Journal of Machine Learning Research 21 (2020) 1-28 Submitted 1/18; Revised 3/20; Published 6/20

# Memoryless Sequences for General Losses

**Rafael Frongillo** RAF@COLORADO.EDU
*CU Boulder*

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

**Editor:** Manfred Warmuth

## Abstract

One way to define the randomness of a fixed individual sequence is to ask how hard it is to predict relative to a given loss function. A sequence is memoryless if, with respect to average loss, no continuous function can predict the next entry of the sequence from a finite window of previous entries better than a constant prediction. For squared loss, memoryless sequences are known to have stochastic attributes analogous to those of truly random sequences. In this paper, we address the question of how changing the loss function changes the set of memoryless sequences, and in particular, the stochastic attributes they possess. For convex differentiable losses we establish that the statistic or property elicited by the loss determines the identity and stochastic attributes of the corresponding memoryless sequences. We generalize these results to convex non-differentiable losses, under additional assumptions, and to non-convex Bregman divergences. In particular, our results show that any Bregman divergence has the same set of memoryless sequences as squared loss. We apply our results to price calibration in prediction markets.

**Keywords:** algorithmic randomness, property elicitation, prediction markets.

## 1. Introduction

Since the beginnings of probability theory there has been interest in understanding fixed, complex objects through the lens of randomness. One manifestation of this interest is the problem of assessing and defining the randomness of an individual numerical sequence, a representative question being whether, and in what sense, the digits of $\pi$ are random. One influential approach to this problem is algorithmic randomness, introduced by Martin-Löf (1966), under which a sequence is said to be random if it passes every statistical test performed by some class of algorithms (typically Turing machines).

Perhaps the strongest mathematical model of randomness is that of an independent, identically distributed (i.i.d.) sequence of random variables, a paradigmatic example being independent tosses of a fair coin. Under appropriate moment conditions, i.i.d sequences obey the basic limit laws of probability, including the strong law of large numbers, the central limit theorem, and the law of the iterated logarithm. Sequences of i.i.d. random variables also have the property of being unpredictable: beyond any information they provide about their marginal distribution, the first $n$ elements in the sequence do not convey information about the exact values of the elements that follow.

The unpredictability of an i.i.d. sequence follows from its randomness. In this paper we study the randomness of an individual sequence by inverting this relationship. Following earlier work of Nobel (2004), we regard an individual sequence as being random if, in the limit, the average loss of the best constant predictor is no greater than the average loss of any rule that predicts the next entry from a continuous function of a fixed, finite number of previous entries, and call such sequences *memoryless*. (A precise definition of memoryless sequences in the general setting of this paper is given in Section 2.2.) Consideration of finite-memory predictors is a natural desideratum, for example, in pseudorandomness, where one may only be concerned about the ability of a bystander with limited computational resources 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. We note that the definition of memoryless sequences naturally applies to sequences of real numbers, in contrast to Turing machines which require a careful theory of computation over the reals.

Nobel (2004) studied real-valued memoryless sequences under the squared loss. He established that 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 in this previous work leads to a number of interesting questions. How do the properties of memoryless sequences depend on the loss? For which other loss functions are memoryless sequences related to martingale difference sequences? Does some analog of the martingale difference property hold for memoryless sequences under general losses?

This paper provides answers to these questions for convex loss functions, and establishes close connections between property elicitation and memoryless sequences for general losses. We show 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. We establish that the property elicited by the loss function uniquely determines which sequences are memoryless, and under mild assumptions, the weak limits of these sequences are such that the conditional value of the property is constant. When the loss is a Bregman divergence, it elicits the mean, and the weak limits of memoryless sequences reduce to martingale differences. In particular, we establish that the family of memoryless sequences investigated by Nobel (2004) remains unchanged when the squared loss is replaced by any Bregman divergence. Our results rely on a key lemma concerning the relationship between the optimal sequential prediction and an asymptotic form of orthogonality. We establish the orthogonality lemma for convex differentiable losses and, with further assumptions, extend it to nondifferentiable losses, allowing us to cover many loss functions used in practice, such as those eliciting the median (absolute loss) and quantiles (pinball loss). We conclude the paper with applications to prediction markets, concerning the calibration of forecasts, 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). The survey of Uspenskii et al. (1990) gives an account of this early work. V'yugin (1998) established an ergodic theorem for Martin-Löf random (typical) individual sequences under certain computability conditions. A good overview of algorithmic randomness, including recent advances, can be found in the book of Downey and Hirschfeldt (2010).

Our work is related to the literature 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), in that the definition of memorylessness employs a infinite-horizon regret quantity. For example, Haussler, Kivinen, and Warmuth (1998) study binary individual sequence prediction under potentially nonconvex loss functions, showing that losses exhibit one of two possible finite-time regret bounds. An important distinction is that our "learning algorithm" will be restricted to employ a fixed map from the last $k$ elements of the sequence to predict the next element.

Dawid and Vovk (1999; 2001) studied a game-theoretic setting in which a learner attempts to predict a sequence of outcomes that may adapt to her predictions. The conclusion is that 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 the law of the iterated logarithm hold. A important feature of this work is that the outcome sequence in online learning and game-theoretic probability is allowed to adapt to the choices of the learner. As a consequence, the resulting notion of unpredictability is much stronger than the notion of memorylessness considered here, and does not apply to fixed individual sequences. See Section 7 for further discussion.

In other work, Vovk (1988) established a law of the iterated logarithm for Kolmogorov random binary sequences, namely binary sequences $x_1, x_2, \ldots$ such that the Kolmogorov complexity of the prefix $x_1, \ldots, x_n$ is of order $n$. In general, memoryless sequences need not be Kolmogorov random; many have constant Kolmogorov complexity. For details, and additional references, we refer the reader to Nobel (2004).

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., 2016), which concerns the design of loss functions $L(p, y)$ such that when the outcome $y$ is drawn from some distribution $q$, the optimal prediction is $p = q$. Property elicitation is the partial-information analog, where one only wishes to score predictions about some property or statistic of the distribution. The modern literature begins with Osband (1985) and Lambert et al. (2008) and continues in computer science, economics, and statistics (Lambert and Shoham, 2009; Lambert, 2018; Gneiting, 2011; Abernethy and Frongillo, 2012; Steinwart et al., 2014; Agarwal and Agarwal, 2015; Frongillo and Kash, 2015).

## 2. Memoryless Sequences and Elicitation

This section is devoted to the definition and discussion of memoryless sequences and an overview of elicitation. We begin with the basic components and assumptions of the predictive setting, and then turn our attention to memoryless sequences. We establish close connections between memoryless sequences and elicitation in Section 4.2

### 2.1. Basic Assumptions for Predictions, Outcomes, and Loss

The predictive setting of this paper has three primary components: a fixed prediction space $\mathcal{X} \subseteq \mathbb{R}^d$, a fixed outcome space $\mathcal{Y} \subseteq \mathbb{R}^d$, and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. In particular,

$\ell(x, y)$ is the loss incurred when the element $x \in \mathcal{X}$ is predicted and the outcome $y \in \mathcal{Y}$ occurs. The following assumptions will be made throughout the paper:

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

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

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

A4. the loss function $\ell(x, y)$ 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.

The continuity assumption A4 can be expressed equivalently as follows: If $(x_1, y_1)$, $(x_2, y_2)$, $\ldots \in \mathcal{X} \times \mathcal{Y}$ converge to $(x, y) \in \mathcal{X} \times \mathcal{Y}$ then $\ell(x_n, y_n)$ converges to $\ell(x, y)$. Our assumption that $\mathcal{X}$ is open and $\mathcal{Y}$ is closed ensures that these sets, and their cartesian product $\mathcal{X} \times \mathcal{Y}$ are Borel-measurable. Measurability of the loss $\ell$ then follows by standard arguments from the fact that it is continuous. Similar remarks apply to condition (A5) concerning the gradient of the loss.

In what follows we will consider infinite sequences of predictions $\mathbf{x} = (x_1, x_2, \ldots)$ and outcomes $\mathbf{y} = (y_1, y_2, \ldots)$ taking values, respectively, in the sequence spaces $\mathcal{X}^{\mathbb{N}}$ and $\mathcal{Y}^{\mathbb{N}}$. To reduce notation, we will usually omit the parentheses when specifying infinite or finite sequences.

### 2.2. Memoryless Sequences

Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$, and $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfy A1, A2, and A4 above, and let $\mathbf{y} = y_1, y_2, \ldots \in \mathcal{Y}$ be a fixed individual sequence. Consider the problem of sequentially predicting the entries $y_i$ of $\mathbf{y}$ by elements $x_i \in \mathcal{X}$: each $x_i$ is obtained by applying a prediction rule to the previous entries $y_1, \ldots, y_{i-1}$ of $\mathbf{y}$, and the performance of these predictions at time $n$ is measured by the average loss $n^{-1} \sum_{i=1}^n \ell(x_i, y_i)$. A set of prediction rules, one for each time index $i \geq 1$, is called a prediction scheme. Let us say that the sequence $\mathbf{y}$ is unpredictable with respect to a family of prediction schemes if no scheme in the family can outperform a simple reference scheme that always predicts the same (constant) value. Then, informally, a sequence $\mathbf{y}$ is memoryless if it is unpredictable with respect to the family of finite order continuous Markov prediction schemes.

In more detail, 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 for $1 \leq i \leq j$ let $y_i^j = y_i, y_{i+1}, \ldots, y_j$. Each function $g \in \mathcal{C}_k$ represents a continuous $k$th order Markov prediction scheme for $\mathbf{y}$ that predicts the outcome $y_i$ by $x_i = g(y_{i-k}^{i-1})$, which is the value of $g$ applied to the $k$ previous values in the sequence. The family of continuous finite order Markov prediction schemes corresponds to the union $\cup_{k \geq 1} \mathcal{C}_k$ of the families $\mathcal{C}_k$.

