A complexity theoretic approach to logical uncertainty (Draft)

Vadim Kosoy

June 30, 2015

Abstract

We interpret logical uncertainty as the assignment of probabilities to outcomes of computations within computing resources smaller than sufficient for the computation. To formalize this idea, we introduce the concept of optimal predictors for distributional decision problems. We show that this concept satisfies theorems analogical to theorems of classical probability theory and give a few examples.

1 Introduction

Logical uncertainty is usually thought of as the assignment of probabilities to sentences in a formal theory. Certain formalisms [TBD] were proposed in which these probabilities come from a probability measure over models. Such constructions are uncomputable for the fundamental reason that the promise problem of separating provable from refutable sentences is undecidable. Since uncomputable functions cannot be used in realistic AGI architectures, it was suggested [TBD] to look for computable approximations of these uncomputable functions. Here, we take a different point of view.

The simplest argument for the meaningfulness of logical uncertainty is its use in wagers. A rational agent should only accept a bet at odds which correspond to positive expected gain. The critical odds correspond to the probabilities of the wager outcomes. In classical probability theory there is no prescription for assigning probabilities to results of computations (e.g. the probability the millionth digit of π is 7). For an agent with unbounded computing resources, this kind of wager is deterministic and should be either accepted or rejected regardless of odds. However, for an agent with bounded computing resources, that is unable to evaluate the computation within the allotted time, it stands to reason to treat the result as uncertain and accept the wager starting from certain odds. The determination of these odds would be the subject of logical uncertainty.

This point of view differs from the point of view based on logic since it only calls assigning probabilities to the results of finite computations. Wagers on infinite computations are meaningless for agents with time discount. Agents without time discount suffer from a number of problems (Pascal mugging [TBD], procrastination paradox [TBD], Poincare recurrences [TBD]) and might be meaningless altogether.

2 Background and Notation

$\theta(t)$ will denote the step function:

$$\theta(t) := \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$\eta(t)$ will denote the function

$$\eta(t) := \begin{cases} 1 & \text{if } t \geq 1 \\ t & \text{if } 0 < t < 1 \\ 0 & \text{if } t \leq 0 \end{cases}$$

$\{0,1\}^*$ is the set of all finite binary strings (words). Given $x \in \{0,1\}^*$, $|x|$ denotes the length of x. We also denote

$$\{0,1\}^k := \{x \in \{0,1\}^* \mid |x| = k\}$$

$$\{0,1\}^{\leq k} := \{x \in \{0,1\}^* \mid |x| \leq k\}$$

U^k denotes the uniform probability measure on $\{0,1\}^k$.

A *language* is a subset of $\{0,1\}^*$.

For X a subset of Y, we denote by $\chi_X : Y \to \{0,1\}$ the characteristic function of X i.e.

$$\chi_X(x) := \begin{cases} 1 & \text{if } x \in X \\ 0 & \text{if } x \notin X \end{cases}$$

Definition 2.1. A function $\delta : \mathbb{N} \to \mathbb{R}$ is called *negligible* when

$$\forall k \in \mathbb{N} : \lim_{n \to \infty} n^k \delta(n) = 0$$

A function $f : \mathbb{N} \to \mathbb{R}$ is called *superpolynomial* when $\frac{1}{f}$ is negligible.

2.1 Algorithms and Boolean Circuits

Given sets X and Y we use the notation $A : X \xrightarrow{alg} Y$ to mean a Turing machine halting on every input which receives an element of X as input and produces an element of Y as output. This implicitly assumes a function $f : X \to \{0,1\}^*$ for encoding elements of X and a function $g : \{0,1\}^* \to Y$ for decoding elements of Y. When $Y = [0,1]$, g is assumed to map the bit string $b_1 b_2 \dots b_n$ to the finite binary fraction $0.b_1 b_2 \dots b_n$. When X or Y is defined as a subset of $\mathbb{Q}$, f or g is assumed to interpret the binary string as encoding the nominator and denominator.

When X is finite, we use the notation $B : X \xrightarrow{circ} Y$ to mean a Boolean circuit which receives an element of X as input and produces an element of Y as output. This implicitly assumes a function $f : X \to \{0,1\}^n$ for encoding elements of X and a function $g : \{0,1\}^m \to Y$ for decoding elements of Y. B has n inputs and m outputs. When X is a finite subset of $\{0,1\}^*$, $f(x) = 0^{\max_{y \in X} |y| - |x|} 1x$.

We denote the size of the Boolean circuit by $|B|$.
All assumptions about encoding naturally extend to Cartesian products.

2.2 Average-Case Complexity

This subsection loosely follows [1].

Definition 2.2. A *word ensemble* μ is a sequence of probability measures $\{\mu^k : \{0,1\}^* \to [0,1]\}_{k\in\mathbb{N}}$ s.t. for some polynomial p we have

$$\forall k \in \mathbb{N}, x \in \{0,1\}^* : |x| > p(k) \implies \mu^k(x) = 0$$

We call the above p a *polynomial bound* for μ.

Definition 2.3. A word ensemble μ is called *computable* when there is $M : \mathbb{N} \times \{0,1\}^* \xrightarrow{alg} [0,1]$ s.t. $M(k,x)$ is computed in time polynomial in k and $|x|$ and

$$\forall k \in \mathbb{N}, x^* \in \{0,1\}^* : M(k,x^*) = \mu^k\{x \in \{0,1\}^* \mid x < x^*\}$$

where the order on $\{0,1\}^*$ is lexicographic.

Definition 2.4. A word ensemble μ is called *samplable* when there is a polynomial p and $M : \mathbb{N} \times \{0,1\}^* \xrightarrow{alg} \{0,1\}^*$ s.t. $M(k,r)$ is computed in time polynomial in k and $|r|$ and

$$\forall k \in \mathbb{N}, x \in \{0,1\}^* : \mu^k(x) = Pr_{U^{p(k)}}[M(k,r) = x]$$

A *distributional decision problem* is a pair (D,μ) where D is a language and μ is a word ensemble.

Definition 2.5. Consider (D,μ) a distributional decision problem. A heuristic algorithm for (D,μ) is $M : \{0,1\}^* \xrightarrow{alg} \{0,1\}$ s.t. $\mu^k\{x \in \{0,1\}^* \mid M(x) \neq \chi_D(x)\}$ is a negligible function of k.

The set of distributional decision problems for which there is a polynomial time heuristic algorithm is denoted $Heur_{neg}P$.

Definition 2.6. Consider (D,μ) a distributional decision problem. A heuristic family of circuits for (D,μ) is a family of circuits $\{C^k : \operatorname{supp}\mu^k \xrightarrow{circ} \{0,1\}\}_k$ s.t. $\mu^k\{x \in \operatorname{supp}\mu^k \mid C^k(x) \neq \chi_D(x)\}$ is a negligible function of k.

The set of distributional decision problems for which there is a polynomial size heuristic family of circuits is denoted $Heur_{neg}P/poly$.

2.3 One-Way Functions

Definition 2.7. $f : \{0,1\}^* \xrightarrow{alg} \{0,1\}^*$ is called a *non-uniformly hard one-way function* when it runs in polynomial time and for any polynomial size family of circuits $\{g^k : f(\{0,1\}^k) \xrightarrow{circ} \{0,1\}^k\}_{k\in\mathbb{N}}$, $Pr_{U^k}[g^k(f(x)) = x]$ is a negligible function of of k.

The following theorem is adapted from [2], where it appears as Theorem 7.7:

Theorem 2.1. *Consider f a non-uniformly hard one-way function and a polynomial size family of circuits $\{g^k : f(\{0,1\}^k) \times \{0,1\}^k \xrightarrow{circ} \{0,1\}\}_{k\in\mathbb{N}}$. Then, there is a negligible function δ s.t.*

$$\forall k \in \mathbb{N} : Pr_{U^k \times U^k}[g^k(f(x), r) = x \cdot r] \leq \frac{1}{2} + \delta(k)$$

Here, $x \cdot r$ is defined as $\sum x_i r_i$ by identifying $\{0,1\}$ with $\mathbb{F}_2$.

