Bernoulli sets: a model for modeling sets with random errors
and 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

1 Introduction

Conceptually, an *approximate set* is a set that approximates some other set that is of objective interest. That is to say, the approximation has *errors* with respect to its *member-of* relations. Ideally, no errors would occur, but typically, we either apply lossy data compression techniques to reduce bit rate (or bit length), which introduce errors (rate-distortion), or apply data redundancy techniques to reduce errors caused by noise, which increases bit rate (or bit length).

The *Bloom filter* is a popular example of a data structure and algorithm that models the concpet of generating approximate sets that only includes *false positives* due to *rate distortion*.

In section 2, we define the algebra of sets.

In section 2.1, we provide a formal definition of the *Bernoulli set* model, in which the error rates, such as false positive or false negative rates, are *expectations*.

We describe the axioms of the Bernoulli set model such that, if satisfied, also satisfy the axioms of the approximate algebra of sets.

We further derive the probability distribution of Bernoulli sets entailed by the axioms.

In section 4, we derive the random variables that are fundamental to the Bernoulli set model.

In **??**, we provide a detailed treatment on distributions that are induced by functions that depend on random approximate sets, e.g., in **??** we derive the probability distribution of random approximate sets that are generated from arbitrary set-theoretic operations on random approximate sets and in section 4.2 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*.

In section 6, we provide the probabilistic model for random approximate sets with *uncertain* rate distortions, such as an uncertain false positive rate.

In section 7, 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.

Finally, in **??**, 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.

2 Algebra of sets

A *set* is an unordered collection of distinct elements. If we know the elements in a set, we may denote the set by these elements, e.g., $\{a, c, b\}$ denotes a set whose members are exactly a, b, and c.

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.

A *finite set* has a finite number of elements. For example, $\{1, 3, 5\}$ is a finite set with three elements. When sets $\mathcal{A}$ and $\mathcal{B}$ are *isomorphic*, denoted by $\mathcal{A} \cong \mathcal{B}$, they can be put into a one-to-one correspondence (bijection), e.g., $\{b, a, c\} \cong \{1, 2, 3\}$.

The cardinality of a finite set $\mathcal{A}$ is the number of elements in the set, denoted by $|\mathcal{A}|$, e.g., $|\{1, 3, 5\}| = 3$. A *countably infinite set* is isomorphic to the set of *natural numbers* $\mathbb{N} \coloneqq \{1, 2, 3, 4, 5, \ldots\}$.

Given two elements a and b, an ordered pair of a then b is denoted by $\langle a\,, b\rangle$, where $\langle a\,, b\rangle = \langle c\,, d\rangle$ *if and only if* $a = c$ and $b = d$. Ordered pairs are non-commutative and non-associative, i.e., $\langle a\,, b\rangle \neq \langle b\,, a\rangle$ if $a \neq b$ and $\langle a\,, \langle b\,, c\rangle\rangle \neq \langle\langle b\,, a\rangle\,, c\rangle$.

Related to the ordered pair is the Cartesian product.

Definition 2.1. *The set* $\mathcal{X} \times \mathcal{Y} \coloneqq \big\{\langle x\,, y\rangle \colon x \in \mathcal{X} \wedge y \in \mathcal{Y}\big\}$ *is the Cartesian product of sets* $\mathcal{X}$ *and* $\mathcal{Y}$.

By the non-commutative and non-associative property of ordered pairs, the Cartesian product is non-commutative and non-associative. However, they are isomorphic, i.e., $\mathcal{X} \times \mathcal{Y} \cong \mathcal{Y} \times \mathcal{X}$.

A *tuple* is a generalization of order pairs which can consist of an arbitrary number of elements, e.g., $\langle x_1, x_2, \ldots, x_n\rangle$.

Definition 2.2 (n-fold Cartesian product)**.** *The* n*-ary Cartesian product of sets* $\mathcal{X}_1, \ldots, \mathcal{X}_n$*, is given by* $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n = \big\{\langle x_1, \ldots, x_n\rangle \colon x_1 \in \mathcal{X}_1 \wedge \cdots \wedge x_n \in \mathcal{X}_n\big\}$.

Note that $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \cong \mathcal{X}_1 \times (\mathcal{X}_2 \times \mathcal{X}_3) \cong (\mathcal{X}_1 \times \mathcal{X}_2) \times \mathcal{X}_3$, thus we may implicitly convert between them without ambiguity.

If each set in the n-ary Cartesian product is the same, the power notation may be used, e.g., $\mathcal{X}^3 \coloneqq \mathcal{X} \times \mathcal{X} \times \mathcal{X}$. As special cases, $\mathcal{X}^0 \coloneqq \{\varnothing\}$ and $\mathcal{X}^1 \coloneqq \mathcal{X}$.

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 $\mathcal{A} = \mathcal{B}$.

Definition 2.3. *Set builder notation*

Definition 2.4. *A* function *of type* $\mathcal{X} \mapsto \mathcal{Y}$ *is a binary relation on* $\mathcal{X} \times \mathcal{Y}$ *with the constraint that each* $x \in \mathcal{X}$ *is paired with exactly one* $y \in \mathcal{Y}$.

A function of type $\mathcal{X} \mapsto \mathcal{Y}$ has a domain $\mathcal{X}$ and a codomain $\mathcal{Y}$. Since every $x \in \mathcal{X}$, given a pair $\langle x\,, y\rangle \in \mathrm{f}$, y may also be denoted by $\mathrm{f}(x)$.

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., $2^{\{a,b\}} = \big\{\varnothing, \{a\}, \{b\}, \{a, b\}\big\}$.

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

$$\mathbb{1}_{\mathcal{A}} \colon \mathcal{X} \mapsto \{0, 1\} \tag{2.1}$$

defined as

$$\mathbb{1}_{\mathcal{A}}(x) \coloneqq \begin{cases} 0 & \text{if } x \notin \mathcal{A}\,, \\ 1 & \text{if } x \in \mathcal{A}\,. \end{cases} \tag{2.2}$$

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

A few important operations on sets are *set-union* $\cup \colon 2^{\mathcal{X}} \times 2^{\mathcal{X}} \mapsto 2^{\mathcal{X}}$, *set-intersection*, $\cap \colon 2^{\mathcal{X}} \times 2^{\mathcal{X}} \mapsto 2^{\mathcal{X}}$, and *set-complement* $\overline{} \colon 2^{\mathcal{X}} \mapsto 2^{\mathcal{X}}$, respectively defined as

$$\mathcal{A} \cup \mathcal{B} \coloneqq \{\, x \in \mathcal{X} \mid x \in \mathcal{A} \lor x \in \mathcal{B} \,\}\,, \tag{2.3}$$

$$\mathcal{A} \cap \mathcal{B} \coloneqq \{\, x \in \mathcal{X} \mid x \in \mathcal{A} \land x \in \mathcal{B} \,\}\,, \text{ and} \tag{2.4}$$

$$\overline{\mathcal{A}} \coloneqq \{\, x \in \mathcal{X} \mid x \notin \mathcal{A} \,\}\,. \tag{2.5}$$

where $\lor$ and $\land$ are respectively the logical-connectives *or* and *and*. If $\mathcal{A} \cap \mathcal{B} = \varnothing$, then we say $\mathcal{A}$ and $\mathcal{B}$ are *disjoint* sets.

2.1 Approximate sets

Given an objective set $\mathcal{A}$, any element that is a member of $\mathcal{A}$ is denoted a *positive* (of $\mathcal{A}$) and otherwise the element is denoted a *negative*. Suppose $\hat{\mathcal{A}}$ is used as an *approximation* of $\mathcal{A}$. If the *only* information we have about $\mathcal{A}$ is given by $\hat{\mathcal{A}}$, then we may perform membership tests on $\hat{\mathcal{A}}$ to make *predictions* or *estimations* about $\mathcal{A}$.