**Definition 1** *A sequence* $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ *is* $\ell$-memoryless *if there exists a predictor* $c \in \mathcal{X}$ *such that for every* $k \geq 1$ *and every function* $g \in \mathcal{C}_k$

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

*Here we adopt the convention that $g(y_{i-k}^{i-1})$ is equal to a fixed predictor $x_0 \in \mathcal{X}$ for $i \leq k$. When (1) holds we will say that* $\mathbf{y}$ *is* $\ell$-memoryless for $c$. *Note that neither average in (1) is assumed to converge.*

Thus a sequence is memoryless if no bounded continuous function of finitely many arguments outperforms the "lazy" prediction scheme that ignores the past and always predicts the next value of the sequence by $c$. We note that the finite sample performance of a prediction scheme is measured by its average loss, and that schemes are compared via the limiting difference of their finite sample performances. Markov prediction schemes have been widely studied in information theory, see for example the survey of Merhav and Feder (1998). In terms of standard "no-regret" online learning guarantees, equation (1) says that all Markov predictors have nonnegative regret with respect to the best constant predictor in hindsight.

Memoryless real-valued sequences under the squared loss $\ell(x, y) = (x - y)^2$ were previously defined and studied by Nobel (2004). The main focus of that work was establishing that memoryless sequences exhibit asymptotic behavior analogous to that of i.i.d. sequences, including versions of the law of large numbers, the central limit theorem, and Hoeffding's inequality. Our primary focus here is on general convex loss functions $\ell$, and the close connection between memoryless sequences and properties elicited by $\ell$ (elicitation is defined and discussed in the next section). It follows from the results here that if $\ell$ is a Bregman divergence, then $\ell$-memoryless sequences have a stochastic nature analogous to that under the squared loss. Details and discussion can be found in Section 4.3.

We note that the memoryless property is defined asymptoticly. A finite order prediction scheme $g$ may perfectly predict the initial elements of a sequence $\mathbf{y}$, and yet the sequence could still be memoryless. In particular, a memoryless sequence $\mathbf{y}$ that is padded with a large but finite number of initial 0s (or some fixed element of $\mathcal{Y}$) is still memoryless. In this sense, memorylessness is a somewhat weak notion of randomness, in a manner analogous to online learning: just as online learning algorithms can produce arbitrary outputs for an 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.

### 2.3. Property Elicitation

One of the key conclusions of this work is Theorem 9, which establishes a close connection between $\ell$-memoryless sequences and properties elicited by the loss function $\ell$. This terminology is defined below.

**Definition 2** *Let spaces $\mathcal{X}$ and $\mathcal{Y}$ satisfy A1-A2, let loss function $\ell(x, y)$ satisfy A4, and let $\mathcal{Q}$ be the set of all compactly supported probability measures on $\mathcal{Y}$.*

- *A* property *is a set-valued function $\Gamma : \mathcal{Q} \to 2^{\mathcal{X}}$ that associates each probability measure $Q \in \mathcal{Q}$ with a subset $\Gamma(Q)$ of the prediction space $\mathcal{X}$.*
- *The loss function $\ell$* elicits *property $\Gamma$ if*

$$\Gamma(Q) = \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}_Q \ell(x, Y)$$

*for all $Q \in \mathcal{Q}$, where $\mathbb{E}_Q \ell(x, Y) = \int_{\mathcal{Y}} \ell(x, y)\, dQ(y)$. A property is* elicitable *if it is elicited by some loss.*

It is well known, and easy to check, that the squared loss $\ell(x,y) = (x-y)^2$ elicits the mean, and that the absolute loss $\ell(x,y) = |x-y|$ elicits the median. While the squared loss is differentiable, the absolute loss is not, nor is any loss function that elicits the median (Gneiting, 2011). This fact will play a role in our results (see § 3).

Squared loss is a special case of a broader family of loss functions, called *Bregman divergences*, that measure the error of a linear approximation to a convex function.

**Definition 3** *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be convex, $\mathcal{X} \subseteq \mathcal{D}$ be convex and open, and $\mathcal{Y} \subseteq \mathcal{D}$ be closed. Given a strictly convex function $G : \mathcal{D} \to \mathbb{R}$ which is differentiable on $\mathcal{X}$, its associated* Bregman divergence *is the loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given by*

$$\ell(x,y) = G(y) - G(x) - \langle \nabla G(x), y - x \rangle \ . \tag{2}$$

While much more general than squared loss, it is easy to see that Bregman divergences also elicit the mean. In fact, they are the only such losses, even when removing the differentiability assumption (Frongillo and Kash, 2015).

**Theorem 4 (Savage (1971))** *If $\ell$ is the Bregman divergence associated with $G$ and $\mathbb{E}_Q|\ell(x,Y)|$ is finite for each $x \in \mathcal{X}$ and $Q \in \mathcal{Q}$, then $\ell$ elicits the mean $\Gamma : Q \mapsto \{\mathbb{E}_Q Y\}$.*

**Proof** Letting $x^* = \mathbb{E}_Q Y$, note that $\mathbb{E}_Q \ell(x^*, Y) = \mathbb{E}_Q G(Y) - G(x^*)$. Expanding and simplifying, we find that $\mathbb{E}_Q \ell(x^*, Y) - \mathbb{E}_Q \ell(x, Y) = G(x^*) - G(x) - \langle \nabla G(x), x^* - x \rangle$, which is nonnegative by the subgradient inequality. Uniqueness of the argmin follows from strict convexity of $G$. ■

We briefly mention two other examples of properties elicited by convex losses: quantiles, ratios of expectations, and expectiles.

**Quantiles.** The classic loss function eliciting the $\alpha$-quantile is the so-called "pinball loss" given by $\ell(x,y) = (\mathbb{1}\{x \geq y\} - \alpha)(x-y)$, for which absolute loss is the special case $\alpha = 1/2$. Here $\ell$ is convex, but not differentiable in $x$ when $x = y$.

**Ratios of expectations.** Consider a measurable function $b : \mathcal{Y} \to [0, \infty)$ such that $\mathbb{E}_Q b(Y) > 0$ for all $Q \in \mathcal{Q}$. If $G$ is as in Definition 3, then the transformed Bregman divergence $\ell(x,y) = b(y)\, G(y) - b(y)\, G(x) - \langle \nabla G(x), y - b(y)x \rangle$ elicits the ratio of expectations $\Gamma : Q \mapsto \{\mathbb{E}_Q Y / \mathbb{E}_Q b(Y)\}$ (Gneiting, 2011; Frongillo and Kash, 2015). (To see this, consider the ratio $\mathbb{E}_Q \ell(x,Y)/\mathbb{E}_Q b(Y)$, ignore terms depending on $Y$ but not $x$, and appeal to Theorem 4.)

**Expectiles.** The $\tau$-expectile, introduced by Newey and Powell, is a type of generalized quantile defined as the unique solution $x = \mu_\tau$ to the equation $\mathbb{E}_Q\left[|\mathbb{1}_{x \geq y} - \tau|(x - Y)\right] = 0$, where $\mathcal{Y} = \mathbb{R}$ and $\tau \in (0,1)$. The expectile $\mu_\tau$ is elicited by *asymmetric squared loss* $\ell(x,y) = |\mathbb{1}_{x \geq y} - \tau|(x-y)^2$, which is strictly convex and differentiable (Newey and Powell, 1987; Gneiting, 2011).

As we show in §4.2, our main results can be cast in terms of elicitation. In particular we find that individual sequences that are "hard to predict" under a loss $\ell$ are determined only by the property the loss elicits.

## 3. An Orthogonality Condition

To make headway in characterizing memoryless sequences and understanding their stochastic behavior, our first step will be to reduce the "no-regret" definition of memorylessness to a statement of orthogonality, as captured in this section. Given an outcome sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ and a family $\mathbb{X}$ of bounded prediction sequences, the results of this section show that a sequence $\mathbf{x}^* \in \mathbb{X}$ is optimal for predicting $\mathbf{y}$ under the average loss if and only if for every $\mathbf{x} \in \mathbb{X}$ the difference sequence $\mathbf{x} - \mathbf{x}^*$ is orthogonal, in an appropriate sense, to the gradients sequence $(\nabla \ell(x_i^*, y_i))_{i \geq 1}$. We first show this result for the case of differentiable losses, then demonstrate why it does not extend to the nondifferentiable case without further assumptions, and finally make such an extension under mild assumptions on the weak limits of $\mathbf{y}$. We begin with a short review of weak convergence, which plays an important role in our principal results and proofs.

### 3.1. 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 $\nu$, 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 $\nu_n(K^c) < \epsilon$ for each $n \geq 1$. Thus any sequence of measures supported on a common compact set is tight. Prokhorov's theorem states that if $\{\nu_n\}$ is tight then any subsequence $\{\nu_{n_k}\}$ has a further subsequence $\{\nu_{m_k}\}$ that converges weakly to a limiting measure.

Of particular interest in this paper are the empirical measures of vector sequences. Let $\mathbf{u} = u_1, u_2, \ldots$ a sequence of vectors $\mathbb{R}^p$. For each $n \geq 1$ define a probability measure $\nu_n$ on $\mathbb{R}^p$ by assigning mass $1/n$ to each of the vectors $u_1, \ldots, u_n$. More formally,

$$\nu_n(A) \;=\; \frac{1}{n} \sum_{i=1}^{n} I\{u_i \in A\} \tag{3}$$

for all Borel sets $A \subseteq \mathbb{R}^p$. If the sequence $\mathbf{u}$ is bounded, then it is easy to see that the empirical measures $\{\nu_n\}$ are tight and, as a consequence of Prohorov's theorem, any subsequence of $\{\nu_n\}$ has a further subsequence that converges weakly to a limiting measure on $\mathbb{R}^p$. Although the empirical measures $\nu_n$ are discrete, their weak limits may be absolutely continuous with respect to Lebesgue measure. Finally, we note that any measure $\nu$ on $\mathcal{Y}^m \subseteq \mathbb{R}^{dm}$ corresponds to a sequence of random vectors $Y_1, \ldots, Y_m \in \mathcal{Y}$, defined on a common probability space and having $\nu$ as their joint distribution. Thus we may write $\nu_n \Rightarrow \nu$ equivalently as $\nu_n \Rightarrow (Y_1, \ldots, Y_m)$ where $(Y_1, \ldots, Y_m) \sim \nu$.

