = 1\,. \tag{f}$$

□

[7] As $(\grave{\varepsilon}_1, \grave{\varepsilon}_2) \to (0, 0)$, the absolute efficiency depends on the path taken. In most cases, it goes to 0.

Figure 8.1: Expected lower-bound on efficiency of the union of two approximate sets, neither of which can be enumerated.



See **??** for a contour plot of the expected lower-bound as a function of $\grave{\varepsilon}_1$ and $\grave{\varepsilon}_2$. As $\grave{\varepsilon}_1 \to 0$ or $\grave{\varepsilon}_2 \to 0$, the efficiency goes to 0.

The lower-bound on the efficiency of the union of approximate sets is given by the following corollary.

Corollary 8.1.1. *Given two positive, non-enumerable approximate sets with false positive rates* $\grave{\varepsilon}_1$ *and* $\grave{\varepsilon}_2$*, their union is an approximate set that has an efficiency that is expected to be greater than the lower bound given by*

$$\min \mathtt{E}(\grave{\varepsilon}_1, \grave{\varepsilon}_2) = \frac{\log_2(\grave{\varepsilon}_1 + \grave{\varepsilon}_2 - \grave{\varepsilon}_1\grave{\varepsilon}_2)}{\log_2 \grave{\varepsilon}_1\grave{\varepsilon}_2} \,. \tag{8.18}$$

Corollary 8.1.2. *If* $\varepsilon_1 = \varepsilon_2 = \varepsilon$*, then the absolute efficiency is given by*

$$\mathtt{E}(\grave{\varepsilon}|\alpha) = \left(1 + \frac{\log_2(2 - \grave{\varepsilon})}{\log_2 \grave{\varepsilon}}\right)\left(1 - \frac{\alpha}{2}\right) , \tag{8.19}$$

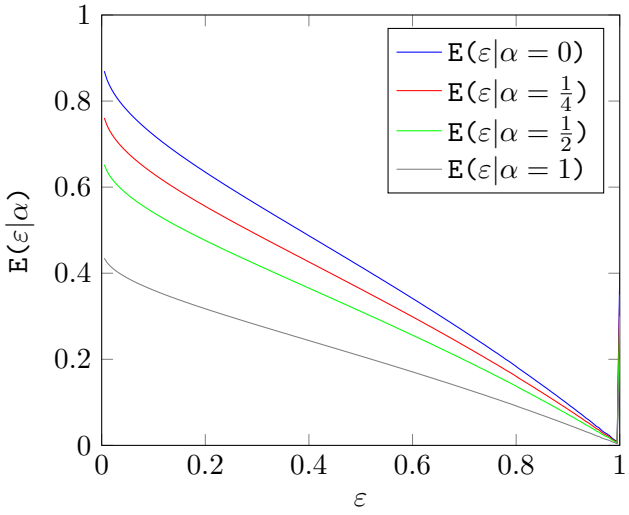
where

$$0 \leq \alpha = \frac{|\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \leq 1 \,, \tag{8.20}$$

which is a monotonically decreasing function with respect to ε *and* α *with limits given by* $\lim_{\grave{\varepsilon}\to 0} \mathtt{E}(\varepsilon) = 1$ *and* $\lim_{\varepsilon \to 1} \mathtt{E}(\varepsilon) = 0$.

See fig. 8.2 for a graphic illustration.

Figure 8.2: Expected lower-bound on efficiency of the union of two approximate sets with the same false positive rate ε, neither of which can be enumerated.



Chapter 9

Application: Encrypted set-theoretic Boolean search

TODO: let's do the subset relation instead! just to make it a little more sophisticated/interesting. refer to oblivious types also, i.e., a hidden query is an oblivious type and the secure indexes are oblivious sets of the oblivious types. in this case, the oblivious types represent elements of the powerset of keys up to k-graphs?

An information retrieval process begins when a user submits a *query* to an information system, where a query represents an *information need*. In response, the information system returns a set of relevant documents that satisfy the query.

Boolean search is an information retrieval model given by the following definition.

Definition 9.1. *A document in the collection is either* relevant *or* non-relevant *to a Boolean query.*

We consider queries over the Boolean algebra $Q = \left(2^{\mathcal{K}}, \wedge, \vee, \neg, \epsilon, \mathcal{K}\right)$, where $\mathcal{K}$ denotes a set of *search keys*, e.g., units of information like English words. This is isomorphic to the Boolean algebra $Q' = \left(\{0,1\}^k, \wedge, \vee, \neg, 0^k, 1^k\right)$ where $k = |\mathcal{K}|$.

