

Supermartingales in Prediction with Expert Advice

Alexey Chernov, Yuri Kalnishkan, Fedor Zhdanov,
Vladimir Vovk

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London,
Egham, Surrey TW20 0EX, UK
{chernov, yura, fedor, vovk}@cs.rhul.ac.uk

March 12, 2010

Abstract

We apply the method of defensive forecasting, based on the use of game-theoretic supermartingales, to prediction with expert advice. In the traditional setting of a countable number of experts and a finite number of outcomes, the Defensive Forecasting Algorithm is very close to the well-known Aggregating Algorithm. Not only the performance guarantees but also the predictions are the same for these two methods of fundamentally different nature. We discuss also a new setting where the experts can give advice conditional on the learner's future decision. Both the algorithms can be adapted to the new setting and give the same performance guarantees as in the traditional setting. Finally, we outline an application of defensive forecasting to a setting with several loss functions.

1 Introduction

The framework of prediction with expert advice was introduced in the late 1980s. In contrast to statistical learning theory, the methods of prediction with expert advice do not require statistical assumptions about the source of data. The role of the assumptions is played by a “pool of experts”: the forecaster, called Learner, bases his predictions upon the predictions and performance of the experts. For details and references, see the monograph [6].

Many methods for prediction with expert advice are known. This paper deals with two of them: the Aggregating Algorithm [24] and defensive forecasting [26]. The Aggregating Algorithm (the AA for short) is a member of the family of exponential-weights algorithms and implements a Bayesian-type aggregation; various optimality properties of the AA have been established [25]. Defensive forecasting is a recently developed technique that combines the ideas of game-theoretic probability [21] with Levin and Gács's ideas of neutral measure [10, 16] and Foster and Vohra's ideas of universal calibration [8].

The idea of defensive forecasting comes from an interpretation of probability with the help of perfect information games. The Learner develops his strategy modeling a game where a probability forecaster plays on the actual data against an imaginary opponent, Sceptic, that represents a law of probability. The capital of Sceptic tends to infinity (or becomes large) if the players' moves lead to violation of this law. The capital of a strategy for Sceptic as a function of other players' moves is called a (game-theoretic) supermartingale. It is known (see Lemma 4 in this paper) that for any supermartingale there is a forecasting strategy that prevents this supermartingale from growing ("defending" against this strategy of Sceptic), thereby forcing the corresponding law of probability. The older versions of defensive forecasting (see, e.g., [26]) minimize Learner's actual loss with the help of the following trick: a forecasting strategy is constructed so that the actual losses (Learner's and experts') are close to the (one-step-ahead conditional) expected losses; at each step Learner minimizes the expected loss (that is, the law of probability used in this case is the conjunction of several laws of large numbers). This paper gives a self-contained description of a different version of the defensive forecasting method. We use certain supermartingales and do not need to talk about the underlying laws of probability.

Defensive forecasting, as well as the AA, can be used for competitive online prediction against "pools of experts" consisting of all functions from a large function class (see [27, 28]). However, the loss bounds proved so far are generally incomparable: for large classes (such as many Sobolev spaces), defensive forecasting is better, whereas for smaller classes (such as classes of analytical functions), the AA works better. Note that the optimality results for the AA are obtained for experts that are free agents, not functions from a given class; thus we need to evaluate the algorithms anew. This general task requires a deeper understanding of the properties of defensive forecasting.

In this paper, the AA and defensive forecasting are discussed in the simple case of a finite number of outcomes. Learner competes with a countable pool of Experts Θ . Experts and Learner give predictions and suffer some loss at each step. A game is a specification what predictions are admissible and what losses a prediction incur for each outcome. For every game, we are interested in performance guarantees of the form

$$\forall \theta \in \Theta \forall N \quad L_N \leq cL_N^\theta + a^\theta,$$

where L_N is the cumulative loss of Learner and L_N^θ is the cumulative loss of Expert θ over the first N steps, c is some constant and a^θ depends on θ only. Section 2 recalls the AA and its loss bound (Theorem 1) and introduces notation used in the paper.

Section 3 presents the main results of the paper. Subsection 3.1 describes the Defensive Forecasting Algorithm (DFA), which is based on the use of game-theoretic supermartingales, and its loss bound (Theorem 5). It turns out that if the AA and the DFA are both applicable to a game, they guarantee the same loss bound. Subsections 3.3–3.6 discuss when the DFA and the AA are applicable. Loosely speaking, if the DFA is applicable then the AA is applicable as well (Theorem 9); and for games satisfying some additional assumptions, if the AA is applicable then the DFA is applicable (Theorems 13 and Theorem 20). Subsection 3.7 gives a criterion of the AA realizability in terms of supermartingales (Theorem 22) using a rather awkward variant of the DFA. The construction of

the supermartingales used in this paper involves a parameterization of the game with the help of a proper loss function. Proper loss functions play an important role in Bayesian statistics, and their meaning in our context is discussed in Subsections 3.4 and 3.6.

The rest of the paper is devoted to modifications of the standard setting. Subsection 3.8 applies the DFA in an extended setting where the outcomes form a finite-dimensional simplex. Section 4 introduces a new setting for prediction with expert advice, where the experts are allowed to “second-guess”, that is, to give “conditional” predictions that are functions of the future Learner’s decision (cf. the notion of internal regret [9]). If the dependence is regular enough (namely, continuous), the DFA works in the new setting virtually without changes (Theorem 26). The AA with some modification based on the fixed point theorem can be applied in the new setting too (Theorem 29). Section 5 briefly outlines one more application of the DFA: a setting with several loss functions.

Some results of the paper appeared in [29] and in ALT’08 proceedings [4].

2 Games of Prediction and the Aggregating Algorithm

We begin with formulating the setting of prediction with expert advice. A *game of prediction* consists of three components: a non-empty set Ω of possible outcomes, a non-empty set Γ of possible decisions, and a function $\lambda: \Gamma \times \Omega \rightarrow [0, \infty]$ called the *loss function*. In this paper we assume that the set Ω is finite.

The set $\Lambda = \{g \in [0, \infty]^\Omega \mid \exists \gamma \in \Gamma \forall \omega \in \Omega g(\omega) = \lambda(\gamma, \omega)\}$ is called the set of *predictions* of the game. In this paper, we will identify each decision $\gamma \in \Gamma$ with the function $\omega \mapsto \lambda(\gamma, \omega)$ (and also with a point in a $|\Omega|$ -dimensional Euclidean space with pointwise operations). A loss function can be considered as a parameterization of Λ by elements of Γ . To study the properties of a game, we do not need to know the decision set Γ and the loss function; we can forget about them and consider the prediction set Λ only. *From now on, a game will be specified by a pair (Ω, Λ) , where $\Lambda \subseteq [0, \infty]^\Omega$. We will use the letter γ (as well as g) with indices to denote elements of $[0, \infty]^\Omega$ (rather than decisions).*

However, loss functions remain a convenient method to specify a game, and we will use them in examples. Also an important technical tool will be a kind of canonical parameterization of Λ given by the so called proper loss functions. Also loss functions are unavoidable in Section 5, where we consider games with several simultaneous losses.

The game of prediction with expert advice is played by Learner, Experts, and Reality; the set (“pool”) of Experts is denoted by Θ . We will assume that Θ is (finite or) countable. There is no loss of generality in assuming that Reality and all Experts are cooperative, since we are only interested in what can be achieved by Learner alone; therefore, we essentially consider a two-player game. The game is played according to Protocol 1.

The goal of Learner is to keep L_n smaller or at least not much greater than L_n^θ , at each step n and for all $\theta \in \Theta$.

To analyze the game, we need some additional notation. A point $g \in [0, \infty]^\Omega$ is called a *superprediction* in the game (Ω, Λ) if there is $\gamma \in \Lambda$ such that

Protocol 1 Prediction with Expert Advice

$L_0 := 0$.
 $L_0^\theta := 0$, for all $\theta \in \Theta$.
for $n = 1, 2, \dots$ **do**
 All Experts $\theta \in \Theta$ announce $\gamma_n^\theta \in \Lambda$.
 Learner announces $\gamma_n \in \Lambda$.
 Reality announces $\omega_n \in \Omega$.
 $L_n := L_{n-1} + \gamma_n(\omega_n)$.
 $L_n^\theta := L_{n-1}^\theta + \gamma_n^\theta(\omega_n)$, for all $\theta \in \Theta$.
end for

$\gamma(\omega) \leq g(\omega)$ for all $\omega \in \Omega$. It is convenient to write the last condition as $\gamma \leq g$. In the sequel, we will use pointwise relations and operations for the elements of $[0, \infty]^\Omega$ without special mentioning.

For a game (Ω, Λ) , denote by Σ_Λ the set of all superpredictions. Using operations on sets, this definition can be written as $\Sigma_\Lambda = \Lambda + [0, \infty]^\Omega = \{\gamma + g \mid \gamma \in \Lambda, g \in [0, \infty]^\Omega\}$.

The *Aggregating Algorithm* is a strategy for Learner. It has four parameters: reals $c \geq 1$ and $\eta > 0$, a *distribution* P_0 on Θ (that is, $P_0(\theta) \in [0, 1]$ for every $\theta \in \Theta$ and $\sum_{\theta \in \Theta} P_0(\theta) = 1$), and a *substitution function* $\sigma: \Sigma_\Lambda \rightarrow \Lambda$ such that $\sigma(g) \leq g$ for any $g \in \Sigma_\Lambda$.

At step N , the AA computes $g_N \in [0, \infty]^\Omega$ by the formula

$$g_N(\omega) = -\frac{c}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta \in \Theta} P_{N-1}(\theta)} \exp(-\eta \gamma_N^\theta(\omega)) \right),$$

where

$$P_{N-1}(\theta) = P_0(\theta) \prod_{n=1}^{N-1} \exp(-\eta \gamma_n^\theta(\omega_n))$$

is the (posterior) distribution on Θ . Then, $\gamma_N = \sigma(g_N)$ is announced as Learner's prediction.

The step N of the AA can be performed if and only if g_N is a superprediction ($g_N \in \Sigma_\Lambda$), that is, if

$$\exists \gamma_N \in \Lambda \quad \forall \omega \quad \gamma_N(\omega) \leq -\frac{c}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta' \in \Theta} P_{N-1}(\theta')} \exp(-\eta \gamma_N^\theta(\omega)) \right). \quad (1)$$

We say that the AA is (c, η) -*realizable* (for the game (Ω, Λ)) if condition (1) is true regardless of Θ , N , $\gamma_N^\theta \in \Lambda$ and P_{N-1} (that is, regardless of P_0 , the history of the previous moves, and the opponents' moves at the last step). This requirement can be restated in several equivalent forms: for any finite set $G \subseteq \Lambda$ and for any distribution ρ on G , it holds that

$$\exists \gamma \in \Lambda \quad \gamma \leq -\frac{c}{\eta} \ln \left(\sum_{g \in G} \rho(g) \exp(-\eta g) \right); \quad (2)$$

or equivalently, for any finite $G \subseteq \Sigma_\Lambda$ and any distribution ρ on G , it holds that

$$\exists \gamma \in \Sigma_\Lambda \quad \gamma \leq -\frac{c}{\eta} \ln \left(\sum_{g \in G} \rho(g) \exp(-\eta g) \right); \quad (3)$$

equivalently, in the last formula \leq can be replaced by $=$. Indeed, the condition (1) implies (2) since γ_N^θ and P_0 are arbitrary; $G \subseteq \Lambda$ can be replaced by $G \subseteq \Sigma_\Lambda$ since the right-hand side of (2) increases when elements of G increase; by definition, (2) means that its right-hand side belongs to Σ_Λ , and we get (3) with $=$ instead of \leq . Clearly, (1) follows from (3), if we allow countably infinite G as well (then we can take $\{\gamma_N^\theta \mid \theta \in \Theta\}$ for G), which is possible due to the following property of convex sets.

For a given η , the *exp-convex hull* of Σ_Λ is the set $\Sigma_\Lambda^\eta \supseteq \Sigma_\Lambda$ that consists of all points in $[0, \infty]^\Omega$ of the form

$$\log_{(e^{-\eta})} \left(\sum_{g \in G} \rho(g) (e^{-\eta})^g \right) = -\frac{1}{\eta} \ln \left(\sum_{g \in G} \rho(g) \exp(-\eta g) \right), \quad (4)$$

where G is a finite subset of Σ_Λ and ρ is a distribution on G . Actually, $\exp(-\eta \Sigma_\Lambda^\eta)$ is the convex hull of $\exp(-\eta \Sigma_\Lambda)$. As known from convex analysis, we get the same definition if we allow infinite G (see e. g. [2, Theorem 2.4.1]). With this notation, the condition (3) says that $\Sigma_\Lambda \supseteq c \Sigma_\Lambda^\eta$.

Let us state some properties of the set Σ_Λ^η . First, $\Sigma_\Lambda^\eta = \Sigma_\Lambda + [0, \infty]^\Omega$, that is, if Σ_Λ^η is a prediction set then its superprediction set is Σ_Λ^η itself. (Indeed, if a point g_0 of the form (4) belongs to Σ_Λ^η as a combination of $g_i \in G \subseteq \Sigma_\Lambda$ then, for any $g \in [0, \infty]^\Omega$, the point $g_0 + g$ belongs to Σ_Λ^η as the combination of $g_i + g$.) The set $\exp(-\eta \Sigma_\Lambda^\eta)$ is convex (clearly, the points of the form (4) belong to Σ_Λ^η also if we allow $G \subseteq \Sigma_\Lambda^\eta$). The convexity of exponent implies that the set Σ_Λ^η is convex as well (if $g_1, g_2 \in \Sigma_\Lambda^\eta$ then $\alpha g_1 + (1 - \alpha)g_2 \geq -\frac{1}{\eta} \ln(\alpha \exp(-\eta g_1) + (1 - \alpha) \exp(-\eta g_2))$ and hence $\alpha g_1 + (1 - \alpha)g_2 \in \Sigma_\Lambda^\eta$ too).

The game (Ω, Λ) is called *η -mixable* if the AA is $(1, \eta)$ -realizable, that is, if $\Sigma_\Lambda = \Sigma_\Lambda^\eta$. The game is *mixable* if it is η -mixable for some $\eta > 0$. The mixable games are of special interest. In a sense, the AA works with mixable games only, and to any non-mixable game (Ω, Λ) the AA assigns the η -mixable game $(\Omega, \Sigma_\Lambda^\eta)$ and then simply transfers the loss bound (at the price of a constant factor). Standard examples of mixable games are the square loss game [25, Example 4], which is η -mixable for $\eta \in (0, 2]$, and the logarithmic loss game [25, Example 5], which is η -mixable for $\eta \in (0, 1]$; see Subsection 3.2. A standard example of a non-mixable game is the absolute loss game [25, Example 3] with the loss function $\lambda(p, \omega) = |p - \omega|$, $p \in [0, 1]$, $\omega \in \{0, 1\}$ (its prediction set Λ is $\{(x, y) \in [0, 1]^2 \mid x + y = 1\}$); for the absolute loss game, the AA is (c, η) -realizable for $\eta > 0$ and $c \geq \eta / (2 \ln(2 / (1 + e^{-\eta})))$.

A detailed survey of the AA, its properties, attainable bounds and realizability conditions for a number of games can be found in [25]. Here we reproduce the proof of the main loss bound in the form that motivates our further study.

Theorem 1 ([24]). *If the AA is (c, η) -realizable then the AA with parameters c, η, P_0 , and σ guarantees that, at each step N and for all experts θ , it holds*

$$L_N \leq c L_N^\theta + \frac{c}{\eta} \ln \frac{1}{P_0(\theta)}.$$

Proof. We need to deduce the performance bound from the condition (1). To this end, we will rewrite (1) and get a semi-invariant of the AA—a value that does not grow. Indeed, the inequality (1) is equivalent to

$$\sum_{\theta \in \Theta} P_{N-1}(\theta) \geq \sum_{\theta \in \Theta} P_{N-1}(\theta) \exp(-\eta \gamma_N^\theta(\omega)) \exp\left(\frac{\eta}{c} \gamma_N(\omega)\right).$$

Multiplying both sides by $\prod_{n=1}^{N-1} \exp\left(\frac{\eta}{c} \gamma_n(\omega_n)\right)$ (which is independent of θ and hence can be placed under the sum), and expanding P_{N-1} , we get

$$\begin{aligned} \sum_{\theta \in \Theta} P_0(\theta) \prod_{n=1}^{N-1} \exp(-\eta \gamma_n^\theta(\omega_n)) \prod_{n=1}^{N-1} \exp\left(\frac{\eta}{c} \gamma_n(\omega_n)\right) \\ \geq \sum_{\theta \in \Theta} P_0(\theta) \prod_{n=1}^{N-1} \exp(-\eta \gamma_n^\theta(\omega_n)) \prod_{n=1}^{N-1} \exp\left(\frac{\eta}{c} \gamma_n(\omega_n)\right) \\ \times \exp(-\eta \gamma_N^\theta(\omega)) \exp\left(\frac{\eta}{c} \gamma_N(\omega)\right), \end{aligned}$$

that is,

$$\sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta) \geq \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta) \exp\left(\eta \left(\frac{\gamma_N(\omega)}{c} - \gamma_N^\theta(\omega)\right)\right)$$

where Q_{N-1} is defined by the formula:

$$Q_{N-1}(\theta) = \exp\left(\eta \sum_{n=1}^{N-1} \left(\frac{\gamma_n(\omega_n)}{c} - \gamma_n^\theta(\omega_n)\right)\right).$$

That is, the condition (1) is equivalent to

$$\exists \gamma_N \in \Lambda \forall \omega \sum_{\theta \in \Theta} P_0(\theta) \tilde{Q}_N(\theta) \leq \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta), \quad (5)$$

where \tilde{Q}_N is the result of substituting ω for ω_N in Q_N .

In other words, the AA (when it is (c, η) -realizable) guarantees that after each step n the value $\sum_{\theta \in \Theta} P_0(\theta) Q_n(\theta)$ does not increase whatever ω_n is chosen by Reality. Since $\sum_{\theta \in \Theta} P_0(\theta) Q_0(\theta) = \sum_{\theta \in \Theta} P_0(\theta) = 1$, we get $\sum_{\theta \in \Theta} P_0(\theta) Q_N(\theta) \leq 1$ and $Q_N(\theta) \leq 1/P_0(\theta)$ for each step N . To complete the proof it remains to note that

$$Q_N(\theta) = \exp\left(\eta \left(\frac{L_N}{c} - L_N^\theta\right)\right).$$

□

For $c = 1$, the value $\frac{1}{\eta} \ln(\sum_{\theta} P_0(\theta) Q_N(\theta))$ is known as the exponential potential (see [6, Sections 3.3, 3.5]) and plays an important role in the analysis of weighted average algorithms. In the next section we show that the reason why condition (5) can be satisfied is essentially that the function $\sum_{\theta} P_0(\theta) Q_N(\theta)$ is a supermartingale.

3 Supermartingales and the AA

Let $\mathcal{P}(\Omega)$ be the set of all distributions on Ω . Note that since Ω is finite we can identify $\mathcal{P}(\Omega)$ with a $(|\Omega| - 1)$ -dimensional simplex in Euclidean space $\mathbb{R}^{|\Omega|}$ equipped with the standard distance and topology. Let E be any non-empty set. A real-valued function S defined on $(E \times \mathcal{P}(\Omega) \times \Omega)^*$ is called a (game-theoretic) *supermartingale* if for any N , for any $e_1, \dots, e_N \in E$, for any $\pi_1, \dots, \pi_N \in \mathcal{P}(\Omega)$, for any $\omega_1, \dots, \omega_{N-1} \in \Omega$, it holds that

$$\sum_{\omega \in \Omega} \pi_N(\omega) S(e_1, \pi_1, \omega_1, \dots, e_{N-1}, \pi_{N-1}, \omega_{N-1}, e_N, \pi_N, \omega) \leq S(e_1, \pi_1, \omega_1, \dots, e_{N-1}, \pi_{N-1}, \omega_{N-1}). \quad (6)$$

For $N = 1$, the argument of S in the right-hand side is the empty sequence, and we treat $S()$ as a real constant. The intuition behind the definition is the following: there is a sequence of events ω_n , each event is generated according to its own distribution π_n selected (or revealed) at each step anew; when the event happens we compute the next value of S depending on the outcomes of the previous events, the previous distributions and some side information e_n ; the supermartingale property of S means that the expectation of the next value (when the distribution π_n has been selected but the outcome is not known yet) never exceeds the previous value of S .

Remark 2. The notion of a supermartingale is well-known in the probability theory. Let X_1, X_2, \dots be a sequence of random elements with values in Ω . Denote by x_n some realization of X_n , $n = 1, 2, \dots$, and let π_n be a conditional distribution of X_n given $X_1 = x_1, \dots, X_{n-1} = x_{n-1}$. If we fix some values for e_n and substitute X_n for ω_n in S , we can rewrite condition (6) as

$$ES(x_1, \dots, x_{N-1}, X_N) \leq S(x_1, \dots, x_{N-1})$$

(the parameters e_n and π_n in S are omitted). We get the usual definition of a (probabilistic) supermartingale $S_N = S(X_1, \dots, X_N)$, $N = 1, 2, \dots$, with respect to the sequence X_1, X_2, \dots :

$$E[S_N \mid X_1, \dots, X_{N-1}] \leq S_{N-1}.$$

In a sense, a game-theoretic supermartingale is a family of probabilistic supermartingales parameterized by some e_n and also by probabilistic distributions π_n , where the latter serve as conditional probabilities of the underlying random process.

Remark 3. A reader familiar with the supermartingales in algorithmic probability theory may also find helpful the following connection. Let $\mu: \Omega^* \rightarrow [0, 1]$ be a measure on Ω^∞ (where Ω^* and Ω^∞ are the sets of finite and infinite sequences of elements from Ω). As defined in e.g. [17, p. 296], a function $s: \Omega^* \rightarrow \mathbb{R}_+$ is called a supermartingale with respect to μ if for any N and any $\omega_1, \dots, \omega_{N-1} \in \Omega$ it holds that

$$\sum_{\omega \in \Omega} \mu(\omega \mid \omega_1, \dots, \omega_{N-1}) s(\omega_1, \dots, \omega_{N-1}, \omega) \leq s(\omega_1, \dots, \omega_{N-1}),$$

where $\mu(\omega \mid \omega_1, \dots, \omega_{N-1}) = \frac{\mu(\omega_1, \dots, \omega_{N-1}, \omega)}{\mu(\omega_1, \dots, \omega_{N-1})}$ (and $\mu(\omega_1, \dots, \omega_n)$ means the measure of the set of all infinite sequences with the prefix $\omega_1 \dots \omega_n$). Let e_n be any functions of $\omega_1, \dots, \omega_{n-1}$. Let $\pi_n(\omega)$ be $\mu(\omega \mid \omega_1, \dots, \omega_{n-1})$. Having substituted these functions in any game-theoretic supermartingale S , we get a supermartingale with respect to μ in the algorithmic sense.

A supermartingale S is called *forecast-continuous* if for any N , for any $e_1, \dots, e_N \in E$, for any $\pi_1, \dots, \pi_{N-1} \in \mathcal{P}(\Omega)$, for any $\omega_1, \dots, \omega_{N-1}, \omega_N \in \Omega$, the function $S(e_1, \pi_1, \omega_1, \dots, e_N, \pi, \omega_N)$ is continuous as the function of $\pi \in \mathcal{P}(\Omega)$.

The main use of forecast-continuous supermartingales in this paper is explained by the following lemma.

Lemma 4. *Suppose that S is a forecast-continuous supermartingale. Then for any N , for any $e_1, \dots, e_N \in E$, for any $\pi_1, \dots, \pi_{N-1} \in \mathcal{P}(\Omega)$, for any $\omega_1, \dots, \omega_{N-1} \in \Omega$, it holds that*

$$\exists \pi \in \mathcal{P}(\Omega) \forall \omega \in \Omega \quad S(e_1, \pi_1, \omega_1, \dots, e_N, \pi, \omega) \leq S(e_1, \pi_1, \omega_1, \dots, e_{N-1}, \pi_{N-1}, \omega_{N-1}).$$

Note that the property provided by this lemma is similar to the condition (5), where the role of S with the first $N - 1$ triples of the arguments is played by $\sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta)$, the role of $S(\dots, e_N, \pi, \omega)$ (the left-hand side) is played by $\sum_{\theta \in \Theta} P_0(\theta) \tilde{Q}_N(\theta)$, the variable π corresponds to γ_N , and for $n = 1, \dots, N - 1$, the parameters π_n and e_n are represented by γ_n and the vector of γ_n^θ , $\theta \in \Theta$, respectively.

A variant of this lemma was originally proved by Levin [16] in the context of algorithmic theory of randomness. We will prove this lemma later (see Lemma 8), and in the next subsection we consider the Defensive Forecasting Algorithm, the main application of this lemma in our paper.

3.1 Defensive Forecasting

The *Defensive Forecasting* Algorithm (DFA) is another strategy for Learner in the game of prediction with expert advice. Let (Ω, Λ) be a game. The DFA has five parameters: reals $c \geq 1$, $\eta > 0$, a (canonic) loss function $\lambda: \mathcal{P}(\Omega) \rightarrow \Sigma_\Lambda$, a distribution P_0 on Θ , and a substitution function $\sigma: \Sigma_\Lambda \rightarrow \Lambda$ such that $\sigma(\gamma) \leq \gamma$ for all $\gamma \in \Sigma_\Lambda$.

Given λ , c and η , let us define the following function on $(\Sigma_\Lambda \times \mathcal{P}(\Omega) \times \Omega)^*$:

$$Q(g_1, \pi_1, \omega_1, \dots, g_N, \pi_N, \omega_N) = \exp \left(\eta \sum_{n=1}^N \left(\frac{\lambda(\pi_n, \omega_n)}{c} - g_n(\omega_n) \right) \right). \quad (7)$$

To simplify notation, here and in the sequel we consider λ as a function from $\mathcal{P}(\Omega) \times \Omega$ to $[0, \infty]$, that is, we write $\lambda(\pi, \omega)$ instead of $(\lambda(\pi))(\omega)$ and $\lambda(\pi, \cdot)$ instead of $\lambda(\pi)$. For $N = 0$, we let $Q() = 1$ in accordance with the usual agreement that the sum of zero number of terms equals 0. Note that Q is similar to $Q_N(\theta)$ from the proof of Theorem 1, with g_n standing for γ_n^θ and $\lambda(\pi_n, \cdot)$ standing for γ_n .

Given also P_0 , let us define the function Q^{P_0} on $((\Sigma_\Lambda)^\Theta \times \mathcal{P}(\Omega) \times \Omega)^*$ as the following weighted sum of Q :

$$Q^{P_0}(\{\gamma_1^\theta\}_{\theta \in \Theta}, \pi_1, \omega_1, \dots, \{\gamma_N^\theta\}_{\theta \in \Theta}, \pi_N, \omega_N) = \sum_{\theta \in \Theta} P_0(\theta) Q(\gamma_1^\theta, \pi_1, \omega_1, \dots, \gamma_N^\theta, \pi_N, \omega_N). \quad (8)$$

At step N , the DFA chooses any $\pi_N \in \mathcal{P}(\Omega)$ such that

$$\forall \omega \in \Omega \quad Q^{P_0}(\{\gamma_1^\theta\}_{\theta \in \Theta}, \pi_1, \omega_1, \dots, \{\gamma_N^\theta\}_{\theta \in \Theta}, \pi_N, \omega) \leq Q^{P_0}(\{\gamma_1^\theta\}_{\theta \in \Theta}, \pi_1, \omega_1, \dots, \{\gamma_{N-1}^\theta\}_{\theta \in \Theta}, \pi_{N-1}, \omega_{N-1}), \quad (9)$$

stores this π_N for use at later steps, and announces $\gamma_N = \sigma(\lambda(\pi_N, \cdot))$ as Learner's prediction.

Assume that the function Q defined by (7) is a forecast-continuous supermartingale. Clearly, this implies that Q^{P_0} defined by (8) is also a forecast-continuous supermartingale for any P_0 . Then Lemma 4 guarantees that the DFA can choose π_N with the required property.

Theorem 5. *If Q defined by (7) is a forecast-continuous supermartingale for certain c , η , and λ then the DFA with parameters c , η , λ , P_0 , and σ guarantees that, at each step N and for all experts θ , it holds*

$$L_N \leq cL_N^\theta + \frac{c}{\eta} \ln \frac{1}{P_0(\theta)}.$$

Proof. The step of the DFA guarantees that at each step N the value of Q^{P_0} does not increase independent of the outcome ω_N . Thus, the value of Q^{P_0} at each step N is not greater than its initial value, 1. Since Q is always non-negative and Q^{P_0} as the sum of non-negative values can be bounded from below by any of its terms, we get

$$P_0(\theta) \exp\left(\eta \sum_{n=1}^N \left(\frac{\lambda(\pi_n, \omega_n)}{c} - \gamma_n^\theta(\omega_n)\right)\right) \leq 1,$$

and therefore

$$\sum_{n=1}^N \lambda(\pi_n, \omega_n) \leq cL_N^\theta + \frac{c}{\eta} \ln \frac{1}{P_0(\theta)}.$$

It remains to recall that $\gamma_n = \sigma(\lambda(\pi_n, \cdot)) \leq \lambda(\pi_n, \cdot)$, thus summing up we get $L_N \leq \sum_{n=1}^N \lambda(\pi_n, \omega_n)$. \square

In Subsections 3.3–3.6 we discuss general conditions when Q defined by (7) is a supermartingale. In the next subsection we begin with examples for two widely used games of prediction.

3.2 Two Examples of Supermartingales

The *logarithmic loss game* is defined by the loss function

$$\lambda^{\log}(p, \omega) := \begin{cases} -\ln p & \text{if } \omega = 1, \\ -\ln(1-p) & \text{if } \omega = 0, \end{cases}$$

where $\omega \in \{0, 1\}$ is the outcome and $p \in [0, 1]$ is the decision (notice that the loss function is allowed to take value ∞). It is known [25, Example 5] that this game is η -mixable for $\eta \in (0, 1]$. The corresponding prediction set is $\Lambda^{\log} = \{(x, y) \in \mathbb{R}^2 \mid e^{-x} + e^{-y} = 1\}$. The losses in the game are $L_N := \sum_{n=1}^N \lambda^{\log}(p_n, \omega_n)$ for Learner who predicts p_n and $L_N^\theta := \sum_{n=1}^N \lambda^{\log}(p_n^\theta, \omega_n)$ for Expert θ who predicts p_n^θ . Consider the following function:

$$\exp \left(\eta \sum_{n=1}^N \left(\lambda^{\log}(p_n, \omega_n) - \lambda^{\log}(p_n^\theta, \omega_n) \right) \right). \quad (10)$$

This function is actually Q defined by (7), where $c = 1$ and $\lambda^{\log}(p_n^\theta, \cdot)$ stands for g_n . The only difference is that p_n is not an element of $\mathcal{P}(\Omega)$. To fix this, let us assign Learner's decision $p \in [0, 1]$ (and thereby prediction $(-\ln(1-p), -\ln p) \in \Lambda^{\log}$) to each distribution $\pi = (1-p, p)$ on $\{0, 1\}$. With this identification $\pi \mapsto p$, the expression (10) specifies a function on $([0, 1] \times \mathcal{P}(\{0, 1\}) \times \{0, 1\})^*$ with the arguments p_n^θ , π_n (represented by $p_n = \pi_n(1)$) and ω_n .

Lemma 6. *For $\eta \in (0, 1]$, the function (10) is a forecast-continuous supermartingale.*

Proof. The continuity is obvious. For the supermartingale property, it suffices to check that

$$p_n e^{\eta(-\ln p_n + \ln p_n^\theta)} + (1-p_n) e^{\eta(-\ln(1-p_n) + \ln(1-p_n^\theta))} \leq 1 \quad (11)$$

i.e., that $p_n^{1-\eta} (p_n^\theta)^\eta + (1-p_n)^{1-\eta} (1-p_n^\theta)^\eta \leq 1$ for all $p_n, p_n^\theta, \eta \in [0, 1]$. The last inequality immediately follows from the generalized inequality between arithmetic and geometric means: $u^\alpha v^{1-\alpha} \leq \alpha u + (1-\alpha)v$ for any $u, v \geq 0$ and $\alpha \in [0, 1]$, which after taking the logarithm just expresses that logarithm is concave. (Remark: The left-hand side of (11) is a special case of what is known as the Hellinger integral in probability theory.) \square

In the *square loss game*, the outcomes are $\omega \in \{0, 1\}$ and the decisions are $p \in [0, 1]$ as before, and the loss function is $\lambda^{\text{sq}}(p, \omega) = (p - \omega)^2$. It is known [25, Example 4] that this game is η -mixable for $\eta \in (0, 2]$. The corresponding prediction set is $\Lambda^{\text{sq}} = \{(x, y) \in [0, 1]^2 \mid \sqrt{x} + \sqrt{y} = 1\}$. The losses of Learner and Expert θ are $L_N := \sum_{n=1}^N (p_n - \omega_n)^2$ and $L_N^\theta := \sum_{n=1}^N (p_n^\theta - \omega_n)^2$, respectively. With the same identification $\pi \mapsto p$, the following expression specifies a function on $([0, 1] \times \mathcal{P}(\{0, 1\}) \times \{0, 1\})^*$:

$$\exp \left(\eta \sum_{n=1}^N \left((p_n - \omega_n)^2 - (p_n^\theta - \omega_n)^2 \right) \right) \quad (12)$$

(again, note that it is a special case of Q defined by (7)).

Lemma 7. *For $\eta \in (0, 2]$, the function (12) is a forecast-continuous supermartingale.*

Proof. It is sufficient to check that

$$p_n e^{\eta((p_n-1)^2 - (p_n^\theta-1)^2)} + (1-p_n) e^{\eta((p_n-0)^2 - (p_n^\theta-0)^2)} \leq 1$$

for all $p_n, p_n^\theta \in [0, 1]$ and $\eta \in [0, 2]$. To simplify notation, let us substitute p for p_n and $p + x$ for p_n^θ . Then after trivial transformations we get:

$$pe^{2\eta(1-p)x} + (1-p)e^{-2\eta px} \leq e^{\eta x^2}, \quad \forall x \in [-p, 1-p].$$

The last inequality is a simple corollary of the following well-known variant of Hoeffding's inequality [15, 4.16]:

$$\ln \mathbb{E} e^{sX} \leq s\mathbb{E}X + \frac{s^2(b-a)^2}{8},$$

which is true for any random variable X taking values in $[a, b]$ and for any $s \in \mathbb{R}$; see [6, Lemma A.1] for a proof. Indeed, applying the inequality to the random variable X that is equal to 1 with probability p and to 0 with probability $(1-p)$, we obtain $p \exp(s(1-p)) + (1-p) \exp(-sp) \leq \exp(s^2/8)$. Substituting $s := 2\eta x$, we have $p \exp(2\eta(1-p)x) + (1-p) \exp(-2\eta px) \leq \exp(\eta^2 x^2/2) \leq \exp(\eta x^2)$, the last inequality assuming $\eta \leq 2$. \square

3.3 Supermartingales and the Realizability of the AA

Our next goal is to find when Q defined by (7) is a supermartingale, depending on the parameters c , η and λ . