# LEARNING AND LONG-RUN FUNDAMENTALS IN STATIONARY ENVIRONMENTS

NABIL I. AL-NAJJAR AND ERAN SHMAYA

Abstract. We prove that a Bayesian agent observing a stationary process learns to make predictions as if the fundamental, defined as the ergodic component or the long run empirical frequencies of the process, was known to him. We interpret the ergodic representation as a decomposition of a stationary belief into risk and uncertainty.

## 1. Introduction

Dynamic economic models often assume that agents make decisions with prior knowledge of the "true" or "objective" stochastic structure of their environment. This assumption is invoked in dynamic stochastic models in macroeconomics, asset pricing, and industrial organization to, among other things, endogenize expectation formation. The idea that agents know the true probabilities presumes the existence of meaningful objective probabilities that could be learnt as part of a long-run learning process.

To make these ideas precise we consider an agent who faces a stationary stochastic process with values in a finite set of outcomes. The stationarity assumption guarantees that long-run frequencies exist. We view these frequencies as the fundamental, objective properties of the process. Stationarity reflects the premises that there is nothing remarkable about the point in time when the agent started observing the process, and that the fact that it is being observed has no impact on the process. This is a natural modeling assumption for an agent who is observing the stock market or is involved in a strategic interaction with many players. Stationarity is also of great practical importance in many econometric methodologies.

*Date*: First draft: February 2013; This version: December 23, 2013.
We thank Ehud Kalai, Ehud Lehrer and Rann Smorodinsky for helpful discussions.

An agent's decisions at time $t$ depends on his subjective predictive distribution $\mu(\cdot|h^{t-1})$ about that period's outcome given observed history $h^{t-1}$ and prior belief $\mu$. The contribution of this paper is to formalize and prove the intuition that as data accumulates predictive distributions become the 'correct' predictions, i.e., the predictions that are based on the objective long-run frequencies of outcomes. We show that, almost surely, the agent's predictive distribution becomes arbitrarily close to the predictive distribution conditioned on true-long run frequencies of the process in most periods. We demonstrate that the various qualifications in the formal theorem cannot be dropped.

The formal concept connecting probabilities to frequencies is ergodicity. Ergodic processes are processes for which the objective realized frequency of every finite sequences of outcomes equals its probability. Thus, for ergodic processes, predictive distributions of an agent who observes the process equals the prediction he would have made if he knew the objective empirical frequencies of outcomes in the infinite realization of the process. The ergodic representation theorem states that every stationary belief can be represented as a belief about such ergodic processes. The ergodic representation therefore has a natural interpretation in terms of long-run fundamentals of a stochastic environment: an agent's belief can be decomposed into (1) risk that remains even conditional on knowledge of the true ergodic parameters, and (2) uncertainty about long-run fundamentals, represented by a non-degenerate belief about the value of that parameter. Risk is objective in the sense that it corresponds to objectively measurable long-run empirical frequencies. Uncertainty cannot be similarly connected to frequencies, and may thus be interpreted as the agent's ex-ante subjective assessment about the stochastic structure of his environment. Our theorem then says that the agent's uncertainty about fundamentals eventually dissipates. The agent continues to face unpredictable outcomes, but this unpredictability corresponds to known objective long-run frequencies. For more on this point, see our paper Al-Najjar and Shmaya (2012).

Our results connect several literatures on the structure of stochastic environments. The first literature centers around the concept of merging of beliefs. The seminal papers are

Blackwell and Dubins (1962) on the strong merging of opinions and its application to learning in games by Kalai and Lehrer (1993). More directly relevant to our purpose are the weaker notions of merging introduced by Kalai and Lehrer (1994) and Lehrer and Smorodinsky (1996), which focus on closeness of near-horizon predictive distributions. While strong merging obtains only under stringent assumptions, weak merging can be more easily satisfied. In our setting, for example, it is not true that the posteriors strongly merge with the true parameter, no matter how much data accumulates. On the other hand, in models where players discount the future, the relevant object is near-horizon predictive distributions thus strong merging is not needed.

Another line of enquiry focuses on representations $\mu = \int \mu_\theta \, \mathrm{d}\lambda(\theta)$ where a probability measure $\mu$ is expressed as a convex combination of of "simple, elementary" distributions $\{\mu_\theta\}_{\theta\in\Theta}$ indexed by a set of parameters $\Theta$. Two seminal theorems are de Finetti's representation of exchangeable distributions and the ergodic decomposition theorem for stationary processes. Exchangeability rules out many interesting patterns of intertemporal correlation, so it is natural for us to focus on the larger class of stationary distributions. For this class, the canonical representation is in terms of the ergodic distributions. This is the finest representation possible using parameters that are themselves stationary. And as noted earlier, the ergodic distributions can also be identified with long-run frequencies via the ergodic theorem (the formal connection appears in the body of the paper).

Representations is a natural way to think about learning. The familiar intuition that "in the long-run agents learn the true process" draws its appeal from results, such as Doob's consistency theorem (1949), that Bayesian posteriors weakly converge to the true parameter. However, decisions in many economic contexts are determined not by the agents' belief about the true parameter but by their subjective predictions about near-horizon events. Although the two concepts are related, they are not the same. The difference is seen in the following example from Jackson, Kalai, and Smorodinsky (1999, Example 5): Assume that the outcomes Heads and Tails are known to be generated by tossing a fair coin. If we take the set of all dirac measures on infinite sequences of Heads-Tails outcomes as "parameters",

then Doob's theorem implies that the posterior about the parameter converges weakly to a belief that is concentrated on the true realization. On the other hand the agent's predictions about next period's outcome is constant and never approach the predictions given the true "parameter". This example highlights that convergence of posterior beliefs to the true parameters may have little relevance to an agent's predictions and behavior.

Every process can be represented in an infinite number of ways, many of which, like the representation of the coin toss process in the last paragraph, are not very sensible. Jackson, Kalai, and Smorodinsky (1999) study the question of what makes a particular representation of a stochastic process sensible in economic and game theoretic applications. One requirement is for the process to be learnable, in the sense that an agent's predictions about near-horizon events become close to what he would have predicted had he known the true parameter. Given the close connection between ergodic distributions and long-run frequencies, the most natural representation $\mu = \int \mu_\theta \, \mathrm{d}\lambda(\theta)$ of a stationary process is where the $\theta$'s index the ergodic distributions. We show that Jackson, Kalai, and Smorodinsky results do not apply to the class of stationary distributions and to their ergodic representation. We show, however, that this representation is learnable in a weaker, yet meaningful sense as described below. We discuss the relation with Jackson, Kalai, and Smorodinsky's work in greater detail below.

