Proceedings of Machine Learning Research vol 65:1–15, 2017

Memoryless Sequences for Differentiable Losses

Rafael Frongillo RAF@COLORADO.EDU
CU Boulder

Andrew Nobel NOBEL@EMAIL.UNC.EDU
UNC Chapel Hill

Abstract

One way to define the "randomness" of a fixed individual sequence is to ask how hard it is to predict. When prediction error is measured via squared loss, it has been established that memoryless sequences (which are, in a precise sense, hard to predict) have some of the stochastic attributes of truly random sequences. In this paper, we ask how changing the loss function used changes the set of memoryless sequences, and in particular, the stochastic attributes they possess. We answer this question for differentiable convex loss functions using tools from property elicitation, showing that the property elicited by the loss determines the stochastic attributes of the corresponding memoryless sequences. We apply our results to price calibration in prediction markets.

Keywords: Algorithmic randomness, property elicitation, loss functions, prediction markets

1. Introduction

Since the dawn of probability theory in the 17th century, there has been interest in the randomness of fixed objects, a representative question being whether the digits of π are random in some sense. Questions of this sort lead naturally to the problem of how to define the "randomness" of a fixed individual sequence. Building on the work of Von Mises (1919) and Kolmogorov (1965), Martin-Löf (1966) introduced the notion of *algorithmic randomness*, in which a sequence is random if it passes every statistical test performed by some class of algorithms (typically Turing machines).

In this paper we focus on a notion of randomness under which a sequence is random if its entries cannot be predicted well from previous elements of the sequence. This is a natural desideratum, for example, in pseudorandomness, where one may only be concerned about a bystander's ability to guess the next bit in sequence, and not concerned with whether an algorithm having access to the entire sequence could distinguish it from a random one. Nobel (2004) calls a sequence *memoryless* if its entries cannot be predicted well by any continuous function applied to a fixed-width sliding window of previous entries.[1] It is shown that when prediction error is measured using squared loss, memoryless sequences exhibit a number of stochastic properties, including a law of large numbers and a version of the central limit theorem. These and other results follow from the fact that the weak limits of the empirical distributions of a memoryless sequence are stationary martingale difference sequences. The central role of the squared loss leads to a number of questions. How does the set of memoryless sequences depend on the loss? Does some analog of the martingale difference property hold for memoryless sequences under general losses?

1. An appealing quality of the definition of memoryless sequences is that it naturally applies to sequences of real numbers, in contrast to Turing machines which require a careful theory of computation over the reals.

This paper provides answers to these questions for convex differentiable losses, using ideas from property elicitation. We establish that, in a manner reminiscent of Blackwell approachability, the one-shot statistical attributes of the loss function alone determine which sequences are memoryless with respect to that loss. In particular, we show that the set of memoryless sequences studied by Nobel (2004) remain exactly the same when replacing squared loss by any other loss eliciting the mean (i.e. for which the mean minimizes the expected loss), namely the class of Bregman divergences (Savage, 1971; Frongillo and Kash, 2015). More generally, the property/statistic Γ elicited by the loss function uniquely determines which sequences are memoryless, and under mild assumptions, the weak limits of these sequences are "Γ-centered", in the sense that the conditional value of the statistic Γ is constant; when the loss is a Bregman divergence, Γ is the mean, and Γ-centered sequences (centered at 0) reduce to standard martingale differences. We conclude with applications to prediction markets and future work.

1.1. Related Work

The literature on the randomness of individual sequences began with Von Mises (1919), Kolmogorov (1965), and Martin-Löf (1966). In a survey of the area, Uspenskii et al. (1990) gives an account of this early work. V'yugin (1998) shows an ergodic theorem for Martin-Löf random (typical) individual sequences under certain conditions of computability. While aside from Section 4.1 our results rely on convexity of the loss, Haussler, Kivinen, and Warmuth 1998 study individual sequences for general loss functions. We refer the reader to Nobel 2004 for additional references.

Aside from this classical perspective, the literatures on no-regret online learning algorithms (Foster and Vohra, 1999; Cesa-Bianchi et al., 1999; Cesa-Bianchi and Lugosi, 2006) and game-theoretic probability (Shafer and Vovk, 2005) are both related. In particular, Vovk (2001) studies a game-theoretic setting which is similar to ours, wherein a learner attempts to predict an adaptive sequence of outcomes. The conclusion is that, much like our Lemma 9 and Theorem 10, either the learner can achieve low loss (squared or logarithmic), or the outcome sequence is "random" in the sense that the martingale law of large numbers and law of the iterated logarithm hold. A major distinction, however, is that the outcome sequence in both online learning and game-theoretic probability is allowed to adapt to the choices of the learner, which does not allow these results to apply to fixed individual sequences. See Section 6 for further discussion.

Finally, the literature on property elicitation extends that of proper scoring rules (Brier, 1950; Good, 1952; McCarthy, 1956; Savage, 1971; Gneiting and Raftery, 2007) and proper losses (Reid and Williamson, 2010; Vernet et al., 2011). The modern literature begins with Osband (1985) and Lambert et al. (2008) and continues (Lambert and Shoham, 2009; Lambert, 2011; Abernethy and Frongillo, 2012; Steinwart et al., 2014; Agarwal and Agarwal, 2015; Frongillo and Kash, 2015).

2. Setting and Definitions

We begin with key definitions and an overview of elicitation.

2.1. Spaces and loss functions

The main notation we use throughout the paper is summarized below. We will denote elements of the spaces $\mathcal{X}, \mathcal{Y}$ by x, y, and elements of the sequence spaces $\mathbb{X}, \mathbb{Y}$ by $\mathbf{x}, \mathbf{y}$.

- $\mathcal{X} \subseteq \mathbb{R}^d$ is the prediction space;

- $\mathcal{Y} \subseteq \mathbb{R}^d$ is the outcome space;
- $\mathbb{X} \subseteq \mathcal{X}^{\mathbb{N}}$ is the prediction sequence space;
- $\mathbb{Y} \subseteq \mathcal{Y}^{\mathbb{N}}$ is the outcome sequence space;
- $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the loss function.

We are interested in the task of sequentially predicting an outcome sequence $\mathbf{y} \in \mathbb{Y}$ by a sequence $\mathbf{x} \in \mathbb{X}$, with predictive performance at stage n measured by the average value of the entrywise loss. The following assumptions on the spaces and loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ will be made throughout the paper.

A1. $\mathcal{X} \subseteq \mathbb{R}^d$ is open and convex;

A2. $\mathcal{Y} \subseteq \mathbb{R}^d$ is closed;

A3. $\ell(\cdot, y)$ is convex for each fixed $y \in \mathcal{Y}$;

A4. $\ell(x, y)$ is jointly continuous in (x, y).

