Agnostic insurability of model classes

Narayana Santhanam nsanthan@hawaii.edu
Dept of Electrical Engineering
University of Hawaii at Manoa
Honolulu, HI 96822

Venkat Anantharam ananth@eecs.berkeley.edu
Dept of EECS
University of California, Berkeley
Berkeley, CA 94720

Editor:

Abstract

Motivated by problems in insurance, our task is to predict finite upper bounds on a future draw from an unknown distribution p over the set of natural numbers. We can only use past observations generated independently and identically distributed according to p. While p is unknown, it is known to belong to a given collection $\mathcal{P}$ of probability distributions on the natural numbers.

The support of the distributions $p \in \mathcal{P}$ may be unbounded, and the prediction game goes on for *infinitely* many draws. We are allowed to make observations without predicting upper bounds for some time. But we must, with probability 1, start and then continue to predict upper bounds after a finite time irrespective of which $p \in \mathcal{P}$ governs the data.

If it is possible, without knowledge of p and for any prescribed confidence however close to 1, to come up with a sequence of upper bounds that is never violated over an infinite time window with confidence at least as big as prescribed, we say the model class $\mathcal{P}$ is *insurable*.

We completely characterize the insurability of any class $\mathcal{P}$ of distributions over natural numbers by means of a condition on how the neighborhoods of distributions in $\mathcal{P}$ should be, one that is both necessary and sufficient.

Keywords: insurance, ℓ_1 topology of probability distributions over countable sets, non-parametric approaches, prediction of quantiles of distributions, universal compression.

1. Introduction

Insurance is a means of managing risk by transfering a potential sequence of losses to an *insurer* for a price paid on a regular basis, the *premium*. The insurer attempts to break even by balancing the possible loss that may be suffered by a few with the guaranteed premiums of many. We aim to study the fundamentals of this problem when the losses can be unbounded and a precise model for the probability distribution of the aggregate loss in each period either does not exist or is infeasible to get.

A systematic, theoretical, as opposed to empirical, study of insurance goes back to 1903 when Filip Lundberg (see Englund and Martin-Löf (2001)) defined a natural probabilistic setting as part of his thesis. In particular, Lundberg formulated a collective risk problem pooling together the risk of all the insured parties into a single entity, which we call the insured. Typically, studies of insurance derived from the approach in Englund and Martin-Löf (2001) depend on working with specific models for the loss distribution, e.g. compound Poisson models, after which questions of interest in practice, such as the relation between the size of the premiums charged and the probability of the insurer going bankrupt,

can be analyzed. A rather comprehensive theory of insurance along these lines has evolved, see Cramer (1969) and more recently in Asmussen and Albrecher (2010). This theory is able to incorporate several model classes for the distribution of the losses over time other than compound Poisson processes, including some heavy tailed distribution classes.

We depart from the existing literature on insurance in two important respects.

No upper bound on loss The first departure relates to the practice among insurers to limit payments to a predetermined ceiling, even if the loss suffered by the insured exceeds this ceiling. In both the insurance industry and the legal regulatory framework surrounding it, this is assumed to be common sense. But is it always necessary to impose such ceilings? Moreover, in scenarios such as reinsurance, a ceiling on compensation is not only undesirable, but may also limit the very utility of the business. As we will see, we may be able to handle scenarios where the loss can be unbounded.

Universal approach The second aspect of our approach arises from our motivation to deal with several new settings for which some sort of insurance is desirable, but where insurers are hesitant to enter the market due to lack of sufficient data. Examples of such settings include insuring against network outages or attacks against future smart grids, where the cascade effect of outages or attacks could be catastrophic. In these settings, it is not clear today what should constitute a reasonable risk model because of the absence of usable information about what might cause the outages or motivate the attacks.

We address the second issue by working with a *class* of models, *i.e.,* a set of probability laws over loss sequences that adheres to any assumptions the insurer may want to make or any information it may already have. In this paper we will only consider loss models that are independent and identically distributed (*i.i.d.*) from period to period, so we can equivalently think of a model class as defined in terms of its one dimensional marginals.

As an example, we may want to consider the set of all finite moment probability distributions over the nonnegative integers as our class of possible models for the loss distribution in each period. Now, we ask the question: what classes of models are the ones on which the insurer can learn from observations and set premiums so as to remain solvent? In this paper, we completely answer this question by giving a necessary and sufficient condition that characterizes what classes of models lend themselves to this insurance task.

This is very reminiscent of the universal compression/estimation/prediction approaches (see Shtarkov (1987); Fittingoff (1972); Rissanen (1984); Ryabko (2008))—we will have more to say on this shortly. There is also extensive work regarding learning from experts that has a related flavor, see Cesa-Bianchi and Lugosi (2006) for a survey.

Formulation Formally, we adopt the collective risk approach, namely, we abstract the problem to include just two agents, the insurer and the insured. Losses incurred by the insured are considered to form a discrete time sequence of random variables, with the sequence of losses denoted by $\{X_i,\ i \geq 1\}$, and we assume that $X_i \in \mathbb{N}$ for all $i \geq 1$, where $\mathbb{N}$ denotes the set of natural numbers, $\{0, 1, 2, \ldots\}$.

A *model class* $\mathcal{P}^\infty$ is a collection of measures on infinite length loss sequences, and is to be thought of as the set of all potential probability laws governing the loss sequence. Each element of $\mathcal{P}^\infty$ is a *model* for the sequence of losses. Any prior knowledge on the structure of the problem is accounted for in the definition of $\mathcal{P}^\infty$. We focus on measures corresponding to *i.i.d.* samples, i.e. each member of $\mathcal{P}^\infty$ induces marginals that are product distributions. We denote by $\mathcal{P}$ the set of distributions on $\mathbb{N}$ obtained as one dimensional marginals of $\mathcal{P}^\infty$. Since there is no risk of confusion, we will also refer to the distributions in $\mathcal{P}$ as models and to $\mathcal{P}$ as the model class.

The actual model in $\mathcal{P}$ governing the law of the loss in each period remains unknown to the insurer. We assume no ceiling on the loss, and require the insurer to compensate the insured in full for the loss in each period at the end of that period. The insurer is assumed to start with some initial capital $\Pi_0 \in \mathbb{R}^+$, a nonnegative real number. The insurer then sets a sequence of premiums based on the past losses—at time i, the insurer collects a premium $\Pi(X_1^{i-1})$ at the beginning of the period, and pays out full compensation for loss X_i at the end of the period. If the built up capital till step i (including $\Pi(X_1^{i-1})$, and after having paid out all past losses) is less than X_i, the insurer is said to be *bankrupted*.

Given a class $\mathcal{P}^\infty$ of loss models, we ask if, for every prescribed upper bound $\eta > 0$ on the probability of bankruptcy, the insurer can set (finite) premiums at every time step based only on the loss sequence observed thus far and with no further knowledge of which law $p \in \mathcal{P}^\infty$ governs the loss sequence, while simultaneously ensuring that the insurer remains solvent with probability bigger than $1 - \eta$ under p irrespective of which $p \in \mathcal{P}^\infty$ is in effect. If the probability of the insurer ever going bankrupt over an infinite time window can be made arbitrarily small in this sense, the class of *i.i.d.* loss measures $\mathcal{P}^\infty$ is said to be *insurable*.

A couple of clarifications are in order here. First, to make the problem non-trivial, we allow the insurer to observe the loss sequence for some arbitrary finite length of time without having to provide compensations. We require that the insurer has to eventually provide insurance with probability 1 no matter which $p \in \mathcal{P}^\infty$ is in effect. The insurer cannot quit providing insurance once it has entered into the insurance contract with the insured. Premiums set before the entry time can be thought of as being 0 and the question of bankruptcy only arises after the insurer has entered into the contract. Secondly, at this point of research, we do not concern ourselves with incentive compatibility issues on the part of the insured and assume that the insured will accept the contract once the insurer has entered, agreeing to pay the premiums as set by the insurer.

It turns out that the fact that the capital available to the insurer at any time is built up from past premiums does not play any role in whether a model class is insurable or not. In fact, the problem is basically one of finding a sequence of finite upper bounds $\Phi(X_1^{i-1})$ on the loss X_i for all $i \geq 1$. We refer to the sequence $\{\Phi(X_1^{i-1}),\ i \geq 1\}$ as the *loss dominating sequence* and call $\Phi(X_1^{i-1})$ the *loss dominant* at step i.

The notion of insurability of a model class $\mathcal{P}$ comes down to whether for each $\eta > 0$ there is a way of choosing the loss dominants such that the probability of the loss X_i ever exceeding the loss dominant $\Phi(X_1^{i-1})$ is smaller than η irrespective of which model p in the model class $\mathcal{P}^\infty$ is in effect. Here again we allow some initial finite number of periods for which the loss dominant can be set to ∞, but it must become finite with probability 1 under each $p \in \mathcal{P}^\infty$ and stay finite from that point onwards.

Results For a model class to be insurable, roughly speaking, close distributions must have comparable percentiles. Distributions in the model class that, in every neighborhood, have some other distribution with arbitrarily different percentiles are said to be *deceptive*. In Section 3, we define what it means for distributions to be close, and what it means for distributions to have comparable percentiles. In Section 4, we provide several examples of insurable and non-insurable model classes. Our main result is Theorem 1 of Section 3, which states that that $\mathcal{P}^\infty$ is insurable iff it has no deceptive distributions. We prove this theorem in Sections 5 and 6. In the rest of this section, we discuss the problem of insurability in the broader contexts of uniform and pointwise convergence of estimators and universal compression.

1.1 Pointwise vs uniform convergence

Theoretically, the flexibility we have permitted regarding when to start proposing finite loss dominants allows us to categorize the insurance problem formulated above as one that admits what we call *useful* pointwise convergent estimators. Indeed, we are particularly interested in cases where uniform learnability is not possible. In such cases, often guarantees on estimators are provided that hold *pointwise* for all models. Typically, the results proven in such cases are of consistency, or bounds on rate of convergence of estimators that depend on the parameters of the model in force—of course, the model by itself is usually not known *a-priori*. Therefore there is a practical issue with such pointwise guarantees—it says little about what is happening with the specific sample at hand. Even if we know that such an estimator is consistent, or have model dependent bounds, for an unknown model and a given sample there may be no way of telling how good or bad the estimate is.

Roughly speaking, the insurance problem is equivalent to learning all the percentiles of an unknown distribution from $\mathcal{P}$, using only *i.i.d.* draws generated from the distribution. However, as the sample size increases, the estimate of any given percentile need not converge to the true value (according to some predefined metric) uniformly over the entire class $\mathcal{P}$.

This poses a conundrum, since when dealing with large alphabets or high dimensions, it may be too restrictive to only deal with model classes or problem formulations that admit uniformly convergent estimators. Conversely, what if we want to work with a model class which is sufficiently complex that uniformly convergent estimators are impossible?

We will then have estimators that only converge pointwise over the model class. Yet, we can still salvage the situation if for any given finite sample, we had some way to tell from the sample if the estimate was doing well or not relative to the true unknown model. To illustrate this concept simply, we provide a simple running example below that we will also use to demonstrate connections with compression.

Example 1 [Birthday problem] Consider the problem of estimating the size of a discrete, finite, set $S \subset \mathbb{N}$ (specifically, there is no upper bound on the size of S) if we can draw as many random samples from S as needed. If we have independent, uniform draws from S, one simple way to estimate the size of S is keep sampling till some element from S is drawn twice. This is the first *repeat*. A simple back of the envelope calculation analogous to the Birthday Problem shows that if N_1 is the sample size when the first repeat occured, then a good estimate of the set size is $N_1^2/2$. One can then provide PAC-learning kind of bounds that, with some confidence, the above estimate based on the first repeat is accurate to a certain level.

This is an example where there can be no uniformly convergent estimator of the set size. Given a fixed sample of size n, if the size of S is $\gg \Omega(n^2)$ the sample consists of n distinct symbols with probability close to 1. If we can assume no further structure on S, there is no way to distinguish between samples obtained from any two sets with size $\gg \Omega(n^2)$—thus no estimator can distinguish between these large sets. It is therefore futile, with a finite sample size, to expect an estimator that can estimate the set size to any non-trivial accuracy no matter what the set is. Equivalently, there can be no uniformly convergent estimator of the set size.

The simple "Birthday Problem" estimator above only converges pointwise. It may take an arbitrarily larger sample for some models to give an answer. That is the nature of the problem. But the estimator is imbued with a very useful property—with a guaranteed confidence, the estimator does not make a mistake even though it may not always have an answer. If the sample has no repeats yet, the estimator does not overreach and volunteer a wrong answer. Hence, we can tell from the sample if we can do well or not. □

This is what our insurance formulation capitalizes on as well, by providing a finite (but not fixed) observation window before the insurer enters the game. For a theoretical understanding of the insurance scenario that we aim for here, confining ourselves to only classes $\mathcal{P}$ of loss models that have uniformly convergent estimators for percentiles of a distribution is too restrictive for the unbounded loss case. We want to deal with broader classes of models, and the flexibility about the start point for prediction allows us to consider significantly richer model classes. For other kinds of such "useful" pointwise estimation, particularly in relation to Markov processes, see Asadi et al. (2013).

1.2 Universal compression

The approach we take is not unconnected with the universal compression literature, as well as certain learning formulations involving regret with log-loss. There are many variations in how compression problems are formulated, and the following example illustrates the main variants studied. It also serves to provide an example of the "useful" pointwise estimation introduced in the last section, and hence places insurability in context of the universal compression literature.

Example 2 We will study the so-called "Birthday problem" from the previous example in a little more depth. Let $\mathcal{B}$ denote the collection of all distributions p_M ($M \geq 1$), where p_M is a uniform distribution over $\{0, \ldots, M\}$. We use this example to distinguish between uniformly good compressors (strongly universal) and compressors that are only good pointwise (weakly universal).

Suppose we consider one draw from an unknown distribution $p \in \mathcal{B}$. The worst case *redundancy* quantifies the minimum possible excess codelength of a universal distribution q over $\mathbb{N}$ over the (unknown) distribution p

$$\inf_q \sup_{p \in \mathcal{B}} \sup_{x \in \mathbb{N}} \log \frac{p(x)}{q(x)}. \tag{1}$$

Note that the redundancy of any class of distributions is always ≥ 0. We could, of course, consider a sequence of n independent draws from an unknown $p \in \mathcal{B}$, and ask now for a measure q over infinite sequences of numbers that is universal for all the *i.i.d.* measures corresponding to distributions in $\mathcal{B}$. We then concentrate on the *redundancy*

$$R_n(\mathcal{B}) \stackrel{\text{def}}{=} \inf_q \sup_{p \in \mathcal{B}} \sup_{x^n \in \mathbb{N}^n} \frac{1}{n} \log \frac{p(x^n)}{q(x^n)},$$

where we abuse notation and write for all $p \in \mathcal{B}$

$$p(x^n) = \prod_{j=1}^{n} p(x_j)$$

to be the probability assigned to x^n by independent draws from p. Of course, we could similarly define redundancy for length n sequences for any collection of measures over infinite sequences from a countable alphabet. *Strongly compressible* classes are those sets $\mathcal{P}^\infty$ of measures over infinite sequences satisfying

$$\lim_{n \to \infty} R_n(\mathcal{P}^\infty) \to 0.$$

For the single-letter formulation in (1), clearly the optimal universal distribution gives any number x a probability proportional to the highest probability that number gets from any model in $\mathcal{B}$, followed by a normalization. But the highest probability a model in $\mathcal{B}$ gives any $x \in \mathbb{N}$ is $1/(x+1)$, which is

not summable over x. Thus the redundancy is infinite here—or equivalently, no matter what universal distribution q we choose and no matter how large a number M we pick, there is a $p' \in \mathcal{B}$ and a number $x' \in \mathbb{N}$ such that

$$\log \frac{p'(x')}{q(x')} > M.$$

In this case, we will therefore not have redundancy bounds holding uniformly for the model class. We say $\mathcal{B}$ is not strongly compressible. With a very similar argument, it is easy to see that $R_n(\mathcal{B}^\infty) = \infty$ for all n, where we use $\mathcal{B}^\infty$ to denote the set of *i.i.d.* measures over infinite sequences constructed as above from marginals in $\mathcal{B}$.

But we can say something more. Consider again compressing sequences of numbers drawn *i.i.d.* from an unknown distribution in $\mathcal{B}$. Noting that $\mathcal{B}$ is countable, we focus on a measure q over infinite sequences that gives a sequence x^n the probability

$$q_w(x^n) = \sum_{p_i \in \mathcal{B}} \frac{1}{i(i+1)} p_i(x^n).$$

It is easy to verify that q_w above satisfies

$$\sup_{p \in \mathcal{B}^\infty} \lim_{n \to \infty} \sup_{x^n \in \mathbb{N}^n} \frac{1}{n} \log \frac{p(x^n)}{q_w(x^n)} = 0, \tag{2}$$

or that q_w matches every p *pointwise* over the model class $\mathcal{B}$. Such classes of sources are *weakly universal*. The codelength of the universal measure q_w matches that of p for every $p \in \mathcal{B}$, but at arbitrarily slower rates for some sources (since the class cannot be strongly compressed).

A couple of points. Note that admittedly it has been easy to define q_w here since $\mathcal{B}$ was countable to begin with. If not, a condition reminiscent of countability above is necessary and sufficient for a class to be weakly universal as shown in (Kieffer, 1978). To emphasize, q_w is guaranteed to satisfy (2), implying that it does not underestimate relative to any p for sufficiently long sequences. But how long is "sufficiently long" depends on p, and for a given length-n sequence without knowing p, it may not be possible to say if q_w is doing well or not. This second aspect of knowing when an estimator is good is crucial to insurance formulations.

One could also replace the sup over x^n with expectation, and get average case versions for both strong and weak' universality. □

Strong compression is well known and is the more studied version of universal compression and regret formulations involving log loss. As one might expect, we show in (Santhanam and Anantharam, 2012) that strong compression implies insurability but not vice-versa.

However, the insurance formulation has more in common with weak compression and pointwise convergence, rather than strong compression and uniform convergence. The connection between insurability and weak compression turns out to be rather interesting. In (Santhanam and Anantharam, 2012), we show classes of models that can be weakly compressed but are not insurable. At the same time, we also construct classes of models in (Santhanam and Anantharam, 2012) that are insurable, but cannot be weakly compressed.

To summarize, our formulation is interesting precisely in cases where the strong notions of (worst-case or average-case) redundancy fail. Namely, classes of distributions whose redundancy is not finite. The universal compression formulation closer to our notion of insurability here is in the idea of weak universal compression in Kieffer (1978). However, weak compression formulations thus far have never included the aspect of determining from the data at hand *when* a compressor is doing well—a crucial part of our problem.

There may be one insight that we conjecture can be generalized beyond insurability to all problems with the flavor of useful pointwise convergence of estimates. What matters seems to be the local complexity as opposed to global complexity of model classes. Insurability of model classes does not depend on global complexity measures of model classes—as with the (strong) redundancy of model classes (which is determined by the integral of the square root of absolute Fisher information over the entire model class) or the Rademacher complexity. Instead, insurability is related to how local neighborhoods look; in particular it depends on *local tightness* as we will see in Section 3.

2. Precise formulation of the problem

We model the loss at each time by a random variable taking values in $\mathbb{N} = \{0, 1, \ldots\}$. Denote the sequence of losses by $X_1, X_2 \ldots$ where $X_i \in \mathbb{N}$. Let $\mathbb{N}^*$ be the set of all finite length sequences from $\mathbb{N}$, including the empty sequence. We will write x^n for the sequence $x_1, \ldots, x_n$. Where it appears, x^0 denotes the empty sequence. A loss distribution is a probability distribution on $\mathbb{N}$. Let $\mathcal{P}$ be a set of loss distributions. $\mathcal{P}^\infty$ is the collection of *i.i.d.* measures over infinite sequences of symbols from $\mathbb{N}$ such that the set of one dimensional marginals over $\mathbb{N}$ they induce is $\mathcal{P}$.

We write $\mathbb{R}^+$ for the set of nonnegative real numbers and use $:=$ for equality by definition.

Consider an insurer with an *initial capital* $\Pi_0 \in \mathbb{R}^+$. An *insurance scheme* for $\mathcal{P}$ is comprised of a pair (τ, Π).

Here $\tau : \mathbb{N}^* \mapsto \{0, 1\}$ satisfies $\tau(x_1, \ldots, x_n) = 1 \implies \tau(x_1, \ldots, x_{n+1}) = 1$ for all x^n and also $p(\sup_n \tau(X^n) = 1) = 1$ for all $p \in \mathcal{P}^\infty$. τ should be thought of as defining an *entry time* for the insurer with the property that once the insurer has entered it stays entered and that the insurer enters with probability 1 irrespective of which $p \in \mathcal{P}^\infty$ is in effect. Here we say the insurer enters after seeing the sequence $x^n \in \mathbb{N}^*$ (possibly the empty sequence) if $\tau(x^n) = 1$. The other ingredient of an insurance scheme is the *premium setting scheme* $\Pi : \mathbb{N}^* \to \mathbb{R}^+$, satisfying $\Pi(x^n) = 0$ if $\tau(x^n) = 0$, with $\Pi(x^n)$ being interpreted as the premium demanded by the insurer from the insured after the loss sequence $x^n \in \mathbb{N}^*$ is observed.