We denote the set of documents by $\mathcal{D}$. *Search indexes* may be used to quickly compute which subset of $\mathcal{D}$ is relevant to a given query. Since the queries are a Boolean algebra over the powerset of $\mathcal{K}$, we have chosen a simple model given by the following:

1. Search indexes model *sets* in $2^{\mathcal{K}}$, i.e., search indexes are over the Boolean algebra $S = (2^{\mathcal{K}}, \cap, \cup, \overline{\ }, \varnothing, \mathcal{K})$. In particular, let $\texttt{index}\colon \mathcal{D} \mapsto 2^{\mathcal{K}}$ be a function that maps documents to search indexes with a definition given by

$$\texttt{index}(d) \coloneqq \{\, k \in \mathcal{K} \mid k \text{ is relevant to } d \,\}. \tag{9.1}$$

2. There is a one-to-one correspondence between $\mathcal{D}$ and the set of search indexes $\{\, \texttt{index}(d) \in 2^{\mathcal{K}} \mid d \in \mathcal{D} \,\}$.

3. A bijection $\mathrm{f}\colon Q \mapsto S$ is given by the mappings $\wedge \mapsto \cap$, $\vee \mapsto \cup$, $\neg \mapsto \overline{\ }$, and $\epsilon \mapsto \varnothing$.

4. A document $d \in \mathcal{D}$ is relevant to a Boolean query q in Q if the the query maps to a *subset* of $\texttt{index}(d)$, e.g., document d is relevant to $(a \wedge b) \vee c$ if $\{a, b\} \subseteq \texttt{index}(d)$ or $\{c\} \subseteq \texttt{index}(d)$.

5. Let $\texttt{find}\colon Q \mapsto 2^{\mathcal{D}}$ be a function that maps a query q in Q to the *subset* of $\mathcal{D}$, denoted the *result set*, that is relevant to q. In particular,

$$\texttt{find}(q) \coloneqq \{\, d \in \mathcal{D} \mid \mathrm{f}(q) \subseteq \texttt{index}(d) \,\}, \tag{9.2}$$

which forms the Boolean algebra $R = (2^{\mathcal{D}}, \cap, \cup, \overline{\ }, \varnothing, \mathcal{D})$, i.e.,

An implementation of Boolean search is given by algorithm 1, which implements a function `find`: [query] $\mapsto 2^{\mathcal{D}}$ where $2^{\mathcal{D}}$ is the result set.

Algorithm 1: Pseudo-code for `find`

params : $\mathcal{D}$, the collection of documents.
input : q, a Boolean query of type Q in *disjunctive normal form*.
output : $\mathcal{R}_q$, the subset of documents in $\mathcal{D}$ relevant to query q.

```
1  function find(q)
       // head gets the outer most operation or terminal key.
2      h ← head(q);
3      if h = not then
4          t ← tail(q);
5          R_t ← find(t);
6          return complement(R_t);
7      else if h = or then
8          t ← tail(q);
9          R_l ← find(left(t));
10         R_r ← find(right(t));
11         return R_l ∪ R_r;
12     else
           // h must be a key.
13         R_h ← ∅;
14         for d ∈ D do
15             if h ∈ index(d) then
16                 R_h ← R_h ∪ {d};
17         return R_h;
```

9.1 Secure indexes based on the random approximate set model

We consider an *approximation* of the set-theoretic *Boolean search* model where the Boolean search indexes are *secure indexes* based on *random approximate sets*, which is an appropriate abstract data type in *Encrypted Search* [**?**] where typical search indexes reveal too much information to untrusted third-parties. We denote the resulting find function that searches over the random approximate sets by $\texttt{find}^{\sigma}$ as opposed to the objective function `find`.

This replacement *induces* random approximate *result sets*, i.e., $\texttt{find}^{\sigma}(q)$ maps to a random approximate set of the set that $\texttt{find}(q)$ maps to. Note that $\texttt{find}^{\sigma}$ is a deterministic algorithm that always generates the same output (result set) for the same input (query), i.e., $\texttt{find}^{\sigma}$ is still a function rather than a distribution, but it is compatible with the random approximate set model described in **??**.