### 3.2. Differentiable Losses

We begin our study of orthogonality in the case where assumption (A5) holds, namely the loss $\ell(x, y)$ is differentiable with respect to $x$ for each $y \in \mathcal{Y}$. The following definitions will be used in Lemma 7 below.

**Definition 5** *A sequence* $\mathbf{u} = u_1, u_2, \ldots$ *with values* $u_i \in \mathbb{R}^d$ *is* bounded *if there exists a constant* $L < \infty$ *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* $\{u_1, u_2, \ldots\}$ *containing its elements. We will say that* $\mathbf{u}$ *is* interior to *an open set* $U \subseteq \mathbb{R}^d$ *if* $\mathrm{cl}(\mathbf{u})$ *is contained in* $U$.

**Definition 6** *Let* $U$ *be a subset of a vector space. The* star interior *of* $U$ *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* $U$.

**Lemma 7 (Orthogonality, differentiable case)** *Let* $\mathcal{X}, \mathcal{Y}$ *be subsets of* $\mathbb{R}^d$ *satisfying assumptions A1 and A2, and let* $\ell(x, y)$ *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:*

$$\liminf_{n \to \infty} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^*, y_i) \right] \geq 0 \ \ \textit{for all} \ \ \mathbf{x} \in \mathbb{X} \tag{4}$$

*and*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \rangle \ = \ 0 \ \ \textit{for all} \ \ \mathbf{x} \in \mathbb{X} \ . \tag{5}$$

**Proof** Let $\mathbf{x}^* \in \mathrm{starint}(\mathbb{X})$ be fixed. To show that (5) implies (4), we use convexity of $\ell$ 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 inferior, inequality (4) follows from (5).

To establish the converse, suppose that (5) fails to hold. Then there exists a sequence $\mathbf{x} \in \mathbb{X}$ and $\delta > 0$ such that the average in (5) has limit inferior less than $-\delta$ or limit superior greater than $\delta$. 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 $\alpha$ and each $i \geq 1$,

$$\ell(x_i^\alpha, y_i) - \ell(x_i^*, y_i) \ = \ G_\alpha(y_i, x_i, x_i^*) \ + \ \alpha \, \langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \rangle \tag{6}$$

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

$$G_\alpha(u, v, w) \ = \ \ell(w + \alpha(v - w), u) \ - \ \ell(w, u) \ - \ \alpha \, \langle v - w, \nabla \ell(w, u) \rangle \tag{7}$$

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

$$\lim_{l \to \infty} \frac{1}{n_l} \sum_{i=1}^{n_l} \langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \rangle \ \leq \ -\delta. \tag{8}$$

As $K$ is compact the sequence $\{\nu_n\}$ is tight, there is a subsequence $\{n_k\}$ of $\{n_l\}$ such that $\nu_{n_k}$ converges weakly to some probability measure $\nu$ on $K$. Using equation (6), we find that for each $0 < \alpha < \alpha_0$,

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

The last equality above follows from the weak convergence of $\nu_{n_k}$ to $\nu$, and the final inequality follows from (8).

It suffices to show that the final term of the previous display 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 $G_\alpha$ 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-Schwarz 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. ■

### 3.3. A Counterexample

While Lemma 7 applies to differentiable loss functions such as squared loss, it does not address continuous, but non-differentiable, loss functions. Perhaps the simplest non-differentiable loss function is absolute loss $\ell_1(x, y) = |x - y|$. Given the prominence of absolute loss in statistics and machine learning, it is natural to ask whether the conclusion of Lemma 7 continues to hold for $\ell_1$ with $\nabla \ell_1(x, y)$ replaced by a subgradient of $\ell_1(x, y)$ with respect to $x$ at a fixed value of $y$. As the following example shows, the answer is no in general, essentially because the weak limits of a sequence $\mathbf{y}$ may concentrate on "kinks" of the loss. We will then show in §3.4 that if the weak limits do not concentrate on kinks, the result does go through.

**Example.** Consider a setting with $\mathcal{X} = [-1, 1]$ and $\mathcal{Y} = (0, 2)$, and let $\mathbb{X} = \{c^{\mathbb{N}} : c \in \mathcal{X}\}$ be the set of constant predictors, with $x_i = c$ for all $i$ for some $c \in \mathcal{X}$. Now let $\mathbf{y} = (1/i)_{i \geq 1}$ and $\mathbf{x}^* = \mathbf{0}$. Then for any $\mathbf{x} = (c, c, c, \ldots) \in \mathbb{X}$,

$$\begin{aligned}
\liminf_{n\to\infty} \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^*, y_i) \right] &= \liminf_{n\to\infty} \left[ \frac{1}{n} \sum_{i=1}^{n} |c - 1/i| - \frac{1}{n} \sum_{i=1}^{n} |1/i| \right] \\
&\geq \liminf_{n\to\infty} \frac{1}{n} \left[ \sum_{i \leq |c|^{-1}} |1/i| + \sum_{i > |c|^{-1}} (|c| - 2/i) \right] \\
&\geq |c| - \limsup_{n\to\infty} \frac{1}{n} \left[ 2 \log n \right] = |c| \geq 0 \ ,
\end{aligned}$$

where the first inequality follows from $|c - 1/i| \geq \big| |c| - |1/i| \big|$, and the second from properties of the harmonic numbers. Thus $\mathbf{y}$ and $\mathbf{x}^*$ satisfy the "no-regret" condition (4) of Lemma 7 with respect to absolute loss. They do not satisfy the orthogonality condition (5), however, as we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \langle x_i - x_i^*, \nabla \ell(x_i^*, y_i) \rangle = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (c - x_1^*) \cdot (-1) = x_1^* - c = -c \ ,$$

which is non-zero when $c \neq 0$. In fact, when $c > 0$, it appears from this expression that the loss would decrease by moving $x_1^* = 0$ slightly toward $c$, as the derivative of the loss is strictly negative, but as we know from the above, any movement away from 0 eventually will be bounded away from $\mathbf{y}$. This tells us that the "derivative" of the regret term in (4) does not commute with the limit operator, even when all the derivatives $\nabla \ell(x_i^*, y_i)$ exist in (5).

### 3.4. Non-differentiable Losses

As the counterexample above makes clear, in order to establish a version of Lemma 7 in the non-differentiable setting, additional assumptions are required to compensate for "kinks" in the loss function, specifically $(x, y)$ pairs where $\ell(x, y)$ is not differentiable in its first argument. We formulate one such assumption below, which ensures that the sequences of interest avoid such pairs.

Assume that $\mathcal{X}$ and $\mathcal{Y}$ are subsets of $\mathbb{R}^d$ satisfying assumptions A1 and A2, respectively, and that $\ell(x, y)$ is a loss function satisfying assumptions A3 and A4. It follows from A3 that the function $\ell(x, y)$ is subdifferentiable in $x$ for each $y \in \mathcal{Y}$. Let $\tilde{\nabla}\ell(x, y)$ denote the (non-empty) set of subgradients for $\ell(x, y)$ at $x \in \mathcal{X}$. Furthermore, let $S \subseteq \mathcal{X} \times \mathcal{Y}$ be the set of points $(x, y)$ where the ordinary gradient $\nabla\ell(x, y)$ exists and is jointly continuous in both arguments.

**Lemma 8 (Orthogonality, non-differentiable case)** *Assume the subgradients $\tilde{\nabla}\ell(x, y)$ are bounded on bounded subsets of $\mathcal{X} \times \mathcal{Y}$. Let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be bounded, $\mathbb{X} \subseteq \mathcal{X}^{\mathbb{N}}$ a family of bounded sequences that are interior to $\mathcal{X}$, and $\mathbf{x}^* \in \operatorname{starint}(\mathbb{X})$. If every weak limit $\eta$ of the empirical measures of $\{(x_i^*, y_i) : i \geq 1\}$ is such that $\eta(S) = 1$ then the following two statements are equivalent:*

*(i) For all $\mathbf{x} \in \mathbb{X}$*

$$\liminf_{n\to\infty} \left[ \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i) - \frac{1}{n}\sum_{i=1}^{n} \ell(x_i^*, y_i) \right] \;\geq\; 0; \tag{9}$$

*(ii) For all $\mathbf{x} \in \mathbb{X}$ and all sequences $z_1, z_2, \ldots$ with $z_i \in \tilde{\nabla}\ell(x_i^*, y_i)$*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \langle x_i - x_i^*, z_i\rangle \;=\; 0. \tag{10}$$