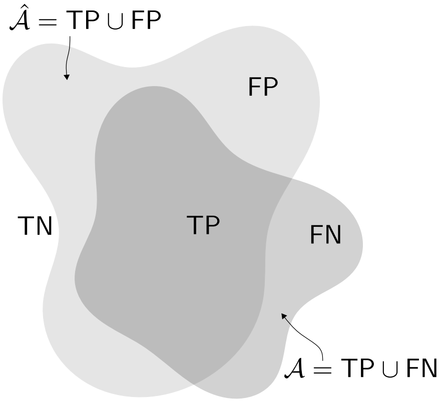
There are two ways a binary prediction can be false.

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*.

If we denote the set of false positives by FP, true positives by TP, false negatives by FN, and true negatives by TN, then the objective set $\mathcal{A} = \mathsf{FN} \cup \mathcal{TP}$ and the approximate set $\hat{\mathcal{A}} = \mathsf{TP} \cup \mathsf{FP}$. See fig. 1 for an illustration.

Figure 1: An approximate set $\hat{\mathcal{A}}$ of an objective set $\mathcal{A}$


$\hat{\mathcal{A}} = \mathsf{TP} \cup \mathsf{FP}$
FP
TN
TP
FN
$\mathcal{A} = \mathsf{TP} \cup \mathsf{FN}$

If we only have access to the approximation $\hat{\mathcal{A}}$, we do not have the ability to partition the universe into the disjoint sets FP, TP, FN, and TN as demonstrated in fig. 1. However, we can quantify the degree of *uncertainty* about the elements that it predicts to be positive or negative. The false positive and true negative rates are given by the following.

Definition 2.5. *The* false positive rate *is the proportion of predictions that are* false positives *as given by*

$$\hat{\varepsilon} = \frac{|\mathsf{FP}|}{|\mathsf{FP}| + |\mathsf{TN}|} \tag{2.6}$$

and the true negative rate *is given by* $\hat{\eta} = 1 - \hat{\varepsilon}$.

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

Definition 2.6. *The* true positive rate *is the proportion of predictions that are* true positives *as given by*

$$\hat{\tau} = \frac{|\mathsf{TP}|}{|\mathsf{TP}| + |\mathsf{FN}|} \tag{2.7}$$

and the false negative rate *is given by* $\hat{\omega} = 1 - \hat{\tau}$.

The *probabilities* of the four possible predictive outcomes are given by table 1.

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

Table 1: The 2 × 2 contingency table of outcomes for approximate sets.

3 Bernoulli set model

In the *Bernoulli* set model, we describe the statistical properties of processes that *generate* approximations of a certain kind that model many existing processes and generalizes to higher-order approximations under algebraic composition.

In what follows, we specify the axioms of the Bernoulli set model.

Theoretically, a process that generates approximations could exhibit correlations of any sort, but Bernoulli sets are constrained by the following axiom.

Axiom 1. *Each element of a Bernoulli set has an independently distributed error rate,*

$$\mathrm{P}\big[\mathbb{1}_{\mathcal{A}^\sigma}(x) \neq \mathbb{1}_{\mathcal{A}}(x) \;\big|\; \mathbb{1}_{\mathcal{A}^\sigma}(y) \neq \mathbb{1}_{\mathcal{A}}(y)\big] = \mathrm{P}\big[\mathbb{1}_{\mathcal{A}^\sigma}(x) \neq \mathbb{1}_{\mathcal{A}}(x)\big]\,. \tag{3.1}$$

The complexity in the Bernoulli set model stems from the fact that different subsets of the universal set may exhibit different error rates.

Axiom 2. *An n-th order random approximate set with random errors* $\Delta_1, \ldots, \Delta_n$ *respectively for partition blocks* $1, \ldots, n$ *are conditionally independent given* R*.*

For instance, a positive-negative approximate set has random false positive rate $\mathcal{E}$ and a random false negative rate $\mathcal{W}$ that are conditionally independent given R.

There are two natural characteristics of the Bernoulli set model, the random false positive and false negative rates conditioned on $\mathrm{R} = \mathcal{A}$. They are respectively given by

$$\mathcal{E} = \frac{1}{|\overline{\mathcal{A}}|} \sum_{x \in \overline{\mathcal{A}}} \mathbb{1}_{\mathcal{A}^\sigma}(x) \tag{3.2}$$

and

$$\mathcal{W} = \frac{1}{|\mathcal{A}|} \sum_{x \in \mathcal{A}} \mathbb{1}_{\mathcal{A}^\sigma}(x)\,. \tag{3.3}$$

Since the random approximate set is *random*, properties like its *false positive rate* and *false negative rate* are also random, respectively modeled by the random false positive rate $\mathcal{E}$ and false negative rate $\mathcal{W}$.

We denote the *first-order* random approximate set generative model by $\mathrm{R}^{\pm}$. The joint distribution of $\mathrm{R}^{\pm}$, $\mathcal{E}$, $\mathcal{W}$, and R given a unviersal set $\mathcal{U}$ has a probability density

$$\mathrm{f}_{\mathrm{R}^{\pm},\mathcal{E},\mathcal{W},\mathrm{R}}(\mathcal{Y}, \varepsilon, \omega, \mathcal{X} \mid \mathcal{U})\,. \tag{3.4}$$

By the axioms of probability theory, this may be decomposed into

$$\mathrm{f}_{\mathrm{R}^{\pm},\mathcal{E},\mathcal{W},\mathrm{R}}(\mathcal{Y}, \varepsilon, \omega, \mathcal{X} \mid \mathcal{U}) = \mathrm{f}_{\mathrm{R}^{\pm}|\mathcal{E},\mathcal{W},\mathrm{R}}\big(\mathcal{Y} \mid \varepsilon, \omega, \mathcal{X}|\mathcal{U}\big)\, \mathrm{f}_{\mathcal{E},\mathcal{W}|\mathrm{R}}(\varepsilon, \omega \mid \mathcal{X})\, \mathrm{f}_{\mathrm{R}}(\mathcal{X} \mid \mathcal{U})\,. \tag{3.5}$$

We typically omit the explicit reference to $\mathcal{U}$, since it may usually be understood as implicit to the model.

The object of central interest is the distribution of $\mathrm{R}^{\pm}$ given R. The conditional distribution of $\mathrm{R}^{\pm}$ given $\mathrm{R} = \mathcal{X}$ is denote by $\mathcal{X}^{\pm}$. By the axioms of probability,

$$\mathrm{f}_{\mathcal{X}^{\pm},\mathcal{E},\mathcal{W}}(\mathcal{Y}, \varepsilon, \omega) = \mathrm{f}_{\mathcal{X}^{\pm}|\mathcal{E},\mathcal{W}}(\mathcal{Y} \mid \varepsilon, \omega)\, \mathrm{f}_{\mathcal{E},\mathcal{W}|\mathrm{R}}(\varepsilon, \omega \mid \mathcal{X})\,. \tag{3.6}$$

There are a few natural partitions.

If the rates happen to pick out a specific set in the support, then the result is a degenerate distribution, e.g., $\mathcal{A}^{\pm}$ given $\mathcal{E} = 0$ and $\mathcal{W} = 0$ is degenerate where all probability mass is assigned to $\mathcal{A}$.

We denote the distributions of $\mathcal{X}^{\pm}$ given $\mathrm{E}[\mathcal{E}] = \varepsilon$ and $\mathcal{X}^{\pm}$ given $\mathrm{E}[\mathcal{W}] = \omega$ respectively by $\mathcal{X}_{\varepsilon}^{-}$ and $\mathcal{X}_{+}^{\omega}$. An object of central interest is the distribution of $\mathcal{X}^{\pm}$ given $\mathrm{E}[\mathcal{E}] = \varepsilon$ and $\mathrm{E}[\mathcal{W}] = \omega$, denoted by

$$\mathcal{X}_{\varepsilon}^{\omega}\,. \tag{3.7}$$

If we *sample* from $\mathcal{A}_{\varepsilon}^{\omega}$, some set $\mathcal{Y} \in 2^{\mathcal{U}}$ with false positive rate a and false negative rate b will be realized with probability $\mathrm{f}_{\mathcal{A}_{\varepsilon}^{\omega}|\mathcal{E},\mathcal{W}}(\mathcal{Y} \mid a, b)$. However, as the number of samples goes to infinity, the mean false positive and false negative rates go to ε and ω respectively.