The formulation given in [2] speaks of uniformly hard f and uniform g but adapting the proof to the non-uniform case is straightforward.

3 Optimal Predictors

Definition 3.1. Given (D, μ) a distributional decision problem, an *optimal predictor* for (D, μ) is a family of polynomial size Boolean circuits

$$\{P^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$$

s.t. for any family of polynomial size Boolean circuits

$$\{Q^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$$

we have

$$E_{\mu^k}[(P^k(x) - \chi_D(x))^2] \leq E_{\mu^k}[(Q^k(x) - \chi_D(x))^2] + \delta(k)$$

where δ is a negligible function.

The set of distributional decision problems for which an optimal predictor exists is denoted $OP_{neg}P/poly$. The set of distributional decision problems for which a uniform optimal predictor exists (that is, P^k can be computed by an algorithm running in time polynomial in k) is denoted $OP_{neg}P$.

Note 3.1. Definition 3.1 uses non-uniform computation (Boolean circuit families) rather than uniform computation (Turing machines). The use of non-uniform Q is crucial to derive the theorems in section 4. The use of non-uniform P makes the concept more general i.e. applicable to a larger class of distributional decision problems. That said, the special cases in which P is uniform (i.e. P^k can be computed in time polynomial in k) are interesting and important.

It is often handy to use the following stronger characterization of optimal predictors.

Lemma 3.1. *Consider (D, μ) a distributional decision problem and*

$$\{P^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$$

a family of polynomial size. Then, P is an optimal predictor if and only if there is a negligible function δ s.t. for any $Q : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]$ we have

$$E_{\mu^k}[(P^k(x) - \chi_D(x))^2] \leq E_{\mu^k}[(Q(x) - \chi_D(x))^2] + |Q|\delta(k)$$

The "if" part is obvious. To show the "only if" part we need the following propositions.

Proposition 3.1. *Consider $\Delta : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$. Assume that for any polynomial $p : \mathbb{N} \to \mathbb{N}$ there is δ negligible s.t. $\forall k : \Delta(p(k), k) \leq \delta(k)$. Then, there is $f : \mathbb{N} \to \mathbb{N}$ superpolynomial s.t. $\forall k : \Delta(f(k), k) \leq f(k)^{-1}$.*

Proof. For any $i \geq 1$, define $k_i := min\{k \mid \forall j \leq i, l \geq k : l^i \Delta(l^j, l) \leq \frac{1}{i}\}$. For any k, define $d(k) := \min(k, \max\{i \mid k_i \leq k\})$, $f(k) := k^{d(k)}$.

d is non-decreasing (obviously) and unbounded since $d(\max(i, k_i)) \geq i$. Therefore f is superpolynomial.

Since the k_i are a non-decreasing sequence and $d(k) \leq \max\{i \mid k_i \leq k\}$, $k_{d(k)} \leq k$. Hence $k^{d(k)} \Delta(k^{d(k)}, k) \leq \frac{1}{d(k)}$ and therefore $\Delta(f(k), k) \leq \frac{f(k)^{-1}}{d(k)}$.

□

Proposition 3.2. *Consider $\Delta : \mathbb{N} \times \mathbb{N} \to (-\infty, 1]$, non-decreasing in the first argument. Assume that for any polynomial $p : \mathbb{N} \to \mathbb{N}$ there is δ negligible s.t. $\forall k : \Delta(p(k), k) \leq \delta(k)$. Then there is δ^* negligible s.t. $\forall n \geq 1, m : \Delta(n, m) \leq n\delta^*(m)$.*

Proof. Apply Proposition 3.1 and take $\delta^*(k) := f(k)^{-1}$. For $n \leq f(m)$, $\Delta(n, m) \leq \Delta(f(m), m) \leq f(m)^{-1} \leq n\delta^*(m)$. For $n > f(m)$, $\Delta(n, m) \leq 1 = f(m)f(m)^{-1} \leq nf(m)^{-1} = n\delta^*(m)$.

□

Proof of Lemma 3.1. Define

$$\Delta(n, k) := \max_{|Q| \leq n} \{E_{\mu^k}[(P^k(x) - \chi_D(x))^2] - E_{\mu^k}[(Q(x) - \chi_D(x))^2]\}$$

Applying Proposition 3.2 yields the desired result.

□

4 Parallel to Probability Theory

In this section we derive a number of theorems which demonstrate a parallel between optimal predictors and classical probability theory.

In classical probability theory, a calibrated predictor has to assign probability within a certain interval with frequency which lies within the same interval. The analogous property of optimal predictors is as follows.

Theorem 4.1. *Consider (D, μ) a distributional decision problem and P an optimal predictor for (D, μ). Suppose $\{p_k \in [0, 1]\}_{k \in \mathbb{N}}$, $\{q_k \in [0, 1]\}_{k \in \mathbb{N}}$ are s.t. $\mu^k\{x \in \{0, 1\}^* \mid p_k \leq P^k(x) \leq q_k\}^{-1}$ is bounded by a polynomial in k. Then, $E_{\mu^k}[P^k(x) - \chi_D(x) \mid p_k \leq P^k(x) \leq q_k]$ is negligible in k.*

The relation of the theorem to calibration is made explicit by the following corollary:

Corollary 4.1. *Consider (D, μ) a distributional decision problem and P an optimal predictor for (D, μ). Suppose $\{p_k \in [0, 1]\}_{k \in \mathbb{N}}$, $\{q_k \in [0, 1]\}_{k \in \mathbb{N}}$ are s.t. $\mu^k\{x \in \{0, 1\}^* \mid p_k \leq P^k(x) \leq q_k\}^{-1}$ is bounded by a polynomial in k. Then, there is a negligible function ϵ s.t.*

$$\forall k : p_k - \epsilon(k) \leq Pr_{\mu^k}[x \in D \mid p_k \leq P^k(x) \leq q_k] \leq q_k + \epsilon(k)$$

Proof. Theorem 4.1 tells us that $\epsilon(k) := E_{\mu^k}[P^k(x) - \chi_D(x) \mid p_k \leq P^k(x) \leq q_k]$ is negligible. Thus

$$E_{\mu^k}[\chi_D(x) \mid p_k \leq P^k(x) \leq q_k] = E_{\mu^k}[P^k(x) \mid p_k \leq P^k(x) \leq q_k] - \epsilon(k)$$

$$Pr_{\mu^k}[x \in D \mid p_k \leq P^k(x) \leq q_k] = E_{\mu^k}[P^k(x) \mid p_k \leq P^k(x) \leq q_k] - \epsilon(k)$$

Since $p_k \leq E_{\mu^k}[P^k(x) \mid p_k \leq P^k(x) \leq q_k] \leq q_k$, it follows that

$$p_k - |\epsilon(k)| \leq Pr_{\mu^k}[x \in D \mid p_k \leq P^k(x) \leq q_k] \leq q_k + |\epsilon(k)|$$

□

In order to proof theorem 4.1, we are going to need the following lemma:

Lemma 4.1. *Consider (D, μ) a distributional decision problem and P a corresponding optimal predictor. Then, there is a polynomial in two variables p and a negligible function δ s.t. for all $k \in \mathbb{N}$, $Q : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]$ and $w : \operatorname{supp} \mu^k \xrightarrow{circ} \mathbb{Q}^{\geq 0}$ we have*

$$E_{\mu^k}[w(x)(P^k(x) - \chi_D(x))^2] \leq E_{\mu^k}[w(x)(Q(x) - \chi_D(x))^2] + (\max w) p(|Q|, |w|)\delta(k)$$

Proof. Given $t \in [0, \max w]$, denote

$$\alpha(t) := min\{s \geq t \mid \exists x \in \operatorname{supp} \mu^k : w(x) = s\}$$