**Proof** The proof follows that for the differentiable case, with minor changes. We sketch the argument below. The argument that (10) implies (9) is identical to the differentiable case. To establish the converse, we may assume that for some $\mathbf{x} \in \mathbb{X}$ and some sequence $z_i \in \tilde{\nabla}\ell(x_i^*, y_i)$, there exists $\delta > 0$ such that the average in (10) has limit inferior less than $-\delta$. Note that if $(x_i^*, y_i) \in S$ then $z_i = \nabla\ell(x_i^*, y_i)$. Let $\mathbf{x}^\alpha$ be defined as before. For sufficiently small $\alpha > 0$ and each $i \geq 1$, we have

$$\ell(x_i^\alpha, y_i) - \ell(x_i^*, y_i) \;=\; H_\alpha(y_i, x_i, x_i^*, z_i) \,+\, \alpha\,\langle x_i - x_i^*, z_i\rangle$$

where $H_\alpha : \mathcal{Y} \times \mathcal{X}^2 \times \mathbb{R}^d \to \mathbb{R}$ is defined by

$$H_\alpha(u, v, w, z) \;=\; \begin{cases} G_\alpha(u, v, w) & \text{if } (w, u) \in S \\ \ell(w + \alpha(v - w), u) \,-\, \ell(w, u) \,-\, \alpha\,\langle v - w, z\rangle & \text{if } (w, u) \in S^c, \end{cases}$$

and $G_\alpha$ is defined as in (7). By assumption, the sequence $(y_i, x_i, x_i^*, z_i)$ takes values in a compact set $K \subseteq \mathcal{Y} \times \mathcal{X}^2 \times \mathbb{R}^d$. Moreover, $H_\alpha$ is continuous and bounded on the set $K_S = \{(u, v, w, z) \in K : (w, u) \in S\}$. By arguments like those in the differentiable case, there is a subsequence $\{n_k\}$ of the positive integers with the following properties: (i) the limit of $n_k^{-1}\sum_{i=1}^{n_k}\langle y_i - y_i^*, g_i\rangle$ exists and is at most $-\delta$; (ii) the empirical measures $\gamma_{n_k}$ of the sequence $(y_i, x_i, x_i^*, z_i)$ converge weakly to a measure $\gamma$ supported on $K$; and (iii) for all sufficiently small $\alpha > 0$,

$$\liminf_{n\to\infty} \left[ \frac{1}{n}\sum_{i=1}^{n} \ell(x_i^\alpha, y_i) - \frac{1}{n}\sum_{i=1}^{n} \ell(x_i^*, y_i) \right] \;\le\; \liminf_{k\to\infty} \left[ \int H_\alpha \, d\gamma_{n_k} + \frac{\alpha}{n_k} \sum_{i=1}^{n_k} \langle x_i - x_i^*, z_i \rangle \right].$$

Let $\eta$ be the $(w, z)$-marginal of $\gamma$ on $\mathcal{X} \times \mathbb{R}^d$. It is easy to see that $\eta$ is a weak limit of the sequence $(x_i^*, y_i)$, and therefore $\gamma(K_S^c) \le \eta(S^c) = 1 - \eta(S) = 0$ by assumption. It follows from the Portmanteau Theorem that

$$\lim_{k\to\infty} \int H_\alpha \, d\gamma_{n_k} \;=\; \int H_\alpha \, d\gamma \;=\; \int_{K_S} H_\alpha \, d\gamma \;=\; \int_{K_S} G_\alpha \, d\gamma.$$

By arguments like those in the differentiable case, the final integral above is $o(\alpha)$, and the result follows as before. ■

## 4. Weak Limits of Memoryless Sequences

We now turn to our principal result, which establishes that memoryless sequences are characterized by an optimality type property of their limiting empirical measures. We begin by noting that any infinite sequence $\mathbf{y} = y_1, y_2, \ldots \in \mathcal{Y}^{\mathbb{N}}$ gives rise, for each $m \ge 1$, to a sequence of empirical measures on $\mathcal{Y}^m$ that are obtained by placing mass on overlapping blocks of $m$ successive terms in the sequence. Any weak limit of these measures can be represented as a sequence of random vectors $Y_1, \ldots, Y_m$ with $Y_i \in \mathcal{Y}$. Using the orthogonality lemma, we establish in Theorem 9 that $\mathbf{y}$ is $\ell$-memoryless for $c$ if and only if $c$ is the best predictor of $Y_{k+1}$ given $Y_1, \ldots, Y_k$ for $k = 1, \ldots, m-1$. In Section 4.2 we show that this optimality can be expressed equivalently in terms of the property $\Gamma$ elicited by the loss $\ell$. In particular, $\ell$-memoryless sequences are characterized and determined by the property $\Gamma$ elicited by the loss $\ell$. For losses that elicit the mean, the limiting random vectors $Y_1, \ldots, Y_m$ are closely related to martingale difference sequence, and this fact can be used to extend results of Nobel (2004) on the stochastic behavior of memoryless sequences to these loss functions (see Section 4.3).

### 4.1. Principal Result

We require several preliminaries concerning measures derived from an individual sequence taking values in the outcome space $\mathcal{Y}$. Let $\mathbf{y} = y_1, y_2, \ldots \in \mathcal{Y}^{\mathbb{N}}$ and let $m \ge 1$ be a block-length. For each and $n \ge 1$ define the $n$-sample $m$-dimensional empirical measure of $\mathbf{y}$ by

$$\mu_{n,m}(A) \;=\; \frac{1}{n}\sum_{i=0}^{n-1} I\{(y_{i+1}, \ldots, y_{i+m}) \in A\} \tag{11}$$

for all Borel sets $A \subseteq \mathcal{Y}^m$. The measure $\mu_{n,m}(A)$ places mass $1/n$ at each of the $n$ successive $m$-tuples $y_1^m$, $y_2^{m+1}$, $\ldots$, $y_{n+1}^{n+m}$ in $\mathcal{Y}^m$ obtained by sliding a window of width $m$ along the sequence $\mathbf{y}$ one component at a time.

If $\mathbf{y}$ is bounded then the empirical measures $\{\mu_{n,m} : n \geq 1\}$ are tight, and thus every subsequence has a further subsequence that converges weakly to a limiting measure $\mu_m$ on $\mathcal{Y}^m$. We will express the convergence $\mu_{n_k,m} \Rightarrow \mu_m$ in the equivalent form $\mu_{n_k,m} \Rightarrow (Y_1, \ldots, Y_m)$, where $Y_1, \ldots, Y_m \in \mathcal{Y}$ are random vectors with joint distribution $\mu_m$. Recall that a sequence $Y_1, \ldots, Y_m$ 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 9** *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 $c \in \mathcal{X}$ and let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be bounded. The following are equivalent:*

*(i) The sequence $\mathbf{y}$ is $\ell$-memoryless for 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 satisfies*

$$c \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}[\, \ell(x, Y_{k+1}) \,|\, Y_1^k \,] \quad wp1 \tag{12}$$

*for $1 \leq k \leq n - 1$.*

**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 (11). 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 random variable $Y_k$ 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 $\gamma$ 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 7 are satisfied with $\mathbf{x}^*$ identically equal to $c$, and therefore

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \langle x_i - c, \nabla \ell(c, y_i) \rangle \;=\; 0 \;\text{ for all }\; \mathbf{x} \in \mathbb{X}. \tag{13}$$

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 (13) 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, giving condition (ii).

Suppose now that $\mathbf{y}$ fails to satisfy (i). It follows from Lemma 7 (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

$$\lim_{r \to \infty} \frac{1}{n_r} \sum_{i=1}^{n_r} \left\langle g(y_{i-k}^{i-1}) - c, \nabla \ell(c, y_i) \right\rangle \; < \; 0. \tag{14}$$

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 (14) that $\mathbb{E} \left\langle g(Y_1^k) - c, \nabla \ell(c, Y_{k+1}) \right\rangle$ 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

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

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 $\mathbb{E}[\, \ell(x_\alpha, Y_{k+1}) \,|\, Y_1^k \,] \; < \; \mathbb{E}[\, \ell(c, Y_{k+1}) \,|\, Y_1^k \,]$. Thus condition (ii) fails to hold, and the proof is complete. ■

Theorem 9 can be extended to non-differentiable loss functions $\ell$ satisfying A3-A4 by using Lemma 8 instead of Lemma 7 in equation 13 and the subsequent display. For $c \in \mathcal{X}$ let $S_c$ be the (open) set of points $y$ where the derivative $\nabla \ell(c, y)$ exists and is continuous in a neighborhood of $(c, y)$. We state the following result without proof.

**Theorem 10** *Suppose that the subgradients $\tilde{\nabla} \ell(x, y)$ are bounded on bounded subsets of $\mathcal{X} \times \mathcal{Y}$, and let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be a bounded sequence such that every weak limit of $\{\mu_{n,1}\}$ assigns probability one to $S_c$. Then $\mathbf{y}$ is $\ell$-memoryless for $c$ if and only if for each $m \geq 1$ every weak limit $(Y_1, \ldots, Y_m)$ of $\{\mu_{n,m}\}$ is stationary, bounded, and satisfies $c \in \mathrm{argmin}_{x \in \mathcal{X}} \mathbb{E}[\, \ell(x, Y_{k+1}) \,|\, Y_1^k \,]$ with probability one for $1 \leq k \leq m - 1$.*

### 4.2. Memoryless Sequences and Property Elicitation

Theorem 9 establishes a close connection between memoryless sequences and the property elicited by the loss function $\ell$. Recall that $\ell$ elicits the property $\Gamma_\ell : \mathcal{Q} \to 2^{\mathcal{X}}$ defined by $\Gamma_\ell(Q) = \operatorname{argmin}_{x\in\mathcal{X}} \int \ell(x,y)\, dQ(y)$. Condition (12) in Theorem 9 can be stated equivalently as

$$c \in \Gamma_\ell(Y_{k+1}|Y_1^k) \;\; \text{wp1} \tag{15}$$

where $Y_{k+1}|Y_1^k$ denotes the conditional distribution of $Y_{k+1}$ given $Y_1, \ldots, Y_k$ when $k \geq 1$, and the distribution of $Y_1$ when $k = 0$. Thus Theorem 9 implies that *$\ell$-memoryless sequences are characterized by the property $\Gamma_\ell$ elicited by $\ell$.* As an immediate 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 loss functions satisfying A3-A5. If $\ell$ and $\ell'$ elicit the same property $\Gamma : \mathcal{Q} \to 2^{\mathcal{X}}$, then a bounded sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ is $\ell$-memoryless for $c$ if and only if it is $\ell'$-memoryless for $c$.*

Thus any convex loss eliciting the mean, which must be a Bregman divergence, will have the same memoryless sequences as squared loss. As discussed in § 4.3 below, the weak limits of such a sequence are martingale difference sequences, and this can be used to establish stochastic properties of the sequence.