Random *positive* and *negative* approximate sets are special cases respectively given by the following definitions.

Definition 3.1. *A random approximate set $\mathcal{A}^{+}_{0}$ is a random* positive *approximate set denoted by $\mathcal{A}^{+}$.*

Definition 3.2. *A random approximate set $\mathcal{A}^{0}_{-}$ is a random* negative *approximate set denoted by $\mathcal{A}^{-}$.*

By these definitions, every realization of $\mathcal{A}^{+}$ and $\mathcal{A}^{-}$ are respectively *supersets* or *subsets* of $\mathcal{A}$. The complement of a random positive (negative) approximate set is a random negative (positive) approximate set.

By axiom 2 and by the axioms of probability,

$$\mathrm{f}_{\mathcal{X}^{\pm},\mathcal{E},\mathcal{W}}(\mathcal{Y},\varepsilon,\omega) = \mathrm{f}_{\mathcal{X}^{\pm}|\mathcal{E},\mathcal{W}}(\mathcal{Y} \mid \varepsilon,\omega)\, \mathrm{f}_{\mathcal{E}|\mathrm{R}}(\varepsilon \mid \mathcal{X})\, \mathrm{f}_{\mathcal{W}|\mathrm{R}}(\omega \mid \mathcal{X})\,. \tag{3.8}$$

Every statistical property of the first-order random approximate set model is entailed by **????**. Furthermore, these assumptions generally hold in practice, e.g., the Bloom filter[1] and Perfect hash filter[4] are two separate implementations[1] of the random positive approximate set in which these assumptions hold.

Suppose the first-order random approximate sets are over the universal set $\mathcal{U}$. Compositions of first-order random approximate sets over the Boolean algebra $(2^{\mathcal{U}}, \cup, \cap, \overline{\ }, \varnothing, \mathcal{U})$, or random approximate sets of random approximate sets, are not closed over the *first-order* model. These subject is beyond the scope of this paper.

3.1 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 $\mathrm{P}[\mathcal{A}]$.

These are the *elementary events* of the probability space. The random approximate set model given $\mathrm{R} = \mathcal{Y}$ is given by the *probability space*

$$\left(\Omega = 2^{\mathcal{U}}, 2^{\Omega}, \mathrm{P}\right)\,, \tag{3.9}$$

where Ω is the *sample space*, 2^{Ω} is the set of all events, and $\mathrm{P}\colon 2^{\Omega} \mapsto [0,1]$ is the probability set function.

Consider an objective set $\mathcal{A}$ and a random approximate set 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 $\mathbf{A} = \langle \mathrm{A}_1, \ldots, \mathrm{A}_{\mathrm{u}} \rangle$ where $\mathrm{A}_{\mathrm{j}} = \mathbb{1}_{\mathcal{A}}\left(x_{(j)}\right)$ for $j = 1, \ldots, u$. Similarly, we model the uncertainty of the elements of $\mathcal{A}^{\pm}$ by $\mathbf{A}^{\varepsilon}_{\tau} = \langle \mathrm{A}^{\pm}_{1}, \ldots, \mathrm{A}^{\pm}_{\mathrm{u}} \rangle$.

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

$$\mathrm{P}[\mathbf{A}^{\varepsilon}_{\tau} = \mathbf{x}, \mathbf{A} = \mathbf{y}] = \mathrm{P}[\mathbf{A}^{\varepsilon}_{\tau} = \mathbf{x} \mid \mathbf{A} = \mathbf{y}]\, \mathrm{P}[\mathbf{A} = \mathbf{y}]\,. \tag{3.10}$$

By **????**, $\mathrm{A}^{\pm}_{\mathrm{j}}$ is only dependent on A_{j} for $j = 1, \ldots, u$ and thus by the axioms of probability

$$\mathrm{P}[\mathbf{A}^{\varepsilon}_{\tau} = \mathbf{x}, \mathbf{A} = \mathbf{y}] = \mathrm{P}[\mathbf{A} = \mathbf{y}] \prod_{j=1}^{u} \mathrm{P}\left[\mathrm{A}^{\pm}_{\mathrm{j}} = x_j \,\middle|\, \mathrm{A}_{\mathrm{j}} = y_j\right]. \tag{3.11}$$

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

$$\mathrm{P}[\mathbf{A}^{\varepsilon}_{\tau} = \mathbf{x} \mid \mathbf{A} = \mathbf{y}] = \prod_{j=1}^{u} \mathrm{P}\left[\mathrm{A}^{\pm}_{\mathrm{j}} = x_j \,\middle|\, \mathrm{A}_{\mathrm{j}} = y_j\right] \tag{3.12}$$

where $\varepsilon = \mathrm{P}\left[\mathrm{A}^{\pm}_{\mathrm{j}} = 1 \,\middle|\, \mathrm{A}_{\mathrm{j}} = 0\right]$ and $\tau = \mathrm{P}\left[\mathrm{A}^{\pm}_{\mathrm{j}} = 1 \,\middle|\, \mathrm{A}_{\mathrm{j}} = 1\right]$.

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

Consider the following example.

Example 1 *Suppose the universal set is $\{x_1, x_2\}$ and consider the distribution of the first-order random approximate set $\{x_1\}^{\omega}_{\varepsilon}$. The probability mass function $\mathrm{p}_{\{x_1\}^{\omega}_{\varepsilon}}$ is given by*

$$\mathrm{p}_{\{x_1\}^{\tau}_{\varepsilon}}(\mathcal{X}) = \begin{cases} \omega(1-\varepsilon) & \mathcal{X} = \varnothing\,, \\ \omega\varepsilon & \mathcal{X} = \{x_2\}\,, \\ (1-\omega)(1-\varepsilon) & \mathcal{X} = \{x_1\}\,, \\ (1-\omega)\varepsilon & \mathcal{X} = \{x_1, x_2\}\,. \end{cases} \tag{a}$$

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

3.2 Higher-order models

There are $k!/n_1!n_2!\ldots n_k!$ partitioning schemes in the n-th order model.

, any of which may be realized by an appropriate composition of first-order and positive-negative approximations.

In the first-order model with error rate ϵ, the error rate is ϵ over every subset.

The *first-order* model has only one partition possible, a single block consisting of the universal set. In the second-order model, the universal set may be partitioned into two subsets $\mathcal{A}$ and $\mathcal{B}$ with non-identical error rates ϵ_A and ϵ_B. If the subset of interest is a subset of $\mathcal{A}$, then the error rate over that subset is Bernoulli distributed with error rate ϵ_A and similarly if it is a subset of $\mathcal{B}$. For example, in the positive-negative random (PNR) approximate set model, the error rate over the positives is ω, denoted the *false negative rate*, and the error rate over the negatives is ε, denoted the *false positive rate.* Thus, if either the positives or negatives (or their respective subsets) are of interest, the error rates are completely characterized by either ω or ε.

The second-order model has 2^k partitioning schemes, where k is the cardinality of the universal set.

One could imagine, say, a *guarded approximate model* which, say, ensures that some special subset of elements have a reduced error rate ϵ_1 and the rest have some error rate $\epsilon_2 > \epsilon_1$.

A natural set to guard is the positive set with respect to some objective input set. We denote this second-order model the *positive-negative* approximate set model.