Consider circuit $Q_t : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]$ computing the following function:

$$Q_t(x) := \begin{cases} Q(x) & \text{if } w(x) \geq \alpha(t) \\ P^k(x) & \text{if } w(x) < \alpha(t) \end{cases}$$

There is a polynomial q s.t. $|Q_t| \leq q(k, |Q|, |w|)$. By Lemma 3.1,

$$E_{\mu^k}[(P^k(x) - \chi_D(x))^2] \leq E_{\mu^k}[(Q_t(x) - \chi_D(x))^2] + q(k, |Q|, |w|)\delta(k)$$

for δ negligible.

$$E_{\mu^k}[(P^k(x) - \chi_D(x))^2 - (Q_t(x) - \chi_D(x))^2] \leq q(k, |Q|, |w|)\delta(k)$$

$$E_{\mu^k}[\theta(w(x) - t)(P^k(x) - \chi_D(x))^2 - (Q(x) - \chi_D(x))^2] \leq q(k, |Q|, |w|)\delta(k)$$

Integrating the inequality with respect to t from 0 to $\max w$, we get

$$E_{\mu^k}[\int_0^{\max w} \theta(w(x) - t)\, dt\, ((P^k(x) - \chi_D(x))^2 - (Q(x) - \chi_D(x))^2] \leq (\max w) q(k, |Q|, |w|)\delta(k)$$

$$E_{\mu^k}[w(x)(P^k(x) - \chi_D(x))^2 - (Q(x) - \chi_D(x))^2] \leq (\max w) q(k, |Q|, |w|)\delta(k)$$

The dependence of q on k can be eliminated by redefining δ, yielding the desired result.

□

The proof of Theorem 4.1 is going to use only the special case of Lemma 4.1 when w takes values in $\{0, 1\}$. However, we will need the general case later.

Proof of Theorem 4.1. Define

$$\phi_k := E_{\mu^k}[\chi_D(x) - P^k(x) \mid p_k \leq P^k(x) \leq q_k]$$

Assume to the contrary that there is a positive polynomial $q(k)$ and an infinite set $I \subseteq \mathbb{N}$ s.t.

$$\forall k \in I : |\phi_k| \geq q(k)^{-1}$$

Define $\{w^k : \operatorname{supp} \mu^k \xrightarrow{circ} \{0, 1\}\}_{k \in \mathbb{N}}$ as the circuits computing

$$w^k(x) := \theta(P^k(x) - p_k)\theta(q_k - P^k(x))$$

$|w^k|$ is bounded by a polynomial since P^k produces binary fractions of polynomial size therefore it is possible to compare them to the fixed numbers p_k, q_k using a polynomial size circuit even if the latter have infinite binary expansions.

We have

$$\phi_k = \frac{E_{\mu^k}[w^k(x)(\chi_D(x) - P^k(x))]}{E_{\mu^k}[w^k(x)]}$$

Define ψ_k to be ϕ_k truncated to the first significant binary digit. Define $\{Q^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0, 1]\}_{k \in \mathbb{N}}$ as the circuits computing

$$Q^k(x) := \eta(P^k(x) + \psi_k)$$

By the assumption, ψ_k has binary notation of at most logarithmic size in k, therefore $|Q^k|$ is bounded by a polynomial.

Applying Lemma 4.1 we get

$$\forall k \in I : E_{\mu^k}[w^k(x)(P^k(x) - \chi_D(x))^2] \leq E_{\mu^k}[w^k(x)(Q^k(x) - \chi_D(x))^2] + \delta(k)$$

for some negligible δ.

$$\forall k \in I : E_{\mu^k}[w^k(x)((P^k(x) - \chi_D(x))^2 - (Q^k(x) - \chi_D(x))^2)] \leq \delta(k)$$

$$\forall k \in I : E_{\mu^k}[w^k(x)((P^k(x) - \chi_D(x))^2 - (\eta(P^k(x) + \psi_k) - \chi_D(x))^2)] \leq \delta(k)$$

Obviously $(\eta(P^k(x) + \psi_k) - \chi_D(x))^2 \leq (P^k(x) + \psi_k - \chi_D(x))^2$, therefore

$$\forall k \in I : E_{\mu^k}[w^k(x)((P^k(x) - \chi_D(x))^2 - (P^k(x) + \psi_k - \chi_D(x))^2)] \leq \delta(k)$$

$$\forall k \in I : \psi_k E_{\mu^k}[w^k(x)(2(\chi_D(x) - P^k(x)) - \psi_k)] \leq \delta(k)$$

The expression on the left hand side is a quadratic polynomial in ψ_k which attains its maximum at ϕ_k and has roots at 0 and $2\phi_k$. ψ_k is between 0 and ϕ_k, but not closer to 0 than $\frac{\phi_k}{2}$. Therefore, the inequality is preserved if we replace ψ_k by $\frac{\phi_k}{2}$.

$$\forall k \in I : \frac{\phi_k}{2} E_{\mu^k}[w^k(x)(2(\chi_D(x) - P^k(x)) - \frac{\phi_k}{2})] \leq \delta(k)$$

Substituting the equation for ϕ_k we get

$$\forall k \in I : \frac{1}{2}\frac{E_{\mu^k}[w^k(x)(\chi_D(x) - P^k(x))]}{E_{\mu^k}[w^k(x)]} E_{\mu^k}[w^k(x)(2(\chi_D(x) - P^k(x)) - \frac{1}{2}\frac{E_{\mu^k}[w^k(x)(\chi_D(x) - P^k(x))]}{E_{\mu^k}[w^k(x)]})] \leq \delta(k)$$

$$\forall k \in I : \frac{3}{4}\frac{E_{\mu^k}[w^k(x)(\chi_D(x) - P^k(x))]^2}{E_{\mu^k}[w^k(x)]} \leq \delta(k)$$

$$\forall k \in I : \frac{3}{4} E_{\mu^k}[w^k(x)]\phi_k^2 \leq \delta(k)$$

$$\forall k \in I : \phi_k^2 \leq \frac{4}{3} E_{\mu^k}[w^k(x)]^{-1}\delta(k)$$

$$\forall k \in I : \phi_k^2 \leq \frac{4}{3}\mu^k\{x \in \{0,1\}^* \mid p_k \leq P^k(x) \leq q_k\}^{-1}\delta(k)$$

Thus ϕ_k is negligible on I, which is a contradiction.

□

The probability of the disjunction of mutually exclusive events is the sum of their probabilities. The analogous property of optimal predictors is as follows.

Theorem 4.2. *Consider μ a word ensemble and D_1, D_2 disjoint languages. Suppose P_1 is an optimal predictor for (D_1, μ) and P_2 is an optimal predictor for (D_2, μ). Then, $P := \eta(P_1 + P_2)$ is an optimal predictor for $(D_1 \cup D_2, \mu)$.*

In order to prove this theorem, we will use the following lemma.