Let $1(\cdot)$ denote the indicator function of its argument. The event that the insurer goes bankrupt is the event that

$$\Pi_0 + \sum_{i=1}^{n} (\Pi(X^{i-1}) - X_i) 1(\tau(X^{i-1}) = 1) < 0 \text{ for some } n \geq 1 \ .$$

In words, this is the event that in some period $n \geq 1$ after the insurer has entered, the loss X_n incurred by the insured exceeds the built up capital of the insurer, namely the sum of its initial capital and all the premiums it has collected after it has entered (including the currenly charged premium $\Pi(X^{n-1})$) less all the losses paid out so far.

Definition 1 A class $\mathcal{P}^\infty$ of laws on loss sequences is called insurable by an insurer with initial capital $\Pi_0 \in \mathbb{R}^+$ if $\forall\ \eta > 0$, there exists an insurance scheme (τ, Π) such that $\forall\ p \in \mathcal{P}^\infty$,

$$p((\tau, \Pi) \text{ goes bankrupt }) < \eta \ .$$

We should remark that despite the apparent role of the initial capital of the insurer in this definition, it plays no role from a mathematical point of view. To see this note first that if a model class $\mathcal{P}^\infty$ is insurable by an insurer with capital Π_0 it is clearly insurable by all insurers with initial captial at least Π_0, since such an insurer can use the same entry time and premium setting scheme as the insurer with initial capital Π_0. On the other hand, an insurer with initial capital less than Π_0 can use the same

entry time as an insurer with initial capital Π_0 and simply charge an additional premium at the time of entry which in effect builds up its initial capital to Π_0, and then proceed with the same premium setting scheme as that used by the insurer with initial captial Π_0. This feature is an artifact of the complete flexibility we give the insurer in setting premiums; for more on this see the concluding remarks in Section 7.

As indicated in the introductory Section 1, we will first show that whether a model class of loss distributions is insurable is equivalent to whether we can find suitable loss-domination sequences for the sequence of losses. We next make this connection and the associated terminology precise.

Definition 2 A *loss-domination scheme* for $\mathcal{P}$ is a mapping $\Phi : \mathbb{N}^* \mapsto \mathbb{R}^+ \cup \{\infty\}$, where for $x^n \in \mathbb{N}^*$, we interpret $\Phi(x^n)$ as an estimated upper bound on x_{n+1}. We call $\{\Phi(X^{i-1}),\ i \geq 1\}$ the loss-domination sequence and $\Phi(X^{i-1})$ the loss dominant at step i. We require for all $x^n \in \mathbb{N}^*$ that

$$\Phi(x_1, \ldots, x_n) < \infty \Longrightarrow \Phi(x_1, \ldots, x_{n+1}) < \infty$$

and also that for all $p \in \mathcal{P}^\infty$,

$$p(\inf_{n \geq 1} \Phi(X^n) < \infty) = 1. \qquad \square$$

We think of $\Phi(x^n) = \infty$ as saying that the scheme has not yet committed to proposing finite loss dominants after having seen the sequence x^n, while if $\Phi(x^n) < \infty$ it has. Once the scheme commits to proposing finite loss dominants it has to continue to propose finite loss dominants from that point onwards. Further, with probability 1 under every $p \in \mathcal{P}^\infty$, the scheme has to eventually start proposing finite loss dominants.

Definition 3 Given our motivation from the insurance problem, we will say the loss-domination scheme Φ goes *bankrupt* if $\Phi(X^{n-1}) < X_n$ for some $n \geq 1$. $\square$

The connection between the insurance problem and the problem of selecting loss dominants can now be made precise as follows.

Observation 1 Let $\mathcal{P}^\infty$ be a model class and $\eta > 0$. Let $\Pi_0 \in \mathbb{R}^+$. An insurer with initial capital Π_0 can find an insurance scheme (τ, Π) such that the probability of remaining solvent is bigger than $1 - \eta$ irrespective of which $p \in \mathcal{P}^\infty$ is in effect if and only if there is a loss-domination scheme Φ such that the probability of it going bankrupt is less than η irrespective of which $p \in \mathcal{P}^\infty$ is in effect.

Proof Given an insurance scheme (τ, Π) consider the loss-domination scheme Φ that has $\Phi(x^n) := \infty$ iff $\tau(x^n) = 0$ and

$$\Phi(X^{n-1}) := \Pi_0 + \sum_{i=1}^{n-1} (\Pi(X^{i-1}) - X_i) 1(\tau(X^{i-1}) = 1) \ + \ \Pi(X^{n-1}) \ ,$$

if $\tau(X^n) = 1$. Since τ enters (become equal to 1) with probability 1 under each $p \in \mathcal{P}^\infty$ and stays equal to 1 once it has become 1, Φ becomes finite with probability 1 under each $p \in \mathcal{P}^\infty$ and stays finite once it has become finite. Thus Φ is indeed a loss-domination scheme. It is straightforward to check that if the insurance scheme (τ, Π) stays solvent with probability bigger than $1 - \eta$ irrespective of which $p \in \mathcal{P}^\infty$ is in effect then the loss-domination scheme Φ becomes bankrupt with probability less than η irrespective of which $p \in \mathcal{P}^\infty$ is in effect.

Conversely, given a loss-domination scheme Φ define the insurance scheme (τ, Π) by setting $\tau(x^n) := 0$ iff $\Phi(X^n) = \infty$ (and $\tau(x^n) := 1$ iff $\Phi(x^n) < \infty$) and defining $\Pi(x^n) := 0$ if $\Phi(x^n) = \infty$ and $\Pi(x^n) := \Phi(x^n)$ if $\Phi(x^n) < \infty$.

One sees that τ as defined becomes 1 with probability 1 under each $p \in \mathcal{P}^\infty$ and stays equal to 1 once it becomes 1. Further, the premiums set at each time are finite and equal to 0 till the entry time. Thus (τ, Π) as defined is indeed an insurance scheme.

It is straightforward to check if Φ becomes bankrupt with probability less than η irrespective of which $p \in \mathcal{P}^\infty$ is in effect, then (τ, Π) stays solvent with probability bigger than $1 - \eta$ irrespective of which $p \in \mathcal{P}^\infty$ is in effect. Hence the above observation. □

We may therefore conclude that a model class $\mathcal{P}^\infty$ is insurable iff for all $\eta > 0$ there is a loss-domination scheme Φ such that the probability of going bankrupt under Φ is less than η irrespective of which $p \in \mathcal{P}^\infty$ is in effect. In the rest of the paper we will therefore focus mainly on whether the model class $\mathcal{P}^\infty$ is such that for every $\eta > 0$ a loss-domination sequence Φ exists with its probability of bankruptcy being less than η irrespective of which model in the model class governs the sequence of losses.

In Theorem 1, we provide a condition on $\mathcal{P}$ that is both necessary and sufficient for insurability.

3. Statement of the main result

We go through a few technical points before spelling out the results in detail in 3.3.

3.1 Close distributions

Insurability of $\mathcal{P}^\infty$ depends on the neighborhoods of the probability distributions among its one dimensional marginals $\mathcal{P}$. The relevant measure of closeness between distributions in $\mathcal{P}$ that decides the neighborhoods is

$$\mathcal{J}(p, q) := D\left(p || \frac{p+q}{2}\right) + D\left(q || \frac{p+q}{2}\right).$$

Note that the above is the Jensen-Shannon divergence (with an additional factor of 2) and is *not* a true distance. Here $D(p||q)$ denotes the relative entropy of p with respect to q, where p and q are probability distributions on $\mathbb{N}$, defined by

$$D(p||q) := \sum_{y \in \mathbb{N}} p(y) \log \frac{p(y)}{q(y)} \ .$$

The logarithm is assumed to be taken to base 2 (we use ln for the logarithm to the natural base).

The reason for choosing the Jensen-Shannon (JS) divergence is that it has two convenient properties—(i) for the necessary part, it becomes easy to quantify how "close" distributions yield very similar measures on sequences, (ii) for the sufficient part, we bound the JS divergence with the ℓ_1 norm in Lemma 4 which in turn lets us work with the ℓ_1 topology induced on the class $\mathcal{P}$ of distributions. Specifically, we show that if p and q are probability distributions on $\mathbb{N}$, then

$$\frac{1}{4 \ln 2} |p - q|_1^2 \leq \mathcal{J}(p, q) \leq \frac{1}{\ln 2} |p - q|_1 \ .$$

3.2 Cumulative distribution function

Since we would like to discuss percentiles, it is convenient to use a non-standard definition for the cumulative distribution function of a probability distribution on $\mathbb{N}$.

For our purposes, the cumulative distribution function of any probability distribution p on $\mathbb{N}$ is a function $F_p : \mathbb{R}^+ \cup \{\infty\} \to [0,1]$ defined in an unconventional way. We obtain F_p by first defining F_p on points in the support of p in the way cumulative distribution functions are normally defined. We define F_p for all other nonnegative real numbers by linearly interpolating between the values in the support of p. Finally, $F_p(\infty) := 1$.

Let $F_p^{-1} : [0,1] \mapsto \mathbb{R}^+ \cup \{\infty\}$ denote the inverse function of F_p. Then $F_p^{-1}(x) = 0$ for all $0 \leq x < F_p(0)$. If p has infinite support then $F_p^{-1}(1) = \infty$, else $F_p^{-1}(1)$ is the smallest natural number y such that $F_p(y) = 1$.

Two simple and useful observations can now be made. Consider a probability distribution p with support $\mathcal{A} \subset \mathbb{N}$. For $\delta > 0$, let (T for tail)

$$T_{p,\delta} := \{y \in \mathcal{A} : y \geq F_p^{-1}(1-\delta)\},$$

and let (H for head)

$$H_{p,\delta} := \{y \in \mathcal{A} : y \leq 2F_p^{-1}(1-\delta/2)\}.$$

It is easy to see that

$$p(T_{p,\delta}) > \delta \tag{3}$$

and that

$$p(H_{p,\delta}) > 1-\delta. \tag{4}$$

Suppose that for some $\delta > 0$ we have $F_p^{-1}(1-\delta) > 0$ and the loss dominant at the beginning of period $i \geq 1$ happens to be set to $F_p^{-1}(1-\delta)$, then the probability under p of the loss in period i exceeding the loss dominant is bigger than δ. If the loss dominant at the beginning of period i happens to be set to $2F_p^{-1}(1-\delta/2)$, then the probability that the loss in period i exceeds the loss dominant is less than δ. We will use these observations in the proofs to follow.