It is isomorphic to the binary symmetric channel model. Suppose we have a communications channel over which we transmit 1s and 0s and due to *noise* or *rate-distortion* flips 0s and 1s respectively with probabilities ω and ε.

If we serialize a set as a bit string where the j-th bit is 1 if the j-th element is a member of the set and otherwise 0, then the channel induces a *positive-negative* approximation of any such set transmitted over the channel.

Typically, the communications channel is a storage medium and ω and ε are rate distortions caused by *lossy* compression *algorithms* that construct approximations of input sets. Data structures like the Bloom filter are a practical example which models the concept of a *positive* approximate set, where $\varepsilon > 0$ and $\varepsilon = 0$. If we take the *complement* of a positive approximate set with a false positive rate ε, the result is a *negative* approximate set with a false negative rate $\omega = \varepsilon$.

Positive-negative approximate sets may be conditioned on $\mathrm{E}[\mathcal{E}] = \varepsilon$ or $\mathrm{E}[\mathcal{W}] = \omega$, which results in an approximations that are expected to obtain the indicated false positive and false negative rates.

4 Distributions of binary classification measures

The error rate on some identically distributed subset is an *expectation.* However, some *error measures* span multiple subsets, such as positives and negatives. Any such error measure may be modeled as a *Bernoulli mixture.*

We are typically interested in the error rates of *special* subsets. For example, if we have a universal set $\mathcal{X}$ and with probability $\mathrm{P}(x)$ an element $x \in \mathcal{X}$ is tested for membership in a collection of sets over $\mathcal{X}$, then to reduce the expected error rate on membership tests elements with higher probability of begin tested should be assigned smaller error rates.

The first-order random approximate sets are *parameterized* 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.

Definition 4.1. *The uncertain number of* negatives *is a random variable denoted by* N *and is statistically dependent on* R*,*

$$\mathrm{N} \coloneqq |\overline{\mathrm{R}}| \,. \tag{4.1}$$

Definition 4.2. *The uncertain number of false positives is a random variable denoted by* FP *and is statistically dependent on* N *and* $\mathcal{E}$*,*

$$\mathrm{FP} \coloneqq \mathrm{N}\mathcal{E} \,. \tag{4.2}$$

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

Theorem 4.1. *The random number of false positives* FP *given* $\mathrm{N} = n$ *in the first-order random approximate set* R^{ε} *is given by*

$$\mathrm{FP}_n \coloneqq n\mathcal{E} \tag{4.3}$$

with a distribution given by

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

Proof. TODO: Look up the def. given earlier instead for $\mathcal{E}$.

By **??**, 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 realization of a random approximate set $\mathcal{S}^{\varepsilon}$ is *uncertain.*

Theorem 4.2. *The random false positive rate $\mathcal{E}$ conditioned on $\mathrm{R} = n$ is denoted by $\mathcal{E}_n$ and has a distribution given by*

$$\mathcal{E}_n = \frac{\mathrm{FP}_n}{n}\,, \tag{4.5}$$

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

$$\mathrm{f}_{\mathcal{E}_n}(\hat{\varepsilon} \mid \varepsilon) = \mathrm{f}_{\mathrm{FP}_n}(\hat{\varepsilon} n \mid \varepsilon)\,. \tag{4.6}$$

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

Proof. By definition 2.5, 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

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

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*

$$\mathrm{TN}_n = n - \mathrm{FP}_n \sim \mathrm{BIN}(n, 1-\varepsilon)\;. \tag{4.7}$$

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. 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 section A 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 uncertain number of* false negatives *in random approximate sets with a false negative rate ω is modeled as a binomial distributed random variable denoted by FP_p,*

$$\mathrm{FN}_p \sim \mathrm{BIN}(p, \omega)\,. \tag{4.8}$$

Proof. By **??**, 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*

$$\mathcal{W}_p = \frac{\mathrm{FN}_p}{p} \tag{4.9}$$

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 **??**, 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*

$$\mathrm{TP}_p \sim \mathrm{BIN}(p, \tau)\,. \tag{4.10}$$

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, the distribution of $|\mathcal{A}^\tau_\varepsilon|$ given p positives is

$$|\mathcal{A}^\tau_\varepsilon| = \mathrm{TP}_p + \mathrm{FP}_{u-p} \tag{4.11}$$

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

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

$$\widehat{p} = \frac{|\mathcal{B}| - u\varepsilon}{\tau - \varepsilon}\,. \tag{4.12}$$

If the universal set $\mathcal{U}$ is 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*

$$\mathcal{E}_n \xrightarrow{d} \mathcal{N}(\varepsilon, \varepsilon(1-\varepsilon)/n)\,. \tag{4.13}$$

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*

$$\mathcal{T}_p \xrightarrow{d} \mathcal{N}(\tau, \tau(1-\tau)/p)\,. \tag{4.14}$$

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

$$\mathcal{E}_n = \frac{\mathrm{X_1}}{n} + \cdots + \frac{\mathrm{X_n}}{n}\,, \tag{a}$$

where $\mathrm{X_1}, \ldots, \mathrm{X_n}$ 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. (4.13) and (4.14),

$$\mathcal{N}_n \xrightarrow{d} \mathcal{N}\big(1-\varepsilon, \varepsilon(1-\varepsilon)/n\big) \text{ and } \mathcal{W}_n \xrightarrow{d} \mathcal{N}\big(1-\tau, \tau(1-\tau)/p\big)\,. \tag{4.15}$$

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., $\mathcal{E}_n = \hat{\varepsilon}$ and $\mathcal{T}_p = \hat{\tau}$. 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 $\hat{\varepsilon}$ and $\hat{\tau}$ may be recorded. If these outcomes cannot be observed, e.g., it is too costly to compute or the information to compute $\hat{\varepsilon}$ or $\hat{\tau}$ 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 $\hat{\varepsilon}$ and the true true positive rate $\hat{\tau}$ 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*

$$\varepsilon \pm \sqrt{\frac{\varepsilon(1-\varepsilon)}{n}} \Phi^{-1}(\alpha/2) \tag{4.16}$$

and

$$\tau \pm \sqrt{\frac{\tau(1-\tau)}{p}} \Phi^{-1}(\alpha/2)\,, \tag{4.17}$$

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. (4.16) and (4.17).

4.2 Other binary performance measures

Suppose we have some other function $\mathrm{g} \colon 2^{\mathcal{X}} \mapsto \mathcal{Y}$ that is not a *constant*, then the composition $\mathrm{g} \circ \mathrm{id}_{\varepsilon}^{\tau}$ is some probability distribution over the codomain $\mathcal{Y}$. That is, $(\mathrm{g} \circ \mathrm{id}_{\varepsilon}^{\tau})\,(\mathcal{A})$ is a *random variable.*

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

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

The composition $\mathrm{f} \circ \mathrm{id}_{.25}^{.75}$ generates Bernoulli distribution random variables, e.g., $\left(\mathrm{f} \circ \mathrm{id}_{.25}^{.75}\right)(\{0\}) \sim \mathrm{BER}(0.25)$.