Remark. *The function* $\texttt{find}^{\sigma}$ *as described is not by itself an implementation of Encrypted Search. A weak implementation may be based off of a simple substitution cipher from keys to* trapdoors *using a one-way cryptographic hash function* $\mathrm{h}\colon \mathcal{B}^* \mapsto \mathcal{B}^k$. *Suppose we have a function* $\mathrm{H}\colon [q] \mapsto [q']$ *that maps set-theoretic queries on keys to equivalent set-theoretic queries on trapdoors using* h.

Then, Encrypted Search's find function is the composition given by

$$\texttt{hidden\_find}^{\sigma} = \texttt{find}^{\sigma} \circ \mathrm{H} \colon [q] \mapsto 2^{\mathcal{D}} \,. \tag{9.3}$$

Generally, H *is computed on a* trusted machine *and the output, the set-theoretic query of trapoors, is sent to an untrusted machine that executes* $\texttt{find}^{\sigma}$*.*

The simple substitution cipher is not a sophisticated approach since the set-theoretic queries on trapdoors have the same entropy as the set-theoretic queries on keys. Thus, using entropy as a measure of confidentiality, this solution does not increase the confidentiality, especially against an adversary with a sufficiently large sample of queries.[1]

△

Theorem 9.1. *The result set* $\mathcal{R}^{x}$ *that is relevant to key* x *is a* random approximate set *of the* objective *result* $\mathcal{R}_{x}$*.*[2]

Proof. A search index $\mathcal{S}^{\pm}$ with a false positive rate ε and false negative rate ω is relevant to a key x if the key x tests positive in $\mathcal{S}^{\pm}$. A false positive occurs if the key x is not in $\mathcal{S}$ but is in $\mathcal{S}^{\pm}$, which occurs with some probability $\varepsilon \geq 0$. A false negative occurs if a key x is in $\mathcal{S}$ but is not in $\mathcal{S}^{\pm}$, which occurs with some probability $\omega \geq 0$. □

We have established that the result sets of a single atomic key are approximate result sets. We may now apply the set-theoretic results in **??** to implement the full set-theoretic model for approximate sets.

Example 6 *Suppose the search indexes are positive approximate sets each with a false positive rate* ε*. A common type of Boolean query is the* intersection *(conjunction) of atomic keys. Consider a conjunctive query of* k *keys,* $x_1, \ldots, x_k$*. The result set* $\mathcal{R}^{-}\left(\{x_1\} \cap \cdots \cap \{x_k\}\right) = \texttt{find}(\overline{\overline{x_1} \cup \cdots \cup \overline{x_k}})$ *is a positive approximate set with an uncertain false positive rate* $[\varepsilon_k] = [\varepsilon^k, \varepsilon]$*.*

Proof. Let the approximate result set for key x_j be denoted by $\mathcal{R}^{-}_{x_j}$. The result set is given by

$$\mathcal{R}^{-}_{\mathcal{X}} = \bigcap_{j=1}^{k} \mathcal{R}^{-}_{j} \,. \tag{a}$$

By **??**, $\mathcal{R}^{-}_{j}$ has a false positive rate ε for $j \in [1, \ldots, k]$. By **??**, $\mathcal{R}^{-}_{1} \cap \mathcal{R}^{-}_{2}$ has a false positive rate $[\varepsilon^2, \varepsilon]$. Similarly,

$$\left(\mathcal{R}^{-}_{1} \cap \mathcal{R}^{-}_{2}\right) \cap \mathcal{R}^{-}_{3} = \mathcal{R}^{-}_{1} \cap \mathcal{R}^{-}_{2} \cap \mathcal{R}^{-}_{3} \tag{b}$$

has a false positive rate $[\varepsilon^3, \varepsilon]$. Continuing in this fashion, we see that $\mathcal{R}^{-}_{\mathcal{X}} = \mathcal{R}^{-}_{1} \cap \cdots \cap \mathcal{R}^{-}_{k}$ has a false positive rate $[\varepsilon^k, \varepsilon]$. □

To quantify the performance measure of the information retrieval system, we may use the binary classification results in section 5.4.

Appendix

[1] For instance, it may be highly suspectible to known-plaintext attacks.

[2] Disregarding the notion that the approximate result sets may also have obfuscated references that are homomorphic to the objective result sets.