A5. The derivative $\nabla \ell(x, y)$ of the loss $\ell(x, y)$ with respect to its first argument exists for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and is jointly continuous.

2.2. Property elicitation and property-centered sequences

We will frame our results in the language of property elicitation, which studies losses that "elicit" a particular statistic, or *property*, of interest. The following definitions make this precise.

Definition 1 *Let $\mathcal{P}$ be the set of all probability measures on $\mathcal{Y}$. A* property *is a function $\Gamma : \mathcal{P} \to \mathcal{X}$ that associates a value to each distribution on $\mathcal{Y}$.*

A loss elicits a property if, for every distribution in $\mathcal{P}$, the expected loss is (uniquely) minimized by the value of the property for that distribution.

Definition 2 *A loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$* elicits *a property $\Gamma : \mathcal{P} \to \mathcal{X}$ if for all $p \in \mathcal{P}$, $\{\Gamma(p)\} = \operatorname{argmin}_{x \in \mathcal{X}} \mathbb{E}_p \ell(x, Y)$ where $Y \sim p$. A property is* elicitable *if it is elicited by some loss.*

The most well-known example of elicitation is the fact that squared loss $\ell(x, y) = (x-y)^2$ elicits the mean: in this case one easily checks that, if relevant expectations are well-defined, $\mathbb{E}_p \ell(x, Y) = (x - \mathbb{E}_p[Y])^2 + \mathbb{E}_p[Y]^2 - \mathbb{E}_p[Y^2]$, which is minimized at $x = \mathbb{E}_p[Y]$. Squared loss is a special case of a broader class of loss functions, called *Bregman divergences*, that measure the error of the linear approximation of a convex function.

Definition 3 *Given a differentiable convex function $G : \mathcal{Y} \to \mathbb{R}$, its associated* Bregman divergence *is the loss function*

$$\ell(\hat{y}, y) = G(y) - G(\hat{y}) - \langle \nabla G(\hat{y}), y - \hat{y} \rangle \ . \tag{1}$$

Theorem 4 (Savage (1971)) *If ℓ is a Bregman divergence and $\mathbb{E}_p |\ell(\hat{y}, Y)| < \infty$ for each $\hat{y} \in \mathcal{Y}$ and $p \in \mathcal{P}$, then ℓ elicits the mean $\Gamma : p \mapsto \mathbb{E}_p Y$.*

Proof Letting $y^* = \mathbb{E}_p[Y]$, note that $\mathbb{E}_p \ell(y^*, Y) = \mathbb{E}_p G(Y) - G(y^*)$. Expanding and simplifying, we find that $\mathbb{E}_p \ell(y^*, Y) - \mathbb{E}_p \ell(\hat{y}, Y) = G(y^*) - G(\hat{y}) - \langle \nabla G(\hat{y}), y^* - \hat{y} \rangle$, which is nonnegative as a Bregman divergence (or alternatively, by the subgradient inequality). ■

By applying a property Γ to conditional distributions, we may gain insight into the sequential prediction of a finite or infinite sequence of random variables. Of particular interest are sequences for which the optimal sequential predictor is a constant.

Definition 5 *A sequence of random vectors $Y_1, \ldots, Y_m \in \mathcal{Y}$ is* Γ-centered *if there is a (fixed) vector $c \in \mathcal{X}$ such that $c = \Gamma(Y_{k+1}|Y_1^k)$ with probability 1 for each $k = 0, \ldots, m-1$. Here $Y_{k+1}|Y_1^k$ denotes the conditional probability distribution of Y_{k+1} given $Y_1, \ldots, Y_k$, and when $k = 0$, the distribution of Y_1. The vector c will be called the* center *of sequence.*

2.3. Bounded and interior sequences

The following definitions will be used in what follows.

Definition 6 *Let $\mathbf{u} = u_1, u_2, \ldots$ be a sequence with values in $\mathbb{R}^d$. The sequence $\mathbf{u}$ is* bounded *if there exists a finite constant L such that $||u_i|| \leq L$ for all $i \geq 1$. The* closure $\mathrm{cl}(\mathbf{u})$ *of $\mathbf{u}$ is the (ordinary) closure in $\mathbb{R}^d$ of the countable set $\{u_1, u_2, \ldots\}$. We will say that $\mathbf{u}$ is* interior to *an open set $U \subseteq \mathbb{R}^d$ if the closure of $\mathbf{u}$ is contained in U.*

Definition 7 *Given a subset U of a vector space, its* star interior *is given by*

$$\mathrm{starint}(U) = \{u \in U : \forall v \in U \; \exists \alpha_0 > 0 \; \forall \alpha \in [-\alpha_0, \alpha_0], \; u + \alpha(v - u) \in U\}.$$

Note that the star interior of U is a subset of the relative interior of U.

2.4. Memoryless sequences

Given a sequence of vectors $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$, let y_i denote the ith element of $\mathbf{y}$, and let $y_i^j = y_i, y_{i+1}, \ldots, y_j$ when $i \leq j$. For each $k \geq 1$ let $\mathcal{C}_k = C_b(\mathcal{Y}^k : \mathcal{X})$ be the family of bounded continuous functions $g : \mathcal{Y}^k \to \mathcal{X}$, and let $\mathcal{C}_0$ be the family of constant functions on $\mathcal{Y}$ with values in $\mathcal{X}$.

Definition 8 *A sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ is* memoryless under ℓ *if there exists a vector $c \in \mathcal{X}$ such that for every $k \geq 0$ and every function $g \in \mathcal{C}_k$*

$$\liminf_{n \to \infty} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(g(y_{i-k}^{i-1}), y_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(c, y_i) \right] \; \geq \; 0 \tag{2}$$

where we adopt the convention that $g(y_{i-k}^{i-1}) = 0$ for $i \leq k$. In this case we will say that $\mathbf{y}$ is memoryless under ℓ with respect to c. *Note that neither average is assumed to converge.*

Each function $g \in \mathcal{C}_k$ represents a continuous, kth order Markov prediction scheme for $\mathbf{y}$ that predicts the next value in the sequence by applying g to the k previous values. A sequence is memoryless if there is a constant prediction scheme that is as good as any continuous Markov

prediction scheme of finite order. The constant prediction scheme in the definition ignores the past and always predicts the next value of $\mathbf{y}$ by c.