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

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.,

$$\mathrm{P}\big[\mathbb{1}_{\mathcal{A}_{\varepsilon}^{\tau}}(x) \,\big|\, \mathbb{1}_{\mathcal{A}}(x)\big] = \tau\,. \tag{4.18}$$

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

$$\mathrm{ppv} = \frac{t_p}{t_p + f_p} \tag{4.19}$$

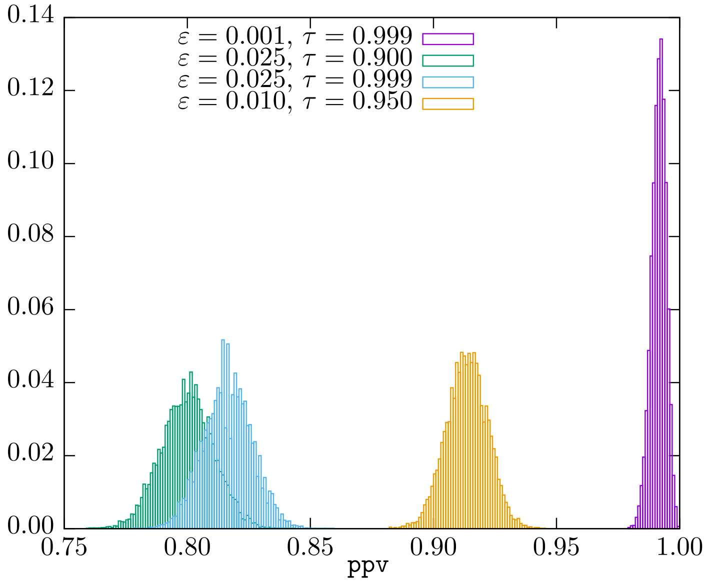
where t_p is the number of true positives *and f_p is the number of* false positives.

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

Theorem 4.7. *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*

$$\mathrm{PPV} = \frac{\mathrm{TP}_p}{\mathrm{TP}_p + \mathrm{FP}_n} \tag{4.20}$$

Figure 2: 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.



with an expectation *given* approximately *by*

$$\texttt{ppv}(\tau, \varepsilon, p, n) \approx \frac{\overline{t}_p}{\overline{t}_p + \overline{f}_p} + \frac{\overline{t}_p \sigma^2_{f_p} - \overline{f}_p \sigma^2_{t_p}}{\left(\overline{t}_p + \overline{f}_p\right)^3}\,, \tag{4.21}$$

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

See section B for a proof of theorem 4.7.

We make the following observations about eq. (4.21):

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

Accuracy is given by the following definition.

Definition 4.4. *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.

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

Theorem 4.8. *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*

$$\mathrm{ACC}_{p+n} = \lambda \mathcal{T}_p + (1-\lambda)\mathcal{N}_n\,. \tag{4.22}$$

has an expected *accuracy*

$$\texttt{acc}(\tau, \varepsilon, n, p) = \lambda\tau + (1-\lambda)\eta \tag{4.23}$$

with a variance

$$\frac{\lambda\omega\tau + (1-\lambda)\varepsilon\eta}{p+n}\,, \tag{4.24}$$

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

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

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

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

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

The variance

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

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

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

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

□

Negative predictive value is given by the following definition.

Definition 4.5.

$$\texttt{npv} = \frac{t_n}{t_n + f_n} \tag{4.25}$$

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 4.9. *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

$$\mathrm{NPV} = \frac{\mathrm{TN}_n}{\mathrm{TN}_n + \mathrm{FN}_p} \tag{4.26}$$

with an expectation *given approximately by*

$$\texttt{npv}(\tau, \varepsilon, p, n) \approx \frac{\bar{t}_n}{\bar{t}_n + \bar{f}_n} + \frac{\bar{t}_n \sigma_{f_n}^2 - \bar{f}_n \sigma_{t_n}^2}{\left(\bar{t}_n + \bar{f}_n\right)^3}\,, \tag{4.27}$$

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

The proof for theorem 4.9 follows the same pattern as the proof for theorem 4.7.

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

$$J = \frac{t_p}{t_p + f_n} + \frac{t_n}{t_n + f_p} - 1\,, \tag{4.28}$$

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

$$\mathrm{J} = \mathcal{T}_p - \mathcal{E}_n\,, \tag{4.29}$$

which has an *expectation*

$$\mathrm{E}[\mathrm{J}] = \tau - \varepsilon\,. \tag{4.30}$$

Table 2: Various *expected* performance measures.

measure	**parameter**	**expected value**
true positive rate	$\mathtt{tpr}(\tau)$	τ
false positive rate	$\mathtt{fpr}(\varepsilon)$	ε
false negative rate	$\mathtt{fnr}(\tau)$	$1-\tau$
true negative rate	$\mathtt{tnr}(\varepsilon)$	$1-\varepsilon$
accuracy	$\mathtt{acc}$	eq. (4.23)
positive predictive value	$\mathtt{ppv}$	eq. (4.21)
negative predictive value	$\mathtt{npv}$	eq. (4.27)
false discovery rate	$\mathtt{fdr}$	$1-\mathtt{ppv}$
false omission rate	$\mathtt{for}$	$1-\mathtt{npv}$

Table 2 may be used to reparameterize an approximate set.

Example 3 *Suppose we seek a* positive approximate set *with an expected accuracy* γ*. By table 2,*

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

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 Relational probabilities

The relational *member-of* predicate $\in \colon \mathcal{U} \times 2^{\mathcal{U}}$ is defined as $x \in \mathcal{A} \coloneqq \mathbb{1}_{\mathcal{A}}(x)$. By **??**, given $x \in \mathcal{A}$, $x \in \mathcal{A}^{\tau}_{\varepsilon}$ with probability τ and given $x \notin \mathcal{A}$, $x \in \mathcal{A}^{\tau}_{\varepsilon}$ with probability ε. These probabilities are axiomatic in the first-order random approximate set model.

Other kinds of relational predicates are functions of the *member-of* predicate, in particular the subset $\subseteq \colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto \{0,1\}$, equality $= \colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto \{0,1\}$, and proper subset $\subset \colon 2^{\mathcal{U}} \times 2^{\mathcal{U}} \mapsto \{0,1\}$ predicates respectively defined as

$$\mathcal{A} \subseteq \mathcal{B} \coloneqq \prod_{x \in \mathcal{A}} \mathbb{1}_{\mathcal{B}}(x)\,, \tag{5.1}$$

$$\mathcal{A} = \mathcal{B} \coloneqq (\mathcal{A} \subseteq \mathcal{B}) \wedge (\mathcal{B} \subseteq \mathcal{A})\,, \text{ and} \tag{5.2}$$

$$\mathcal{A} \subset \mathcal{B} \coloneqq (\mathcal{A} \neq \mathcal{B}) \wedge (\mathcal{A} \subseteq \mathcal{B}) \tag{5.3}$$

where complementary relations are naturally given by complements, e.g., $\mathcal{A} \neq \mathcal{B} \coloneqq 1 - (\mathcal{A} = \mathcal{B})$.

Theorem 5.1. *Given* $\mathcal{X} \subseteq \mathcal{Y}$, $\mathcal{X}^{\tau_1}_{\varepsilon_1} \subseteq \mathcal{Y}^{\tau_2}_{\varepsilon_2}$ *with probability*

$$(1-\tau_1\omega_2)^{|\mathcal{X}|}\,(1-\varepsilon_1\omega_2)^{|\mathcal{Y}|-|\mathcal{X}|}\,(1-\varepsilon_1\eta_2)^{|\mathcal{U}|-|\mathcal{Y}|}\;. \tag{5.4}$$

Proof. Let

$$\gamma = \mathrm{P}\big[\mathcal{X}^{\pm} \subseteq \mathcal{Y}^{\pm} \,\big|\, \mathcal{X} \subseteq \mathcal{Y}\big]\,. \tag{a}$$

Note that in the Boolean vector representation, $\mathcal{X}$ is a subset of $\mathcal{Y}$ if $x_j \implies y_j$, i.e., if x_j then y_j, otherwise y_j can be either true or false. An equivalent expression for $x_j \implies y_j$ is $\neg(x_j \wedge \neg y_j)$.

Switching to the Boolean vector representation, we may rewrite γ as

$$\gamma = \mathrm{P}\left[\bigcap_{j=1}^{u} \neg\left(\mathrm{X}^{\pm}_{\mathrm{j}} \wedge \neg \mathrm{Y}^{\pm}_{\mathrm{j}}\right) \,\middle|\, E\right] \tag{b}$$

where E is the set of Boolean vectors satisfying $x_k \implies y_k$ for $k = 1, \ldots, u$.