Appendix A

Additional proofs

In what follows, we show additional or more rigorous proofs that are not necessarily central to the paper but warrant mention.

A.1 Proof of corollary 4.2.2

By corollary 4.2.2, given *countably infinite* negatives, a random approximate set with a false positive rate ε is *certain* to obtain ε.

Proof. Hoeffding's inequality [**?**] provides that FP_n is concentrated around its mean $n\varepsilon$ as given by

$$\mathrm{P}\big[(\varepsilon-\epsilon)n \leq \mathrm{FP}_n \leq (\varepsilon+\epsilon)n\big] \geq 1-2\exp\Big(-2\epsilon^2 n\Big)\,, \tag{a}$$

where $\epsilon > 0$. We are interested in the limiting probability

$$\begin{aligned}\lim_{n\to\infty} \Pr\big[(\varepsilon-\epsilon)n \leq \mathrm{FP}_n \leq (\varepsilon+\epsilon)n\big] = \\ \lim_{n\to\infty}\left\{1-2\exp\Big(-2\epsilon^2 n\Big)\right\} = 1\,.\end{aligned} \tag{b}$$

As ϵ goes to 0, $\lim_{n\to\infty}\mathrm{FP}_n$ converges almost surely to εn and therefore $\lim_{n\to\infty}\mathrm{FP}_n/n$ converges almost surely to ε. □

A.2 Proof of theorem 5.6

Given p positives and n negatives, by **??** an approximate set with a false positive rate ε and a false negative rate ω has an *expected* precision given *approximately* by

$$\mathtt{ppv}(\omega,\varepsilon;n,p) \approx \frac{\bar{t}_p}{\bar{t}_p+\overline{f}_p} + \frac{\bar{t}_p\sigma_{f_p}^2 - \overline{f}_p\sigma_{t_p}^2}{\Big(\bar{t}_p+\overline{f}_p\Big)^3}\,, \tag{?? revisited}$$

where $\bar{t}_p = p\tau$ is the *expected* number of *true positives*, $\overline{f}_p = n\varepsilon$ is the *expected* number of *false positives*, $\sigma_{t_p}^2 = p\omega\tau$ is the variance of the number of *true positives*, and $\sigma_{f_p}^2 = n\varepsilon\omega$ is the variance of the number of *false positives*.

Proof. The positive predictive value is a random variable given by

$$\frac{\mathrm{TP}_p}{\mathrm{TP}_p + \mathrm{FP}_n}\,. \tag{a}$$

We are interested in the *expected* positive predictive value,

$$\mathtt{ppv}(\varepsilon, \tau) = \mathrm{E}\left[\frac{\mathrm{TP}_p}{\mathrm{TP}_p + \mathrm{FP}_n}\right]. \tag{b}$$

This expectation is of a non-linear function of random variables, which is problematic so we choose to approximate the expectation.

Let the *positive predictive value* function be denoted by

$$\mathrm{f}(t_p, f_p) = \frac{t_p}{t_p + f_p}\,, \tag{c}$$

where t_p is the number of true positives and f_p is the number of false positives. We approximate this function with a second-order Taylor series. The gradient of f is given by

$$\nabla\mathrm{f}(t_p, f_p) = \frac{1}{(t_p + f_p)^2}\begin{bmatrix} f_p \\ -t_p \end{bmatrix} \tag{d}$$

and the Hessian of f is given by

$$\mathcal{H}(t_p, f_p) = \frac{1}{(t_p + f_p)^3}\begin{bmatrix} -2f_p & t_p - f_p \\ t_p - f_p & 2t_p \end{bmatrix}. \tag{e}$$

A linear approximation g of f that is reasonably accurate near the expected value of TP_p, denoted by $\bar{t}_p$, and the expected value of FP_n, denoted by $\overline{f}_p$, is given by

$$\mathrm{g}(t_p, f_p) = \mathrm{f}\left(\bar{t}_p, \overline{f}_p\right) + \nabla\mathrm{f}(\bar{t}_p, \overline{f}_p])^\intercal \begin{bmatrix} t_p - \bar{t}_p \\ f_p - \overline{f}_p \end{bmatrix} + \frac{1}{2}\begin{bmatrix} t_p - \bar{t}_p \\ f_p - \overline{f}_p \end{bmatrix}^\intercal \mathcal{H}(\bar{t}_p, \overline{f}_p) \begin{bmatrix} t_p - \bar{t}_p \\ f_p - \overline{f}_p \end{bmatrix}. \tag{f}$$