Another common loss function is the absolute loss $\ell_1(x,y) = |x-y|$, which is convex and elicits the median, rather than the mean. Theorem 10 implies that a real-valued individual sequence whose weak limits assign no probability to the point 0 is $\ell_1$-memoryless with $c = 0$ if and only if its weak limits have conditional medians equal to 0 with probability one. Linton and Whang (2007) and Coudin and Dufour (2009) have studied such sequences, which they term *mediangale* differences, as well as more general *quantilegale* difference sequences. [1] Following this terminology, sequences $Y_1, \ldots, Y_m$ satisfying condition (15) might be termed $\Gamma$-gale differences for an elicitable property $\Gamma$, or lossingale differences for condition (12).

Similar statements can be made for expectiles and ratios of expectations. For example, as we saw in § 2.3, a ratio of expectations $\Gamma(Q) = \mathbb{E}_Q Y / \mathbb{E}_Q b(Y)$ is elicited by the transformed Bregman divergence $\ell(x,y) = b(y)G(y) - b(y)G(x) - \langle \nabla G(x), y - b(y)x\rangle$, which is convex whenever the original divergence is convex. For such losses, Theorem 9 together with eq. (15) imply that a sequence is $\ell$-memoryless for $c$ if and only if its weak limits satisfy $c = \Gamma(Y_{k+1}|Y_1^k) = \mathbb{E}[Y_{k+1}|Y_1^k]/\mathbb{E}[b(Y_{k+1})|Y_1^k]$. In particular, a sequence is $\ell$-memoryless at $c$ if and only if $\{Y_k - c\, b(Y_k)\}_{k=1}^m$ is a martingale difference sequence.

### 4.3. Mean Elicitation, Martingale Differences, and Stochastic Properties

When the loss $\ell$ elicits the mean Theorem 9 establishes a connection between the weak limits of memoryless sequences and multivariate martingale differences, generalizing earlier work of Nobel (2004), who considered the case of squared loss with $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}$. As its proof of the following result is routine, we omit the details.

[1] Coudin and Dufour use the term "mediangale" to refer to the difference sequence. Medians in higher-dimensional spaces are often defined in terms of minimizing $\ell'(x,y) = \|x-y\|_1$ (Fekete et al., 2005), in which case the same result would apply: sequences whose weak limits avoid $0 \in \mathbb{R}^k$ will be $\ell'$-memoryless if and only if their weak limits are (generalized) mediangale differences.

**Proposition 12** *Let $\ell$ be a loss function satisfying A3-A5 that elicits the mean in the sense of Definition 2, and let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be bounded and $\ell$-memoryless for $c \in \mathcal{X}$. Then $c$ is unique, and the centered sequence $\mathbf{z} = \mathbf{y} - \mathbf{c}$ with $z_i = y_i - c$ is $\ell$-memoryless for $0$. In particular, every weak limit $Z_1, \ldots, Z_m$ of the $m$-dimensional empirical measures of $\mathbf{z}$ satisfies $\mathbb{E}[Z_{k+1} \,|\, Z_1^k] = 0$ for $1 \le k \le n-1$.*

Nobel (2004) showed how the martingale difference property of Proposition 12 could be leveraged to establish stochastic properties of $\ell$-memoryless sequences. These properties include a Hoeffding-type inequality and a central limit type theorem. These arguments extend without modification to any loss that elicits the mean and satisfies the conditions of Theorem 9 or Theorem 10. We state the Hoeffding inequality and central limit theorem without proof. In Section 5 we show how to relax convexity and differentiability of the loss, thereby extending the results below to all Bregman divergences.

**Proposition 13** *Let $\ell$ be a loss function satisfying A3-A5 that elicits the mean. If the sequence $\mathbf{y} = y_1, y_2, \ldots \in [a,b]$ is $\ell$-memoryless for $0$, then for every $m \ge 1$ and every $t > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I\left\{ \left| \frac{1}{m} \sum_{j=1}^{m} y_{i+j} \right| > t \right\} \;\le\; 2e^{-mt^2/2(b-a)^2}.$$

**Remark:** As an easy corollary of Proposition 13 we obtain a law of large numbers for $\mathbf{y}$, namely $n^{-1}\sum_{i=1}^{n} y_i \to 0$ as $n$ tends to infinity.

Let the sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be given. For each $i, m \ge 1$ define the root-normalized partial sums $s_{i,m} = m^{-1/2}\sum_{j=1}^{m} y_{i+j}$. For each $n \ge 1$ let $\nu_{n,m}$ be the empirical measure of $s_{1,m}, \ldots, s_{n,m}$, that is,

$$\nu_{n,m}(A) \;=\; \frac{1}{n}\sum_{i=1}^{n} I\{s_{i,m} \in A\} \;\; \text{for each set } A \in \mathcal{B}$$

where $\mathcal{B}$ denotes the Borel subsets of $\mathbb{R}$. Note that $\nu_{n,m}$ depends only on $x_1, \ldots, x_{n+m}$. Let $\rho(\cdot,\cdot)$ be any metric for probability measures on $\mathbb{R}$ that is compatible with weak convergence, in the sense that $\nu_n \Rightarrow \nu$ if and only if $\rho(\nu_n; \nu) \to 0$. One example is the Prokhorov metric

$$\rho(\nu;\eta) \;=\; \inf\{\epsilon > 0 \;\text{ s.t. }\; \nu(A) \le \eta(A^\epsilon) \text{ for every } A \in \mathcal{B}\},$$

where $A^\epsilon = \{u : |u - v| < \epsilon \text{ for some } v \in A\}$ is the $\epsilon$-blow-up of $A$. Let $\mathcal{N}(0,\sigma^2)$ denote the normal distribution with mean 0 and variance $\sigma^2$.

**Theorem 14** *Let $\mathbf{y} = y_1, y_2, \ldots \in \mathbb{R}$ be a bounded sequence, and let $\ell$ be a loss satisfying A3-A5 that elicits the mean. If $\mathbf{y}$ is $\ell$-memoryless for $0$ and $\mathbf{y}^2 := y_1^2, y_2^2, \ldots$ is $\ell$-memoryless for $\sigma^2 > 0$ then*

$$\lim_{m\to\infty} \limsup_{n\to\infty} \rho(\,\nu_{n,m}\,,\, \mathcal{N}(0,\sigma^2)\,) \;=\; 0.$$

*Equivalently, for every $\epsilon > 0$ there exists a block length $m_0$ (depending on $\epsilon$) and a sample size $n_0$ (depending on $\epsilon$ and $m_0$) such that $\rho(\,\nu_{n,m_0}, \mathcal{N}(0,\sigma^2)\,) < \epsilon$ for each $n \ge n_0$.*

**Remark.** Theorem 14 is expressed in terms of double limits: the first is taken as the number $n$ of sliding blocks increases, and the second is taken with increasing block size $m$. The first limit accounts for the stochastic behavior of the sequence, essentially replacing the $m$-dimensional distribution of a random sequence with the limiting empirical distribution of the $m$ blocks of $\mathbf{x}$. The second limit corresponds to increasing sample size.

When $\mathcal{Y}$ is finite and $\mathcal{X}$ is a subset of distributions on $\mathcal{Y}$, an important special case of Bregman divergences is the class of *proper losses*, which elicit the full distribution on $\mathcal{Y}$ when this distribution is in $\mathcal{X}$. Other than squared loss, the most common proper loss is log loss $\ell_{\log} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, given by $\ell_{\log}(x, y) = -\log x(y)$, where $\mathcal{X}$ is the relative interior of the probability simplex $\Delta(\mathcal{Y})$, and $x(y)$ denotes the probability $x$ assigns to $y$. Up to terms depending only on $y$, log loss can be written as the Kullback–Leibler divergence, which is the Bregman divergence with respect to (negative) Shannon entropy. For proper losses, we can interpret the stochastic results established below as giving certain calibration guarantees; see § 6 for how to reformulate proper losses to satisfy assumptions A1-A5 in the context of prediction markets, where proper losses correspond to the complete market case.

### 4.4. Memoryless Sequences from Random Processes

The characterization of memoryless sequences in Theorem 9 suggests that the sample paths of appropriate sequences of random variables should be memoryless. This is the conclusion of the following result, the proof of which can be found in Appendix A.

**Proposition 15** *Let $\mathcal{X}$, $\mathcal{Y}$, and $\ell$ satisfy assumptions A1-A5, and let $\Gamma_\ell$ be the property elicited by $\ell$. Let $\mathbf{Y} = Y_1, Y_2, \ldots$ be a sequence of random vectors defined on a common probability space, and taking values in a fixed compact subset of $\mathcal{Y}$ such that*

$$c \in \bigcap_{k \geq 1} \Gamma_\ell(Y_{k+1} | Y_1^k) \quad wp1 \tag{16}$$

*for some predictor $c \in \mathcal{X}$. Then with probability one the sequence $\mathbf{Y}$ is $\ell$-memoryless for $c$.*

Proposition 15 relies on the following elementary characterization of conditional loss optimality. Note that one could view this result, and equation (17) in particular, in terms of identification functions from the property elicitation literature, see for example Steinwart et al. (2014); Lambert (2018).

**Lemma 16** *Let $Y$ be a random vector defined on a probability space $(\Omega, \mathcal{G}, \mathbb{Q})$ and taking values in a compact subset of $\mathcal{Y}$, and let $\mathcal{G}_0$ be a sub-sigma field of $\mathcal{G}$. For any fixed $c \in \mathcal{X}$ the following statements are equivalent:*

$$\mathbb{E}[\, \nabla\ell(c, Y) \,|\, \mathcal{G}_0 \,] \;=\; 0 \text{ with probability 1 }, \tag{17}$$

$$\mathbb{E}[\, \ell(c, Y) \,|\, \mathcal{G}_0 \,] \;\leq\; \mathbb{E}[\, \ell(x, Y) \,|\, \mathcal{G}_0 \,] \text{ with probability 1 for every } x \in \mathcal{X} \; . \tag{18}$$

## 5. Extensions to Non-Convex and Non-Differentiable Losses

We now show that, under appropriate conditions, one may extend Theorem 9 to certain loss functions $\ell$ that do not satisfy condition A3 (convexity) or A5 (differentiability), including some non-differentiable losses that are not covered by Theorem 10. The key idea is to express a non-convex or non-differentiable loss function in a composite form via a continuous invertible link function $\psi$ that effectively replaces the native space $\mathcal{X}$ by a surrogate space $\hat{\mathcal{X}}$ on which the loss is convex and differentiable. Suitable assumptions on the link function ensure that the set of memoryless sequences is preserved by the link function, and that Theorem 9 applies. In particular, we argue that essentially all Bregman divergences are composite in the sense above, and we can thereby show that Theorem 9 and the results of Nobel (2004) extend to the full class of losses eliciting the mean.