A third related literature, which traces to Cover (**?**), is non-Bayesian estimation of stationary processes. See Morvai and Weiss (Morvai and Weiss 2005) and the reference therein. This literature looks for an algorithm that will make near-horizon predictions which are accurate for every stationary process. Our proofs of Theorem 3.1 and Example 3.2 rely on techniques that were developed in this literature. There is however a major difference between that literature and ours: We are interested in a specific algorithm, namely Bayesian updating. Our agents are fully Bayesian, and as such they are born with some prior belief and they update it as time goes by using Bayesian updating. Their predictions and behavior are derived from this updating process. We show how to apply the mathematical apparatus developed for the non-Bayesian estimation in our Bayesian setup.

## 2. Formal model

2.1. **Preliminaries.** An agent (a decision maker, player, or an econometrician) observes a stochastic process $(\zeta_0, \zeta_1, \zeta_2, \dots)$ that takes values in a finite set of *outcomes* $A$. Time is indexed by $n$ and the agent starts observing the process at $n = 0$. Let $\Omega = A^{\mathbb{N}}$ be the space of *realizations*, with generic element denoted $\omega = (a_0, a_1, \dots)$. Endow $\Omega$ with the product topology and the induced Borel structure $\mathcal{F}$. Let $\Delta(\Omega)$ be the set of probability distributions over $\Omega$. A standard way to represent uncertainty about the process is in terms of an index set of "parameters:"

**Definition 2.1.** Let $\mu \in \Delta(\Omega)$. A *representation* of $\mu$ is given by a quadruple $(\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ where: $(\Theta, \mathcal{B}, \lambda)$ is a standard probability space of *parameters* and $\mu_\theta \in \Delta(\Omega)$ for every $\theta \in \Theta$ such that the map $\theta \mapsto \mu_\theta(A)$ is $\mathcal{B}$-measurable and

$$\mu(A) = \int_\Theta \mu_\theta(A)\ \lambda(\mathrm{d}\theta) \tag{1}$$

for every $A \in \mathcal{F}$. ▲

A representation captures a certain way in which a Bayesian agent arranges his beliefs: The agent views the process as a two stages randomization. First a parameter $\theta$ is chosen according to $\lambda$ and then the outcomes are generated according to $\mu_\theta$. Beliefs can be represented in many ways. The two extreme representations are:

- *The Trivial Representation.* Take $\Theta = \{\bar{\theta}\}$, $\mathcal{B}$ is trivial, and $\mu_{\bar{\theta}} = \mu$.
- *Dirac's Representation.* Take $\Theta = A^{\mathbb{N}}$, $\mathcal{B} = \mathcal{F}$, and $\lambda = \mu$. A "parameter" in this case is just a measure $\delta_\omega$ that assigns probability 1 to the realization $\omega$.

We are interested in representations that identify "useful" patterns shared by many realizations. These patterns capture our intuition of fundamentals of a process. The two extreme cases mentioned above are usually unsatisfactory for our purposes. In Dirac's representation, there are as many parameters as there are realizations; parameters simply copy realizations.

In the trivial representation, there is a single parameter and thus cannot discriminate between different interesting patterns. In the following example, the representation does seem to capture some fundamental properties of the process.

**Example 2.2.** The set of outcomes is $A = \{0, 1\}$ and the agent's belief is given by

$$\mu\left(\zeta_n = a_0, \ldots, \zeta_{n+k-1} = a_{k-1}\right) = \frac{1}{(k+1) \cdot \binom{k}{d}}$$

for every $n, k \in \mathbb{N}$ and $a_0, \ldots, a_{k-1} \in A$ where $d = a_0 + \cdots + a_{k-1}$. Thus, the agent believes that if he observes the process $k$ consecutive days then the number $d$ of good days (days with outcome 1) is distributed uniformly in $[0, k]$ and all configuration with $d$ good outcomes are equally likely.

In this example, the celebrated representation is that provided by de Finetti's Theorem. This is the representation given by $(\Theta, \mathcal{B}, \lambda)$ where $\Theta = [0, 1]$ is the interval equipped standard Borel structure $\mathcal{B}$ and Lebesgue's measure $\lambda$, and, for $\theta \in \Theta$ $\mu_\theta \in \Delta(\Omega)$ is the distribution of i.i.d coin tosses with probability $\theta$ for success:

$$\mu_\theta\left(\zeta_n = a_0, \ldots, \zeta_{n+k-1} = a_k\right) = \theta^d (1-\theta)^{k-d}$$

▲

2.2. **Learning.** In this section we formulate two concepts of learning. The main definition in this section, Definition 2.6, is concerned with learning to make next-period prediction. We explain how this is different from the idea of consistency of representation which appears in Bayesian statistics literature and can be thought of as learning the parameter.

2.2.1. *Learning the parameter.* The following definition captures the idea that the parameter in the representation represents some basic property of the process, in the sense that it can be deduced from the realization.

**Definition 2.3.** Let $\mu \in \Delta(\Omega)$ and let $\mathcal{R} = (\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ be a representation of $\mu$. A function $f : \Omega \to \Theta$ (asymptotically) *identifies* $\mathcal{R}$ if

$$\mu_\theta \left( \{\omega : f(\omega) = \theta\} \right) = 1 \tag{2}$$

for $\lambda$-almost every $\theta$. A representation $\mathcal{R} = (\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ is *identifiable* if there exists some function $f : \Omega \to \Theta$ that identifies $\mathcal{R}$. ▲

De-Finetti's representation in Example 2.2 is identifiable using

$$f(\omega) = \liminf_{k \to \infty} \frac{1}{k} \left( \omega_m + \cdots + \omega_{m+k-1} \right). \tag{3}$$

As an example of a nonidentifiable representation, consider the representation of i.i.d. tosses of a fair coin that is given by the parameter space $\Theta = [0,1]^{\mathbb{N}}$ equipped with the product of uniform distribution over $[0,1]$ and such that, for every $\theta = (\theta_0, \theta_1, \dots) \in \Theta$, the belief $\mu_\theta$ corresponds to a sequence of independent coin tosses when the probability of success in the $n$-th toss is $\theta_n$.