As a function of random variables TP_p and FP_n, $\mathrm{g}(\mathrm{TP}_p, \mathrm{FP}_n)$ is a random variable. Since $\mathrm{E}[\mathrm{TP}_p - \bar{t}_p] = 0$ and $\mathrm{E}\left[\mathrm{FP}_p - \overline{f}_p\right] = 0$, we immediately simplify the expectation of g to

$$\mathrm{E}[\mathrm{g}(\mathrm{TP}_p, \mathrm{FP}_n)] = \frac{\bar{t}_p}{\bar{t}_p + \overline{f}_p} + \frac{\mathrm{E}[A]}{(\bar{t}_p + \overline{f}_p)^3} \tag{g}$$

where

$$A = \frac{1}{2}\begin{bmatrix} \mathrm{TP}_p - \bar{t}_p \\ \mathrm{FP}_n - \overline{f}_p \end{bmatrix}^\intercal \begin{bmatrix} -2\overline{f}_p & \bar{t}_p - \overline{f}_p \\ \bar{t}_p - \overline{f}_p & 2\bar{t}_p \end{bmatrix} \begin{bmatrix} \mathrm{TP}_p - \bar{t}_p \\ \mathrm{FP}_n - \overline{f}_p \end{bmatrix}. \tag{h}$$

Multiplying the right column matrix by the Hessian matrix in A results in

$$A = \frac{1}{2}\begin{bmatrix} \mathrm{TP}_p - \bar{t}_p \\ \mathrm{FP}_n - \overline{f}_p \end{bmatrix}^\intercal \begin{bmatrix} -2\overline{f}\,(\mathrm{TP}_p - \bar{t}_p) + \left(\bar{t}_p - \overline{f}_p\right)\left(\mathrm{FP}_n - \overline{f}_p\right) \\ \left(\bar{t}_p - \overline{f}_p\right)(\mathrm{TP}_p - \bar{t}_p) + 2\bar{t}_p\left(\mathrm{FP}_n - \overline{f}_p\right) \end{bmatrix} \tag{i}$$

Multiplying the left column matrix by the right column matrix in A results in

$$A = -\overline{f}_p \left(\mathrm{TP}_p - \bar{t}_p\right)^2 + \left(\bar{t}_p - \overline{f}_p\right)\left(\mathrm{TP}_p - \bar{t}_p\right)\left(\mathrm{FP}_n - \overline{f}_p\right) + \bar{t}_p \left(\mathrm{FP}_n - \overline{f}_p\right)^2 . \tag{j}$$

As a linear operator, the expectation of A is equivalent to

$$\mathrm{E}[A] = -\overline{f}_p \,\mathrm{E}\big[\mathrm{TP}_p - \bar{t}_p\big]^2 + \left(\bar{t} - \overline{f}_p\right)\mathrm{E}\left[\left(\mathrm{FP}_n - \overline{f}_p\right)\left(\mathrm{TP}_p - \bar{t}_p\right)\right] + \bar{t}_p \,\mathrm{E}\left[\mathrm{FP}_n - \overline{f}_p\right]^2 . \tag{k}$$

By definition, $\mathrm{E}\big[\mathrm{TP}_p - \bar{t}_p\big]^2$ is the variance of TP_p, $\mathrm{E}\left[\mathrm{FP}_n - \overline{f}_p\right]^2$ is the variance of FP_n, and $\mathrm{E}\left[\left(\mathrm{FP}_n - \overline{f}_p\right)\left(\mathrm{TP}_p - \bar{t}_p\right)\right]$ is the covariance of TP_p and FP_n, which is 0 since they are independent. Making these substitutions results in

$$\mathrm{E}[A] = \bar{t}_p \,\mathrm{var}[\mathrm{FP}_n] - \overline{f}_p \,\mathrm{var}\big[\mathrm{TP}_p\big] . \tag{l}$$

Substituting this result into eq. (g) yields

$$\mathrm{E}\big[\mathrm{g}(\mathrm{TP}_p, \mathrm{FP}_n)\big] = \frac{\bar{t}_p}{\bar{t}_p + \overline{f}_p} + \frac{-\overline{f}_p \,\mathrm{var}\big[\mathrm{TP}_p\big] + \bar{t}_p \,\mathrm{var}[\mathrm{FP}_n]}{(\bar{t}_p + \overline{f}_p)^3} \tag{m}$$