3.3 Necessary and sufficient conditions for insurability

Existence of close distributions with very different quantiles is what kills insurability. A loss-domination scheme could be "deceived" by some process $p \in \mathcal{P}^\infty$ into setting low loss dominants, while a close enough distribution hits the scheme with too high a loss. The conditions for insurability of $\mathcal{P}^\infty$ are phrased in terms of the set of its one dimensional marginals, $\mathcal{P}$.

Formally, a distribution p in $\mathcal{P}$ is *not deceptive* if some neighborhood around p is tight. Specifically, $\exists\ \epsilon_p > 0$, such that $\forall\ \delta > 0$, $\exists\ f(\delta) \in \mathbb{R}$, such that all distributions $q \in \mathcal{P}$ with

$$\mathcal{J}(p,q) < \epsilon_p$$

satisfy

$$F_q^{-1}(1-\delta) \leq f(\delta).$$

Equivalently, a probability distribution p in $\mathcal{P}$ is *deceptive* if no neighbourhood of $\mathcal{P}$ around p is tight. Specifically, $\forall\ \epsilon > 0$, $\exists\ \delta > 0$ such that that no matter what $f(\delta) \in \mathbb{R}^+$ is chosen, $\exists$ a (bad) distribution $q \in \mathcal{P}$ such that

$$\mathcal{J}(p,q) < \epsilon$$

and

$$F_q^{-1}(1-\delta) > f(\delta).$$

In the above definition, $f(\delta)$ is simply an arbitrary nonnegative real number. However, it is useful to think of this number as the evaluation of a function $f : (0,1) \to \mathbb{R}$ at δ.

Our main theorem is the following, which we prove in Sections 5 and 6.

Theorem 1 $\mathcal{P}^\infty$ is insurable, iff no $p \in \mathcal{P}$ is deceptive. □

4. Examples

Consider $\mathcal{U}$, the collection of all uniform distributions over finite supports of form $\{m, m+1, \ldots, M\}$ for all positive integers m and M with $m \leq M$. Let the sequence of losses be *i.i.d.* samples from distributions in $\mathcal{U}$—call the resulting model class over infinite loss sequences $\mathcal{U}^\infty$.

Note that no distribution in $\mathcal{U}$ is deceptive. Around each distribution in $\mathcal{U}$ is a neighborhood that that contains no other distribution of $\mathcal{U}$.

Example 3 $\mathcal{U}^\infty$ is insurable.
Proof If the threshold probability of ruin is η, choose the loss-domination scheme Φ as follows. For all sequences x^n with $n \leq \log \frac{1}{\eta} + 1$ set $\Phi(x^n) = \infty$. For all sequences x^n with $n > \log \frac{1}{\eta} + 1$, the loss dominant $\Phi(x^n)$ is set to be twice the largest loss observed thus far. It is easy to see that this scheme is bankrupted with probability less than η irrespective of which $p \in \mathcal{U}^\infty$ is in effect. □

Consider the set $\mathcal{N}_1^\infty$ of all *i.i.d.* processes such that the one dimensional marginals have finite first moment. Namely, $\forall p \in \mathcal{N}_1^\infty$, $\mathbb{E}_p X < \infty$ where $X \in \mathbb{N}$ is distributed according to the single letter marginal of p. If $\mathcal{N}_1$ is the collection of single letter marginals from $\mathcal{N}_1^\infty$, it is easy to verify as below that every distribution in $\mathcal{N}_1$ is deceptive.

Example 4 $\mathcal{N}_1^\infty$ is not insurable.
Proof Note that the loss process that puts probability 1 on the all zero sequence exists in $\mathcal{N}_1^\infty$, since it corresponds to the one dimensional marginal loss distribution that produces loss 0 in each period. Since every loss-domination scheme enters with probability 1 no matter which $p \in \mathcal{N}_1^\infty$ is in force, every loss-domination scheme must enter after seeing some finite number of zeros. Fix any loss-domination scheme Φ. Suppose the scheme starts to set finite loss dominants after seeing N losses of size 0. To show that $\mathcal{N}_1^\infty$ is not insurable, we show that $\exists \eta > 0$ and $\exists p \in \mathcal{N}_1^\infty$ such that

$$p(\ \Phi \text{ goes bankrupt }) \geq \eta.$$

Fix $\delta = 1 - \eta$. Let ϵ be small enough that

$$(1-\epsilon)^N > 1 - \delta/2,$$

and let M be a number large enough that

$$(1-\epsilon)^M < \delta/2.$$

Note that since $1 - \delta/2 \geq \delta/2$, we have $N < M$. Let L be greater than any of the loss dominants set by Φ for the sequences $0^N, 0^{N+1}, \ldots 0^M$. Let $p \in \mathcal{N}_1^\infty$ satisfy, for all i,

$$p(X_i) = \begin{cases} 1-\epsilon & \text{if } X_i = 0 \\ \epsilon & \text{if } X_i = L. \end{cases}$$

For the *i.i.d.* loss process having the law p, the insurer is bankrupted on all sequences that contain loss L in between the N-th and M-th steps. These sequences, $0^N L, 0^{N+1}L, \ldots, 0^{M-1}L$, have respective probabilities (under p)

$$(1-\epsilon)^N\epsilon, (1-\epsilon)^{N+1}\epsilon, \ldots, (1-\epsilon)^{M-1},$$

and they also form a prefix free set. Therefore, summing up the geometric series and using the assumptions on ϵ above,

$$p(\ \Phi \text{ is bankrupted }) \geq (1-\epsilon)^N - (1-\epsilon)^M \geq 1 - \delta/2 - \delta/2 = \eta. \qquad \square$$

A reading of the proof above shows that we can say something much stronger. The distributions that break insurability have *all* their moments finite. Suppose $\mathcal{N}_*^\infty$ is the collection of measures whose single letter marginal has all moments finite. Namely for all $p \in \mathcal{N}_*^\infty$ and all finite $r \geq 0$, $\mathbb{E}_p X_1^r < \infty$. It follows that

Example 5 $\mathcal{N}_*^\infty$ is not insurable. $\square$

Consider the collection of all *i.i.d.* loss distributions with monotone one dimensional marginals. A monotone probability distribution p on $\mathbb{N}$ is one that satisfies $p(y+1) \leq p(y)$ for all $y \in \mathbb{N}$. Let $\mathcal{M}^\infty$ be the set of all *i.i.d.* loss processes, with one dimensional marginal distribution from $\mathcal{M}$, the collection of all monotone probability distributions over $\mathbb{N}$.

Again, it is easily shown that every distribution in $\mathcal{M}$ is deceptive. It follows from Theorem 1 that

Example 6 $\mathcal{M}^\infty$ is not insurable.
Proof To see that any distribution $p \in \mathcal{M}$ is deceptive, consider distributions of form $p' = (1-\epsilon)p + \epsilon q$, where $q \in \mathcal{U}$ is a monotone uniform distribution and $\epsilon > 0$.

Clearly, the ℓ_1 distance between p' and q is $\leq 2\epsilon$ (and therefore so is the JS divergence, up to a constant factor). But for all $M > 0$ and $\delta < \epsilon$, we can pick $q \in \mathcal{U}$ over a sufficiently large support that the $1-\delta-$percentile of p' can be made $\geq M$. Therefore, no neighborhood around p' is tight. $\square$

Note also that if p has finite entropy, so does every p' obtained by the above construction. Let $\mathcal{M}_* \subset \mathcal{M}$ be the collection of finite entropy monotone distributions. The above example also implies that

Example 7 $\mathcal{M}_*^\infty$ is not insurable. $\square$

Now for $h > 0$, we consider the set $\mathcal{M}_h \subset \mathcal{M}$ of all monotone distributions over $\mathbb{N}$ whose entropy is bounded above by h. Let $\mathcal{M}_h^\infty$ be the set of all *i.i.d.* loss processes with one dimensional marginals from $\mathcal{M}_h$. Then

Example 8 $\mathcal{M}_h^\infty$ is insurable.
Proof From Markov inequality, if $p \in \mathcal{M}_h$ and $X \sim p$,

$$p(X > M) = p(\log(X+1) > \log(M+1)) < \frac{E_p \log(X+1)}{\log(M+1)} \leq \frac{E_p \log \frac{1}{p(X)}}{\log(M+1)} \leq \frac{h}{\log(M+1)}.$$

To see the second inequality above, note that p is monotone therefore for $i \in \mathbb{N}$, $p(i) \leq \frac{1}{i+1}$. Therefore, for all $p \in \mathcal{M}_h$,

$$F_p^{-1}(1-\delta) \leq 2^{\frac{h}{\delta}}.$$

Thus no $p \in \mathcal{M}_h$ is deceptive, and $\mathcal{M}_h^\infty$ is insurable. $\square$

In the class $\mathcal{U}$ above, there was a neighborhood around each distribution $p \in \mathcal{U}$ with no other model from $\mathcal{U}$, Hence $\mathcal{U}$ trivially satisfied the local tightness condition that we will prove is necessary and sufficient for insurability. The above case is another extreme—the entire model class $\mathcal{M}_h$ is tight. The following example illustrates a insurable class of models where neither extreme holds.

For a distribution q over $\mathbb{N}$, let $q^{(R)}(i+R) = q(i)$ for all $i \in \mathbb{N}$. Furthermore let the *span* of any finite support probability distribution over naturals be the largest natural number which has non-zero probability. Then, let

$$\mathcal{F}_h = \left\{ (1-\epsilon)p_1 + \epsilon p_2^{(\mathrm{span}(p_1)+1)} : \forall p_1 \in \mathcal{U}, p_2 \in \mathcal{M}_h \text{ and } 1 > \epsilon > 0 \right\}$$

As always $\mathcal{F}_h^\infty$ is the set of measures on infinite sequences formed by *i.i.d.* sampling from distributions in $\mathcal{F}_h$.

Example 9 $\mathcal{F}_h^\infty$ is insurable.