Note that memorylessness is by definition an asymptotic notion. For example, the "learner" g could correctly predict the first N elements of $\mathbf{y}$ correctly, and yet the sequence could still be memoryless. In fact, padding any memoryless sequence by N initial 0's would preserve its memorylessness. In this sense, our notion of randomness (memorylessness) is relatively weak, in a manner analogous to "no-regret" in online learning: just as online learning algorithms can produce arbitrary outputs for any initial block of time and still achieve no regret, here a learner can perform well for a finite amount of time and still fail to predict the sequence $\mathbf{y}$ in an asymptotic sense.

2.5. Weak convergence

Several of the key results and proofs in this paper rely on the notion of weak convergence of probability measures, which we briefly review here. A succinct treatment of weak convergence can be found in Chapter 2 of van der Vaart (2000). A sequence $\{\nu_n : n \geq 1\}$ of probability measures on $\mathbb{R}^p$ is said to *converge weakly* to a limiting probability measure ν, written $\nu_n \Rightarrow \nu$, if $\int f \, d\nu_n \to \int f \, d\nu$ for every bounded continuous function $f : \mathbb{R}^p \to \mathbb{R}$. A sequence $\{\nu_n : n \geq 1\}$ of probability measures on $\mathbb{R}^p$ is *tight* if for every $\epsilon > 0$ there exists a compact set $K \subseteq \mathbb{R}^p$ such that $\mu_n(K^c) < \epsilon$ for each $n \geq 1$. Thus any sequence of measures supported on a common compact set is tight. Prokhorov's theorem states that if $\{\mu_n\}$ is tight then any subsequence $\{\mu_{n_k}\}$ has a further subsequence $\{\mu_{m_k}\}$ that converges weakly to a limiting measure.

3. An Orthogonality Condition

A critical ingredient in our analysis, and one that is interesting in its own right, is the following orthogonality lemma. Roughly speaking the lemma shows that, for a given outcome sequence $\mathbf{y} \in \mathbb{Y}$, a sequence $\mathbf{x}^*$ in $\mathbb{X}$ is optimal under the average loss if and only if for every $\mathbf{x}$ in $\mathbb{X}$ the difference $\mathbf{x} - \mathbf{x}^*$ is orthogonal, in an appropriate sense, to the gradients of $\ell(\cdot, y_i)$ at x_i^*. One may view this result as a kind of first-order optimality condition in which equation (4) acts as the "derivative" of eq. (3).

Lemma 9 (General loss, differentiable case) *Let $\mathcal{X}, \mathcal{Y}$ be subsets of $\mathbb{R}^d$ satisfying assumptions A1 and A2, and let $\ell()$ be a loss function satisfying assumptions A3-A5. Let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be a bounded sequence, and let $\mathbb{X} \subseteq \mathcal{X}^{\mathbb{N}}$ be a family of bounded sequences $\mathbf{x}$ that are interior to $\mathcal{X}$. Then for all $\mathbf{x}^* \in \mathrm{starint}(\mathbb{X})$, the following two statements are equivalent:*

$$\liminf_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^*, y_i) \right] \;\geq\; 0 \;\; \textit{for all} \;\; \mathbf{x} \in \mathbb{X} \tag{3}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \rangle \;=\; 0 \;\; \textit{for all} \;\; \mathbf{x} \in \mathbb{X} \, . \tag{4}$$

Proof Let $\mathbf{x}^* \in \mathrm{starint}(\mathbb{X})$ be fixed. To show that (4) implies (3), we make use convexity of ℓ in its first coordinate. By the subgradient inequality

$$\ell(x_i, y_i) - \ell(x_i^*, y_i) \;\geq\; \langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \rangle \, .$$

Summing over $1 \leq i \leq n$, dividing by n, and taking the limit infimimum, inequality (3) follows from (4).

To establish the converse, suppose that (4) fails to hold. Then there exists a sequence $\mathbf{x} \in \mathbb{X}$ and $\delta > 0$ such that the average in (4) has limit infimum less than $-\delta$ or limit supremum greater than δ. We consider the former case; the argument for the latter is similar. For each $\alpha \in \mathbb{R}$ define $\mathbf{x}^\alpha$ by $x_i^\alpha = x_i^* + \alpha(x_i - x_i^*) = \alpha x_i + (1-\alpha)x_i^*$. As $\mathbf{x}^*$ is in the star interior of $\mathbb{X}$ by assumption, there exists $0 < \alpha_0 \leq 1$ such that $\mathbf{x}^\alpha \in \mathbb{X}$ for all $\alpha \in [0, \alpha_0]$. Note that for each such α and each $i \geq 1$,

$$\ell(x_i^\alpha, y_i) - \ell(x_i^*, y_i) \;=\; G_\alpha(y_i, x_i, x_i^*) \;+\; \alpha \left\langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \right\rangle \tag{5}$$

where $G_\alpha : \mathcal{Y} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by

$$G_\alpha(u, v, w) \;=\; \ell(w + \alpha(v-w), u) \;-\; \ell(w, u) \;-\; \alpha \left\langle v - w, \nabla \ell(w, u) \right\rangle \tag{6}$$

and is equal to the Bregman divergence of $\ell(\cdot, u)$ evaluated at $w + \alpha(v-w)$ and w. In particular, G_α is non-negative.

Define the set $K = \mathrm{cl}(\mathbf{y}) \times \mathrm{cl}(\mathbf{x}) \times \mathrm{cl}(\mathbf{x}^*)$. As $\mathcal{Y}$ is closed and $\mathbf{x}, \mathbf{x}^*$ are interior to $\mathcal{X}$ by assumption, K is a subset of $\mathcal{Y} \times \mathcal{X} \times \mathcal{X}$. Moreover, the boundedness of $\mathbf{y}, \mathbf{x}, \mathbf{x}^*$ implies that K is a compact subset of $(\mathbb{R}^d)^3$. Our assumptions on $\ell()$ ensure that G_α is continuous and bounded on K. For $n \geq 1$ let $\nu_n(\cdot) = n^{-1} \sum_{i=1}^n \mathbb{I}((y_i, x_i, x_i^*) \in \cdot)$ be the empirical measure on K of the finite sequence of triples $(y_1, x_1, x_1^*), \ldots, (y_n, x_n, x_n^*)$. By assumption, there is a subsequence $\{n_l\}$ of the positive integers such that

$$\lim_{l \to \infty} \frac{1}{n_l} \sum_{i=1}^{n_l} \left\langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \right\rangle \;\leq\; -\delta. \tag{7}$$