Lemma 4.2. *Consider (D, μ) a distributional decision problem. If $\{P^k : \operatorname{supp}\mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$ is an optimal predictor for (D, μ) then there is a polynomial p^* and a negligible function δ^* s.t. for any $Q : \operatorname{supp}\mu^k \xrightarrow{circ} \mathbb{Q}$ we have*

$$|E_{\mu^k}[Q(x)(P^k(x) - \chi_D(x))]| \leq p^*(|Q|, E_{\mu^k}[Q(x)^2])\delta^*(k)$$

Conversely, suppose $M \in \mathbb{Q}$ and $\{P^k : \operatorname{supp}\mu^k \xrightarrow{circ} \mathbb{Q} \cap [-M, +M]\}_{k\in\mathbb{N}}$ is a polynomial size family for which there is a polynomial p^* and a negligible function δ^* s.t. for any $Q : \operatorname{supp}\mu^k \xrightarrow{circ} \mathbb{Q} \cap [-M-1, +M]\}_{k\in\mathbb{N}}$ we have

$$|E_{\mu^k}[Q(x)(P^k(x) - \chi_D(x))]| \leq p^*(|Q|)\delta^*(k)$$

Define $\{\tilde{P}^k : \operatorname{supp}\mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$ *to be s.t. computing* $\tilde{P}^k(x)$ *is equivalent to computing* $\eta(P^k(x))$ *rounded to* k *digits after the binary point. Then,* $\tilde{P}$ *is an optimal predictor.*

Note 4.1. Lemma 4.2 is analogous to the fact that given V a Hilbert space, $W \subseteq V$ a closed subspace and $x \in V$, the vector $y^* := \arg\min_{y\in W} \|y - x\|^2$ satisfies $\forall z \in W : \langle z, y^* - x\rangle = 0$

Proof. Assume P is an optimal predictor. Consider $Q : \operatorname{supp}\mu^k \xrightarrow{circ} \mathbb{Q}$ and $t = \sigma 2^{-a}$ where $\sigma \in \{\pm 1\}$ and $a \in \mathbb{N}$. The function $\eta(P^k(x) + tQ(x))$ can be approximated by a circuit of size $p(a, |Q|)$ for some fixed polynomial p, within rounding error $\epsilon_k(x)$ s.t. $\forall x \in \operatorname{supp}\mu^k : |\epsilon_k(x)| \leq 2^{-k}$. By Lemma 3.1,

$$E_{\mu^k}[(P^k(x)-\chi_D(x))^2] \leq E_{\mu^k}[(\eta(P^k(x)+tQ(x))+\epsilon_k(x)-\chi_D(x))^2]+p(a,|Q|)\delta(k)$$

where δ is negligible. ϵ is bounded by a negligible function and therefore can be ignored by redefining p and δ. As in the proof of Theorem 4.1, η can be dropped.

$$E_{\mu^k}[(P^k(x) - \chi_D(x))^2 - (P^k(x) + tQ(x) - \chi_D(x))^2] \leq p(a, |Q|)\delta(k)$$

The expression on the left hand side is a quadratic polynomial in t. Explicitly:

$$-E_{\mu^k}[Q(x)^2]t^2 - 2E_{\mu^k}[Q(x)(P^k(x) - \chi_D(x))]t \leq p(a, |Q|)\delta(k)$$

Moving $E_{\mu^k}[Q(x)^2]t^2$ to the right hand side and dividing both sides by $2|t| = 2^{1-a}$ we get

$$-E_{\mu^k}[Q(x)(P^k(x) - \chi_D(x))]\sigma \leq 2^{a-1}p(a, |Q|)\delta(k) + E_{\mu^k}[Q(x)^2]2^{-a-1}$$

$$|E_{\mu^k}[Q(x)(P^k(x) - \chi_D(x))]| \leq 2^{a-1}p(a, |Q|)\delta(k) + E_{\mu^k}[Q(x)^2]2^{-a-1}$$

Take $a := -\frac{1}{2}\log\delta(k) + \phi(k)$ where $\phi(k) \in [-\frac{1}{2}, +\frac{1}{2}]$ is the rounding error. We get

$$|E_{\mu^k}[Q(x)(P^k(x)-\chi_D(x))]| \leq 2^{\phi(k)-1}p(-\frac{1}{2}\log\delta(k)+\phi(k), |Q|)\delta(k)^{1/2}+E_{\mu^k}[Q(x)^2]2^{-\phi(k)-1}\delta(k)^{1/2}$$

Since $\delta(k)$ is negligible, so is $\delta(k)^{1/2}$ and factors polynomial in $\log\delta(k)$ don't interfere with negligibility. We get the required bound.

Conversely, assume that for any $R : \operatorname{supp}\mu^k \xrightarrow{circ} \mathbb{Q} \cap [-M-1, +M]$

$$|E_{\mu^k}[R(x)(P^k(x) - \chi_D(x))]| \leq p^*(|R|)\delta^*(k)$$

Consider $Q : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]$. We have

$$E_{\mu^k}[(Q(x) - \chi_D(x))^2] = E_{\mu^k}[(Q(x) - P^k(x) + P^k(x) - \chi_D(x))^2]$$

$$E_{\mu^k}[(Q(x)-\chi_D(x))^2] = E_{\mu^k}[(Q(x)-P^k(x))^2]+E_{\mu^k}[(P^k(x)-\chi_D(x))^2]+2E_{\mu^k}[(Q(x)-P^k(x))(P^k(x)-\chi_D(x)]$$

$$2E_{\mu^k}[(P^k(x)-Q(x))(P^k(x)-\chi_D(x)] = E_{\mu^k}[(P^k(x)-\chi_D(x))^2]-E_{\mu^k}[(Q(x)-\chi_D(x))^2]+E_{\mu^k}[(Q(x)-P^k(x))^2]$$

$P^k(x) - Q(x)$ can be computed by a circuit R of size polynomial in $|Q|$ and k. Applying the assumption we get

$$E_{\mu^k}[(P^k(x)-\chi_D(x))^2]-E_{\mu^k}[(Q(x)-\chi_D(x))^2]+E_{\mu^k}[(Q(x)-P^k(x))^2] \leq \tilde{p}(|Q|)\tilde{\delta}(k)$$

where $\tilde{p}$ is a polynomial and $\tilde{\delta}$ is negligible. Here, we absorbed the polynomial dependence on k into the definition of $\tilde{\delta}$. Noting that $E_{\mu^k}[(Q(x) - P^k(x))^2] \geq 0$ and $(\eta(P^k(x)) - \chi_D(x))^2 \leq (P^k(x) - \chi_D(x))^2$ we get

$$E_{\mu^k}[(\eta(P^k(x)) - \chi_D(x))^2] - E_{\mu^k}[(Q(x) - \chi_D(x))^2] \leq \tilde{p}(|Q|)\tilde{\delta}(k)$$

Observing that $\tilde{P} - \eta(P)$ is bounded by a negligible function, we get the desired result.

□

Proof of Theorem 4.2. Consider $Q : \operatorname{supp} \mu^k \xrightarrow{circ} \mathbb{Q}$. We have

$$E_{\mu^k}[Q(x)(P_1^k(x)+P_2^k(x)-\chi_{D_1\cup D_2}(x))] = E_{\mu^k}[Q(x)(P_1^k(x)-\chi_{D_1}(x))]+E_{\mu^k}[Q(x)(P_2^k(x)-\chi_{D_2}(x))]$$

Using Lemma 4.2:

$$|E_{\mu^k}[Q(x)(P_1^k(x) - \chi_{D_1}(x))]| \leq p_1(|Q|, E_{\mu^k}[Q(x)^2]))\delta_1(k)$$

$$|E_{\mu^k}[Q(x)(P_2^k(x) - \chi_{D_2}(x))]| \leq p_2(|Q|, E_{\mu^k}[Q(x)^2])\delta_2(k)$$

Therefore

$$|E_{\mu^k}[Q(x)(P_1^k(x)+P_2^k(x)-\chi_{D_1\cup D_2}(x))]| \leq (p_1(|Q|, E_{\mu^k}[Q(x)^2])+p_2(|Q|, E_{\mu^k}[Q(x)^2]))(\delta_1(k)+\delta_2(k))$$

Using Lemma 4.2 again we get the desired result.

□

Similarly, we have

Theorem 4.3. *Consider μ a word ensemble and D_1, D_2 disjoint languages. Suppose P_1 is an optimal predictor for (D_1, μ) and P is an optimal predictor for $(D_1 \cup D_2, \mu)$. Then, $P_2 := \eta(P - P_1)$ is an optimal predictor for (D_2, μ).*

Proof. Consider $Q : \operatorname{supp} \mu^k \xrightarrow{circ} \mathbb{Q}$. We have

$$E_{\mu^k}[Q(x)(P^k(x)-P_1^k(x)-\chi_{D_2}(x))] = E_{\mu^k}[Q(x)(P^k(x)-\chi_{D_1\cup D_2}(x))]-E_{\mu^k}[Q(x)(P_1^k(x)-\chi_{D_1}(x))]$$

Using Lemma 4.2:

$$|E_{\mu^k}[Q(x)(P^k(x) - \chi_{D_1\cup D_2}(x))]| \leq p_1(|Q|, E_{\mu^k}[Q(x)^2])\delta_1(k)$$

$$|E_{\mu^k}[E_{\mu^k}[Q(x)(P_1^k(x) - \chi_{D_1}(x))]| \leq p_2(|Q|, E_{\mu^k}[Q(x)^2])\delta_2(k)$$

Therefore

$$|E_{\mu^k}[Q(x)(P^k(x)-P_1^k(x)-\chi_{D_2}(x))]| \leq (p_1(|Q|, E_{\mu^k}[Q(x)^2])+p_2(|Q|, E_{\mu^k}[Q(x)^2]))(\delta_1(k)+\delta_2(k))$$

Using Lemma 4.2 again we get the desired result.

□

The probability of the conjunction of independent events is the products of their probabilities. The analogous property of optimal predictors is as follows.

Theorem 4.4. *Consider (D_1, μ_1), (D_2, μ_2) distributional decision problems with respective optimal predictors P_1 and P_2. Consider the language $D_1 \times D_2 := \{(x_1, x_2) \in \{0,1\}^* \mid x_1 \in D_1, x_2 \in D_2\}$ where pairs are understood to be encoded in some standard way. Consider the word ensemble $\mu_1 \times \mu_2$ defined by $(\mu_1 \times \mu_2)^k((x_1, x_2)) := \mu_1^k(x_1)\mu_2^k(x_2)$, whereas for any word y which cannot be decoded to a pair $(\mu_1 \times \mu_2)^k(y) := 0$. Define $\{P^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$ as the family of circuits computing $P^k((x_1, x_2)) := P_1^k(x_1)P_2^k(x_2)$. Then, P is an optimal predictor for $(D_1 \times D_2, \mu_1 \times \mu_2)$.*

Proof. We have

$$P^k((x_1, x_2))-\chi_{D_1\times D_2}((x_1, x_2)) = (P_1^k(x_1)-\chi_{D_1}(x_1))\chi_{D_2}(x_2)+P_1^k(x_1)(P_2^k(x_2)-\chi_{D_2}(x_2))$$

Therefore, for any $Q : \operatorname{supp} (\mu_1 \times \mu_2)^k \xrightarrow{circ} \mathbb{Q} \cap [-1, +1]$

$$|E_{(\mu_1\times\mu_2)^k}[Q(x)(P^k(x) - \chi_{D_1\times D_2}(x))]| \leq |E_{\mu_1^k\times\mu_2^k}[Q((x_1, x_2))(P_1^k(x_1) - \chi_{D_1}(x_1))\chi_{D_2}(x_2)]|+ |E_{\mu_1^k\times\mu_2^k}[Q((x_1, x_2))P_1^k(x_1)(P_2^k(x_2) - \chi_{D_2}(x_2))]|$$

By Lemma 4.2, it is sufficient to show an appropriate bound for each of the terms on the right hand side. For the first term, we have

$$|E_{\mu_1^k\times\mu_2^k}[Q((x_1, x_2))(P_1^k(x_1)-\chi_{D_1}(x_1))\chi_{D_2}(x_2)]| \leq E_{\mu_2^k}[|E_{\mu_1^k}[\chi_{D_2}(x_2)Q((x_1, x_2))(P_1^k(x_1)-\chi_{D_1}(x_1))]|]$$

For any given x_2, $\chi_{D_2}(x_2)Q((x_1, x_2))$ can be computed by a circuit with input x_1 of size polynomial in $|x_2|$ and $|Q|$. Applying Lemma 4.2 to P_1, we get

$$|E_{\mu_1^k\times\mu_2^k}[Q((x_1, x_2))(P_1^k(x_1) - \chi_{D_1}(x_1))\chi_{D_2}(x_2)]| \leq E_{\mu_2^k}[p_1(|x_2|, |Q|)\delta_1(k)]$$

where p_1 is a polynomial and δ_1 is negligible. Since $|x_2|$ is bounded by a polynomial in k for $x_2 \in \operatorname{supp} \mu_2^k$, we get the bound we need.

For the second term, we have

$$|E_{\mu_1^k \times \mu_2^k}[Q((x_1,x_2))P_1^k(x_1)(P_2^k(x_2) - \chi_{D_2}(x_2))]| \leq E_{\mu_1^k}[|E_{\mu_1^k}[Q((x_1,x_2))P_1^k(x_1)(P_2^k(x_2) - \chi_{D_2}(x_2))]|]$$

For any given x_1, $Q((x_1,x_2))P_1^k(x_1)$ can be computed by a circuit with input x_1 of size polynomial in k, $|x_1|$ and $|Q|$. Applying Lemma 4.2 to P_2, we get

$$|E_{\mu_1^k \times \mu_2^k}[Q((x_1,x_2))P_1^k(x_1)(P_2^k(x_2) - \chi_{D_2}(x_2))]| \leq E_{\mu_1^k}[p_2(k,|x_1|,|Q|)\delta_2(k)]$$

Again, we got the required bound.

□

In classical probability theory, conditional probability is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

It is natural to regard an optimal predictor for $(C, \mu \mid D)$ as the conditional probability of $x \in C$ given $x \in D$. This point of view can be connected to the above equation via the following theorems.

Theorem 4.5. *Consider $C, D \subseteq \{0,1\}^*$ and μ a word ensemble. Assume P_D is an optimal predictor for (D, μ) and $P_{C|D}$ is an optimal predictor for $(C, \mu \mid D)$. Then $P_D P_{C|D}$ is an optimal predictor for $(C \cap D, \mu)$.*

Proof. Consider $Q : \operatorname{supp} \mu^k \xrightarrow{circ} \mathbb{Q} \cap [-1,+1]$. We have

$$E_{\mu^k}[Q(x)(P_D^k(x)P_{C|D}^k(x) - \chi_{C\cap D}(x))] = E_{\mu^k}[Q(x)(P_D^k(x)P_{C|D}^k(x) - \chi_C(x)\chi_D(x)]$$

$$P_D^k(x)P_{C|D}^k(x) - \chi_C(x)\chi_D(x) = (P_{C|D}^k(x) - \chi_C(x))\chi_D(x) + P_{C|D}^k(x)(P_D^k(x) - \chi_D(x))$$

$$|E_{\mu^k}[Q(x)(P_D^k(x)P_{C|D}^k(x) - \chi_{C\cap D}(x))]| \leq |E_{\mu^k}[Q(x)(P_{C|D}^k(x) - \chi_C(x))\chi_D(x)]| + |E_{\mu^k}[Q(x)P_{C|D}^k(x)(P_D^k(x) - \chi_D(x)]|$$

By Lemma 4.2, it is sufficient to show an appropriate bound for each of the terms on the right hand side. For the first term, we have

$$E_{\mu^k}[Q(x)(P_{C|D}^k(x) - \chi_C(x))\chi_D(x)] = E_{\mu^k}[\chi_D(x)]E_{\mu^k|D}[Q(x)(P_{C|D}^k(x) - \chi_C(x))]$$

This gives the desired bound by applying Lemma 4.2 for $P_{C|D}$.

Define $R : \operatorname{supp} \mu^k \xrightarrow{circ} \mathbb{Q} \cap [0,1]$ to be the circuit computing $Q(x)P_{C|D}^k(x)$. The second term can be written as $E_{\mu^k}[R(x)(P_D^k(x) - \chi_D(x)]$. Applying Lemma 4.2 for P_D and using the fact $|P_{C|D}^k|$ is polynomial in k we get the desired result.

□

Theorem 4.6. *Consider $C, D \subseteq \{0,1\}^*$ and μ a word ensemble. Assume $\mu^k(D)^{-1}$ is bounded by a polynomial. Assume P_D is an optimal predictor for (D, μ) and $P_{C\cap D}$ is an optimal predictor for $(C \cap D, \mu)$. Define $P_{C|D}$ as the circuit family computing*

$$P^k_{C|D}(x) := \begin{cases} 1 & \text{if } P^k_D(x) = 0 \\ \eta(\frac{P^k_{C\cap D}(x)}{P^k_D(x)}) & \text{rounded to } k \text{ binary places if } P^k_D(x) > 0 \end{cases}$$

Then, $P_{C|D}$ is an optimal predictor for $(C, \mu \mid D)$.

The proof of Theorem 4.6 uses Theorem 5.1 and will therefore be given in the next section. It also uses the following. In classical probability theory, $A \subseteq B$ implies $Pr(A) \leq Pr(B)$. The analogous property of optimal predictors is as follows.

Lemma 4.3. *Consider $C \subseteq D \subseteq \{0,1\}^*$ and μ a word ensemble. Assume P_C is an optimal predictor for (C, μ) and P_D is an optimal predictor for (D, μ). Define*

$$\epsilon^k(x) := \theta(P^k_C(x) - P^k_D(x))(P^k_C(x) - P^k_D(x))$$

Then, $E_{\mu^k}[\epsilon^k(x)]$ is negligible.

Proof. By Theorem 4.3 and Lemma 4.2 there are a polynomial p and a negligible function δ such that for any $Q : \operatorname{supp} \mu^k \xrightarrow{circ} \mathbb{Q} \cap [-2, +1]$ we have

$$|E_{\mu^k}[Q(x)(P^k_D(x) - P^k_C(x) - \chi_{D\setminus C}(x))]| \leq p(|Q|)\delta(k)$$

Take Q to be the circuit computing $\theta(P^k_C(x) - P^k_D(x))$. Its size is polynomial in k therefore

$$|E_{\mu^k}[\theta(P^k_C(x) - P^k_D(x))(P^k_D(x) - P^k_C(x) - \chi_{D\setminus C}(x))]| \leq \delta^*(k)$$

where δ^* is negligible.

$$|E_{\mu^k}[\epsilon^k(x)] + E_{\mu^k}[\theta(P^k_C(x) - P^k_D(x))\chi_{D\setminus C}(x)]| \leq \delta^*(k)$$

Since both terms inside the absolute value are non-negative we get the desired result.

□

5 Uniqueness

The optimal predictor cannot be literally unique, since its definition only depends on asymptotic properties. However, we will show it to be unique up to the following equivalence relation.

Definition 5.1. Consider μ a word ensemble and $\{Q^k_{1,2} : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$ two circuit families. We say Q_1 is *similar* to Q_2 relative to μ (denoted $Q_1 \overset{\mu}{\sim} Q_2$) when $E_{\mu^k}[(Q^k_1(x) - Q^k_2(x))^2]$ is negligible in k.

Theorem 5.1. *Consider (D, μ) a distributional decision problem, P an optimal predictor for (D, μ) and $\{Q^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$ a polynomial size family. Then, Q is an optimal predictor for (D, μ) if and only if $P \overset{\mu}{\sim} Q$.*

Proof. Assume Q is an optimal predictor. Applying Lemma 4.2 to predictor P and circuits computing $P^k - Q^k$, we get

$$|E_{\mu^k}[(P^k(x) - Q^k(x))(P^k(x) - \chi_D(x))]| \leq \delta(k)$$

for some negligible δ. Applying Lemma 4.2 to predictor Q and circuits computing $P^k - Q^k$, we get

$$|E_{\mu^k}[(P^k(x) - Q^k(x))(Q^k(x) - \chi_D(x))]| \leq \epsilon(k)$$

for some negligible ϵ. We have

$$E_{\mu^k}[(P^k(x)-Q^k(x))^2] = E_{\mu^k}[(P^k(x)-Q^k(x))(P^k(x)-\chi_D(x))]-E_{\mu^k}[(P^k(x)-Q^k(x))(Q^k(x)-\chi_D(x))]$$

$$E_{\mu^k}[(P^k(x)-Q^k(x))^2] \leq |E_{\mu^k}[(P^k(x)-Q^k(x))(P^k(x)-\chi_D(x))]|+|E_{\mu^k}[(P^k(x)-Q^k(x))(Q^k(x)-\chi_D(x))]|$$

$$E_{\mu^k}[(P^k(x) - Q^k(x))^2] \leq \delta(k) + \epsilon(k)$$

Conversely, assume $P \overset{\mu}{\sim} Q$. Consider some $R : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]$. We have

$$E_{\mu^k}[R(x)(Q^k(x) - \chi_D(x))] = E_{\mu^k}[R(x)(Q^k(x) - P^k(x) + P^k(x) - \chi_D(x))]$$

$$E_{\mu^k}[R(x)(Q^k(x)-\chi_D(x))] = E_{\mu^k}[R(x)(Q^k(x)-P^k(x))]+E_{\mu^k}[R(x)(P^k(x)-\chi_D(x))]$$

$$|E_{\mu^k}[R(x)(Q^k(x)-P^k(x))]| \leq E_{\mu^k}[|Q^k(x)-P^k(x)|] \leq \sqrt{E_{\mu^k}[(Q^k(x) - P^k(x))^2]} \leq \delta(k)$$

for some negligible δ, since $P \overset{\mu}{\sim} Q$.

$$|E_{\mu^k}[R(x)(P^k(x) - \chi_D(x))]| \leq p^*(|R|)\delta^*(k)$$

for some polynomial p^* and negligible δ^*, using Lemma 4.2. Combining both inequalities we get

$$|E_{\mu^k}[R(x)(Q^k(x) - \chi_D(x))]| \leq \delta(k) + p^*(|R|)\delta^*(k)$$

Using Lemma 4.2 again we conclude Q is an optimal predictor.

□

We are now ready to prove Theorem 4.6:

Proof of Theorem 4.6. When $P_D^k(x) > 0$ we have

$$P_{C|D}^k(x) = \frac{\min(P_{C\cap D}^k(x), P_D^k(x))}{P_D^k(x)}$$

Define $\tilde{P}_{C\cap D}^k$ to be the circuit computing $\min(P_{C\cap D}^k(x), P_D^k(x))$. Since $C \cap D \subseteq D$, Lemma 4.3 implies that $E_{\mu^k}[P_{C\cap D}^k(x) - \tilde{P}_{C\cap D}^k(x)]$ is negligible. This implies $E_{\mu^k}[(P_{C\cap D}^k(x) - \tilde{P}_{C\cap D}^k(x))^2]$ is negligible and by Theorem 5.1 $\tilde{P}_{C\cap D}$ is an optimal predictor for $(C \cap D, \mu)$.

We have $\tilde{P}_{C\cap D}^k(x) = P_{C|D}^k(x)P_D^k(x)$ (whether $P_D^k(x) > 0$ or $P_D^k(x) = 0$) and therefore

$$\tilde{P}_{C\cap D}^k(x) - \chi_{C\cap D}(x) = (P_{C|D}^k(x) - \chi_C(x))\chi_D(x) + P_{C|D}^k(x)(P_D^k(x) - \chi_D(x))$$

$$(P_{C|D}^k(x) - \chi_C(x))\chi_D(x) = \tilde{P}_{C\cap D}^k(x) - \chi_{C\cap D}(x) - P_{C|D}^k(x)(P_D^k(x) - \chi_D(x))$$

Consider $Q : \operatorname{supp}\mu^k \xrightarrow{circ} \mathbb{Q} \cap [-1, +1]$.

$$E_{\mu^k|D}[Q(x)(P_{C|D}^k(x) - \chi_C(x))] = \mu^k(D)^{-1}E_{\mu^k}[Q(x)(P_{C|D}^k(x) - \chi_C(x))\chi_D(x)]$$

By Lemma 4.2 it is sufficient to prove appropriate bounds on $|E_{\mu^k}[Q(x)(\tilde{P}_{C\cap D}^k(x) - \chi_{C\cap D}(x))]|$ and $|E_{\mu^k}[Q(x)P_{C|D}^k(x)(P_D^k(x) - \chi_D(x))]|$. Both bounds follow from Lemma 4.2 using the facts $\tilde{P}_{C\cap D}$ and P_D are optical predictors and $|P_{C|D}^k|$ is bounded by a polynomial.