Proof Let the *base* of any probability distribution over the naturals be the smallest natural number which has non-zero probability. Consider any distribution $p = (1-\epsilon)p_1 + \epsilon p_2^{(\mathrm{span}(p_1)+1)} \in \mathcal{F}_h$ with $p_1 \in \mathcal{U}$, $p_2 \in \mathcal{M}_h$, and $1 > \epsilon > 0$. Let m denote base(p), and $m + M - 1$ denote span(p_1). Thus $|\mathrm{support}(p_1)| = M$, and we have $M \geq 1$. Of course, we also have base(p_1) = base(p).

Consider any other distribution $q = (1-\epsilon')q_1 + \epsilon' q_2^{(\mathrm{span}(q_1)+1)} \in \mathcal{F}_h$, where $q_1 \in \mathcal{U}$, $q_2 \in \mathcal{M}_h$, and $1 > \epsilon' > 0$. Let m' denote base(q) (which equals base(q_1)) and let $m' + M' - 1$ denote span(q_1), so $|\mathrm{support}(q_1)| = M'$, and we have $M' \geq 1$.

It suffices to show that there is an ℓ_1 ball around p of sufficiently small radius, such that for all $\delta > 0$ we can find a uniform bound on the $(1-\delta)$-th percentile of all q in this ball.

If $m' > m$, then the ℓ_1 distance between p and q is at least $\frac{1-\epsilon}{M}$. Hence, whenever the ℓ_1 distance between p and q is strictly less than $\frac{1-\epsilon}{M}$ we must have $m' \leq m$. Thus we may assume that $m' \leq m$.

Suppose $m' + M' - 1 \geq m + \frac{2M}{1-\epsilon}$. Then $\frac{M}{M'} \leq \frac{1-\epsilon}{2}$, from which, because support$(p_1) \subseteq$ support(q_1), we can conclude that the ℓ_1 distance between q and p is at least $\frac{1-\epsilon}{2}$. Thus we may assume that $m' + M' - 1 < m + \frac{2M}{1-\epsilon}$.

Now, for any $i \geq 0$, we have

$$q(m' + M' + i) = \epsilon' q_2(i) \leq \epsilon' \frac{1}{i+1} \leq \frac{1}{i+1} \ .$$

Thus for any $K \geq 0$ we have

$$q(X > m + \frac{2M}{1-\epsilon} + K) \leq q(X > m' + M' + K) \leq \frac{h}{\log(K+1)}$$

by an argument similar to that in the preceding example, which gives the desired conclusion that no $p \in \mathcal{F}_h^\infty$ is deceptive, and hence that $\mathcal{F}_h^\infty$ is insurable. □

5. Necessary condition for insurability

In this section we prove one direction of Theorem 1, as stated next.

Theorem 2 If $\mathcal{P}^\infty$ is insurable, then no $p \in \mathcal{P}$ is deceptive.

Proof To keep notation simple, we will denote by p (or q) both a measure in $\mathcal{P}^\infty$ as well as the corresponding one dimensional marginal distribution, which is a member of $\mathcal{P}$. The context will clarify

which of the two is meant. We prove the contrapositive of the theorem: if some $p \in \mathcal{P}$ is deceptive, then $\mathcal{P}^\infty$ is not insurable.

Pick $\alpha > 0$. Suppose $p \in \mathcal{P}$ is deceptive. Let Φ be any loss-domination scheme. Recall that Φ enters on p with probability 1, in the sense that the loss dominants set by Φ will eventually become finite with probability 1 under p. For all $n \geq 1$, let

$$R_n := \{x^n : \Phi(x^n) < \infty\}$$

be the set of sequences of length n on which Φ has entered and let $N \geq 1$ be a number such that

$$p(R_N) > 1 - \alpha/2. \tag{5}$$

Fix $0 < \eta < \frac{1}{2}(1 - \alpha - \frac{2}{N})(1 - 1/e^2)$. We prove that $\mathcal{P}^\infty$ is not insurable by finding, for each loss-domination scheme Φ, a probability distribution $q \in \mathcal{P}$ such that

$$q(\ \Phi \text{ goes bankrupt }) \geq \eta.$$

The basic idea is that because Φ has to enter with probability 1 under p, it would have been forced to set premiums that are too low for q.

For any sequence x^n, let $A(x^n)$ be the set of symbols that appear in it. Recall that the head of the distribution p, $H_{p,\gamma}$, was defined in Section 3.2 to be the set $\{y \in \mathcal{A}_p : y \leq 2F_p^{-1}(1 - \gamma/2)\}$, where $\mathcal{A}_p$ is the support of p. Further, define for all $\gamma > 0$

$$R_{p,\gamma,n} := \{x^n \in R_n : A(x^n) \subseteq H_{p,\gamma})\}.$$

Given $\delta > 0$, pick $\gamma_p(\delta)$ so small that

$$(1 - \gamma_p(\delta))^{N+1/\delta} \geq 1 - \alpha/2. \tag{6}$$

A word about this parameter $\gamma_p(\delta)$, since it may not be immediately apparent why this should be defined. The advantage of doing so is technical—we will be able to handle p (and q which will be chosen later) as though they were distributions with finite span.

Set[1] $\epsilon = \frac{1}{16(\ln 2)N^8}$. Applying Lemma 5 to distributions over length-N sequences induced by the distributions $p, q \in \mathcal{P}$ such that $\mathcal{J}(p, q) \leq \epsilon$, we have for each $\delta > 0$ that

$$q(X^N \in R_{p,\gamma_p(\delta),N}) \geq 1 - \alpha - \frac{2}{N}\ . \tag{7}$$

This implies that Φ has entered with probability (under q) at least $1 - \alpha - \frac{2}{N}$ for length N sequences. We will find a suitable $\delta > 0$ and a suitable q such that $\mathcal{J}(p, q) \leq \epsilon$, and such that the scheme Φ is bankrupted with probability $> \eta$.

Since p is deceptive, there exists $\delta' > 0$ such that

$$\sup_{\substack{q \in \mathcal{P}: \\ \mathcal{J}(p,q) < \epsilon}} F_q^{-1}(1 - \delta') = \infty.$$

1. Please note that in the interest of simplicity, we have not attempted to provide the best scaling for ϵ or the tightest possible bounds in arguments below

Define

$$\Delta_{p,\epsilon} = \{\delta' : \sup_{q:\mathcal{J}(p,q)<\epsilon} F_q^{-1}(1-\delta') = \infty\}.$$

In connection with this definition, note that if the δ' tails of distributions in the $\epsilon-$neighborhood of p are not bounded, neither are the δ'' tails for all $\delta'' < \delta'$.

Note that if $\sup \Delta_{p,\epsilon} \geq 1/2$, we are done since for some $\delta \geq 1/2$, there is a q satisfying $\mathcal{J}(p,q) \leq \epsilon$ and

$$F_q^{-1}(1-\delta) \geq \max_{x^N \in R_{p,\gamma_p,N}} \Phi(x^N).$$

Therefore, conditioned on the event $\{X^N \in R_{p,\gamma_p(\delta),N}\}$, this q will be bankrupted with probability $\geq 1/2$. From (7) above we thus have

$$q(\ \Phi \text{ goes bankrupt }) \geq \frac{1-\alpha-\frac{2}{N}}{2} > \eta.$$

If not, we have $\Delta_{p,\epsilon} < 1/2$. Pick δ and $r = 2\delta$ such that $\delta \in \Delta_{p,\epsilon}$, but $r \notin \Delta_{p,\epsilon}$.

For convenience, let $M = \lceil \frac{1}{\delta} \rceil$. We consider now a set $\mathcal{S}$ of strings of lengths ranging from N to $N+M$ defined by the following properties: (i) every string in $\mathcal{S}$ has its prefix of length N belonging to $R_{p,\gamma_p(\delta),N}$; and (ii) every string of length k in $\mathcal{S}$, $N+1 \leq k \leq N+M$, has all its symbols at times $N+1$ through to k belonging to $H_{q',2r}$ for some $q' \in \mathcal{P}$ such that $\mathcal{J}(p,q') < \epsilon_p$. To clarify this definition, we recall once again that $H_{q',2r}$ is the $\{y \in \mathcal{A}_{q'} \ : \ y \leq 2F_{q'}^{-1}(1-r)\}$, where $\mathcal{A}_{q'}$ denotes the support of q'. Note that since $r \notin \Delta_{p,\epsilon}$, $\mathcal{S}$ is finite.

Again, we pick q satisfying $\mathcal{J}(p,q) \leq \epsilon$ and

$$F_q^{-1}(1-\delta) \geq \max_{\substack{x^k \in \mathcal{S} \\ N \leq k \leq N+M-1}} \Phi(x^k).$$

Therefore if a symbol from $T_{q,\delta}$ follows any string in $\mathcal{S}$, the scheme goes bankrupt under q. There may be symbols in the complement of $T_{q,\delta}$ that also bankrupt the scheme if they follow a string in $\mathcal{S}$. Taking a different perspective, It could happen that no string in $\mathcal{S}$ contains symbols in $T_{q,\delta}$ (depending on how large $H_{q,2r}$ is), or that strings $\mathcal{S}$ contain symbols from $T_{q,\delta}$. We consider all these variations below.

Let q_1 be the probability (under q) of all sequences in $R_{p,\gamma_p,N}$ under which the scheme Φ has not yet been bankrupted, and let q_2 be the probability (under q) of all sequences in $R_{p,\gamma_p,N}$ where Φ has already been bankrupted. Therefore $q_1 + q_2 = q(R_{p,\gamma_p,N})$.

To continue, we need to consider two cases: (i) no string in $\mathcal{S}$ contains symbols of $T_{q,\delta}$ which happens when $2F_q^{-1}(1-r) \leq F_q^{-1}(1-\delta)$. In this case, $H_{q,2r}$ is contained in the complement of $T_{q,\delta}$ and we show the probability under q with which Φ is bankrupted is bounded below by

$$q_2 + q_1 \left(\delta + (1-r)\delta + \ldots + (1-r)^M \delta\right).$$

This can be seen by defining a sequence in $\mathcal{S}$ to be a *survivor* if the loss-domination scheme Φ has not yet been bankrupted on this sequence under q. Thus, for instance, q_2 is the probability, under q, of survivor sequences of length N.

Given that the sequence seen so far is a survivor sequence, one can classify the next symbol in one of four ways: (a) it is in $T_{q,\delta}$ (which automatically implies bankruptcy); (b) it is in the complement of $T_{q,\delta} \cup H_{q,2r}$ which we ignore; (c) it is in $H_{q,2r}$ and results in bankruptcy; (d) it is in $H_{q,2r}$ and

does not result in bankruptcy. We ignore case (b) since we are only interested in a lower bound. In case (c) the contribution to the conditional probability of bankruptcy given the survivor sequence is 1, but we lower bound it for survivor sequences of length $N+l$, $0 \leq l \leq M-1$, by the running sum $\big(\delta + (1-r)\delta + \ldots + (1-r)^{M-l-1}\delta\big)$.