This notion of identification appears in statistics in many forms. A function $f$ that satisfies (2) is sometimes called a *splif*. See, for example, Weizsacker Weizsäcker (1996) and the references therein. In Weizsacker's terminology, a representation that is asymptotically identifiable from time 0 is said to have property $(\gamma)$. A well-known general implication of identifiability is given by Doob's consistency result for the posterior distributions: The posterior belief over the parameter weakly converges to Dirac atomic distribution over the parameter for $\lambda$-almost every parameter. Thus, using Bayesian updating about the parameter, the agent learns the true parameter.

Assume that $\Theta$ is endowed with some polish topology and let $\Delta(\Theta)$ be the space of beliefs over the parameter set $\Theta$ equipped with the topology of weak converges: For elements $\eta; \eta_0, \eta_1, \dots$ of $\Delta(\Theta)$ we say that $\eta_n \xrightarrow[n\to\infty]{w} \eta$ if $\int h \, \mathrm{d}\eta_n \xrightarrow[n\to\infty]{} \int h \, \mathrm{d}\eta$ for every bounded continuous function $h : \Theta \to \mathbf{R}$.

**Proposition 2.4** (Doob's Theorem)**.** *Let $\mu \in \Delta(\Omega)$ and let $\mathcal{R} = (\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ be a representation of $\mu$ which is identifiable by $f : \Omega \to \Theta$. Then for $\lambda$-almost every $\theta$ and $\mu_\theta$-almost every realization $(a_0, a_1, \dots) \in \Omega$, it holds that*

$$\mu\left(f(\omega) \in \cdot \,|a_0, \dots, a_{n-1}\right) \xrightarrow[n\to\infty]{w} \delta_\theta.$$

Doob's theorem does not necessarily imply that the agent can use his learning to make predictions about future outcomes. For example, consider the example mentioned in the introduction, with Dirac's representation of the process of fair coin tosses. Then the parameter space $\Omega$ is equipped with the product topology. Suppose the true parameter is $\omega^*$ for some $\omega^* = (\omega_0^*, \omega_1^*, \dots)$. After observing the first $n$ outcomes of the process the agent's belief about the parameter is uniform over all $\omega$ that agrees with $\omega^*$ on the first $n$ coordinates. While this belief indeed converges to $\delta_{\omega^*}$ in accordance with Doob's theorem, learning the parameter in this environment is just recording the past. The agent does not gain any new insight about the future of the process from learning the parameter.

2.2.2. *Learning to make predictions.* For every $\mu \in \Delta(\Omega)$ and sequence $(a_0, \dots, a_{n-1}) \in A^n$ with positive $\mu$-probability, the *$n$-day predictive distribution* is the element $\mu(\cdot|a_0, \dots, a_{n-1}) \in \Delta(A)$ representing the agent's prediction about next day's outcomes given a prior $\mu$ and after observing the first $n$ outcomes of the process. Predictive distribution in this paper will always refer to one step ahead predictions. This is for expository simplicity; our analysis covers any finite horizon.

Kalai and Lehrer (1994) and Kalai, Lehrer, and Smorodinsky (1999) introduced the following notions merging. Note that in our setup, where the set of outcomes is the same in every period this definition of merging is the same as 'weak star merging' in **?**.

**Definition 2.5.** Let $\mu, \tilde{\mu} \in \Delta(\Omega)$. Then the belief $\tilde{\mu}$ *merges* to $\mu$ if

$$|\tilde{\mu}(\cdot|a_0, \dots, a_{n-1}) - \mu(\cdot|a_0, \dots, a_{n-1})| \xrightarrow[n\to\infty]{} 0$$

for $\mu$-almost every realization $\omega = (a_0, a_1, \dots) \in A^{\mathbb{N}}$.

The belief $\tilde{\mu}$ *weakly merges* to $\mu$ if

$$|\tilde{\mu}(\cdot|a_0,\dots,a_{n-1}) - \mu(\cdot|a_0,\dots,a_{n-1})| \xrightarrow[n\to\infty]{\text{s.c}} 0.^{1} \tag{4}$$

for $\mu$-almost every realization $\omega = (a_0, a_1, \dots) \in A^{\mathbb{N}}$. ▲

Here and later, for every pair $p, q \in \Delta(A)$ we let $\|p - q\| = \max_{a\in A} |p[a] - q[a]|$. These definitions were inspired by Blackwell and Dubins idea of strong merging, which requires that the prediction of $\tilde{\mu}$ will be similar to the prediction of $\mu$ not just for the next day but for the infinite horizon.

The definition of weak merging is natural: If the truth is $\mu$ and a Bayesian agent is updating according to $\tilde{\mu}$ then his next day predictions are accurate except for rare times. Thus, in terms of optimal decisions, when $\tilde{\mu}$ weakly merges with $\mu$ then $\tilde{\mu}$-optimal strategy in a repeated decision problem will be also approximately $\mu$-optimal if the agent is sufficiently patient. Kalai, Lehrer, and Smorodinsky (1999) provide another motivation by characterizing weak merging in terms of the properties of statistical tests. They showed that $\tilde{\mu}$ weakly merges with $\mu$ if and only if forecasts made by $\tilde{\mu}$ pass all calibration tests under $\mu$. In addition, Lehrer and Smorodinsky (**?**) provides a characterization of weak merging in terms of the relative entropy between $\tilde{\mu}$ and $\mu$[2].

The following definition is the counterpart of the definition of learnable representation in JKS with weak merging instead of merging.

**Definition 2.6.** A representation $(\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ of $\mu \in \Delta(\Omega)$ is *learnable* if $\mu$ merges with $\mu_\theta$ for $\lambda$-almost every $\theta$. The representation is *weakly learnable* if $\mu$ weakly merges with $\mu_\theta$ for $\lambda$-almost every $\theta$. ▲

[1] A bounded sequence of real numbers $a_0, a_1, \dots$ is said to *strongly Cesaro converges* to a real number $a$, denoted $a_n \xrightarrow[n\to\infty]{\text{s.c}} a$, if $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} |a_k - a| = 0$. Equivalently [FIND REFERENCE], $a_n \xrightarrow{\text{s.c}} a$ if there exists a full density set $T \subseteq \mathbb{N}$ such that $\lim_{n\to\infty, n\in T} a_n = a$.