As K is compact the sequence $\{\nu_n\}$ is tight, so there is a subsequence $\{n_k\}$ of $\{n_l\}$ such that ν_{n_k} converges weakly to some probability measure ν on K. Using equation (5), we find that for each $0 < \alpha < \alpha_0$,

$$\begin{aligned}
&\liminf_{n \to \infty} \left[\frac{1}{n} \sum_{i=1}^n \ell(x_i^\alpha, y_i) - \frac{1}{n} \sum_{i=1}^n \ell(x_i^*, y_i) \right] \\
&\quad \leq \liminf_{k \to \infty} \left[\frac{1}{n_k} \sum_{i=1}^{n_k} \ell(x_i^\alpha, y_i) - \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(x_i^*, y_i) \right] \\
&\quad = \liminf_{k \to \infty} \left[\frac{1}{n_k} \sum_{i=1}^{n_k} G_\alpha(y_i, x_i, x_i^*) + \frac{\alpha}{n_k} \sum_{i=1}^{n_k} \left\langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \right\rangle \right] \\
&\quad = \liminf_{k \to \infty} \left[\int G_\alpha \, d\nu_{n_k} + \frac{\alpha}{n_k} \sum_{i=1}^{n_k} \left\langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \right\rangle \right] \\
&\quad = \int G_\alpha \, d\nu \;+\; \alpha \liminf_{k \to \infty} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \right\rangle \\
&\quad \leq \int G_\alpha \, d\nu \;-\; \alpha \, \delta.
\end{aligned}$$

The last equality above follows from the weak convergence of ν_{n_k} to ν, and the final inequality follows from (7).

It suffices to show that the final term of the last inequality above is negative for some $\alpha > 0$, and for this it is enough to show that $\int G_\alpha \, d\nu = o(\alpha)$. Note that for each triple (u, v, w) in K the existence of the gradient $\nabla \ell(w, u)$ implies that $\alpha^{-1} G_\alpha(u, v, w) \to 0$ as $\alpha \to 0$. We show that the functions G_α for $0 < \alpha < \alpha_0$ are dominated by a constant function, and the result then follows from the dominated convergence theorem. To this end, let M be the maximum of the continuous function $||\nabla \ell(w, u)||$ over u in $\mathrm{cl}(\mathbf{y})$ and w in the convex hull W of $\mathrm{cl}(\mathbf{x}) \cup \mathrm{cl}(\mathbf{x}^*)$, which is a compact subset of $\mathcal{X}$. By standard results (Shalev-Shwartz, 2012, Lem 2.6), $|\ell(w_1, u) - \ell(w_2, u)| \leq M \, ||w_1 - w_2||$ for all $u \in \mathrm{cl}(\mathbf{y})$ and $w_1, w_2 \in W$. From this bound and the Cauchy-Schwartz inequality, we find that for all $(u, v, w) \in K$

$$\begin{aligned} \frac{|G_\alpha(u, v, w)|}{\alpha} &\leq \frac{|\ell(w + \alpha(v - w), u) \; - \; \ell(w, u)|}{||\alpha(v - w)||} \cdot ||v - w|| \; + \; |\, \langle v - w, \nabla \ell(w, u) \rangle \,| \\ &\leq \; 2M \, ||v - w|| \; \leq \; 2MD \end{aligned}$$

where D is the diameter of the set $\mathrm{cl}(\mathbf{x}) - \mathrm{cl}(\mathbf{x}^*)$. As $\mathbf{x}$ and $\mathbf{x}^*$ are bounded, D is finite, and the proof is complete. ■

In the next section, Lemma 9 is used to relate memoryless individual sequences to Γ-centered sequences of random variables.

4. Memoryless and Property-Centered Sequences

We now turn to our principal result, which establishes a close connection between memoryless individual sequences and Γ-centered stochastic sequences. Roughly speaking, we show that a sequence is memoryless under a loss ℓ if and only if its limiting empirical distributions are Γ-centered, where Γ is a property elicited by ℓ.

To every measure ν on $\mathcal{Y}^m$ there correspond random vectors $Y_1, \ldots, Y_m \in \mathcal{Y}$, defined on a common probability space and having ν as their joint distribution. Thus we will write $\nu_n \Rightarrow \nu$ equivalently as $\nu_n \Rightarrow (Y_1, \ldots, Y_m)$ where $(Y_1, \ldots, Y_m) \sim \nu$. For $n, m \geq 1$ define the n-sample, m-dimensional empirical measure of $\mathbf{y}$ by

$$\mu_{n,m}(A) \; = \; \frac{1}{n} \sum_{i=0}^{n-1} I\{(y_{i+1}, \ldots, y_{i+m}) \in A\} \tag{8}$$

for all Borel measurable sets $A \subseteq \mathcal{Y}^m$. Note that if $\mathbf{y}$ is bounded then for each m the empirical measures $\{\mu_{n,m} : n \geq 1\}$ are tight, and in particular, every subsequence $\{\mu_{n_l,m}\}$ has a further subsequence that converges weakly to a limiting measure, or equivalently, a jointly distributed sequence $Y_1, \ldots, Y_m \in \mathcal{Y}$. Following standard terminology, we will say that $Y_1, \ldots, Y_m$ is stationary if for each $s, j \geq 1$ with $s + j \leq m$ the sequence $(Y_s, \ldots, Y_{s+j})$ has the same joint distribution as $(Y_1, \ldots, Y_{j+1})$

Theorem 10 *Let $\mathcal{X}, \mathcal{Y}$ be subsets of $\mathbb{R}^d$ satisfying assumptions A1 and A2, and let $\ell()$ be a loss function satisfying assumptions A3-A5 that elicits property $\Gamma : \mathcal{P} \to \mathcal{X}$. Let $c \in \mathcal{X}$ and let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be bounded. The following are equivalent:*

(i) The sequence $\mathbf{y}$ *is memoryless under* ℓ *with respect to* c*;*

(ii) For each $m \geq 1$ *every weak limit* $(Y_1, \ldots, Y_m)$ *of the* m*-dimensional empirical measures* $\{\mu_{n,m} : n \geq 1\}$ *of* $\mathbf{y}$ *is stationary, bounded, and* Γ*-centered with center* c*.*

Proof Let $\mathbf{y}$ be a bounded sequence with values in $\mathcal{Y}$. Suppose that for some $m \geq 2$ and some subsequence $\{n_l\}$ of the positive integers $\mu_{n_l,m} \Rightarrow (Y_1, \ldots, Y_m)$ as $l \to \infty$, where $\mu_{n_l,m}$ are defined as in (8). Then for each $s, j \geq 1$ with $s + j \leq m$, and every bounded continuous function $f : \mathcal{Y}^{j+1} \to \mathbb{R}$,

$$\begin{aligned} \mathbb{E}g(Y_s, \ldots, Y_{s+j}) &= \lim_{l\to\infty} \frac{1}{n_l} \sum_{i=0}^{n_l-1} g(y_{i+s}, \ldots, y_{i+s+j}) \\ &= \lim_{l\to\infty} \frac{1}{n_l} \sum_{i=0}^{n_l-1} g(y_{i+1}, \ldots, y_{i+j+1}) \;=\; \mathbb{E}g(Y_1, \ldots, Y_{j+1}). \end{aligned}$$