In the second case (ii) $2F_q^{-1}(1-r) > F_q^{-1}(1-\delta)$, the complement of $T_{q,\delta}$ is contained in $H_{q,2r}$, and therefore the probability under q with which Φ is bankrupted is bounded below by

$$q_2 + q_1 \left(\delta + (1-\delta)\delta + \ldots + (1-\delta)^M \delta\right).$$

This can be seen, as in the preceding case, by classifying the next symbol following a survivor sequence into three types: (a) it is in $T_{q,\delta}$ (which implies bankruptcy); (b) it is the complement of $T_{q,\delta}$ and results in bankruptcy; (c) it is the complement of $T_{q,\delta}$ and does not result in bankruptcy. In case (b), as before we lower bound the contribution to the conditional probability of bankruptcy given the survivor sequence (which is 1) by the running sum $\big(\delta + (1-\delta)\delta + \ldots + (1-\delta)^{M-l-1}\delta\big)$ for survivor sequences of length $N+l$, $0 \leq l \leq M-1$.

However, we have $1-\delta \geq 1-r$, so once again in case (ii) the probability under q with which Φ is bankrupted is bounded below by

$$q_2 + q_1 \left(\delta + (1-r)\delta + \ldots + (1-r)^M \delta\right).$$

Thus we see that irrespective of which case is in force, under q the loss-domination scheme Φ is bankrupted with probability at least

$$\begin{aligned}
&q_2 + q_1\big(\delta + (1-r)\delta + \ldots + (1-r)^M\delta\big) \\
&= q_2 + q_1\left(\frac{1-(1-2\delta)^{\lceil 1/\delta \rceil}}{2}\right) \\
&\geq \frac{1}{2} q(R_{p,\gamma_p,N})\Big(1-(1-2\delta)^{\lceil 1/\delta \rceil}\Big) \\
&\geq \frac{1}{2}\left(1-\alpha-\frac{2}{N}\right)\left(1-\frac{1}{e^2}\right).
\end{aligned}$$

The theorem follows. □

6. Sufficient condition for insurability

When no $p \in \mathcal{P}$ is deceptive, given any $\eta > 0$ we will construct a loss-domination scheme that goes bankrupt with probability $\leq \eta$.

If no $p \in \mathcal{P}$ is deceptive, there is for each $p \in \mathcal{P}$ a number $\epsilon_p > 0$ such that, for every percentile $\delta > 0$, there is a uniform bound on the δ-percentile over the set of probability distributions in the neighborhood

$$\{p' \in \mathcal{P} : \mathcal{J}(p', p) < \epsilon_p\},.$$

We pick such an ϵ_p for each $p \in \mathcal{P}$ and call it the *reach* of p. For $p \in \mathcal{P}$, the set

$$B_p = \{p' \in \mathcal{P} : \mathcal{J}(p, p') < \epsilon_p\},$$

where ϵ_p is the reach of p, will play the role of the set of probability distributions in $\mathcal{P}$ for which it will be okay to eventually set loss dominants assuming p is in force.

To prove that $\mathcal{P}^\infty$ is insurable if no distribution among its one dimensional marginals $\mathcal{P}$ is deceptive, we will need to find a way to cover $\mathcal{P}$ with countably many sets of the form B_p above. Unfortunately, $\mathcal{J}(p,q)$ is not a metric, so it is not immediately clear how to go about doing this. On the other hand note that $\mathcal{J}(p',p) \leq |p-p'|_1/\ln 2$, where $|p-p'|_1$ denotes the ℓ_1 distance between p and p' (see Lemma 4 in the Appendix). Therefore, we can instead bootstrap off an understanding of the topology induced on $\mathcal{P}$ by the ℓ_1 metric.

6.1 Topology of $\mathcal{P}$ with the ℓ_1 metric

The topology induced on $\mathcal{P}$ by the ℓ_1 metric is Lindelöf, i.e. any covering of $\mathcal{P}$ with open sets in the ℓ_1 topology has a countable subcover (see (Dugundji, 1970, Defn. 6.4) for definitions and properties of Lindelöf topological spaces).

We can show that $\mathcal{P}$ with the ℓ_1 topology is Lindelöf by appealing to the fact that the set of all probability distributions on $\mathbb{N}$ with the ℓ_1 topology, is second countable, i.e. that it has a countable basis. The set of all distributions on $\mathbb{N}$ along with ℓ_1 topology has a countable basis because it has a countable norm-dense set (consider the set of all probability distributions on $\mathbb{N}$ with finite support and with all probablities being rational). Now, $\mathcal{P}$, as a topological subspace of a second countable topological space is also second countable (Dugundji, 1970, Theorem 6.2(2)). Finally, every second countable topological space is Lindelöf (Dugundji, 1970, Thm. 6.3), hence $\mathcal{P}$ is Lindelöf.

6.2 Sufficient condition

We now have the machinery required to prove that if no $p \in \mathcal{P}$ is deceptive, then $\mathcal{P}^\infty$ is insurable, which is the other direction of Theorem 1, as stated next.

Theorem 3 If no $p \in \mathcal{P}$ is deceptive, then $\mathcal{P}^\infty$ is insurable.

Proof The proof is constructive. For any $0 < \eta < 1$, we obtain a loss-domination scheme Φ such that for all $p \in \mathcal{P}^\infty$, $p(\Phi \text{ goes bankrupt }) < \eta$.

For $p \in \mathcal{P}$, let

$$Q_p = \left\{ q : |p-q|_1 < \frac{{\epsilon_p}^2 (\ln 2)^2}{16} \right\},$$

where ϵ_p is the reach of p. We will call Q_p as the *zone* of p. The set Q_p is non-empty when $\epsilon_p > 0$.

For large enough n, the set of loss sequences of length n with empirical distribution in Q_p will ensure that the loss-domination scheme Φ to be proposed enters with probability 1 when p is in force. Note that if $\epsilon_p > 0$ is small enough then $Q_p \cap \mathcal{P} \subset B_p$—we will assume wolog that $\epsilon_p > 0$ is always taken so that $Q_p \cap \mathcal{P} \subset B_p$.

Since no $p \in \mathcal{P}$ is deceptive, none of the zones Q_p are empty and the space $\mathcal{P}$ of distributions can be covered by the sets $Q_p \cap \mathcal{P}$, namely

$$\mathcal{P} = \cup_{p \in \mathcal{P}} (Q_p \cap \mathcal{P}).$$

From Section 6.1, we know that $\mathcal{P}$ is Lindelöf under the ℓ_1 topology. Thus, there is a countable set $\tilde{\mathcal{P}} \subseteq \mathcal{P}$, such that $\mathcal{P}$ is covered by the collection of relatively open sets

$$\{Q_{\tilde{p}} \cap \mathcal{P} : \tilde{p} \in \tilde{\mathcal{P}}\}.$$

We let the above collection be denoted by $\mathcal{Q}_{\tilde{\mathcal{P}}}$. We will refer to $\tilde{\mathcal{P}}$ as the *quantization* of $\mathcal{P}$ and to elements of $\tilde{\mathcal{P}}$ as *centroids* of the quantization, borrowing from commonly used literature in classification.

We index the countable set of centroids, $\tilde{\mathcal{P}}$ (and reuse the index for the corresponding elements of $\mathcal{Q}_{\tilde{\mathcal{P}}}$) by $\iota : \tilde{\mathcal{P}} \to \mathbb{N}$.

We now describe the loss-domination scheme Φ having the property that for all $p \in \mathcal{P}^\infty$,

$$p(\Phi \text{ goes bankrupt }) < \eta.$$

Preliminaries Consider a length-n sequence x^n on which Φ has not entered thus far. Let the empirical distribution of the sequence be q, and let

$$\mathcal{P}'_q := \{p' \in \tilde{\mathcal{P}} : q \in Q_{p'}\}$$

be the set of centroids in the quantization of $\mathcal{P}$ (elements of $\tilde{\mathcal{P}}$) which can potentially *capture* q. Note that q in general need not belong to $\tilde{\mathcal{P}}$ or $\mathcal{P}$.

If $\mathcal{P}'_q \neq \emptyset$, we will further refine the set of distributions that could capture q further to $\mathcal{P}_q \subset \mathcal{P}'_q$ as described below. Refining $\mathcal{P}'_q$ to $\mathcal{P}_q$ ensures that models in $\mathcal{P}'_q$ do not prematurely capture loss sequences.

Let p be the model in force, which remains unknown. The idea is that we want sequences generated by (unknown) p to be captured by those centroids of the quantization $\tilde{\mathcal{P}}$ that have p in their reach. We will require (8) below to ensure that the probability (under the unknown p) of all sequences that may get captured by centroids $p' \in \mathcal{P}_q$ not having p in its reach remains small. In addition, we impose (9) as well to resolve a technical issue since q need not, in general, belong to $\mathcal{P}$.

For $p' \in \mathcal{P}'_q$, let the reach of p' be $\epsilon_{p'}$, and define

$$D_{p'} := \frac{{\epsilon_{p'}}^4 (\ln 2)^4}{256} .$$

In case the underlying distribution p happens to be out of the reach of p' (wrong capture), the quantity $D_{p'}$ will later lower bound the distance of the empirical q in question from the underlying p.

Specifically, we place p' in $\mathcal{P}_q$ if n satisfies

$$\exp\left(-nD_{p'}/18\right) \leq \frac{\eta}{2C(p')\iota(p')^2 n(n+1)}, \tag{8}$$

and

$$2F_q^{-1}(1 - \sqrt{D_{p'}}/6) \leq \log C(p'), \tag{9}$$

where $C(p')$ is

$$C(p') := 2^{2\left(\sup_{r \in B_{p'}} F_r^{-1}(1-\sqrt{D_{p'}}/6)\right)}.$$

Note that $C(p')$ is finite since p' is not deceptive. Comparison with Lemma 7 will give a hint as to why the equations above look the way they do.

Description of Φ For the sequence x^n with type q, if $\mathcal{P}_q = \emptyset$, the scheme does not enter yet. If $\mathcal{P}_q \neq \emptyset$, let p_q denote the distribution in $\mathcal{P}_q$ with the smallest index.