[2] However, we do not know whether their condition can be used to prove our theorem without repeating the whole argument.

As an example of a learnable representation, consider the Bayesian agent of Example 2.2. In this case

$$\mu(1|a_0,\dots,a_{n-1}) = \frac{a_0+\dots+a_{n-1}+1}{n+1}.$$

For every $\theta \in [0,1]$ it follows from the strong law of large numbers that for every parameter $\theta \in [0,1]$ this expression converges $\mu_\theta$-almost surely to $\theta$. Therefore $\mu$ merges with $\mu_\theta$ for every $\theta$, so that De Finetti's representation is learnable (and, a fortiori, weakly learnable). This is a rare case in which the predictions $\mu(\zeta_n \in \cdot|a_0,\dots,a_{n-1})$ and $\mu_\theta(\zeta_n \in \cdot|a_0,\dots,a_{n-1})$ can be calculated explicitly. In general merging and weak merging are difficult to establish, because the Bayesian prediction about the next day is a complicated expression which potentially depends on entire observed past.

2.3. **Stationarity.** A stochastic process $(\zeta_0,\zeta_1,\dots)$ is stationary if, for every natural number $k$, the joint distribution of the $k$-tuple $(\zeta_n,\zeta_{n+1},\dots,\zeta_{n+k-1})$ does not depend on $n$.

The first example of a stationary process is and i.i.d. process. Exchangeable processes, as in Example 2.2, admit a representation with i.i.d. components and are also stationary. Another well-known example is Markov processes in their steady state and mixtures of such processes. As an example of a stationary process which is not a Markov process of any finite memory consider a Hidden Markov model, according to which the outcome at every day is a function (possibly stochastic) of an underlying, unobserved Markov process:

**Example 2.7.** An agent believes that the state of the economy every day is a noisy signal of an underlying "hidden" states that changes according to a Markov chain with memory 1. Formally, let $A = \{\mathrm{B},\mathrm{G}\}$ be the set of outcomes. Let $H = \{\mathrm{B},\mathrm{G}\}$ be the set of hidden (unobserved) states. Let $(\xi_n,\zeta_n)$ be a $(H\times A)$-valued stationary Markov process with transition matrix $\rho : H\times A \to \Delta(H\times A)$ given by

$$\rho(h,a)[h',a'] = \big(p\delta_{h,h'} + (1-p)(1-\delta_{h,h'})\big)\cdot\big(q\delta_{h',a'} + (1-q)(1-\delta_{h',a'})\big),$$

where $1/2 < p,q < 1$. so that if at day $n$ the hidden state was $h$ then at day $n+1$ the hidden state $h'$ remains $h$ with probability $p$ and changes with probability $1-p$, and the observed

state $a'$ of day $n+1$ equals $h'$ with probability $q$ and is different from $h$ with probability $1-q$. Let $\mu_{p,q} \in \Delta(A^{\mathbb{N}})$ be the distribution of $\zeta_0, \zeta_1, \dots$. Then $\mu_{p,q}$ is a stationary process which is not markov of any order. If the agent is uncertain about $p, q$ then his belief $\mu$ about the outcome process is again stationary, and can be represented by some prior over the parameter set $\Theta = (1/2, 1] \times (1/2, 1]$. This representation of $\mu$ will be the ergodic representation, to be defined formally below. ▲

2.4. **The ergodic representation of stationary process.** The set of stationary measures over $\Omega$ is convex and compact in the weak*-topology. Its extreme points are called *ergodic beliefs*. We denote the set of ergodic beliefs by $\mathcal{E}$. Every stationary belief $\mu$ can therefore be uniquely written as a continuous convex combination of ergodic beliefs: $\mu = \int \nu \ \lambda(\mathrm{d}\nu)$ for some belief $\lambda \in \Delta(\mathcal{E})$. This gives rise to a natural representation of a stationary belief in which the parameter set is the set of ergodic beliefs. We call this representation *the ergodic representation*. We proceed to explain how this representation is identifiable using limit of frequencies.

According to the ergodic theorem, for every stationary belief $\mu$ and every block $(\bar{a}_0, \dots, \bar{a}_{k-1}) \in A^k$, the limit frequency

$$\Pi(\omega; \bar{a}_0, \dots, \bar{a}_{k-1}) = \lim_{n\to\infty} \frac{1}{n} \#\big\{0 \le t < n : a_t = \bar{a}_0, \dots, a_{t+k-1} = \bar{a}_{k-1}\big\}$$

exists for $\mu$-almost every realization $\omega = (a_0, a_1, \dots)$. When $\mu$ is ergodic this limit equals the probability $\mu([\bar{a}_0, \dots, \bar{a}_{k-1}])$. Thus, for ergodic processes, the probability of every block equals its (objective) empirical frequency.

The ergodic representation theorem states that for $\mu$-almost every $\omega$, The function $\Pi(\omega; \cdot)$ defined over blocks can be extended to a stationary measure over $\Delta(\Omega)$ which is also ergodic. Moreover, $\mu = \int \Pi(\omega; \cdot) \ \mu(\mathrm{d}\omega)$, so that the function $\omega \to \Pi(\omega; \cdot)$ identifies the ergodic representation. Thus, the parameters $\mu_\theta$ in the ergodic decomposition represent the empirical distribution of finite sequences of outcome along the realization of the stationary process. These parameters capture our intuition of fundamentals of the process.

## 3. Main Theorem

We are now in a position to state our main theorem.

**Theorem 3.1.** *The ergodic representation of every stationary stochastic process is weakly learnable.*

To see the implications of our theorem, consider first the Hidden Markov process of (Example 2.7) under some uncertainty. From Doob's theorem about consistency of Bayesian estimator it follows that the conditional belief over the parameter $(p, q)$ converges almost surely in the weak-topology over $\Delta(\Theta)$ to the belief concentrated on the true parameter. However, because next-day's predictions involve horrible expressions that depend on the entire history of the process, it is not clear whether these predictions merge with the truth. It follows from our theorem that they weakly merge.