By theorem 4.1, FP_n is binomially distributed with a mean $n\varepsilon$ and a variance $n\varepsilon\eta$ and by corollary 4.4.2, TP_p is binomailly distributed with a mean $p\tau$ and a variance $p\omega\tau$. □

A.3 Proof of theorem 5.3

Given $\mathcal{A}^{\pm}(\varepsilon_1, \omega)$ and $\mathcal{B}^{\pm}(\varepsilon_2, \omega_2)$, their union is an *approximate set* with a false negative rate given by

$$\begin{aligned} \omega = {} & \alpha_1\omega_1(1-\varepsilon_2) + \alpha_2\omega_2(1-\varepsilon_1) \\ & + (1-\alpha_1-\alpha_2)\omega_1\omega_2 , \end{aligned} \tag{5.2 revisited}$$

where

$$\begin{aligned} 0 \le \alpha_1 &= \frac{|\mathcal{S}_1 \smallsetminus \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} , \\ 0 \le \alpha_2 &= \frac{|\mathcal{S}_2 \smallsetminus \mathcal{S}_1|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} , \\ \alpha_1 + \alpha_2 &\le 1 . \end{aligned} \tag{5.3 revisited}$$

Proof. Suppose we have two sets $\mathcal{S}_1$ and $\mathcal{S}_2$. The false negative rate ω is a probability conditioned on a positive in the union of sets $\mathcal{S}_1$ and $\mathcal{S}_2$ being a negative in the union of approximate sets $\mathcal{S}^1$ and $\mathcal{S}^2$.

The set of *possible* false negatives is the set of positives, $\mathcal{S}_1 \cup \mathcal{S}_2$, which is equivalent to the union of the disjoint sets $\mathcal{S}_1 \cap \mathcal{S}_2$, $\mathcal{S}_1 \smallsetminus \mathcal{S}_2$ and $\mathcal{S}_2 \smallsetminus \mathcal{S}_1$.

The false negative rate is equivalent to ratio of the *expected* number of false negatives to the maximum possible false negatives $|\mathcal{S}_1 \cup \mathcal{S}_2|$. Since they are disjoint, we may consider each independently to calculate the expected total number of false negatives.

Let A_1 denote the event $\mathrm{X} \in \mathcal{S}^1$, A_2 denote $\mathrm{X} \in \mathcal{S}^2$, B_1 denote $\mathrm{X} \in \mathcal{S}_1$, and B_2 denote $\mathrm{X} \in \mathcal{S}_2$. Suppose we randomly select an element from $\mathcal{S}_1 \cap \mathcal{S}_2$. The probability that X is a negative in $\mathcal{S}^1 \cup \mathcal{S}^2$ given that it is positive in $\mathcal{S}_1 \cap \mathcal{S}_2$ is given by

$$\omega_{1\cap 2} = \mathrm{P}\Big[(A_1 \cup A_2)' \;\Big|\; B_1 \cap B_2\Big]\,. \tag{a}$$

By De Morgan's law, $(A_1 \cup A_2)' \equiv A_1' \cap A_2'$. Making this substitution results in

$$\omega_{1\cap 2} = \mathrm{P}\Big[A_1' \cap A_2' \;\Big|\; B_1 \cap B_2\Big]\,. \tag{b}$$

Since A_1 and A_2 are independent events, by the rules of probability

$$\omega_{1\cap 2} = \mathrm{P}\Big[A_1' \;\Big|\; B_1 \cap B_2\Big]\,\mathrm{P}\Big[A_2' \;\Big|\; B_1 \cap B_2\Big]\,. \tag{c}$$

Since A_1 is independent of B_2 and A_2 is independent of B_1, by the rules of probability

$$\omega_{1\cap 2} = \mathrm{P}\Big[A_1' \;\Big|\; B_1\Big]\,\mathrm{P}\Big[A_2' \;\Big|\; B_2\Big]\,. \tag{d}$$

By definition, $\mathrm{P}\Big[A_j' \;\Big|\; B_j\Big]$ is the false negative rate ω_j. Making this substitution yields

$$\omega_{1\cap 2} = \omega_1\omega_2\,. \tag{e}$$

There are $|\mathcal{S}_1 \cap \mathcal{S}_2|$ elements in $\mathcal{S}_1 \cap \mathcal{S}_2$, where each is an independent Bernoulli trial. Thus, there are expected to be

$$|\mathcal{S}_1 \cap \mathcal{S}_2|\omega_{1\cap 2} = |\mathcal{S}_1 \cap \mathcal{S}_2|\omega_1\omega_2 \tag{f}$$

false negatives in $\mathcal{S}_1 \cap \mathcal{S}_2$.