By the axioms of the approximate set model, each of these events are independent, in which case the probability of the intersection of the events is equal to the product of the probabilities of the events,

$$\gamma = \prod_{j=1}^{u} \mathrm{P}\left[\neg\left(\mathrm{X}^{\pm}_{\mathrm{j}} \land \neg \mathrm{Y}^{\pm}_{\mathrm{j}}\right) \,\middle|\, E\right] \tag{c}$$

$$= \prod_{j=1}^{u} \left(1 - \mathrm{P}\left[\mathrm{X}^{\pm}_{\mathrm{j}} \land \neg \mathrm{Y}^{\pm}_{\mathrm{j}} \,\middle|\, E\right]\right) . \tag{d}$$

By the axioms of the approximate set model, $\mathrm{Y}^{\pm}_{\mathrm{j}}$ is only dependent on y_j and $\mathrm{X}^{\pm}_{\mathrm{j}}$ is only dependent on x_j. Thus, we may rewrite γ as

$$\gamma = \prod_{j=1}^{u} \left(1 - \mathrm{P}\left[\mathrm{X}^{\pm}_{\mathrm{j}} \,\middle|\, x_j\right] \mathrm{P}\left[\neg \mathrm{Y}^{\pm}_{\mathrm{j}} \,\middle|\, y_j\right]\right) , \tag{e}$$

where x_j and y_j are Boolean values satisying E.

Since $\mathcal{X} \subseteq \mathcal{Y}$, an exhaustive, mutually exclusive set of sets is given by $\mathcal{X}$, $\mathcal{Y} \smallsetminus \mathcal{X}$, and $\overline{\mathcal{Y}}$.

Let set $\mathcal{I}$ index the elements in $\mathcal{X}$, set $\mathcal{J}$ index the elements in $\mathcal{Y} \smallsetminus \mathcal{X}$, and set $\mathcal{K}$ index the elements in $\overline{\mathcal{Y}}$.

The elements indexed by $\mathcal{I}$ are members of $\mathcal{X}$ and $\mathcal{Y}$, i.e., x_i and y_i are both true.

The elements indexed by $\mathcal{J}$ are members of $\mathcal{Y}$ but not $\mathcal{X}$, i.e., x_j is false and y_j is true.

The elements indexed by $\mathcal{K}$ are members of neither, i.e., x_j and y_j are false.

$$\begin{aligned}\gamma = &\prod_{i \in \mathcal{I}} \left(1 - \mathrm{P}\left[\mathrm{X}^{\pm}_{\mathrm{i}} \,\middle|\, x_i\right] \mathrm{P}\left[\neg \mathrm{Y}^{\pm}_{\mathrm{i}} \,\middle|\, y_i\right]\right) \prod_{j \in \mathcal{J}} \left(1 - \mathrm{P}\left[\mathrm{X}^{\pm}_{\mathrm{j}} \,\middle|\, \neg x_j\right] \mathrm{P}\left[\neg \mathrm{Y}^{\pm}_{\mathrm{j}} \,\middle|\, y_j\right]\right) \\ &\prod_{k \in \mathcal{K}} \left(1 - \mathrm{P}\left[\mathrm{X}^{\pm}_{\mathrm{k}} \,\middle|\, \neg x_k\right] \mathrm{P}\left[\neg \mathrm{Y}^{\pm}_{\mathrm{k}} \,\middle|\, \neg y_k\right]\right) .\end{aligned} \tag{f}$$

$$\gamma = \prod_{i \in \mathcal{I}} \left(1 - \tau_1(1 - \tau_2)\right) \prod_{j \in \mathcal{J}} \left(1 - \varepsilon_1(1 - \tau_2)\right) \prod_{k \in \mathcal{K}} \left(1 - \varepsilon_1(1 - \varepsilon_2)\right) \tag{g}$$

$$\gamma = \prod_{i \in \mathcal{I}} \left(1 - \tau_1(1 - \tau_2)\right) \prod_{j \in \mathcal{J}} \left(1 - \varepsilon_1(1 - \tau_2)\right) \prod_{k \in \mathcal{K}} \left(1 - \varepsilon_1(1 - \varepsilon_2)\right) \tag{h}$$

By **??**, $\mathrm{P}\left[x \in \mathcal{X}^{\pm} \,\middle|\, x \in \mathcal{X}\right] = \tau_1$, $\mathrm{P}\left[x \in \mathcal{X}^{\pm} \,\middle|\, x \notin \mathcal{X}\right] = \varepsilon_1$, $\mathrm{P}\left[x \in \mathcal{Y}^{\pm} \,\middle|\, x \in \mathcal{Y}\right] = \tau_2$, and $\mathrm{P}\left[x \in \mathcal{Y}^{\pm} \,\middle|\, x \notin \mathcal{Y}\right] = \varepsilon_1$. Making these substitutions yields the result

$$\begin{aligned}\gamma = &\prod_{x \in \mathcal{V}_1} \left(1 - \tau_1(1 - \tau_2)\right) \\ &\prod_{x \in \mathcal{V}_2} \left(1 - \varepsilon_1(1 - \tau_2)\right) \\ &\prod_{x \in \mathcal{V}_3} \left(1 - \varepsilon_1(1 - \varepsilon_2)\right) .\end{aligned} \tag{i}$$

Each of the above products is just repeated multiplication and thus may be replaced by powers,

$$\begin{aligned}\gamma = &\left(1 - \tau_1(1 - \tau_2)\right)^{|\mathcal{V}_1|} \\ &\left(1 - \varepsilon_1(1 - \tau_2)\right)^{|\mathcal{V}_2|} \left(1 - \varepsilon_1(1 - \varepsilon_2)\right)^{|\mathcal{V}_3|} .\end{aligned} \tag{j}$$

□

Corollary 5.1.1. *Given $\mathcal{X} = \mathcal{Y}$, $\mathcal{X}^{\tau_1}_{\varepsilon_1} = \mathcal{Y}^{\tau_2}_{\varepsilon_2}$ with probability*

$$\left(1 - \tau_1\omega_2\right)^{|\mathcal{X}|} \left(1 - \varepsilon_1\omega_2\right)^{|\mathcal{Y}| - |\mathcal{X}|} \left(1 - \varepsilon_1\eta_2\right)^{|\mathcal{U}| - |\mathcal{Y}|} \left(1 - \tau_2\omega_1\right)^{|\mathcal{Y}|} \left(1 - \varepsilon_2\omega_1\right)^{|\mathcal{X}| - |\mathcal{Y}|} \left(1 - \varepsilon_2\eta_1\right)^{|\mathcal{U}| - |\mathcal{X}|} . \tag{5.5}$$

Proof. By **??**,

$$\mathrm{P}\left[\mathcal{X}^{\tau_1}_{\varepsilon_1} \subseteq \mathcal{Y}^{\tau_2}_{\varepsilon_2} \,\middle|\, \mathcal{X} \subset \mathcal{Y}\right] \mathrm{P}\left[\mathcal{Y}^{\tau_2}_{\varepsilon_2} \subseteq \mathcal{X}^{\tau_1}_{\varepsilon_1} \,\middle|\, \mathcal{Y} \subset \mathcal{X}\right] . \tag{a}$$

□

Corollary 5.1.2. *Given $\mathcal{X} = \mathcal{Y}$, $\mathcal{X}^{\tau}_{\varepsilon} = \mathcal{Y}^{\tau}_{\varepsilon}$ with probability*

$$\left(1 - \tau\omega\right)^{|\mathcal{X}| + |\mathcal{Y}|} \left(1 - \varepsilon\eta\right)^{2|\mathcal{U}| - |\mathcal{X}| - |\mathcal{Y}|} . \tag{5.6}$$

Corollary 5.1.3. *Given $\mathcal{X} = \mathcal{Y}$, $\mathcal{X}_{\varepsilon}^{0} = \mathcal{Y}_{\varepsilon}^{0}$ with probability*

$$(1-\varepsilon)^{2|\mathcal{U}|-|\mathcal{X}|-|\mathcal{Y}|}\,. \tag{5.7}$$

Corollary 5.1.4. *Given $\mathcal{X} \subseteq \mathcal{Y}$, $\mathcal{X}_{+}^{\tau_1} \subseteq \mathcal{Y}_{+}^{\omega_2}$ with probability*

$$(1-\tau_1\omega_2)^{|\mathcal{X}|} \tag{5.8}$$

Corollary 5.1.5. *Over an infinite universal set, the probability that two first-order random approximate sets realize the same value is* 0*.*

Corollary 5.1.6. *Over an infinite universal set, the probability that a first-order random approximate set realizes a value that is a subset of another realization of a first-order random approximate set is* 0*.*

TODO: talk about data structures in adt section, in which representational equality implies subset eq and equality.s

6 Uncertain rates

We may not be certain about the *expected* false positive and true positive rates, i.e., we may only have the joint distribution of $\mathcal{A}^{\pm}$, $\mathcal{E}$, and $\mathcal{W}$.

A relatively easy case to analyze is the positive-negative (second-order) random approximate set model. Suppose we are interested in the distribution of $\mathcal{A}^{\pm}$ given R has p positives and n negatives. Since we are primarily interested in the distribution of false positives and true positives (or their corresponding rates), we consider the related random tuple $\langle \mathrm{FP}_n, \mathrm{TP}_p, \mathcal{T}, \mathcal{E} \rangle$ which, assuming $\mathcal{E}$ and $\mathcal{T}$ are independent, has a joint probability density function given by

$$\mathrm{f}_{\mathrm{TP}_p,\mathrm{FP}_n,\mathcal{T},\mathcal{E}}(t,f,\tau,\varepsilon) = \mathrm{f}_{\mathrm{TP}_p,\mathcal{T}}(t,\tau)\,\mathrm{f}_{\mathrm{FP}_n,\mathcal{E}}(f,\varepsilon) \tag{6.1}$$

where

$$\mathrm{f}_{\mathrm{TP}_p,\mathcal{T}}(t,\tau) = \mathrm{f}_{\mathrm{TP}_p|\mathcal{T}}(t \mid \tau)\,\mathrm{f}_{\mathcal{T}}(\tau)\,, \tag{6.2}$$

$$\mathrm{f}_{\mathrm{FP}_n,\mathcal{E}}(f,\varepsilon) = \mathrm{f}_{\mathrm{FP}_p|\mathcal{E}}(f \mid \varepsilon)\,\mathrm{f}_{\mathcal{E}}(\varepsilon)\,. \tag{6.3}$$

When we *marginalize* over the true positives, we get the result

$$\begin{aligned}\mathrm{f}_{\mathrm{TP}_p}(t) &= \int_0^1 \mathrm{f}_{\mathrm{TP}_p|\mathcal{T}}(t \mid \tau)\,\mathrm{f}_{\mathcal{T}}(\tau)\,\mathrm{d}\tau \\ &= \int_0^1 \binom{p}{t}\tau^{t}(1-\tau)^{p-t}\,\mathrm{f}_{\mathcal{T}}(\tau)\,\mathrm{d}\tau\,.\end{aligned} \tag{6.4}$$

If all the probability mass for $\mathcal{T}$ is assigned to a particular point τ, the probability mass function simplifies to

$$\mathrm{f}_{\mathrm{TP}_p}(t) = \binom{p}{t}\tau^{t_p}(1-\tau)^{p-t}\,, \tag{6.5}$$

which is probability mass function of a binomial distribution.

The simplest kind of uncertainty is given by an interval in which the rate is uniformly distributed across the support.

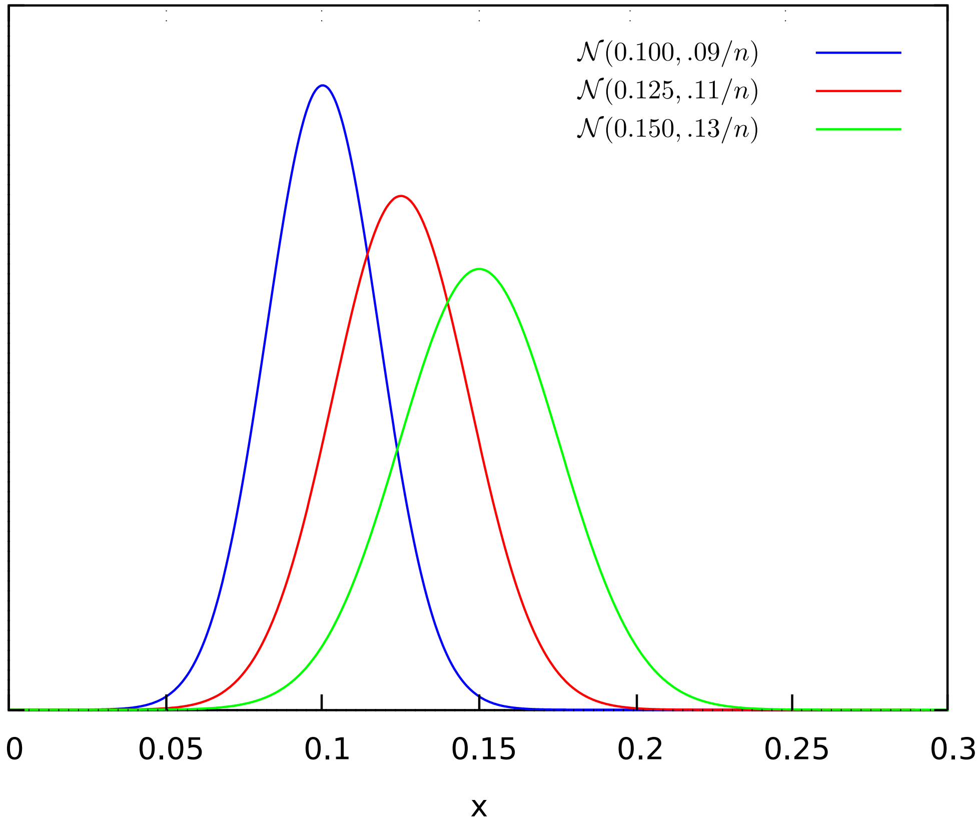
Definition 6.1. *An interval is a convex set of real numbers. We denote by $[\underline{x}, \bar{x}]$ an interval with a lower-bound $\underline{x}$ and an upper-bound $\bar{x}$.*

A confidence interval may be specified in interval 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 2 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.

Figure 3: 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.



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 in the interval*

$$\begin{aligned}\mathtt{acc}([\varepsilon],[\omega];[\lambda]) = \Big[&\mathrm{f}(\overline{\varepsilon},\overline{\omega})(1-\overline{\omega}) + \left(1-\mathrm{f}(\overline{\varepsilon},\overline{\omega})\right)(1-\overline{\varepsilon}),\\ &\mathrm{f}(\underline{\omega},\underline{\varepsilon})(1-\underline{\omega}) + \left(1-\mathrm{f}(\underline{\omega},\underline{\varepsilon})\right)(1-\underline{\varepsilon})\Big],\end{aligned} \tag{a}$$

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* $\mathtt{acc}([\varepsilon],0;[0,1]) = [1-\overline{\varepsilon},1]$.

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 [2].

A set sampled from $\mathcal{A}_{\varepsilon}^{\tau}$ is an approximate set such that the $(1-\alpha)\%$ asymptotic confidence interval for the false negative and false positive rates given respectively by

$$[\omega] = ? \tag{6.6}$$

and

$$[\varepsilon] = ?\,. \tag{6.7}$$

Theorem 6.1. *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* $[\omega]$.

7 Data types that model random approximate sets

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 section 2.1. Suppose T is a data type that overloads the *member-of* predicate $\in : \mathcal{U} \times T \mapsto \{0,1\}$ and has a *value constructor* $\mathrm{F}_{\varepsilon}^{\tau}$ 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 **????** are satisfied, i.e.,

$$\mathrm{P}\left[x \in \mathrm{F}_{\varepsilon}^{\tau}(\mathcal{S}) \,\middle|\, x \notin \mathcal{S}\right] = \varepsilon \tag{7.1}$$

and

$$\mathrm{P}\left[x \in \mathrm{F}_{\varepsilon}^{\tau}(\mathcal{S}) \,\middle|\, x \in \mathcal{S}\right] = \tau\,. \tag{7.2}$$

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 $\mathrm{F}_{\varepsilon}^{\tau}(\mathcal{A})$ generates a sample that converges in distribution to $\mathcal{A}_{\varepsilon}^{\tau}$. However, we also wish to allow for *deterministic* value constructors.[2]

[2] Deterministic algorithms compatible with the random approximate set model are common but frequently have an auxiliary seed which indexes a particular approximation in a family.

7.1 Deterministic value constructors

Value constructors compatible with the random approximate set model may come in many forms.

Suppose $\mathrm{F}^{\varepsilon}_{\varepsilon} \colon 2^{\mathcal{U}} \mapsto T$ (i.e., a deterministic total function) maps sets in $2^{\mathcal{U}}$ to objects of type T that model random approximate sets over the input. Since T models the abstract data type of the set, there is a unique bijection between T and $2^{\mathcal{U}}$, i.e., every value in T models a specific subset of $\mathcal{U}$. Thus, we may view $\mathrm{F}^{\varepsilon}_{\varepsilon}$ as a function $\mathrm{F}^{\tau}_{\varepsilon} \colon 2^{\mathcal{U}} \mapsto 2^{\mathcal{U}}$ with an *image*

$$\texttt{image}(\mathrm{F}^{\tau}_{\varepsilon}) = \{\, \mathrm{F}^{\tau}_{\varepsilon}(\mathcal{A}) \mid \mathcal{A} \in 2^{\mathcal{U}} \,\} \subseteq 2^{\mathcal{U}}\,. \tag{7.3}$$

Since the value constructor $\mathrm{F}^{\tau}_{\varepsilon}$ 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, $\mathrm{F}^{\tau}_{\varepsilon}$ is (possibly) a non-surjective, non-injective function.

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 $\mathrm{F}^{\tau}_{\varepsilon} \colon 2^{\mathcal{U}} \mapsto T$ 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[3] by observing $\mathrm{F}^{\tau}_{\varepsilon}(\mathcal{S})$.

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 $\mathrm{F}^{\tau}_{\varepsilon}(\mathcal{A})$ for each $\mathcal{A} \in 2^{\mathcal{U}}$ and remembering the sets that map to $T(\mathcal{S})$.[4] However, since $\mathrm{F}^{\tau}_{\varepsilon}$ 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.[5]

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 $\mathrm{F}^{\tau}_{\varepsilon}(\mathcal{X})$ is by definition

$$\hat{\varepsilon}(\mathcal{X}) \coloneqq \frac{1}{n} \sum_{x \in \overline{\mathcal{X}}} \mathbb{1}_{\mathrm{F}^{\tau}_{\varepsilon}(\mathcal{X})}(x)\,, \tag{7.4}$$

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

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

$$\overline{\varepsilon} = \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X} \in \mathcal{U}_p} \hat{\varepsilon}(\mathcal{X})\,, \tag{7.5}$$

[3] A posteriori knowledge is dependent on experience.

[4] 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.

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

is an unbiased estimator of ε and the population variance

$$s_\varepsilon^2 = \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X} \in \mathcal{U}_p} \hat{\varepsilon}(\mathcal{X})\,, \tag{7.6}$$

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

Proof. We imagine that the function $\mathrm{F}_\varepsilon^\tau$ 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 $\mathcal{E}_n = \hat{\varepsilon}$. Thus,

$$\begin{aligned} \overline{\varepsilon} &= \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X}_i \in \mathcal{U}_p} \hat{\varepsilon}(\mathcal{X}_i) && \text{(a)} \\ &= \frac{1}{|\mathcal{U}_p|} \sum_{\mathcal{X}_i \in \mathcal{U}_p} \mathrm{E}\left[\mathcal{E}_n^{(i)}\right] = \varepsilon\,. && \text{(b)} \end{aligned}$$

□

7.2 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

$$\mathcal{O}(u) \text{ bits} \tag{7.7}$$

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 7.1. *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*

$$-\,\tau \log_2 \varepsilon \text{ bits/element}\,, \tag{7.8}$$

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,

$$\mathtt{RE}(X, Y) = \frac{\ell(Y)}{\ell(X)}\,, \tag{7.9}$$

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 7.1. *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,*

$$\mathtt{E}(X) = \frac{\theta}{\ell(X)}\,, \tag{7.10}$$

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.[6]

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

$$\mathtt{E}(X) = \frac{-p\tau \log_2 \varepsilon}{\ell(X)}\,. \tag{7.11}$$

A well-known implementation of countably infinite *positive approximate set* is the Bloom filter[1] which has an expected space complexity given by

$$-\frac{1}{\ln 2} \log_2 \varepsilon \text{ bits/element}\,, \tag{7.12}$$

[6] 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.

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*[4], which compares favorably to the Bloom filter in may circumstances.[7]

In **??**, given an approximate set sampled from $\mathcal{A}^{\omega}_{\varepsilon}$, we claimed that the method of moments estimator for $p = |\mathcal{A}|$ 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

$$\widehat{p} = \frac{\ell(x)}{\mathrm{b}(\tau, \varepsilon)}\,, \tag{7.13}$$

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.

A Proof of corollary 4.2.2

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

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

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[3] 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

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

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. □

B Proof of theorem 4.7

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

$$\texttt{ppv}(\omega, \varepsilon; n, p) \approx \frac{\bar{t}}{\bar{t} + \bar{f}} + \frac{\bar{t}\sigma_f^2 - \bar{f}_p \sigma_t^2}{\Big(\bar{t} + \bar{f}\Big)^3}\,, \tag{?? revisited}$$

where $\bar{t} = p\tau$ is the *expected* number of *true positives*, $\bar{f} = n\varepsilon$ is the *expected* number of *false positives*, $\sigma_t^2 = p\omega\tau$ is the variance of the number of *true positives*, and $\sigma_f^2 = n\varepsilon\omega$ 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,

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

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

[7] 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.

Let the *positive predictive value* function be denoted by

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

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

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

and the Hessian of f is given by

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

Substituting this result into eq. (g) yields

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

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 $p\omega\tau$. □

C 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 a 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 $\mathcal{A}^{\pm}_{i\,j}$ such that each random element of the matrix is independently and each column is identically distributed. If we apply g to each row of the matrix,

$$\mathrm{Y}_i = \mathrm{g}\left(\mathcal{A}^{\pm}_{i\,1}, \ldots, \mathcal{A}^{\pm}_{i\,q}\right) \tag{C.1}$$

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

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

$$\overline{\mathrm{Y}}_\mathrm{m} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Y}_\mathrm{i}\,. \tag{C.2}$$

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

$$\lim_{m \to \infty} \overline{\mathrm{Y}}_\mathrm{m} \tag{C.3}$$

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

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

References

[1] Michael Mitzenmacher Andrei Broder. Network applications of bloom filters: A survey. 2005.

[2] T. Hickey, Q. Ju, and M. H. Van Emden. Interval arithmetic: From principles to implementation. *ACM*, page 1038–1068, 2001.

[3] Hoeffing. Hoeffing's inequality.

[4] Alexander Towell. The perfect hash filter: an efficient, immutable implementation of the positive random approximate set abstract data type. 2018.