Consider now the general case. If the agent knew the fundamental $\theta$, then at day $n$, after observing the partial history $(a_0, \ldots, a_{n-1})$, his predictive probability that the next day outcome is $a_n$ would have been

$$\frac{\mu_\theta(a_0, \ldots, a_{n-1}, a_n)}{\mu_\theta(a_0, \ldots, a_{n-1})}. \tag{5}$$

Since the ergodic parameter is identifiable, then again by Doob's theorem it follows that, given uncertainty about the fundamental, the agent's assessment of $\mu_\theta(b)$ becomes asymptotically accurate for every block $b$. However, when the agent has to compute the next-day posterior probability (5), he only had one observation of a block of size $n$ and no observation of the block of size $n + 1$ so at that stage his assessment of the probabilities that appear in (5) may be completely wrong. Our theorem says that the agent would still weakly learn to make these predictions correctly.

Theorem 3.1 states that the agent will make predictions about near-horizon events as if he knew the fundamental of the process. Note, however, that it is not possible to ensure that the agent will learn to predict long-run events correctly, no matter how much data accumulates. For example, consider an agent who faces a sequence of i.i.d. coin tosses with

parameter $\theta \in [0,1]$ representing the probability of Heads. Suppose this agent has a uniform prior over [0,1]. This agent will eventually learn to predict near horizon outcomes as if he knew the true parameter $\theta$, but if he will continue to assign probability 0 to the event that the long-run frequency is $\theta$. In economic models, discounting implies that only near-horizon events matter.

We end this section with an example that in Theorem 3.1 weak learnability cannot be replaced by learnability. The example is a modification of an example given by Ryabko for the forward prediction problem in a non-Bayesian setup (**?**).

**Example 3.2.** Every day there is a probability 1/2 for eruption of war. If no war erupts then the outcome is either bad economy or good economy and is a function of the number of peaceful days since the last war. The function from the number of peaceful days to outcome is an unknown parameter of the process, and the agent has a uniform prior over this parameter.

Formally, let $A = \{\mathrm{W}, \mathrm{B}, \mathrm{G}\}$ be the set of outcomes. We define $\mu \in \Delta(A^{\mathbb{N}})$ through its ergodic representations. Let $\Theta = \{\mathrm{B}, \mathrm{G}\}^{\{1,2,\dots\}}$ be the set of parameters with the standard Borel structure $\mathcal{B}$ and the uniform distribution $\lambda$. Thus, a parameter is a function $\theta : \{1, 2, \dots\} \to \{\mathrm{B}, \mathrm{G}\}$. For every such $\theta$ let $\mu_\theta$ be the distribution of a sequence $\zeta_0, \zeta_1, \dots$ of $A$-valued random variables such that

$$\zeta_n = \begin{cases} \mathrm{W}, & \text{if } \xi_n = 0 \\ \theta(\xi_n), & \text{otherwise.} \end{cases}$$

where $\xi_0, \xi_1, \dots$ is an $\mathbb{N}$-valued stationary Markov process with transition probability

$$\rho(j|k) = \begin{cases} 1/2, \text{ if } j = k+1, \\ 1/2, \text{ if } j = 0, \\ 0, \text{ otherwise.} \end{cases}$$

For every $n, m \in \mathbb{N}$. Here $\xi_n$ is the time that elapsed since the last time a war occurred. Consider a Bayesian agent that observes the process of outcomes. After the first time a war erupts the agent keeps track the state of the process $\xi_n$ at every day. If there is no uncertainty about the parameter, i.e., if the Bayesian agent knew $\theta$, his prediction about the next outcome when $\zeta_n = k$ gives probability 1/2 to outcome W and probability 1/2 to outcome $\theta(k+1)$. On the other hand, if the agent does not know $\theta$ but believes that it is randomized according to $\lambda$, he can deduce the values $\theta(k)$ gradually while he observes the process. However for every $k \in \{1, 2, 3, \dots\}$ there will be a time when the agent will observe $k$ consecutive peaceful day for the first time and at this point the agent's prediction about the next outcome will be $(1/2, 1/4, 1/4)$. Thus there will always be infinitely many occasions in which an agent that predicts according to $\mu$ will differ than an agent who predicts according to $\mu_\theta$. Therefore the representation is not learnable. On the other hand, in agreement with our theorem, these occasions become rarer as time goes by so the representation is weakly learnable. ▲

## 4. Proof of Theorem 3.1

Up to now we assumed that the stochastic process starts at time $n = 0$. When working stationary processes it is natural to extend the index set of the process from $\mathbb{N}$ to $\mathbb{Z}$, i.e. to assume that the process has infinite past. This is without loss of generality: every stationary stochastic process $\zeta_0, \zeta_1, \dots$ admits an extension $\dots, \zeta_{-1}, \zeta_0, \zeta_1, \dots$ to the index set $\mathbb{Z}$ (Kallenberg 2002, Lemma 10.2). We therefore assume hereafter, with harmless contrast with our previous notation, that $\Omega = A^{\mathbb{Z}}$.

Let $\mathcal{D}$ be a $\sigma$-algebra Borel subsets of $\Omega$. The *quotient space* of $(\Omega, \mathcal{F}, \mu)$ by $\mathcal{D}$ is the unique (up to isomorphism of measure spaces) standard probability space $(\Theta, \mathcal{B}, \lambda)$ and a measurable map $\alpha : \Omega \to \Theta$ such that $\mathcal{D}$ is generated by $\alpha$, i.e., for every $\mathcal{F}$-measurable function $f$ from $\Omega$ to some standard probability space there exists a (unique up to equality $\lambda$-almost surely) $\mathcal{B}$-measurable *lifting* $\tilde{f}$ defined over $\Theta$ such that $f = \tilde{f} \circ \alpha \quad \mu - \text{a.s.}$. The

*conditional distributions* of $\mu$ over $\mathcal{D}$ is the unique (up to equality $\lambda$-almost surely) family $\mu_\theta$ of probability measures over $(\Omega, \mathcal{F}, \mu)$ such that:

(1) For every $\theta \in \Theta$ it holds that

$$\mu_\theta\left(\{\omega | \alpha(\omega) = \theta\}\right) = 1. \tag{6}$$

(2) The map $\theta \mapsto \mu_\theta(A)$ is $\mathcal{B}$-measurable and (1) is satisfied for every $A \in \mathcal{F}$.