Suppose we randomly select an element from $\mathcal{S}_1 \smallsetminus \mathcal{S}_2$. The probability that X is a negative in $\mathcal{S}^1 \cup \mathcal{S}^2$ given that it is a positive in $\mathcal{S}_1 \smallsetminus \mathcal{S}_2$ is given by

$$\omega_{1\cap\overline{2}} = \mathrm{P}\Big[A_1' \cap A_2' \;\Big|\; B_1 \cap B_2'\Big]\,. \tag{g}$$

Since A_1 and A_2 are independent events, this may be rewritten as

$$\omega_{1\cap\overline{2}} = \mathrm{P}\Big[A_1' \;\Big|\; B_1 \cap B_2'\Big]\,\mathrm{P}\Big[A_2' \;\Big|\; B_1 \cap B_2'\Big]\,. \tag{h}$$

Since A_1 is independent of B_2 and A_2 is independent of B_1, this may be rewritten as

$$\omega_{1\cap\overline{2}} = \mathrm{P}\Big[A_1' \;\Big|\; B_1\Big]\,\mathrm{P}\Big[A_2' \;\Big|\; B_2'\Big]\,. \tag{i}$$

By definition, $\mathrm{P}\big[A_1' \mid B_1\big]$ is the false negative rate ω_1 and $\mathrm{P}\big[A_2 \mid B_2'\big]$ is the false positive rate ε_2. Thus,

$$\omega_{1\cap\overline{2}} = \omega_1(1-\varepsilon_2)\,. \tag{j}$$

There are $|\mathcal{S}_1 \smallsetminus \mathcal{S}_2|$ elements in $\mathcal{S}_1 \smallsetminus \mathcal{S}_2$, where each is an independent Bernoulli trial. Thus, there are expected to be

$$|\mathcal{S}_1 \smallsetminus \mathcal{S}_2|\omega_{1\cap\overline{2}} = |\mathcal{S}_1 \smallsetminus \mathcal{S}_2|\omega_1(1-\varepsilon_2) \tag{k}$$

false negatives in $\mathcal{S}_1 \smallsetminus \mathcal{S}_2$. A similar argument follows for $\mathcal{S}_2 \smallsetminus \mathcal{S}_1$ where there are expected to be

$$|\mathcal{S}_2 \smallsetminus \mathcal{S}_1|\omega_2(1-\varepsilon_2) \tag{l}$$

false negatives.

The false negative rate is given by the ratio of the total expected number of false negatives given by eqs. (f), (k) and (l) to the total number of possible false negatives $|\mathcal{S}_1 \cup \mathcal{S}_2|$, which is given by

$$\omega = \frac{|\mathcal{S}_1 \setminus \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|}\omega_1(1-\varepsilon_2) + \frac{|\mathcal{S}_2 \setminus \mathcal{S}_1|}{|\mathcal{S}_1 \cup \mathcal{S}_2|}\omega_2(1-\varepsilon_1) + \frac{|\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|}\omega_1\omega_2 \,. \tag{m}$$

If we let

$$\alpha_1 = \frac{|\mathcal{S}_1 \setminus \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \text{ and } \alpha_2 = \frac{|\mathcal{S}_2 \setminus \mathcal{S}_1|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \,, \tag{n}$$

then

$$1 - \alpha_1 - \alpha_2 = \frac{|\mathcal{S}_1 \cap \mathcal{S}_2|}{|\mathcal{S}_1 \cup \mathcal{S}_2|} \,. \tag{o}$$

Making these substitutions into eq. (m) yields the result

$$\begin{aligned} \omega = \alpha_1\omega_1(1-\varepsilon_2) + \alpha_1\omega_2(1-\varepsilon_1) + \\ (1-\alpha_1-\alpha_2)\omega_1\omega_2 \,. \end{aligned} \tag{p}$$

□