It follows that $(Y_s, \ldots, Y_{s+j})$ has the same joint distribution as $(Y_1, \ldots, Y_{j+1})$, and as this is true for each choice of s, j above, the sequence $Y_1, \ldots, Y_m$ is stationary. By the Portmanteau theorem (see, for example, Lemma 2.2 of Vaart (2000))

$$P(Y_k \in \mathrm{cl}(\mathbf{y})) \;\geq\; \limsup_{l\to\infty} \frac{1}{n_l} \sum_{i=0}^{n_l-1} \mathbb{I}(y_i \in \mathrm{cl}(\mathbf{y})) \;=\; 1,$$

and since $\mathrm{cl}(\mathbf{y})$ is bounded by assumption, each of the random variables Y_k is bounded as well.

Suppose that $\mathbf{y}$ satisfies condition (i) of the theorem. As $\mathcal{X}$ is open, there exists $\delta > 0$ such that $B(c, 2\delta) \subseteq \mathcal{X}$ where $B(c, \gamma) = \{x : ||c - x|| < \gamma\}$ is the open ball of radius γ centered at c. Let $\mathbb{X}$ be the set of all infinite sequences $\mathbf{x} = x_1, x_2, \ldots \in \mathcal{X}$ such that, for some $k \geq 0$ and some continuous function $g : \mathcal{Y}^k \to B(c, \delta)$, $x_1 = \cdots = x_k = c$ and $x_i = g(y_{i-k}^{i-1})$ for $i \geq k + 1$. One may easily verify that the conditions of Lemma 9 are satisfied with $\mathbf{x}^*$ identically equal to c, and therefore

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \langle x_i - c, \nabla\ell(c, y_i)\rangle \;=\; 0 \;\text{ for all }\; \mathbf{x} \in \mathbb{X}. \tag{9}$$

Let $1 \leq k < m$ and let $g : \mathcal{Y}^k \to B(c, \delta)$ be any continuous function. Then the function $f : \mathcal{Y}^{k+1} \to \mathbb{R}$ defined by $f(u_1^{k+1}) = \langle g(u_1^k) - c, \nabla\ell(c, u_{k+1})\rangle$ is continuous and is bounded on the compact set $\mathrm{cl}(\mathbf{y})^{k+1}$ supporting $(Y_1, \ldots, Y_{k+1})$. By appropriate choice of $\mathbf{x} \in \mathbb{X}$, the relation (9) implies that

$$\mathbb{E}\left\langle g(Y_1^k) - c, \nabla\ell(c, Y_{k+1})\right\rangle \;=\; \lim_{l\to\infty} \int f \, d\mu_{n_l,m} \;=\; \lim_{l\to\infty} \frac{1}{n_l} \sum_{i=0}^{n_l-1} f(y_{i+1}^{i+k+1}) \;=\; 0.$$

As the function $g : \mathcal{Y}^k \to B(c, \delta)$ was arbitrary, a routine argument shows that $\mathbb{E}[\nabla\ell(c, Y_{k+1}) \,|\, Y_1^k] = 0$. Now fix $x \in \mathcal{X}$. As $\ell(u, v)$ is convex in its first argument, $\ell(x, y) - \ell(c, y) \geq \langle y - c, \nabla\ell(c, y)\rangle$. Replacing y by Y_{k+1} and taking the conditional expectation with respect to Y_1^k, we find that $\mathbb{E}[\,\ell(x, Y_{k+1}) \,|\, Y_1^k\,] - \mathbb{E}[\,\ell(c, Y_{k+1}) \,|\, Y_1^k\,] \geq 0$ with probability 1. As ℓ elicits Γ, this implies that $c = \Gamma(\, Y_{k+1} \,|\, Y_1^k\,)$ with probability 1, and therefore condition (ii) is satisfied.

Suppose now that $\mathbf{y}$ fails to satisfy (i). It follows from Lemma 9 (or the subgradient inequality) that there exists $k \geq 0$, $g \in \mathcal{C}_k$, and a subsequence $\{n_r\}$ of the positive integers such that

$$\lim_{r\to\infty} \frac{1}{n_r} \sum_{i=1}^{n_r} \big\langle g(y_{i-k}^{i-1}) - c, \nabla\ell(c, y_i)\big\rangle \;<\; 0. \tag{10}$$

Let $\{n_l\}$ be a further subsequence of $\{n_r\}$ such that $\mu_{n_l,k+1}$ converges in law to a sequence $(Y_1, \ldots, Y_{k+1})$. It follows from (10) that $\mathbb{E}\big\langle g(Y_1^k) - c, \nabla\ell(c, Y_{k+1})\big\rangle$ is non-zero, and therefore the conditional expectation $\mathbb{E}[\nabla\ell(c, Y_{k+1}) \,|\, Y_1^k]$ is non-zero with positive probability. Thus there exists $\gamma \in \mathbb{R}^d$ and $\delta > 0$ such that

$$\mathbb{E}[\,\langle\gamma, \nabla\ell(c, Y_{k+1})\rangle \,|\, Y_1^k\,] \;=\; \Big\langle \gamma, \mathbb{E}[\,\nabla\ell(c, Y_{k+1}) \,|\, Y_1^k\,]\Big\rangle \;=\; -\delta.$$

Our assumptions on $\ell(\cdot,\cdot)$ ensure that for each compact set $K \subseteq \mathcal{Y}$ the supremum

$$\sup_{y\in K} |\ell(x, y) - \ell(c, y) - \nabla\ell(c, y)(x - c)| = o(||x - c||)$$

as $x \to c$. Replacing y by Y_{k+1} and x by $x_\alpha = \alpha\gamma + c$, it follows from the previous two displays that

$$\mathbb{E}[\,\ell(x_\alpha, Y_{k+1}) \,|\, Y_1^k\,] \;=\; \mathbb{E}[\,\ell(c, Y_{k+1}) \,|\, Y_1^k\,] \;-\; \alpha\delta \;+\; o(\alpha).$$

Thus for $\alpha > 0$ sufficiently small, we find that $\mathbb{E}[\,\ell(x_\alpha, Y_{k+1}) \,|\, Y_1^k\,] < \mathbb{E}[\,\ell(c, Y_{k+1}) \,|\, Y_1^k\,]$. As ℓ elicits Γ, condition (ii) fails to hold, and the proof is complete. ■

Theorem 10 shows that memoryless sequences under a loss ℓ are characterized by the property that ℓ elicits. As a consequence, two losses eliciting the same property have the same family of memoryless sequences.