We call $(\Theta, \mathcal{B}, \lambda, \mu_\theta)$ the *representation of $\mu$ induced by $\mathcal{D}$*. For every belief $\mu \in \Delta(\Omega)$, the trivial representation of $\mu$ is generated by the trivial sigma-algebra $\{\emptyset, \Omega\}$, Dirac's representation is generated by the sigma-algebra of all Borel subsets of $\Omega$ and the ergodic representation is induced. The ergodic representation of a stochastic process is generated by the $\sigma$-algebra $\mathcal{I}$ of all of all *invariant* Borel sets of $\Omega$, i.e. all Borel sets $S \subseteq \Omega$ such that $S = T^{-1}(S)$ where $T : \Omega \to \Omega$ is the left shift.

We will prove a more general theorem, which is interesting in its own. Let $T : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ be the left shift so that $T(\omega)_n = \omega_{n+1}$ for every $n \in \mathbb{Z}$. A sigma-algebra $\mathcal{D}$ of Borel subsets of $\Omega$ is *shift-invariant* if $S \in \mathcal{D} \leftrightarrow T(S) \in \mathcal{D}$ for every Borel subset $S$ of $A^{\mathbb{Z}}$.

**Theorem 4.1.** *Let $\mu$ be a stationary distribution over $\Omega$ and let $\mathcal{D}$ be a shift invariant $\sigma$-algebra of subsets of $\Omega$ such that $\mathcal{D} \subseteq \mathcal{F}^0_{-\infty}$. Then the representation of $\mu$ induced by $\mathcal{D}$ is weakly learnable.*

Theorem 3.1 follows immediately from Theorem 4.1 since the sigma-algebra of invariant sets $\mathcal{I}$ which induces the ergodic representation satisfies the assumption of the Theorem 4.1. We will prove Theorem 4.1 using Lemma 4.2

**Lemma 4.2.** *Let $\mu$ be a stationary distribution over $A^{\mathbb{Z}}$ and let $\mathcal{D}$ be a shift invariant $\sigma$-algebra of Borel subsets of $A^{\mathbb{Z}}$. Then*

$$\|\mu(\zeta_n = \cdot | \mathcal{F}^n_0 \vee \mathcal{D}) - \mu(\zeta_n = \cdot | \mathcal{F}^n_{-\infty} \vee \mathcal{D})\| \xrightarrow[n\to\infty]{s.c} 0 \ \mu\text{-a.s.} \tag{7}$$

Consider the case in which $\mathcal{D} = \{\emptyset, \Omega\}$ is trivial. Then Lemma 4.2 says that a Bayesian agent who observes a stationary process from time $n = 0$ onwards will make predictions in the long run as if he knew the infinite history of the process.

*Proof of Lemma 4.2.* For every $n \leq 0$ let $f_n : \Omega \to \Delta(A)$ be a version of the conditional distribution of $\zeta_0$ according to $\mu$ given the finite history $\zeta_{-1}, \ldots, \zeta_{-n}$ and $\mathcal{D}$:

$$f_n = \mu(\zeta_0 = \cdot | \mathcal{F}^0_{-n} \vee \mathcal{D}),$$

and let $f_{-\infty} : \Omega \to \Delta(A)$ be a version of the conditional distribution of $\zeta_0$ according to $\mu$ given the infinite history $\zeta_{-1}, \ldots$ and $\mathcal{D}$:

$$f_{-\infty} = \mu(\zeta_0 = \cdot | \mathcal{F}^0_{-\infty} \vee \mathcal{D}).$$

Let $g_n = \|f_n - f_{-\infty}\|$. By the martingale convergence theorem $\lim_{n\to-\infty} f_n = f_{-\infty}$ $\mu$-a.s and therefore

$$\lim_{n\to-\infty} g_n = 0 \ \mu\text{-a.s} \tag{8}$$

It follows from the stationarity of $\mu$ and the fact that $\mathcal{D}$ is shift invariant that

$$\|\mu\left(\zeta_n = \cdot | \mathcal{F}^n_0 \vee \mathcal{D}\right) - \mu\left(\zeta_n = \cdot | \mathcal{F}^n_{-\infty} \vee \mathcal{D}\right)\| = \|f_n \circ T^n - f_{-\infty} \circ T^n\| = g_n \circ T^n \ \mu\text{-a.s} \tag{9}$$

Therefore

$$\frac{1}{N}\sum_{n=0}^{N-1} \|\mu\left(\zeta_n = \cdot | \mathcal{F}^n_0 \vee \mathcal{D}\right) - \mu\left(\zeta_n = \cdot | \mathcal{F}^n_{-\infty} \vee \mathcal{D}\right)\| = \frac{1}{N}\sum_{n=0}^{N-1} g_n \circ T^n \xrightarrow[N\to\infty]{} 0 \ \mu\text{-a.s}$$

where the equality follows from (9) and the limit follows from Maker's generalization of the individual ergodic theorem (Kallenberg 2002, Corollary 10.8)

**Maker's Ergodic Theorem.** *Let $\mu \in \Delta(\Omega)$ be such that $T\mu = \mu$ and let $h_0, h_1, \cdots : \Omega \to \mathbf{R}$ be such that $\sup_n |h_n| \in L^1(\mu)$ and $h_n \to h_\infty \quad \mu - a.s.$ Then*

$$\frac{1}{N}\sum_{n=0}^{N-1} h_n \cdot T^n \xrightarrow[N\to\infty]{} E(h_\infty|\mathcal{I}) \quad \mu - a.s.$$

□

*Proof of Theorem 4.1.* From $\mathcal{D} \subseteq \mathcal{F}^0_{-\infty}$ it follows that $\mathcal{F}^n_{-\infty} \vee \mathcal{D} = \mathcal{F}^n_{-\infty}$. Therefore, from Lemma 4.2 we get that

$$\|\mu(\zeta_n = \cdot|\mathcal{F}^n_0 \vee \mathcal{D}) - \mu(\zeta_n = \cdot|\mathcal{F}^n_{-\infty})\| \xrightarrow[n\to\infty]{\text{s.c}} 0 \ \mu\text{-a.s.}$$

By the same lemma (with $\mathcal{D}$ trivial)

$$\|\mu(\zeta_n = \cdot|\mathcal{F}^n_0) - \mu(\zeta_n = \cdot|\mathcal{F}^n_{-\infty})\| \xrightarrow[n\to\infty]{\text{s.c}} 0 \ \mu\text{-a.s.}$$