Let $\mathcal{X}$ and $\mathcal{Y}$ satisfy conditions A1 and A2, respectively, and let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function of interest that elicits a property $\Gamma$. Suppose that $\ell$ fails to satisfy one or both of conditions A3 (convexity) or A5 (differentiability), but that $\ell$ can be written in the form

$$\ell(x, y) = \hat{\ell}(\psi(x), y) \tag{19}$$

where $\hat{\ell} : \hat{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$ satisfies conditions A3–A5, $\hat{\mathcal{X}} \subseteq \mathbb{R}^d$ satisfies condition A1, and the link function $\psi : \mathcal{X} \to \hat{\mathcal{X}}$ is a homeomorphism (a continuous bijection with a continuous inverse) of $\mathcal{X}$ and $\hat{\mathcal{X}}$. Then Theorem 9 applies to $\ell$ and the property $\Gamma$.

**Proposition 17** *Under the conditions above, a bounded sequence $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ is $\ell$-memoryless for $c \in \mathcal{X}$ if and only if the weak limits of $\mathbf{y}$ are stationary, bounded, and satisfy the optimality relation (12).*

**Proof** Let $\mathbf{y} \in \mathcal{Y}^{\mathbb{N}}$ be bounded, and let $\hat{\Gamma}$ be the property elicited by $\hat{\ell}$. As $\psi : \mathcal{X} \to \hat{\mathcal{X}}$ is a homeomorphism, it is easy to see that every bounded continuous function $\hat{g} : \mathcal{Y}^k \to \hat{\mathcal{X}}$ can be written in the form $\hat{g} = \psi \circ g$ for some bounded continuous function $g : \mathcal{Y}^k \to \mathcal{X}$. Similarly, every bounded continuous function $g : \mathcal{Y}^k \to \mathcal{X}$ can be written in the form $g = \psi^{-1} \circ \hat{g}$ for some bounded continuous function $\hat{g} : \mathcal{Y}^k \to \hat{\mathcal{X}}$. It follows from (19) that $\mathbf{y}$ is $\ell$-memoryless for $c \in \mathcal{X}$ if and only if $\mathbf{y}$ is $\hat{\ell}$-memoryless for $\psi(c) \in \hat{\mathcal{X}}$ which, by Theorem 9, holds if and only if the weak limits of $\mathbf{y}$ are stationary, bounded, and satisfy (12) with $c$ replaced by $\psi(c)$. A straightforward argument using (19) shows that $\Gamma = \psi^{-1} \circ \hat{\Gamma}$ and the result follows. ■

Note that Proposition 17 allows us to characterize memoryless sequences for certain nondifferentiable losses without needing the assumptions of Theorem 10. In particular, if $\psi : \mathcal{X} \to \hat{\mathcal{X}}$ is a nondifferentiable homeomorphism, then in general $\ell(x, y) = \hat{\ell}(\psi(x), y)$ will be nondifferentiable, yet Proposition 17 will still apply.

Now let $\ell_G(x, y) = G(y) - G(x) - \langle \nabla G(x), y - x \rangle$ be the Bregman divergence of a convex function $G : \mathcal{D} \to \mathbb{R}$. The results of Section 4 imply that the weak limits of an $\ell_G$-memoryless sequences are shifted martingale difference sequences if conditions A3-A5 hold. However, the convexity condition A3 is somewhat restrictive: while $\ell_G$ is always convex in $y$, it is generally *not* convex in $x$. For example, when $G(x) = e^x$ the corresponding divergence $\ell_G(x, y) = e^y - e^x(1 + y - x)$ is nonconvex in $x$ for all $y$ (see Figure 1). However, for suitable convex functions $G$ the Bregman divergences $\ell_G$ can be written in the composite form (19) using the *mixed Bregman divergence* (Gordon, 1999). This leads to the following extension of Theorem 9, which relies on Proposition 17. For more details, and the proof of the result, see Appendix B.

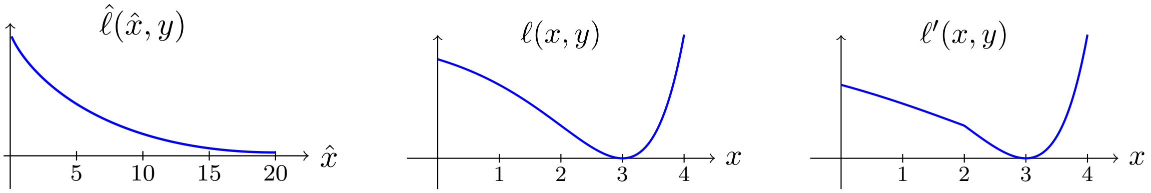

Figure 1: Three loss functions evaluated at $y = 3$. The first, given by $\hat{\ell}(\hat{x}, y) = e^y + \hat{x}\log\hat{x} - \hat{x} - y\hat{x}$, satisfies the conditions A3-A5, and thus Theorem 9 applies. The second, $\ell(x, y) = e^y - e^x(1 + y - x)$, is a nonconvex Bregman divergence $\ell_G$ for $G(x) = e^x$, but can be made convex via the invertible link $\psi(x) = e^x$ as for all $x, y$ we have $\ell(x, y) = \hat{\ell}(\psi(x), y)$, so Proposition 17 applies. The third is both nonconvex and nondifferentiable, but again we may apply Proposition 17 as $\ell'(x, y) = \hat{\ell}(\psi'(x), y)$ for $\psi'(x) = e^x$ for $x \geq 2$ and $1 + e^x/2$ for $x < 2$.

**Proposition 18** *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be convex, $\mathcal{Y} \subseteq \mathcal{D}$ be closed, and $\mathcal{X} \subseteq \mathcal{D}$ be open and convex. If $G : \mathcal{D} \to \mathbb{R}$ is strictly convex, closed, and differentiable on $\mathcal{X}$, then the characterization of Theorem 9 applies to the Bregman divergence $\ell_G$.*

## 6. 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 of a sequence of markets are memoryless with respect to the market prices. 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, our setting 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 vector-valued function $\phi : \mathcal{Z} \to \mathbb{R}^d$, where the component $\phi_i(z)$ denotes the payoff of security $i$ upon outcome $z$. The prediction space $\mathcal{X} = \mathbb{R}^d$ will represent vectors of *shares* in these $d$ 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 seeking to maximize expected payoff) who buys 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 $\phi$. An important special case is when $|\mathcal{Z}| = d$ and $\phi_i(z) = \mathbb{1}\{z = z_i\}$ is the indicator of the element $z_i \in \mathcal{Z}$, corresponding to a *complete* market. In this case the expected value of $\phi(Z)$ is simply the distribution $p$ over $\mathcal{Z}$.

```
Market maker initializes state $x_0 \leftarrow 0 \in \mathbb{R}^d$
for all traders $t = 1, \dots, T$ 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 $t = 1, \dots, T$
```

**Algorithm 1:** The cost-function-based market maker

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 market maker chooses a convex potential function $C : \mathbb{R}^d \to \mathbb{R}$, and maintains a vector $x_t = \sum_{i=1}^t r_i$ describing the total number of shares bought and sold of each security up to time $t$. The cost of purchasing a bundle $r_t \in \mathcal{X}$ of shares at time $t \in \mathbb{N}$ is then given by the difference in the potential, $C(x_{t-1} + r_t) - C(x_{t-1})$, at which point we set $x_t = x_{t-1} + r_t$. 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 $\nabla C$ 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 $p$ over $\mathcal{Z}$, then monotonicity of $\nabla C$ 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 the set of possible gradients $\nabla C(\mathbb{R}^d)$ should be equal to $\mathcal{D} := \operatorname{relint}(\operatorname{conv}(\phi(\mathcal{Z})))$, the relative interior of the convex hull of the security payoffs (Abernethy et al., 2013). This is equivalent to being able to write $C(x) = \sup_{\mu \in \mathcal{D}} \langle \mu, x \rangle - G(\mu)$ for some differentiable and strictly convex function $G : \mathcal{D} \to \mathbb{R}$ with $\nabla G(\mathcal{D}) = \mathbb{R}^d$ (Abernethy et al., 2013).

Now consider running this market, from initalization to the final outcome revelation, many times for many events. One can ask, if the final market prices are "correct", in the sense that the "expected value" of the securities really matched the price vector, i.e., $\nabla C(x) = \mathbb{E}_p \phi(Z)$ where $p$ is the true "distribution" over $Z$? In attempting to answer this question one quickly encounters issues concerning the nature of probability and whether or not the market outcome is truly random. A convenient way around these issues is to appeal to individual sequences, as we do in this paper, and simply ask whether a given (infinite) series of market prices are memoryless with respect to the corresponding outcomes. To do so, we will need to translate the cost-function-based market setting to our own.