All sequences with prefix x^n (namely sequences obtained by concatenating x^n with by any other sequence of symbols) are then said to be *trapped* by p_q—namely, loss dominants will be based on p_q. The loss dominant assigned for a length-m sequence trapped by p_q is

$$2g_{p_q}\left(\frac{\eta}{4n(n+1)}\right) := 2 \sup_{r \in B_{p_q}} F_r^{-1}\left(1 - \frac{\eta}{4n(n+1)}\right).$$

Φ enters with probability 1 First, we verify that the scheme enters with probability 1, no matter what distribution $p \in \mathcal{P}$ is in force. Every distribution $p \in \mathcal{P}$ is contained in at least one of the elements of the cover $\mathcal{Q}_{\tilde{\mathcal{P}}}$.

Recall the enumeration of $\tilde{\mathcal{P}}$. Let p' be centroid with the smallest index among all centroids in $\tilde{\mathcal{P}}$ whose zones contain p. Let Q be the zone of p'. There is thus some $\gamma > 0$ such that the neighborhood around p given by

$$I(p, \gamma) := \{q : |p - q|_1 < \gamma\}$$

satisfies $I(p, \gamma) \subseteq Q$. Note in particular that p is in the reach of p'.

With probability 1, sequences generated by p will have their empirical distribution within $I(p, \gamma)$ (see Chung (1961) or Lemma 7 for an alternate proof). Next (8) will hold for all sequences whose empirical distributions that fall in $I(p, \gamma)$ whose length n is large enough—since $C(p')$ and $\iota(p')$ do not change with n, the right hand side diminishes to zero polynomially with n while the left hand side diminishes exponentially to zero. Thus we conclude (8) will be satisfied with probability 1.

Next, (9) will also hold almost surely, for if q is the empirical probability of sequences generated by p, then (with a little abuse of notation)

$$F_q^{-1}(1 - \sqrt{D_{p'}}/6) \to F_p^{-1}(1 - \sqrt{D_{p'}}/6)$$

with probability 1. Note that the quantity on the left is actually a random variable that is sequence dependent (since q is the empirical distribution of the sequence). Furthermore, we also have

$$\begin{aligned} 2F_p^{-1}(1 - \sqrt{D_{p'}}/6) &\leq 2\left(\sup_{r \in B_{p'}} F_r^{-1}(1 - \sqrt{D_{p'}}/6)\right) \\ &= \log C(p'), \end{aligned}$$

where the first inequality follows since p is in the reach of p'.

Thus the scheme enters with probability 1 no matter which $p \in \mathcal{P}$ is in force.

Probability of bankruptcy $\leq \eta$ We now analyze the scheme. Consider any $p \in \mathcal{P}$. Among sequences on which Φ has entered, we will distinguish between those that are in *good* traps and those in *bad* traps. If a sequence x^n is trapped by p' such that $p \in B_{p'}$, p' is a good trap. Conversely, if $p \notin B_{p'}$, p' is a bad trap.

(Good traps) Suppose a length-n sequence x^n is in a good trap, namely, it is trapped by a distribution p' such that $p \in B_{p'}$. Recall that the loss dominant assigned is

$$2g_{p'}\left(\frac{\eta}{4n(n+1)}\right) \geq 2F_p^{-1}\left(1 - \frac{\eta}{4n(n+1)}\right),$$

where the inequality follows because p' is not deceptive, and p is within the reach of p'. Therefore from (4), given any sequence in a good trap the scheme is bankrupted with conditional probability at most $\delta' = \eta/2n(n+1)$ in the next step. Therefore, summing over all n, sequences in good traps contribute at most $\eta/2$ to the probability of bankruptcy.

(Bad traps) We will show that the probability with which sequences generated by p fall into bad traps $\leq \eta/2$. Pessimistically, the conditional probability of bankruptcy in the very next step given a sequence falls into a bad trap is going to be bounded above by 1. Thus the contribution to bankruptcy by sequences in bad traps is at most $\eta/2$.

Let q be any length-n empirical distribution trapped by $\tilde{p}$ with reach $\tilde{\epsilon}$ such that $p \notin B_{\tilde{p}}$.

If p is "far" from $\tilde{p}$ (because p is not in $\tilde{p}$'s reach), namely

$$\mathcal{J}(\tilde{p}, p) \geq \tilde{\epsilon},$$

but q is "close" to $\tilde{p}$ (because q has to be in $\tilde{p}$'s zone to be captured by it), namely

$$|\tilde{p} - q|_1 < \frac{\tilde{\epsilon}^2 (\ln 2)^2}{16},$$

then we would like q to be far from p. That is exactly what we obtain from the triangle-inequality like Lemma 6, namely that

$$\mathcal{J}(p, q) \geq \frac{\tilde{\epsilon}^2 \ln 2}{16}$$

and hence, for all q trapped by $\tilde{p}$ that

$$|p - q|_1^2 \geq \mathcal{J}^2(p, q)(\ln 2)^2 \geq \frac{\tilde{\epsilon}^4 (\ln 2)^4}{256} = D_{\tilde{p}}^2.$$

We need not be concerned that the right side above depends on $\tilde{p}$, and there may be actually no way to lower bound the rhs as a function of just p. Rather, we take care of this issue by setting the entry point appropriately via (8).

Thus, for $p \in \mathcal{P}^\infty$, the probability length-n sequences with empirical distribution q is trapped by a bad $\tilde{p}$ is, using (8) and (9)

$$\begin{aligned}
&\leq p\left(|q - p|^2 \geq D_{\tilde{p}} \text{ and } 2F_q^{-1}(1 - \frac{\sqrt{D_{\tilde{p}}}}{6}) \leq \log C(\tilde{p})\right) \\
&\overset{(a)}{\leq} (C(\tilde{p}) - 2) \exp\left(-\frac{nD_{\tilde{p}}}{18}\right) \\
&\overset{(b)}{\leq} \frac{\eta(C(\tilde{p}) - 2)}{2C(\tilde{p})\iota(\tilde{p})^2 n(n+1)} \\
&\leq \frac{\eta}{2\iota(\tilde{p})^2 n(n+1)},
\end{aligned}$$

where the inequality (a) follows from Lemma 7 and (b) from (8). Therefore, the probability of sequences falling into bad traps

$$\leq \sum_{n \geq 1} \sum_{\tilde{p} \in \tilde{\mathcal{P}}} \frac{\eta}{2\iota(\tilde{p})^2 n(n+1)} \leq \eta/2$$

since $\sum_{\tilde{p} \in \tilde{\mathcal{P}}} \frac{1}{\iota(\tilde{p})^2} \leq \sum_{n \geq 1} \frac{1}{n(n+1)} = 1$. The theorem follows. □

7. Concluding remarks

We conclude with a few observations about insurability that, while evident from the proofs and approaches we have taken, are interesting in themselves and worth highlighting. This formulation opens up new directions of work, both in the context of insurance as well as other problems where the uniformly convergent estimators are too restrictive.

7.1 A few additional observations

Finite unions of insurable classes The first is relative simple—finite unions of insurable model classes are insurable in themselves. If $\mathcal{P}_1, \ldots, \mathcal{P}_m$ are m insurable model collections, then $\cup_{i=1}^{m} \mathcal{P}_i$ is insurable.

Countable unions of insurable classes The union of countably infinitely many classes of models each of which is insurable need not be insurable. As we have seen, the collection of monotone distributions with entropy $\leq h$, $\mathcal{M}_h$, is insurable for all $h \in \mathbb{N}$. However, the collection $\mathcal{M}_*^{\infty} = \cup_{h \in \mathbb{N}} \mathcal{M}_h$ is not insurable.

Countable unions of tight sets Note that from our arguments while proving the sufficiency criterion, it follows that every insurable model class is a countable union of tight sets. The converse is not however true. Note that $\mathcal{M}_h$ is a tight set for any $h > 0$, yet $\mathcal{M}_*^{\infty} = \cup_{h \in \mathbb{N}} \mathcal{M}_h$ is not insurable.

7.2 General remarks

The loss-domination problem formulated and solved in this paper appears to be of natural interest. However, there are several features of the insurance problem formulated here that might appear troubling even to the casual reader. In practice an insured party entering into an insurance contract would expect some stability in the premiums that are expected to be paid. A natural direction for further research is therefore to study how the notion of insurability of a model class changes when one imposes restrictions on how much the premium set by the insurer can vary from period to period. Another obvious shortcoming of the formulation of the insurance problem studied here is the assumption that the insured will accept any contract issued by the insurer. Since the insured in our model represents an aggregate of individual insured parties, a natural direction to make the framework more realistic would be to think of the insured parties as being of different *types*. This would in effect make the total realized premium from the insured (the aggregate of the insured parties) and the distribution of the realized loss in each period a function of the size of the premium per insured party set by the insurer in that period. Characterizing which model classes are insurable when the realized premium and the realized loss are functions of a set premium per insured party would be of considerable interest.

Both for the loss-domination problem and for the insurance problem, working with model classes for the loss sequence that allow for dependencies in the loss from period to period, for instance Markovian dependencies, would be another interesting direction for further research. Considering models with multiple, possibly competing insurers, as well as considering an insurer operating in multiple markets, where losses in one market can be offset by gains in another, also seem to be useful directions to investigate.

Acknowledgments

We are very grateful to anonymous reviewers whose insightful comments has made this paper much better. We thank C. Nair (Chinese Univ of Hong Kong) and K. Viswanathan (HP Labs) for helpful discussions. N. Santhanam was supported by NSF Grants CCF-1065632, CCF-1018984 and EECS-1029081. V. Anantharam was supported by the ARO MURI grant W911NF- 08-1-0233, Tools for the Analysis and Design of Complex Multi-Scale Networks, the NSF grant CNS-0910702, the NSF Science & Technology Center grant CCF-0939370, Science of Information, Marvell Semiconductor Inc., and the U.C. Discovery program.