By the last two limits and the triangular inequality

$$\|\mu(\zeta_n = \cdot|\mathcal{F}^n_0) - \mu(\zeta_n = \cdot|\mathcal{F}^n_0 \vee \mathcal{D})\| \leq \tag{10}$$

$$\|\mu(\zeta_n = \cdot|\mathcal{F}^n_0) - \mu(\zeta_n = \cdot|\mathcal{F}^n_{-\infty})\| + \|\mu(\zeta_n = \cdot|\mathcal{F}^n_0 \vee \mathcal{D}) - \mu(\zeta_n = \cdot|\mathcal{F}^n_{-\infty})\| \xrightarrow[n\to\infty]{\text{s.c}} 0 \ \mu\text{-a.s}$$

Let $(\Theta, \mathcal{B}, \lambda)$ be the quotient of $(\Omega, \mathcal{F}, \mu)$ over $\mathcal{D}$ and let $(\mu_\theta)$ be the corresponding conditional distributions. Let $S$ be the set of all realizations $\omega \in \Omega$ such that $omega = (\dots, a_{-1}, a_0, a_1, \dots)$ such that

$$\|\mu(\zeta_n = \cdot|a_{n-1}, \dots, a_0) - \mu_\omega(\zeta_n = \cdot|a_{n-1}, \dots, a_0) \xrightarrow[n\to\infty]{\text{s.c}} 0.$$

Then $\mu(S) = 1$ by (10). But $\mu(S) = \int \mu_\theta(S)\lambda(\mathrm{d}\theta)$. It follows that $\mu_\theta(S) = 1$ for $\lambda$-almost every $\theta$, a desired. □

## 5. Ergodicity and mixing

Mixing conditions formalize the intuition that observing a sequence of outcomes of a process does not change one's belief about events in the far future. Many ergodic stochastic processes that are common in economic modeling are mixing, such as i.i.d. processes and Markov processes. In this section we recall a mixing condition that was called "sufficiency for prediction" in JKS, show that the ergodic representation is not necessarily sufficient for prediction and show that a finer representation than the ergodic representation is sufficient for prediction and also weakly learnable.

Let $\overrightarrow{\mathcal{T}} = \bigwedge_{m\geq 0} \mathcal{F}_m^\infty$ be the *future tail* sigma-algebra where $\mathcal{F}_m^\infty$ the $\sigma$-algebra of $\Omega$ that is generated by $(\zeta_m, \zeta_{m+1}, \dots)$. A probability distribution (not necessarily stationary) $\nu \in \Delta(\Omega)$ is *mixing* if it is $\overrightarrow{\mathcal{T}}$-trivial, i.e., if $\nu(B) \in \{0,1\}$ for every $B \in \overrightarrow{\mathcal{T}}$.[3] The following well-known example shows that the ergodic components of a stationary process needs not be mixing.

**Example 5.1.** Let $A = \{\mathrm{B}, \mathrm{G}\}$ and let $\alpha \in [0,1]$ be irrational. Let $\xi_0, \xi_1, \dots$ be the $[0,1]$-valued stationary process of rotation by $\alpha$, so that $\xi_0$ has uniform distribution and $\xi_n = (\xi_0 + n\alpha) \bmod 1$. Let

$$\zeta_n = \begin{cases} \mathrm{G}, \text{ if } \xi_n \bmod 1 > 1/2, \\ \mathrm{B}, \text{ if } \xi_n \bmod 1 < 1/2. \end{cases}$$

The process $\zeta_0, \zeta_1, \dots$ is ergodic, so that its ergodic representation is the trivial, but every Borel set is measurable with respect to the tail. ▲

It follows from Example 5.1 that if we want the components of the representation to be mixing we need a finer representation than the ergodic representation. This representation is the representation that is induced by the tail $\overrightarrow{\mathcal{T}}$ as shown in the following proposition.

[3] An equivalent way to write this condition is that for every $n$ and $\epsilon$, there is $m$ such that

$$|\nu(B|a_0, \dots, a_{n-1}) - \nu(B)| < \epsilon$$

for every $B \in \mathcal{F}_m^\infty$ and partial history $(a_0, \dots, a_{n-1}) \in A^n$. JKS call such belief *sufficient for prediction*. They establish the equivalence with the mixing condition in their proof of their Theorem 1

**Proposition 5.2.** *Let $(\Theta, \mathcal{B}, \lambda, (\mu_\theta))$ be the representation of a belief $\mu \in \Delta(\Omega)$ that is induced by the tail $\overrightarrow{\mathcal{T}}$. Then $\mu_\theta$ is mixing for $\lambda$-almost every $\theta$.*

*Proof.* This proposition is KJS' Theorem 1. We repeat the argument here to clarify a non-trivial point in the proof.

The proposition follows from the fact that the conditional distributions of every probability distribution $\mu \in \Delta(\Omega)$ over the tail are almost surely tail-trivial (i.e., mixing). This fact was recently proved in Berti and Rigo (2007, Theorem 15)[4]. We note that it is not true for every sigma-algebra $\mathcal{D}$ that the conditional distributions of $\mu$ over $\mathcal{D}$ are almost surely $\mathcal{D}$-trivial. This property is very intuitive (and indeed, easy to prove) when $\mathcal{D}$ is generated by a finite partition, or more generally when $\mathcal{D}$ is countably generated, but the tail is not countably generated, which is why Berti and Rigo's result is required. □

Consider again Example 5.1. As we have seen, the ergodic representation in this example is the trivial representation. In contrast, the tail representation is Dirac's representation. Note that the components of the tail representation are not stationary.

We end this section by showing that Lemma 4.2 can be used to show that the tail representation is also weakly learnable. In particular, Theorem 5.3 implies that the ergodic representation does not capture all the learnable properties of a stationary process.

**Theorem 5.3.** *The tail representation of a stationary stochastic process is weakly learnable.*

*Proof.* From Lemma 4.2 it follows that the representation induced by the past tail $\overleftarrow{\mathcal{T}}$ is learnable, since the past tail is shift invariant. The theorem now follows from the fact that for every stationary belief $\mu$ over a finite set of outcomes it holds that $\overleftarrow{\mathcal{T}}_\mu = \overrightarrow{\mathcal{T}}_\mu$ where $\overleftarrow{\mathcal{T}}_\mu$ and $\overrightarrow{\mathcal{T}}_\mu$ are the completions of the left and future tails under $\mu$. The equality of the left and future tails of a stationary process is not trivial, it relies on finiteness of the set of outcomes $A$, and the proof relies on the notion of entropy (See (Weiss 2000, Section 7)). □

[4]It is taken for granted in the first sentence of KJS's proof of Their Theorem 1

We conclude with further comments on the relationship with JKS. Their main result is a characterization of the class of distributions which admit a representation which is both learnable and sufficient for prediction. They dub these processes 'asymptotically reverse mixing'. In particular, they prove that, for every asymptotically reverse mixing $\mu$, the representation of $\mu$ induced by the future tail is learnable and sufficient to prediction. In our Example 3.2, the tail representation equals the ergodic representation, and, as we have shown, is not learnable. This shows that stationary processes needs not be asymptotic reverse mixing. On the other hand, the class of asymptotically reverse mixing processes is of course larger than the class of stationary processes. For example, the Dirac atomic measure $\delta_\omega$ is asymptotically reverse mixing for every realization $\omega \in \Delta(\Omega)$.

## 6. Extensions

In this section we discuss to what extent the theorems and tools of this paper extend to a larger class of process. Our purpose is mainly to highlight the assumptions made in our framework.

6.1. **Infinite set of outcomes.** The definitions of merging and weak merging can be extended to the case in which the outcome set $A$ is a compact metric space[5]: Let $\phi$ be the Prohorov Metric over $\Delta(A)$. Say that the belief $\tilde{\mu} \in \Delta(A^{\mathbb{N}})$ *merges* to $\mu \in \Delta(A^{\mathbb{N}})$ if

$$\phi\left(\mu(\cdot|a_0,\dots,a_{n-1}),\mu(\cdot|a_0,\dots,a_{n-1})\right) \xrightarrow[n\to\infty]{} 0$$

for $\mu$-almost every realization $\omega = (a_0, a_1, \dots) \in A^{\mathbb{N}}$ and that $\tilde{\mu}$ *weakly merges* to $\mu$ if the limit holds in strong Cesaro sense. Theorem 3.1 extends to the case of infinite set $A$ of outcomes. However, Theorem 5.3 is not true for infinite set $A$ of outcomes. We used the finiteness in the proof when we used the equality of the left and right tail of the process. The following example shows the problem where $A$ is infinite:

[5]Also for the case that $A$ is a separable metric space, but then there are several possible non-equivalent definitions (**?**)

**Example 6.1.** Let $A = \{0,1\}^{\mathbb{N}}$ equipped with the standard Borel structure. Thus an element $a \in A$ is given by $a = (a[0], a[1], \dots)$ where $a[k] \in \{0,1\}$ for every $k \in \mathbb{N}$. Let $\mu$ be the belief over $A^{\mathbb{Z}}$ such that $\{\zeta_n[0]\}_{n\in\mathbb{Z}}$ are i.i.d. fair coin tosses and $\zeta_n[k] = \zeta_{n-k}[0]$ for every $k \geq 1$. Note that in this case $\overrightarrow{\mathcal{T}} = \mathcal{B}$ (so the right tail contains the entire history of the process) while $\overleftarrow{\mathcal{T}} = \mathcal{R}$ (the left tail is empty). The tail representation in this case will be Dirac's representation. However, this representation is not learnable: an agent who predict according to $\mu$ will at every day $n$ will be completely in the dark about $\zeta_{n+1}[0]$. ▲

6.2. **Relaxing stationarity.** As we have argued earlier, stationary beliefs are useful to model situations where there is nothing remarkable about the point in time in which the agent started to keep track of the processes (so that other agents, who start observing the process at different times, have the same beliefs) and where the fact that the agent observes the process has no impact on the process. The first assumption is rather strong, and can be somewhat relaxed. In particular, consider a belief that is the posterior of some stationary prior conditioned on the occurrence of some event, as in the case of an agent whose interest in the stock market is sparked by a recent crash and starts observing the return process at that point. This agent will not hold a stationary belief about the market – his belief about the first periods will be typical to what happens after a crash. A similar situation is an agent who observes a finite state markov process that starts at a given state rather than the stationary distribution. Let us say that a belief $\nu \in A^{\mathbb{N}}$ is *conditioned stationary* if $\nu = \mu(\cdot|B)$ for some Borel subset $B$ of $A^{\mathbb{N}}$ such that $\mu(B) > 0$. While such processes are not stationary, they still admits an ergodic decomposition. they exhibit the same tail behavior of stationary processes. In particular, our theorems extend to such processes. We omit the details.

## References

Al-Najjar, N. I., and E. Shmaya (2012): "Uncertainty and Disagreement in Equilibrium Models," Northwestern University.

Berti, P., and P. Rigo (2007): "0-1 Laws for Regular Conditional Distributions," *The Annals of Probability*, 35, 649–662.

Blackwell, D., and L. Dubins (1962): "Merging of opinions with increasing information," *Ann. Math. Statist.*, 33, 882–886.

Doob, J. L. (1949): "Application of the theory of martingales," *Le calcul des probabilites et ses applications*, pp. 23–27.

Jackson, M. O., E. Kalai, and R. Smorodinsky (1999): "Bayesian Representation of Stochastic Processes under Learning: de Finetti Revisited," *Econometrica*, 67, 875–893.

Kalai, E., and E. Lehrer (1993): "Rational Learning Leads to Nash Equilibrium," *Econometrica*, 61, 1019–1045.

Kalai, E., and E. Lehrer (1994): "Weak and Strong Merging of Opinions," *Journal of Mathematical Economics*, 23, 73–86.

Kalai, E., E. Lehrer, and R. Smorodinsky (1999): "Calibrated Forecasting and Merging," *Games and Economic Behavior*, 29(1), 151–159.

Kallenberg, O. (2002): *Foundations of Modern Probability*. Second edn. New York: Springer-Verlag.

Lehrer, E., and R. Smorodinsky (1996): "Compatible Measures and Merging," *Mathematics of Operations Research*, pp. 697–706.

Morvai, G., and B. Weiss (2005): "Forward estimation for ergodic time series," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 41, pp. 859–870. Elsevier.

Weiss, B. (2000): *Single Orbit Dynamics*. AMS Bookstore.

Weizsäcker, H. (1996): "Some reflections on and experiences with SPLIFs," *Lecture Notes-Monograph Series*, pp. 391–399.

Kellogg School of Management, Northwestern University

*E-mail address*: al-najjar@kellogg.northwestern.edu

School of Mathematics, Tel Aviv University and Kellogg School of Management, Northwestern University

*E-mail address*: erans@post.tau.ac.il