To capture this prediction market setting, let $\mathcal{X} = \mathbb{R}^d$ as above 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 $\hat{\ell}_G(x, y) = C(x) + G(y) - \langle x, y \rangle$, with $C$ and $G$ as above, which satisfies assumptions A3-A5 as $C$ and $G$ are both convex and continuously differentiable (see the proof in Appendix B). Note that if the current market state is $x^*$, and a trader moves the state to $x = x^* + r$ by purchasing bundle $r$, then $\hat\ell_G(x,y) - \hat\ell_G(x^*,y) = C(x^*+r) - C(x^*) - \langle r, y\rangle$; this 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 7, 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,

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \left(C(x_i^* + r_i) - C(x_i^*) - \langle r_i, \phi(z_i)\rangle\right) \geq 0 \;\text{ for all }\; \mathbf{x}^* + \mathbf{r} \in \mathbb{X} \tag{20}$$

$$\iff \quad \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \langle r_i, \nabla C(x_i^*) - \phi(z_i)\rangle = 0 \;\text{ for all }\; \mathbf{x}^* + \mathbf{r} \in \mathbb{X}\ , \tag{21}$$

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 $i$th run of the market. Thus, one can interpret the application of Lemma 7 to prediction markets as a version of the efficient market hypothesis: 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. (21), 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. (21) 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 9, we can say something stronger. Observe that the loss $\hat\ell_G$ elicits the property $\Gamma$ such that $\Gamma(p)$ is the set of share vectors whose price matches the expected security payoffs under $p$; that is, $\nabla C(\Gamma(p)) = \{\mathbb{E}_p \phi(Z)\}$. From the discussion in § 4, we find that for any sequence $\mathbf{z} \in \mathcal{Z}^{\mathbb{N}}$, no continuous finite-memory trading strategy can garner (linearly) infinite profits from a series of markets initialized at $x^* \in \mathcal{X}$, if and only if the weak limits of the security payoff sequence $\mathbf{y} = \phi(\mathbf{z})$ are "$\Gamma$-centered" at the initial price $\nabla C(x^*)$, in the sense of eq. (15). In particular, subtracting off $\nabla C(x^*)$, the weak limits of the security payoffs form a multidimensional martingale difference sequence.

**Theorem 19** *Let outcome space $\mathcal{Z}$, securities $\phi : \mathcal{Z} \to \mathbb{R}^d$, strictly convex and differentiable cost function $C : \mathbb{R}^d \to \mathbb{R}$ with $\nabla C(\mathbb{R}^d) = \operatorname{relint}(\operatorname{conv}(\phi(\mathcal{Z})))$, and inital share vector $x^* \in \mathcal{X}$ be given. Then for all $\mathbf{z} \in \mathcal{Z}^{\mathbb{N}}$, we have the following for every $k \geq 1$ and $g \in \mathcal{C}_k$,*

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \left(C(x^* + g(z_{i-k}^{i-1})) - C(x^*) - \left\langle g(z_{i-k}^{i-1}), \phi(z_i)\right\rangle\right) \geq 0\ , \tag{22}$$

*if and only if every weak limit $(Z_1,\ldots,Z_m)$, $m \geq 1$, of the empirical measures $\{\mu_{n,m} : n \geq 1\}$ of $\mathbf{z}$ is stationary, bounded, and satisfies $\mathbb{E}_p[\phi(Z_{k+1})|Z_1^k] = \nabla C(x^*)$ wp1 for $1 \leq k \leq n-1$.*

**Proof** As discussed above, we take $y_i = \phi(z_i)$, $\mathbf{y} = \phi(\mathbf{z})$, and predictors $g' = g + x^*$; note that $g'$ is a continuous Markov predictor if and only if $g$ is. Memorylessness is equivalent to eq. (22) as argued above. For the conditional expectation, from Theorem 9 we have $x^* \in \operatorname{argmin}_{x\in\mathcal{X}} \mathbb{E}[\,\ell(x, Y_{k+1}) \,|\, Y_1^k\,]$ wp1. As we assume $C$ is strictly convex and differentiable, from Lemma 16 we see that $x^*$ achieves the minimum loss if and only if

$$0 = \mathbb{E}[\, \nabla_x \left( C(x^* + x) - C(x^*) - \langle x, \phi(Z_{k+1}) \rangle \right) \mid Z_1^k \,] = \mathbb{E}[\, \nabla C(x^*) - \phi(Z_{k+1}) \mid Z_1^k \,]. \qquad \blacksquare$$

We conclude with some remarks. First, 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, for generic outcome sequences.) Intuitively, one should allow the outcome sequence to be "independent" of the prediction sequence, but this would stray from 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 game-theoretic probability (Shafer and Vovk, 2005; Vovk, 2014). Finally, we note that the connection laid out in this section also applies to the broader family of *scoring rule markets*, by replacing the Bregman divergence with the corresponding score or loss (Lambert et al., 2008; Frongillo and Waggoner, 2018).

## 7. Discussion and Future Work

We have extended the notion of memoryless sequences, introduced in Nobel (2004), to higher dimensions and general convex differentiable losses, as well as some nondifferentiable and nonconvex cases that encompass Bregman divergences. Our results establish that memoryless sequences are characterized by the stochastic behavior of their finite-dimensional weak limits, and that the distribution of these limits is governed by the property elicited by the loss function. In particular, the broad class of Bregman divergences have the same set of memoryless sequences as squared loss, and hence their weak limits form martingale difference sequences. Similarly, the memoryless sequences for absolute loss correspond to mediangales, whose conditional medians are 0. Finally, we applied our results to the question of price calibration in prediction markets, showing that if no trader can make an infinite profit, then prices are calibrated.

It would be interesting to establish similar results in a more online setting, as discussed in Section 6, where the outcome sequence can adapt to the predictions adversarially. Such a setting would be closer to online learning and game-theoretic probability, allowing a more direct comparison to that literature. Another interesting future direction would be to incorporate a dependent variable, and study the relationship between the complexity of the prediction function class and the resulting set of memoryless sequences.

### Acknowledgements

RM Frongillo was supported in part by NSF grant CCF-1657598. AB Nobel was supported in part by NSF grants DMS-1613072 and DMS-1613261, and NIH grant HG009125-01.

## Appendix A. Stochastic Sources of Memoryless Sequences

We assume below that the prediction space $\mathcal{X}$ and outcome space $\mathcal{Y}$ satisfy assumptions A1 and A2, respectively, and that the loss $\ell(x, y)$ satisfies assumptions A3-A5.

**Proof** [of Lemma 16]. Let $c \in \mathcal{X}$ be fixed, and assume without loss of generality that $\mathcal{Y}$ is compact. The subgradient inequality ensures that

$$\ell(x,Y) - \ell(c,Y) \;\geq\; \langle x - c, \nabla \ell(c,Y) \rangle \,.$$

Taking conditional expectations of both sides and using linearity of the conditional expectation, it is clear that (17) implies (18), and we turn our attention to the converse.

As $\mathcal{X}$ is open, there exists $\delta > 0$ such that the closed ball $\overline{B}(c,\delta)$ of radius $\delta$ centered at $c$ is contained in $\mathcal{X}$. For each $x \in \overline{B}(c,\delta)$ and each $y \in \mathcal{Y}$ define $r(x:y)$ by

$$\ell(x,y) = \ell(c,y) + \langle x - c, \nabla \ell(c,y) \rangle + r(x:y).$$

By assumption (A5), for each $y \in \mathcal{Y}$ the ratio $r(x:y)/||x-c|| \to 0$ as $x \to c$. By arguments like those in the proof of Lemma 7 one may readily establish that $r(x:y)/||x-c|| \leq aM$ for all $(x,y) \in \overline{B}(c,\delta) \times \mathcal{Y}$, where $M$ is the maximum of $||\nabla \ell(x,y)||$ over the compact set $\overline{B}(c,\delta) \times \mathcal{Y}$, and $a$ is a constant. Note that $M$ is finite as the gradient is assumed to be continuous.

Now fix $k \geq 1$ and let $x \in B(c,\delta)$. It follows from assumption (18) and the linearity of the conditional expectation that

$$0 \leq \mathbb{E}\big[\, \ell(x,Y) - \ell(c,Y) \,|\, \mathcal{G}_0 \,\big] = \big\langle x - c,\, \mathbb{E}\big[\nabla \ell(c,Y) \,|\, \mathcal{G}_0\big] \big\rangle + \mathbb{E}\big[\, r(x:Y) \,|\, \mathcal{G}_0 \,\big].$$

By the dominated convergence theorem the second expectation above is $o(||x-c||)$, and we conclude that for every $\tilde{x} \in B(0,\delta)$

$$0 \;\leq\; \big\langle \tilde{x},\, \mathbb{E}\big[\nabla \ell(c,Y) \,|\, \mathcal{G}_0\big] \big\rangle \;+\; o(||\tilde{x}||).$$

Let $v \in \mathbb{R}^d$ be any unit vector, and let $\tilde{x} = \alpha v$ for some $0 < \alpha < \delta$. It follows from the previous display that

$$0 \;\leq\; \alpha \, \big\langle v,\, \mathbb{E}\big[\nabla \ell(c,Y) \,|\, \mathcal{G}_0\big] \big\rangle \;+\; o(\alpha).$$

Dividing both sides above by $\alpha$ and letting $\alpha$ tend to zero we find $\big\langle v,\, \mathbb{E}\big[\nabla \ell(c,Y) \,|\, \mathcal{G}_0\big] \big\rangle \geq 0$; replacing $v$ by $-v$ it follows that the inner product is equal to zero. As the unit vector $v$ was a arbitrary, we conclude that $\mathbb{E}\big[\nabla \ell(c,Y) \,|\, \mathcal{G}_0\big] = 0$, as desired. ■

We note that statement (18) of Lemma 16 can be written in the equivalent form

$$c \in \operatorname*{argmin}_{x \in \mathcal{X}} \mathbb{E}[\, \ell(x,Y) \,|\, \mathcal{G}_0 \,] \;\; \text{wp1}$$

**Corollary 20** *Let $Y_1, \ldots, Y_n$ be random vectors taking values in a compact subset of $\mathcal{Y}$. If $c \in \operatorname{argmin}_{x \in \mathcal{X}} \mathbb{E}[\, \ell(x, Y_{k+1}) \,|\, Y_1^k \,]$ for $1 \leq k \leq n-1$, then the gradients $\nabla \ell(c, Y_1), \ldots, \nabla \ell(c, Y_n)$ are a martingale difference sequence with respect to the filtration $\mathcal{F}_k = \sigma(Y_1^k)$.*