□

6 Reductions

In this section we show that optimal predictors are stable with respect to a certain type of reductions. In particular, if a class of languages contains a language complete for such reductions, it is sufficient to prove the existence of an optimal predictor for this language in order to deduce it for the entire class.

Definition 6.1. Consider (C, μ), (D, ν) distributional decision problems, $\{f^k : \operatorname{supp}\mu^k \xrightarrow{circ} \{0,1\}^*\}_{k\in\mathbb{N}}$ a polynomial size family of circuits. f is called a (non-uniform) *pseudo-invertible reduction* of C to D when there is a polynomial $p : \mathbb{N} \to \mathbb{N}$ s.t. the following conditions hold:

(i) $\forall k \in \mathbb{N}, x \in \operatorname{supp}\mu^k : \chi_D(f^k(x)) = \chi_C(x)$

(ii) There is a polynomial d s.t.

$$\forall k \in \mathbb{N}, y \in \{0,1\}^* : \frac{\mu^k((f^k)^{-1}(y))}{\nu^{p(k)}(y)} \leq d(k)$$

(iii) There is a polynomial $q : \mathbb{N} \to \mathbb{N}$ and a family of polynomial size circuits $\{g^k : \operatorname{supp} \nu^{p(k)} \times \{0,1\}^{q(k)} \xrightarrow{circ} \{0,1\}^*\}_{k\in\mathbb{N}}$ s.t.

$$\forall y \in f^k(\operatorname{supp} \mu^k), x^* \in \{0,1\}^* : Pr_{U^{q(k)}}[g^k(y,r) = x^*] = Pr_{\mu^k}[x = x^* | f^k(x) = y]$$

(iv) There are polynomial size circuits $\{R^k : \operatorname{supp} \nu^{p(k)} \xrightarrow{circ} \mathbb{Q}^{\geq 0}\}_{k\in\mathbb{N}}$ s.t.

$$\forall k \in \mathbb{N}, y \in \operatorname{supp} \nu^{p(k)} : R^k(y) = \frac{\mu^k((f^k)^{-1}(y))}{\nu^{p(k)}(y)}$$

Note 6.1. Conditions (i) and (ii) comprise the usual definition of reductions between distributional decision problems in average-case complexity theory. Conditions (iii) and (iv) are special requirement of the theory of optimal predictors.

Note 6.2. A natural special case in which conditions (iii) and (iv) hold is when μ and ν are computable and it is possible to enumerate the f^k-inverse images of any word in (non-uniform) time polynomial in k.

Theorem 6.1. *Consider (C,μ), (D,ν) distributional decision problems, f a pseudo-invertible reduction of (C,μ) to (D,ν) and P_D an optimal predictor for (D,ν). Define $\{P_C^k : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]\}_{k\in\mathbb{N}}$ as the family of circuits computing $P_C^k(x) := P_D^{p(k)}(f^k(x))$. Then, P_C is an optimal predictor for (C,μ).*

Proof. Consider $k \in \mathbb{N}$, $Q_C : \operatorname{supp} \mu^k \xrightarrow{circ} [0,1]$. Define $Q_D : \operatorname{supp} \nu^{p(k)} \times \{0,1\}^{q(k)} \xrightarrow{circ} [0,1]$ to be the circuit computing $Q_D(y,r) := Q_C(g^k(y,r))$. Applying Lemma 4.1, treating r as a constant and using R as the weight circuit, we get

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] \leq E_{\nu^{p(k)}}[R^k(y)(Q_D(y,r) - \chi_D(y))^2] + m(|Q_C|)\delta(k)$$

where m is a polynomial and δ is negligible. We used condition (ii) to get a polynomial bound on $\max R^k$ and condition (iv) to get a polynomial bound on $|R^k|$.

We take the expectation value of both sides with respect to the uniform measure over r:

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] \leq E_{\nu^{p(k)} \times U^{q(k)}}[R^k(y)(Q_D(y,r) - \chi_D(y))^2] + m(|Q_C|)\delta(k)$$

The left hand side can be rewritten as follows

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] = \sum_{y\in\{0,1\}^*} \nu^{p(k)}(y) \frac{\mu^k((f^k)^{-1}(y))}{\nu^{p(k)}(y)} (P_D^{p(k)}(y) - \chi_D(y))^2$$

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] = \sum_{y\in\{0,1\}^*} \mu^k((f^k)^{-1}(y))(P_D^{p(k)}(y) - \chi_D(y))^2$$

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] = \sum_{y \in \{0,1\}^*} \sum_{\substack{x \in \operatorname{supp} \mu^k \\ f^k(x) = y}} \mu^k(x)(P_D^{p(k)}(y) - \chi_D(y))^2$$

Grouping the sum by x, we get

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] = \sum_{x \in \operatorname{supp} \mu^k} \mu^k(x)(P_C^k(x) - \chi_C(x))^2$$

$$E_{\nu^{p(k)}}[R^k(y)(P_D^{p(k)}(y) - \chi_D(y))^2] = E_{\mu^k}[(P_C^k(x) - \chi_C(x))^2]$$

The first term on the right hand side can be rewritten as

$$E_{\nu^{p(k)} \times U^{q(k)}}[R^k(y)(Q_D(y,r) - \chi_D(y))^2] = \sum_{y \in \{0,1\}^*} \sum_{r \in \{0,1\}^{q(k)}} 2^{-q(k)} \mu^k((f^k)^{-1}(y))(Q_D(y,r) - \chi_D(y))^2$$

Grouping the sum by $x := g(y,r)$ we get:

$$E_{\nu^{p(k)} \times U^{q(k)}}[R^k(y)(Q_D(y,r) - \chi_D(y))^2] = \sum_{x \in \{0,1\}^*} \sum_{y \in \{0,1\}^*} \sum_{\substack{r \in \{0,1\}^{q(k)} \\ g^k(y,r) = x}} 2^{-q(k)} \mu^k((f^k)^{-1}(y))(Q_C(x) - \chi_C(x))^2$$

Condition (iii) tells us that $\sum_{\substack{r \in \{0,1\}^{q(k)} \\ g^k(y,r) = x}} 2^{-q(k)}$ is only non-vanishing when $y = f^k(x)$ and that in this case it equals $\frac{\mu^k(x)}{\mu^k((f^k)^{-1}(y))}$. Therefore

$$E_{\nu^{p(k)} \times U^{q(k)}}[R^k(y)(Q_D(y,r) - \chi_D(y))^2] = \sum_{x \in \{0,1\}^*} \mu^k(x)(Q_C(x) - \chi_C(x))^2$$

$$E_{\nu^{p(k)} \times U^{q(k)}}[R^k(y)(Q_D(y,r) - \chi_D(y))^2] = E_{\mu^k}[(Q_C(x) - \chi_C(x))^2]$$

Putting everything together, we get

$$E_{\mu^k}[(P_C^k(x) - \chi_C(x))^2] \leq E_{\mu^k}[(Q_C(x) - \chi_C(x))^2] + m(|Q_C|)\delta(k)$$

□

7 Examples

The trivial example of an optimal predictor is when $(D, \mu) \in Heur_{neg}P/poly$. In this case we can take $P^k(x) = H^k(x)$ where H is a heuristic family of circuits for (D, μ). This example represents the edge case in which essentially all information can be obtained in polynomial time. The following theorem represents the opposite edge case, in which essentially *no* information can be obtained in polynomial time.

Theorem 7.1. *Consider $f : \{0,1\}^* \to \{0,1\}^*$ a one-to-one non-uniformly hard one-way function. Define $D_f := \{(y,r) \mid \exists x \in \{0,1\}^{|r|} : (f(x) = y) \wedge (x \cdot r = 0)\}$. Define $f^{(k)} : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}^*$ by $f^{(k)}(x,r) := (f(x),r)$. Define μ_f^k to be the direct image of $U^k \times U^k$ under $f^{(k)}$. Define $\{P_f^k : \operatorname{supp} \mu_f^k \xrightarrow{circ} [0,1]\}_{k \in \mathbb{N}}$ as the circuits computing the constant function $P_f^k(x) := 1/2$. Then, P_f is an optimal predictor for (D_f, μ_f).*

Proof. Assume to the contrary that P_f is not optimal. Then there is an infinite set $I \subseteq \mathbb{N}$, a polynomial size family of circuits $\{Q^k : \operatorname{supp} \mu_f^k \xrightarrow{circ} [0,1]\}_{k \in I}$ and a positive polynomial r s.t.

$$\forall k \in I : E_{\mu_f^k}[(P_f^k(x) - \chi_{D_f}(x)^2)] \geq E_{\mu_f^k}[(Q^k(x) - \chi_{D_f}(x))^2] + \frac{1}{r(k)}$$

$$\forall k \in I : E_{\mu_f^k}[(Q^k(x) - \chi_{D_f}(x))^2] \leq \frac{1}{4} - \frac{1}{r(k)}$$

Define the functions $\{q^k : \operatorname{supp} \mu_f^k \times [0,1] \to \{0,1\}\}_{k \in I}$ by $q^k(x,t) := \theta(Q^k(x) - t)$. We have

$$\forall k \in I, x \in \operatorname{supp} \mu_f^k : Q^k(x) = \int_0^1 q^k(x,t)\, dt$$

Substituting into the inequality above

$$\forall k \in I : E_{\mu_f^k}[(\int_0^1 q^k(x,t)\, dt - \chi_{D_f}(x))^2] \leq \frac{1}{4} - \frac{1}{r(k)}$$

$$\forall k \in I : E_{\mu_f^k}[|\int_0^1 q^k(x,t)\, dt - \chi_{D_f}(x)|]^2 \leq \frac{1}{4} - \frac{1}{r(k)}$$

$$\forall k \in I : E_{\mu_f^k}[|\int_0^1 (q^k(x,t) - \chi_{D_f}(x))\, dt|] \leq \sqrt{\frac{1}{4} - \frac{1}{r(k)}}$$

For every given x, $q^k(x,t) - \chi_{D_f}(x)$ is either non-negative for all t or non-positive for t. Hence we can move the absolute value inside the integral:

$$\forall k \in I : E_{\mu_f^k}[\int_0^1 |q^k(x,t) - \chi_{D_f}(x)|\, dt] \leq \sqrt{\frac{1}{4} - \frac{1}{r(k)}}$$

$$\forall k \in I : \int_0^1 E_{\mu_f^k}[|q^k(x,t) - \chi_{D_f}(x)|]\, dt \leq \sqrt{\frac{1}{4} - \frac{1}{r(k)}}$$

This implies that we can choose $\{t_k \in Q^k(\operatorname{supp} \mu_f^k) \cup \{0,1\}\}_{k \in I}$ s.t.

$$\forall k \in I : E_{\mu_f^k}[|q^k(x,t_k) - \chi_{D_f}(x)|] \leq \sqrt{\frac{1}{4} - \frac{1}{r(k)}}$$

$$\forall k \in I : Pr_{\mu_f^k}[q^k(x,t_k) \neq \chi_{D_f}(x)] \leq \sqrt{\frac{1}{4} - \frac{1}{r(k)}}$$

$$\forall k \in I : Pr_{\mu_f^k}[q^k(x, t_k) = \chi_{D_f}(x)] \geq 1 - \sqrt{\frac{1}{4} - \frac{1}{r(k)}}$$

Using the fact that the graph of the square root lies below its tangent at any point, this leads to

$$\forall k \in I : Pr_{\mu_f^k}[q^k(x, t_k) = \chi_{D_f}(x)] \geq \frac{1}{2} + \frac{1}{r(k)}$$

Define $\{g^k : f(\{0,1\}^k) \times \{0,1\}^k \xrightarrow{circ} \{0,1\}\}_{k \in \mathbb{N}}$ as the circuits computing $g^k(y, r) := 1 - q^k((y, r), t_k)$. The definitions of q^k and t_k imply that $|g^k|$ is bounded by a polynomial. The inequality above and the definitions of D_f and μ_f imply

$$\forall k \in I : Pr_{U^k \times U^k}[g^k(f(x), r) = x \cdot r] \geq \frac{1}{2} + \frac{1}{r(k)}$$

But this contradicts Theorem 2.1.

□

In particular, we showed that if non-uniformly hard one-way functions exist (which is believed to be the case) then $OP_{neg}P/poly \supsetneq Heur_{neg}P/poly$ and $OP_{neg}P \supsetneq Heur_{neg}P$.

Using Theorems 4.2, 4.3, 4.4 and 6.1, it is possible to build many examples from the two basic types above.

8 Future Research

There are many direction in which the results we presented can be expanded. Here, we attempt to discuss those that seem the most important for FAI.

8.1 Updateless Decision Theory

In order to use optimal predictors to formalize UDT, we need two primary components: conditional probabilities and optimal search.

A concept of conditional probabilities is provided by Theorems 4.5 and 4.6.

In a way, optimal predictors are optimal approximate solution for decision problems. It should be possible to extend this to other kinds of problems, e.g. search problems, maximization problems and counting problems.

Thus we expect to get something like the best algorithm for maximizing a certain quantity within polynomial computing resources. This algorithm can be used for maximizing payoff, where the payoff itself is defined as a logical conditional expectation value.

8.2 Vingean Reflection

It is natural to look for concept of pseudo-invertible Cook reductions which would preserve optimal predictors. This is likely to have applications to Vingean reflection, if the successor agent is assumed to take the shape of an oracle machine with a language admitting an optimal predictor taking the place of the

oracle. Thus, the predecessor agent will be able to use a small amount of computing resources (use the induced optimal predictor) to evaluate the consequences of constructing an agent using large amounts of of computing resources (assuming the extra computing resources are only used to evaluate the oracle).

The other aspect of Vingean reflection, namely the dependence of the future on the environment can be addressed by constructing an efficient approximation of Solomonoff induction. Such an approximation should be possible to define by applying optimal predictors to counting problems. In particular, Solomonoff induction seems to be related to the following class of distributional decision problems:

Definition 8.1. A distributional decision problem (D, μ) is called *verifiable* when there is a negligible function δ, a polynomial p and $M : \mathbb{N} \times \{0,1\}^* \xrightarrow{alg} \{0,1\}^* \times \{0,1\}$ s.t. $M(k,r)$ is computed in time polynomial in k and $|r|$ and the following conditions hold:

(i) $\forall k \in \mathbb{N}, x \in \{0,1\}^* : \mu^k(x) = Pr_{U^{p(k)}}[M(k,r)_1 = x]$

(ii) $\forall k \in \mathbb{N} : Pr_{U^{p(k)}}[M(k,r)_2 \neq \chi_D(M(k,r)_1)] \leq \delta(k)$

Here, $M(k,r) = (M(k,r)_1, M(k,r)_2)$.
The set of verifiable distributional decision problems is denoted $VerNP$.

9 References

References

[1] "Average-Case Complexity", Andrej Bogdanov and Luca Trevisan, 2008 (arXiv:cs/0606037v2).

[2] "Computational Complexity: A Conceptual Perspective", Oded Goldreich