Corollary 11 *Let $\mathcal{X}, \mathcal{Y}$ be subsets of $\mathbb{R}^d$ satisfying assumptions A1 and A2, and let $\ell()$ and $\ell'()$ be a loss functions satisfying assumptions A3-A5. If $\ell()$ and $\ell'()$ elicit the same property $\Gamma : \mathcal{P} \to \mathcal{X}$, then a bounded sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ is memoryless under ℓ with respect to a vector $c \in \mathcal{X}$ if and only if it is memoryless under ℓ' with respect to c.*

The characterization of memoryless sequences in Theorem 10 suggests that the sample paths of a sequence of Γ-centered random variables should be memoryless under an eliciting loss ℓ. This is the conclusion of the following result, which we state without proof.

Proposition 12 *Let $\mathcal{X}$, $\mathcal{Y}$, and ℓ be as in Theorem 10, and suppose that ℓ elicits the property Γ. Let $\mathbf{Y} = Y_1, Y_2, \ldots$ be a Γ-centered sequence of random vectors defined on a common probability space, and taking values in a fixed, compact subset of $\mathcal{Y}$. Then with probability one $\mathbf{Y}$ is memoryless under ℓ.*

In some cases, memoryless sequences exhibit asymptotic behavior similar to that of random sequences. Suppose for the moment that $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$ and that ℓ is a Bregman divergence satisfying the conditions of Theorem 4. If $\mathbf{y}$ is memoryless under ℓ with respect to 0 then it follows from Theorems 4 and 10 that for each $m \geq 1$ every weak limit $(Y_1, \ldots, Y_m)$ of the m-dimensional empirical measures of $\mathbf{y}$ is a stationary bounded martingale difference sequence. Using this fact, one may show that $\mathbf{y}$ obeys an elementary law of large numbers, namely $n^{-1}\sum_{i=1}^n y_i \to 0$ as n tends to infinity, and a sliding-block central limit theorem. Details and further discussion can be found in Nobel (2004).

4.1. Special case: (mixed) Bregman divergences

Combining Corollary 11 and Theorem 4, we can now see that the ℓ-memoryless sequences, with constant $c = 0$, for any *convex* Bregman divergence ℓ are those whose weak limits form martingale difference sequences, thus showing that the results of Nobel (2004) generalize to a wide class of losses. (Standard martingale differences arise when $\mathcal{Y} \subseteq \mathbb{R}$, and are multivariate martingale differences otherwise.) However, the convexity conditions on ℓ are somewhat restrictive: while Bregman divergences $\ell_{\mathcal{Y}}(\hat{y}, y) = G(y) - G(\hat{y}) - \langle \nabla G(\hat{y}), y - \hat{y} \rangle$ are always convex in y, they are generally non-convex in $\hat{y}$. To remedy this situation, we will work with the class of *mixed* Bregman divergences, where we replace $G(\hat{y}) - \langle \nabla G(\hat{y}), \hat{y} \rangle$ by the convex conjugate $F = G^*$ of G, as these losses are always convex in the first argument.

Definition 13 *Let $\mathcal{Y}$ be a convex subset of $\mathbb{R}^d$. Letting $F(x) := \sup_{y \in \mathcal{Y}} \langle x, y \rangle - G(y)$ be the convex conjugate of G, let $\mathcal{X} := \mathrm{dom} F = \{x \in \mathbb{R}^d : F(x) < \infty\}$. The associated* mixed (Bregman) divergence *is the loss $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given by*

$$\ell_{\mathcal{X}}(x, y) = F(x) + G(y) - \langle x, y \rangle \ . \tag{11}$$

The loss $\ell_{\mathcal{X}}(x, y)$ is convex in x, and is closely related to $\ell_{\mathcal{Y}}$; as we now argue, their corresponding memoryless sequences are also closely related. If G is strictly convex and differentiable, and is a closed convex function, then F is also differentiable, and the gradient maps ∇G and ∇F have continuous inverses (and are therefore homeomorphisms). Consider a continuous Markov predictor $g : \mathcal{Y}^k \to \mathcal{Y}$ for $\ell_{\mathcal{Y}}$; as ∇G is a continuous map, $(\nabla G \circ g) : \mathcal{Y}^k \to \mathcal{X}$ is a continuous Markov predictior for $\ell_{\mathcal{X}}$. Moreover, as ∇G is surjective, every continuous Markov predictor for $\ell_{\mathcal{X}}$ can be written this way. Conversely, every continuous Markov predictor for $\ell_{\mathcal{Y}}$ takes the form $(\nabla F \circ g)$ for continuous $g : \mathcal{Y}^k \to \mathcal{X}$. Thus, as continuous functions preserve compactness, a bounded sequence $\mathbf{y}$ is $\ell_{\mathcal{Y}}$-memoryless with respect to $c \in \mathcal{Y}$ if and only if $\mathbf{y}$ is $\ell_{\mathcal{X}}$-memoryless with respect to $\nabla G(c) \in \mathcal{X}$. (This argument extends to any such homeomorphism.)

Putting the above together, we can broaden the scope of Theorem 10 to include any Bregman divergence $\ell_{\mathcal{Y}}(\hat{y}, y)$ defined by a strictly convex and differentiable G, even if ℓ is not convex in $\hat{y}$. In particular, for a bounded sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$, we have the following:

$$\begin{aligned}
& \mathbf{y} \text{ is } \ell_{\mathcal{Y}}\text{-memoryless with respect to } c \in \mathbb{R}^d \\
\iff\ & \mathbf{y} \text{ is } \ell_{\mathcal{X}}\text{-memoryless with respect to } \nabla G(c) \\
\iff\ & \text{the weak limits of } \mathbf{y} \text{ are } \Gamma_{\mathcal{X}}\text{-centered where } \Gamma_{\mathcal{X}}(p) = \nabla G(\mathbb{E}_p Y), \text{ with center } \nabla G(c) \\
\iff\ & \text{the weak limits of } \mathbf{y} \text{ are } \Gamma_{\mathcal{Y}}\text{-centered where } \Gamma_{\mathcal{Y}}(p) = \mathbb{E}_p Y, \text{ with center } c.
\end{aligned}$$

Thus, the conclusion of Theorem 10 extends to any Bregman divergence $\ell_{\mathcal{Y}}$ generated by a strictly convex, closed, differentiable $G : \mathcal{Y} \to \mathbb{R}$. In particular, $\mathbf{y}$ is $\ell_{\mathcal{Y}}$-memoryless with respect to $0 \in \mathbb{R}^d$ if and only if the weak limits of $\mathbf{y}$ form martingale difference sequences.