**Corollary 21** *Let the random vector $Y$ and sigma field $\mathcal{G}_0$ be as in Lemma 16. If the random vector $X_0 \in \mathcal{X}$ is measurable $\mathcal{G}_0$ then $\mathbb{E}[\, \nabla \ell(X_0, Y) \,|\, \mathcal{G}_0 \,] = 0$ with probability one if and only if $X_0 \in \operatorname{argmin}_{x \in \mathcal{X}} \mathbb{E}[\, \ell(x,Y) \,|\, \mathcal{G}_0 \,]$ with probability one.*

**Proof** [of Proposition 15] Let $c \in \mathcal{X}$ be as in the statement of the theorem, and assume without loss of generality that $\mathcal{Y}$ is compact. Recall that $\mathcal{C}_k$ denotes the family of (bounded) continuous functions from $\mathcal{Y}^k$ to $\mathcal{X}$. Using the orthogonality lemma it suffices to show that with probability one, for each $k \geq 1$ and each $g \in \mathcal{C}_k$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=k+1}^{n} \left\langle g(Y_{i-k}^{i-1}) - c, \nabla\ell(c, Y_i) \right\rangle \;=\; 0. \tag{23}$$

As $\mathcal{Y}$ is compact, the family of functions $\mathcal{C}_k$ contains a countable subset $\tilde{\mathcal{C}_k}$ that is dense in the supremum norm. Let $g$ be any function in $\tilde{\mathcal{C}_k}$ and for $m \geq k+1$ define

$$Z_m \;=\; \left\langle g(Y_{m-k}^{m-1}) - c, \nabla\ell(c, Y_m) \right\rangle .$$

Lemma 16 ensures that $\mathbb{E}[Z_m \,|\, Y_1^{m-1}] = 0$, and therefore $Z_{k+1}, Z_{k+2}, \ldots$ is a bounded martingale difference sequence. It follows from the Azuma-Hoeffding inequality and the Borel-Cantelli lemma that $n^{-1}\sum_{m=k+1}^{n} Z_m \to 0$ with probability one as $n$ tends to infinity, which establishes (23) for the function $g$. As $g$ was an arbitrary element of the countable dense family $\tilde{\mathcal{C}_k}$, one may extend (23) to all of $\mathcal{C}_k$ by a straightforward approximation argument. Finally, excluding an exceptional set of probability zero for each family $\mathcal{C}_k$, we find that, with probability one, (23) holds for all $k \geq 1$ and all $g \in \mathcal{C}_k$, as desired. ■

## Appendix B. Proof of Proposition 18

Assume that $\mathcal{D} \subseteq \mathbb{R}^d$ is convex, $\mathcal{Y} \subseteq \mathcal{D}$ is closed, and $\mathcal{X} \subseteq \mathcal{D}$ is open and convex. Let $G : \mathcal{D} \to \mathbb{R}$ be a closed, strictly convex function that is differentiable on $\mathcal{X}$. Recall that the (usual) Bregman divergence $\ell_G$ is given by $\ell_G(x,y) = G(y) - G(x) - \langle \nabla G(x), y - x\rangle$. Let $F(x) := \sup_{\hat{x}\in\mathcal{D}} \langle \hat{x}, x\rangle - G(\hat{x})$ be the convex conjugate of $G$ and $\hat{\mathcal{X}} \subseteq \operatorname{dom} F$. Define the mixed Bregman divergence $\hat{\ell}_G : \hat{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$ by

$$\hat{\ell}_G(\hat{x}, y) = F(\hat{x}) + G(y) - \langle \hat{x}, y\rangle .$$

We claim that $\hat{\ell}_G(\nabla G(c), y) = \ell_G(c, y)$. To see this, we use the fact that $F(\nabla G(c)) = \langle \nabla G(c), c\rangle - G(c)$, which essentially follows from the first-order optimality condition in the definition of $F$ (Urruty and Lemaréchal, 2001, Corollary E.1.4.4). Thus

$$\begin{aligned}
\hat{\ell}_G(\nabla G(c), y) &= F(\nabla G(c)) + G(y) - \langle \nabla G(c), y\rangle \\
&= \langle \nabla G(c), c\rangle - G(c) + G(y) - \langle \nabla G(c), y\rangle \\
&= G(y) - G(c) - \langle \nabla G(c), y - c\rangle \\
&= \ell_G(c, y),
\end{aligned}$$

as claimed.

**Proof** [of Proposition 18] As $G$ is convex and differentiable on $\mathcal{X}$, it must be continuously differentiable (Urruty and Lemaréchal, 2001, Remark D.6.2.6), and thus $\nabla G$ is continuous on $\mathcal{X}$. We conclude that $\ell_G$ satisfies A4 (continuity).

As $G$ is strictly convex and closed, $F$ is continuously differentiable on the interior of its domain, denoted $\operatorname{int}\operatorname{dom} F$ (Urruty and Lemaréchal, 2001, Theorem E.4.1.1). Define $\hat{\mathcal{X}} := \nabla G(\mathcal{X})$; we will now show $\hat{\mathcal{X}} \subseteq \operatorname{int}\operatorname{dom} F$. As $G$ is closed, we have $s \in \partial G(x) \iff x \in \partial F(s)$ (Rockafellar, 1997, Theorem 23.5). Now suppose $s = \nabla G(x) \notin \operatorname{int}\operatorname{dom} F$. By the above, we have $\partial F(s) \neq \emptyset$, and therefore $\partial F(s)$ must be an unbounded set (Rockafellar, 1997, Theorem 23.4). Taking any $x' \in \partial F(s)$, $x' \neq x$, we have again by the above that $\nabla G(x) = s \in \partial G(x')$, which contradicts the strict convexity of $G$.

Now as $\nabla F$ is continuous on $\hat{\mathcal{X}} \subseteq \operatorname{int}\operatorname{dom} F$, $\hat{\mathcal{X}}$ must be an open set as the preimage of $\mathcal{X}$ under $\nabla F$. Moreover, the gradient maps $\nabla G$ and $\nabla F$ (restricted to $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively) are inverses of each other, and therefore homeomorphisms.

We have now established A1 (for $\hat{\mathcal{X}}$) and A2. Letting $\hat{\ell}_G : \hat{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$ be the mixed Bregman divergence with first argument restricted to $\hat{\mathcal{X}}$, we have properties A3–A5 for $\hat{\ell}_G$. (Convexity is immediate, and continuity and differentiability follow from continuity of $G$, $\nabla G$, $F$, and $\nabla F$.) Now Proposition 17 applies with $\ell = \ell_G$, $\hat{\ell} = \hat{\ell}_G$, and $\psi(x) = \nabla G(x)$. ■

## References

Jacob Abernethy and Rafael Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the Conference on Learning Theory*, pp. 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):1–39, 2013.

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of the Conference on Learning Theory*, pp. 4-22, 2015.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, 2006.

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

Elise Coudin and Jean-Marie Dufour. Finite-sample distribution-free inference in linear median regressions under heteroscedasticity and non-linear dependence of unknown form. *The Econometrics Journal*, 12:S19–S49, 2009.

A. Philip Dawid and Vladimir G. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5(1):125–162, 1999.

Rodney G. Downey and Denis R. Hirschfeldt. *Algorithmic Randomness and Complexity.* Springer Science & Business Media, 2010.

Sándor P. Fekete, Joseph S.B. Mitchell, and Karin Beurer. On the continuous Fermat-Weber problem. *Operations Research*, 53(1):61–76, 2005.

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 Conference on Learning Theory*, pp. 710–727, 2015.

Rafael Frongillo and Bo Waggoner. An axiomatic study of scoring rule markets. In *Innovations in Theoretical Computer Science*, 2018.

Tilmann 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 J. Good. Rational decisions. *Journal of the Royal Statistical Society, Series B (Methodological)*, 14(1):107–114, 1952.

Geoffrey J. Gordon. Regret bounds for prediction problems. In *Proceedings of the 12th annual conference on Computational learning theory*, pages 29–40. ACM, 1999.

Robin 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 of information. *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*, pp. 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*, pp. 129–138, 2008.

Nicolas S. Lambert. Elicitation and evaluation of statistical forecasts. *Working paper*, 2019. URL `https://web.stanford.edu/~nlambert/papers/elicitability.pdf`.

Oliver Linton and Yoon-Jae Whang. The quantilogram: With an application to evaluating directional predictability. *Journal of Econometrics*, 141(1):250–282, 2007.

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

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

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

Richard von Mises. Grundlagen der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(191):52–99, 1919.

Whitney K. Newey and James L. Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 819–847, 1987.

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 (Prediction, Uncertainty Elicitation)*. University of California, Berkeley, 1987.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

R. Tyrell Rockafellar. *Convex Analysis.* Princeton University Press, 1997.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* 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 Conference on Learning Theory*, pp. 482–526, 2014.

Jean-Baptiste Hiriart Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis.* Springer, 2001.

Vladimir A. Uspenskii, Alexei L. Semenov, and Alexander Kh. Shen. Can an individual sequence of zeros and ones be random? *Russian Mathematical Surveys*, 45(1):121-189, 1990.

Aad W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, June 2000.

Elodie Vernet, Rorbert C. Williamson, and Mark D. Reid. Composite multiclass losses. *The Journal of Machine Learning Research*, 17(1):7860–7911, 2016.

Vladimir G. Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261(1):57–79, June 2001.

Vladimir G. Vovk. Laws of probabilities in efficient markets. `http://www.probabilityandfinance.com/GTP2014/Slides/Vovk.pdf`. November 2014.

Vladimir G. Vovk. The law of the iterated logarithm for random Kolmogorov, or chaotic, sequences. *Theory of Probability & Its Applications*, 32(3):413–425, 1988.

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

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