Appendix

Lemma 4 Let p and q be probability distributions on $\mathbb{N}$. Then

$$\frac{1}{4\ln 2}|p-q|_1^2 \leq \mathcal{J}(p,q) \leq \frac{1}{\ln 2}|p-q|_1 \ .$$

If, in addition, r is a probability distribution on $\mathbb{N}$, then

$$\mathcal{J}(p,q)+\mathcal{J}(q,r) \geq \mathcal{J}^2(p,r)\frac{\ln 2}{8}.$$

Proof The lower bound in the first statement follows from Pinsker's inequality on KL divergences (see Cover and Thomas (1991) for example):

$$D\left(p||\frac{p+q}{2}\right) \geq \frac{1}{2\ln 2}\frac{1}{4}|p-q|_1^2$$

and similarly for $D\big(q||\frac{p+q}{2}\big)$. Since $\ln(1+z) \leq z$ for all $z \geq 0$, the upper bound in the first statement follows as below:

$$\begin{aligned}\mathcal{J}(p,q)\ln 2 &\leq \sum_{x:p(x)\geq q(x)} p(x)\left(\frac{p(x)-q(x)}{p(x)+q(x)}\right) + \sum_{x':q(x')\geq p(x')} q(x')\left(\frac{q(x')-p(x')}{p(x')+q(x')}\right)\\ &\leq |p-q|_1.\end{aligned}$$

To prove the triangle-like inequality, note that

$$\begin{aligned}\mathcal{J}(p,q)+\mathcal{J}(q,r) &\geq \frac{1}{4\ln 2}\big(|p-q|_1^2+|q-r|_1^2\big)\\ &\geq \frac{1}{8\ln 2}(|p-q|_1+|q-r|_1)^2\\ &\geq \frac{1}{8\ln 2}(|p-r|_1)^2\\ &\geq \frac{\ln 2}{8}\mathcal{J}(p,r)^2,\end{aligned}$$

where the last inequality follows from the upper bound on $\mathcal{J}(p,r)$ already proved. □

Lemma 5 Let p and q be probability distributions on a countable set $\mathcal{A}$ with $\mathcal{J}(p,q) \leq \epsilon$. Let p^N and q^N be distributions over $\mathcal{A}^N$ obtained by *i.i.d.* sampling from p and q respectively (the distribution induced by the product measure). For any $R_N \subset \mathcal{A}^N$ and $\alpha > 0$, if $p^N(R_N) \geq 1-\alpha$, then

$$q^N(R_N) \geq 1-\alpha-2N^3\sqrt{4\epsilon\ln 2}-\frac{1}{N}.$$

Proof Let

$$\mathcal{B}_1 = \left\{i \in \mathcal{A} : q(i) \leq p(i)\left(1-\frac{1}{N^2}\right)\right\},$$

and let

$$\mathcal{B}_2 = \left\{i \in \mathcal{A} : p(i) \leq q(i)\left(1-\frac{1}{N^2}\right)\right\},$$

If $\mathcal{J}(p,q) \leq \epsilon$, then we have

$$\sqrt{\epsilon} \geq \sqrt{\mathcal{J}(p,q)} \geq \frac{|p-q|_1}{\sqrt{4\ln 2}}.$$

It can then be easily seen that

$$p(\mathcal{B}_1 \cup \mathcal{B}_2) \leq 2N^2\sqrt{4\epsilon \ln 2} \text{ and } q(\mathcal{B}_1 \cup \mathcal{B}_2) \leq 2N^2\sqrt{4\epsilon \ln 2} \tag{10}$$

because

$$|p-q|_1 \geq \sum_{x \in \mathcal{B}_1} (p(x) - q(x)) \geq \frac{p(\mathcal{B}_1)}{N^2} \geq \frac{q(\mathcal{B}_1)}{N^2}$$

and similarly

$$N^2|p-q|_1 \geq q(\mathcal{B}_2) \geq p(\mathcal{B}_2).$$

Let $S = \mathcal{A} - \mathcal{B}_1 \cup \mathcal{B}_2$. We have for all $x \in S$,

$$q(x) \geq p(x)\left(1 - \frac{1}{N^2}\right). \tag{11}$$

and from (10) we have $p(S) \geq 1 - 2N^2\sqrt{4\epsilon \ln 2}$. Now, we focus on the set $S_N \subset \mathcal{A}^N$ containing all length-N strings of symbols from S. Clearly

$$p(S_N) \geq 1 - 2N^3\sqrt{4\epsilon \ln 2}.$$

Thus we have

$$p(R_N \cap S_N) \geq 1 - 2N^3\sqrt{4\epsilon \ln 2} - \alpha.$$

From (11), for all $x^N \in S_N$,

$$q(x^N) \geq p(x^N)\left(1 - \frac{1}{N^2}\right)^N \geq p(x^N)\left(1 - \frac{1}{N}\right).$$

Therefore,

$$q(R_N) \geq q(R_N \cap S_N) \geq (1 - 2N^3\sqrt{4\epsilon \ln 2} - \alpha)\left(1 - \frac{1}{N}\right) \geq 1 - \alpha - 2N^3\sqrt{4\epsilon \ln 2} - \frac{1}{N}. \qquad \square$$

Lemma 6 Let $\epsilon_0 > 0$. If

$$|p_0 - q|_1 \leq \frac{\epsilon_0^2(\ln 2)^2}{16} \ ,$$

then for all $p \in \mathcal{P}$ with $\mathcal{J}(p, p_0) \geq \epsilon_0$, we have

$$\mathcal{J}(p,q) \geq \frac{\epsilon_0^2 \ln 2}{16}.$$

Proof Since

$$|p_0 - q|_1 \leq \frac{\epsilon_0^2(\ln 2)^2}{16},$$

Lemma 4 implies that

$$\mathcal{J}(p_0, q) \leq \frac{\epsilon_0^2 \ln 2}{16}.$$

Further, Lemma 4 then implies that

$$\mathcal{J}(p,q) + \frac{\epsilon_0^2 \ln 2}{16} \;\geq\; \mathcal{J}(p,q) + \mathcal{J}(p_0,q) \;\geq\; \frac{\mathcal{J}^2(p,p_0)\ln 2}{8} \;\geq\; \frac{\epsilon_0^2 \ln 2}{8},$$

where the last inequality follows since $\mathcal{J}(p,p_0) \geq \epsilon_0$. □

Lemma 7 Let p be any probability distribution on $\mathbb{N}$. Let $\delta > 0$ and let $k \geq 2$ be an integer. Let X_1^n be a sequence generated *i.i.d.* with marginals p and let $q(X^n)$ be the empirical distribution of X_1^n. Then

$$p\big(|q(X^n) - p| > \delta \text{ and } 2F_q^{-1}(1-\delta/6) \leq k\big) \leq (2^k - 2)\exp\left(-\frac{n\delta^2}{18}\right).$$

Remark There is a lemma that looks somewhat similar in Ho and Yeung (2010). The difference from Ho and Yeung (2010) is that the right side of the inequality above does *not* depend on p, and this property is crucial for its use here. □

Proof The starting point is the following result. Suppose p' is a probability distribution on $\mathbb{N}$ with finite support of size L. Then from Weissman et al. (2005), if we consider length n sequences,

$$p'(|q(X^n) - p'|_1 \leq t) \geq 1 - (2^L - 2)\exp\left(-\frac{nt^2}{2}\right). \tag{12}$$

Since $k \geq 2$, consider the distributions p' and q' with support $A = \{1,\ldots,k-1\} \cup \{-1\}$, obtained as

$$p'(i) = \begin{cases} p(i) & 1 \leq i < k \\ \sum_{j=k}^{\infty} p(j) & i = -1, \end{cases}$$

and similary for q'.

From (12),

$$p'(|p' - q'|_1 > \delta/3) \leq (2^k - 2)\exp\left(-\frac{n\delta^2}{18}\right).$$

We will see that all sequences generated by p with empirical distributions q satisfying

$$|p - q|_1 > \delta \text{ and } 2F_q^{-1}(1-\delta/6) \leq k$$

are now mapped into sequences generated by p' with empirical q' satisfying

$$|p' - q'|_1 > \delta/3 \text{ and } q'(-1) \leq \delta/3. \tag{13}$$

Thus, we will have

$$\begin{aligned} &p(|q(X^n) - p|_1 > \delta \text{ and } 2F_q^{-1}(1-\delta/6) \leq k) \\ &\leq p'(|p' - q'|_1 > \delta/3 \text{ and } q'(-1) \leq \delta/3) \\ &\leq (2^k - 2)\exp\left(-\frac{n\delta^2}{18}\right). \end{aligned}$$

Finally we observe (13) as in Ho and Yeung (2010)

$$\begin{aligned}|p-q|_1 - \sum_{l=1}^{k-1} |p(l)-q(l)| \\ &\leq \sum_{j=k}^{\infty}(p(j)-q(j)) + 2\sum_{j=k}^{\infty} q(j) \\ &\leq |p'(-1)-q'(-1)| + 2\delta/3,\end{aligned}$$

where the last inequality above follows from (4). Since $p(l) = p'(l)$ and $q(l) = q'(l)$ for all $l = 1, \ldots, k-1$, we have

$$|p' - q'|_1 \geq |p-q|_1 - 2\delta/3.$$

If $|p-q|_1 \geq \delta$ in addition, $|p'-q'|_1 \geq \delta/3$. □

References

M. Asadi, R. Paravi, and N. Santhanam. Estimation in slow mixing, long memory channels. 2013. Full version available from arXiv doc id: 1301.6798.

S. Asmussen and H. Albrecher. *Ruin probabilities*. World Scientific Publishing Company, 2nd edition, 2010.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

K.L. Chung. A note on the ergodic theorem of information theory. *Annals of Mathematical Statistics*, 32:612—614, 1961.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and sons., 1991.

H. Cramer. Historical Review of Filip Lundberg's Work on Risk Theory. *Skandinavisk Aktuarietidskrift (Suppl.)*, 52:6–12, 1969. Reprinted in The Collected Works of Harald Cramér edited by Anders Martin-Löf, 2 volumes Springer 1994.

J. Dugundji. *Topology*. Allyn and Bacon Inc., Boston, 1970.

K. Englund and A. Martin-Löf. *Statisticians of the Centuries*, chapter Ernst Filip Oskar Lundberg, pages 308–311. New York: Springer, 2001.

B. Fittingoff. Universal methods of coding for the case of unknown statistics. In *Proceedings of the 5th Symposium on Information Theory*, pages 129—135. Moscow-Gorky, 1972.

S. Ho and R. Yeung. On information divergence measures and joint typicality. *IEEE Transactions on Information Theory*, 56(12):58935905, 2010.

J.C. Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674—682, November 1978.

J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629—636, July 1984.

B. Ryabko. Compression based methods for non-parametric online prediction, regression, classification and density estimation. *Festschrift in Honor of Jorma Rissanen on the occasion of his 75th birthday*, pages 271–288, 2008.

N. Santhanam and V. Anantharam. Agnostic insurance tasks and their relation to compression. In *International conference on signal processing and communications (SPCOM)*, 2012.

Y.M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3—17, 1987.

T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger. Universal discrete denoising: known channel. *IEEE Transactions on Information Theory*, 51(1):5–28, 2005. See also HP Labs Tech Report HPL-2003-29, Feb 2003.