5. Application to Prediction Markets

We now apply our results in the setting of prediction markets, which are markets designed to elicit and aggregate predictions from traders about some future outcome Z in the arena of athletics, finance, entertainment, or politics. Our conclusion will be that, under the "efficient market hypothesis", the outcomes are memoryless with respect to the market prices, which further endows them with stochastic properties. Prediction markets work by offering financial contracts whose payoffs

are contingent in some way on the eventually-observed value of Z; by revealed preference, the choices of traders in such a market can be interpreted as predictions about Z, and the final prices of the market can be viewed as an aggregate, or "consensus" belief of the traders (Hanson, 2003; Wolfers and Zitzewitz, 2004).

Formally, the setting in a prediction market is as follows. The set $\mathcal{Z}$ will represent the possible outcomes, and thus the possible values of Z. The market will support the buying and selling of d different *securities*, whose payoff values are each contingent on which outcome $z \in \mathcal{Z}$ materializes. In particular, we define the payoffs of these securities by a function $\phi : \mathcal{Z} \to \mathbb{R}^d$, where $\phi_i(z)$ denotes the payoff of security i upon outcome z. The prediction space $\mathcal{X} = \mathbb{R}^d$ will represent vectors of *shares* in these k securities, which traders can buy and sell. Thus, if a trader holds a bundle of shares $r \in \mathcal{X}$ and outcome $z \in \mathcal{Z}$ materializes, then the trader is owed $\langle r, \phi(z) \rangle$. Intuitively, a risk-neutral trader (i.e. one who seeks to maximize expected payoff) who is willing to buy a bundle r for a cost of c reveals a belief $\langle r, \mathbb{E}_p \phi(Z) \rangle > c$, that is, the trader must believe the expected value of $\langle r, \phi(Z) \rangle$ to be greater than c. In this way, the market prices are thought to reflect the "consensus" belief about the expected value of the securities ϕ. As an important special case when $|\mathcal{Z}| = d$, a *complete* market has $\phi_i(z) = \mathbb{1}\{z = z_i\}$ be the indicator for each element $z_i \in \mathcal{Z}$, and this case the expected value of $\phi(Z)$ is simply the distribution $p \in \Delta(\mathcal{Z})$.

Due to thin market problems, it is common to employ an *automated market maker* framework, which is simply a central entity in the market through which all transactions must be made. A popular mechanism to determine the cost of each purchase is the *cost-function-based market*, introduced by Abernethy, Chen, and Wortman Vaughan (2013). Here the cost of purchasing at time $t \in \mathbb{N}$ a bundle $r_t \in \mathcal{X}$ of shares is given by $C(x_{t-1} + r_t) - C(x_{t-1})$, where $x_{t-1} \in \mathcal{X}$ is the vector describing the total number of shares bought and sold of each security up to time $t-1$, i.e. $x_{t-1} = \sum_{i=1}^{t-1} r_i$. This procedure is described more formally in Algorithm 1.

Typically one regards the gradient $\nabla C(x)$ of C at the current market state x as the market "price". The reason is that ∇C corresponds to the instantaneous prices of the securities: $\partial C(x)/\partial x_i$ is the price per unit of an infinitessimal quantity of security i. One can check that if a trader believes the outcome to be drawn from some distribution $p \in \Delta(\mathcal{Z})$, then monotonicity of ∇C implies that a risk-neutral trader would have an incentive to buy or sell shares until the market state satisfied $\nabla C(x) = \mathbb{E}_p \phi(Z)$, as discussed above. In this sense, the market is giving traders incentives to predict the value of $\mathbb{E}_p \phi(Z)$. For the market to satisfy standard axioms, C must be strictly convex and differentiable, and $\nabla C(\mathbb{R}^d)$ should be equal to $\mathcal{Y} := \text{relint}(\text{conv}(\phi(\mathcal{Z})))$, the relative interior of the convex hull of the security payoffs (Abernethy et al., 2013). This is equivalent to the existence of a differentiable and strictly convex function $G : \mathcal{Y} \to \mathbb{R}$ with $\nabla G(\mathcal{Y}) = \mathbb{R}^d$ and such that $C(x) = \sup_{\mu \in \mathcal{Y}} \langle \mu, x \rangle - G(\mu)$ (Abernethy et al., 2013; Frongillo and Waggoner, 2017).

Now consider running this market, from initalization to the final outcome revelation, many times for many events. One can ask, were the final market prices "correct" in each market, in the sense that the "probability distribution" of the outcome really matched the price? (As mentioned above, this would mean $\nabla C(x) = \mathbb{E}_p \phi(Z)$.) In attempting to answer this question one quickly approaches deep philosophical waters, about the nature of probability and whether or not true randomness exists. A convenient way out of these waters is to appeal to properties of individual sequences as we do in this paper. To do so, we will need to translate the cost-function-based market setting to our own.

To capture this prediction market setting, let $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{Y} = \{\phi(z) : z \in \mathcal{Z}\} \subseteq \mathbb{R}^d$ be the possible security payoffs. Our loss function will take the form of the mixed Bregman divergence

Market maker initializes state $x_0 \leftarrow 0 \in \mathbb{R}^d$
for *all traders* $t = 1, \ldots, T$ **do**
  Trader t decides to purchase bundle $r_t \in \mathbb{R}^d$
  Market maker updates the state $x_t \leftarrow x_{t-1} + r_t$
  Trader pays the market maker $C(x_t) - C(x_{t-1})$
end
Outcome $z \in \mathcal{Z}$ is revealed and market maker pays $\langle r_t, \phi(z) \rangle$ to trader $t = 1, \ldots, T$

Algorithm 1: The cost-function-based market maker

$\ell(x, y) = C(x) + G(y) - \langle x, y \rangle$, which satisfies assumptions A3-A5 as C is convex and C and G are continuous by assumption. Note that if the current market state is x^*, and a trader moves the state to $x = x^* + r$ by purchasing bundle r, then $\ell(x, y) - \ell(x^*, y) = C(x^* + r) - C(x^*) - \langle r, y \rangle$, which is precisely the net loss of the trader in Algorithm 1, namely the up front cost of bundle r, minus the eventual payoff of the securities $y = \phi(z)$. Now translating Lemma 9, fixing an outcome sequence $\mathbf{z} \in \mathcal{Z}^{\mathbb{N}}$ and set of sequences $\mathbb{X} \subseteq \mathcal{X}^{\mathbb{N}}$ to which the initial market states $\mathbf{x}^*$ belong, we have

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(C(x_i^* + r_i) - C(x_i^*) - \langle r_i, \phi(z_i) \rangle \right) \geq 0 \quad \text{for all} \quad \mathbf{x}^* + \mathbf{r} \in \mathbb{X} \tag{12}$$

$$\iff \quad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \langle r_i, \nabla C(x_i^*) - \phi(z_i) \rangle = 0 \quad \text{for all} \quad \mathbf{x}^* + \mathbf{r} \in \mathbb{X}\,, \tag{13}$$

where now i denotes the run of the market, so that x_i^*, r_i, and z_i represent, respectively, the initial market state, trader's purchase, and outcome in the ith run of the market. Thus, one can interpret the application of Lemma 9 to prediction markets as follows: either the market prices are "calibrated" with respect to the class of trading algorithms whose outputs belong to $\mathbb{X}$, in the sense of eq. (13), or some sequence of trades in $\mathbb{X}$ can make an infinite profit over the course of these market runs. For example, if $\mathbb{X}$ contains all constant sequences, and $\mathbf{x}^*$ is constant, eq. (13) implies a version of the law of large numbers in that the average security payoff must approach the initial market price.

Turning now to Theorem 10, we can say something stronger. Note that the loss ℓ elicits the property $\Gamma(p) = \nabla G(\mathbb{E}_p \phi(Z))$, which is the share value whose price matches the expected security payoffs: $\nabla C(\Gamma(p)) = \mathbb{E}_p \phi(Z)$. From the discussion in Section 4.1, we find that for any sequence $\mathbf{z} \in \mathcal{Z}^{\mathbb{N}}$, no continuous finite-memory trading strategy can garner infinite profits from a series of markets initialized at c if and only if the weak limits of the security payoff sequence $\mathbf{y} = \phi(\mathbf{z})$ are Γ-centered at the initial price $\nabla C(c)$. In particular, if $0 \in \mathcal{Y}$ then initializing the market prices at 0, i.e. letting $c = \nabla G(0)$, would imply that the weak limits of the security payoffs form a martingale difference sequence.

Finally, we note that the class of finite-memory trading algorithms is perhaps restrictive in this setting; ideally, we would allow our trading algorithms to use the entire past history of prices and outcomes. This immediately becomes problematic, however, as it is difficult to exclude algorithms that "know" the outcome sequence $\mathbf{z}$. (The restriction to finite-memory and continuity in the definition of memoryless accomplishes this, at least for some outcome sequences.) Intuitively, one should allow the outcome sequence to be "independent" of the prediction sequence, but this would betray our focus on individual sequences. Nonetheless, it is possible that our techniques can be extended to

such online settings, which could allow for a formal link to similar statements made in the literature on game-theoretic probability (Shafer and Vovk, 2005; Vovk, 2014).

6. Discussion and Future Work

We have generalized the notion of memoryless sequences of Nobel (2004) to higher dimensions and differentiable losses. We conclude that memoryless sequences are characterized by the stochastic behavior of their finite dimensional weak limits, and that the the distribution of these limits is governed by the property elicited by the loss function. In particular, the broad class of Bregman divergences share the same set of memoryless sequences with squared loss, and their weak limits form martingale difference sequences. Finally, we showed how these results can show that prices in prediction markets are calibrated (or traders can make infinite profits).

A promising future direction would be to extend these results to non-differentiable losses, if possible. This would allow for losses eliciting the median, as all losses eliciting the median, such as absolute loss $\ell(x, y) = |x - y|$, are nondifferentiable (Gneiting, 2011). As mentioned in Section 5, it would also be interesting to extend our results to a more online setting, where the outcome sequence can adapt to the predictions adversarially, a setting closer to game-theoretic probability.

References

J. Abernethy and R. Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the 25th Conference on Learning Theory*, pages 1–27, 2012.

Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation*, 1(2):12, 2013.

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *JMLR Workshop and Conference Proceedings*, volume 40, pages 1–19, 2015.

G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950. ISSN 1520-0493.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Gábor Lugosi, and others. On prediction of individual sequences. *The Annals of Statistics*, 27(6):1865–1895, 1999.

Dean P Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1):7–35, 1999.

Rafael Frongillo and Ian Kash. Vector-Valued Property Elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 1–18, 2015.

Rafael Frongillo and Bo Waggoner. An Axiomatic Study of Scoring Rule Markets. *Preprint*, 2017.

T. Gneiting. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

R. Hanson. Combinatorial Information Market Design. *Information Systems Frontiers*, 5(1):107–119, 2003.

David Haussler, Jyrki Kivinen, and Manfred K Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

Andrei N Kolmogorov. Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1):1–7, 1965.

Nicolas S. Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 109–118, 2009.

Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

N.S. Lambert. Elicitation and Evaluation of Statistical Forecasts. *Preprint*, 2011.

Per Martin-Löf. The definition of random sequences. *Information and Control*, 9(6):602–619, December 1966. ISSN 0019-9958.

J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42(9):654, 1956.

R v Mises. Grundlagen der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(1-2):52–99, 1919.

Andrew B. Nobel. Some stochastic properties of memoryless individual sequences. *IEEE Transactions on Information Theory*, 50(7):1497–1505, 2004.

Kent Harold Osband. *Providing Incentives for Better Cost Forecasting*. University of California, Berkeley, 1985.

M.D. Reid and R.C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 9999:2387–2422, 2010.

L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, pages 783–801, 1971.

Glenn Shafer and Vladimir Vovk. *Probability and finance: it's only a game!*, volume 491. John Wiley & Sons, 2005.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and Identification of Properties. In *Proceedings of The 27th Conference on Learning Theory*, pages 482–526, 2014.

Vladimir Andreevich Uspenskii, Alexei L Semenov, and A Kh Shen. Can an individual sequence of zeros and ones be random? *Russian Mathematical Surveys*, 45(1):121, 1990.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, June 2000. ISBN 978-0-521-78450-4. Google-Books-ID: UEuQEM5RjWgC.

E. Vernet, R.C. Williamson, and M.D. Reid. Composite Multiclass Losses. *NIPS*, 2011.

V. Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261(1):57–79, June 2001. ISSN 0304-3975.

Vladimir Vovk. Laws of probabilities in efficient markets, November 2014.

Vladimir V V'yugin. Effective convergence in probability and an ergodic theorem forindividual random sequences. *Theory of Probability & Its Applications*, 42(1):39–50, 1998.

J. Wolfers and E. Zitzewitz. Prediction Markets. *Journal of Economic Perspective*, 18(2):107–126, 2004.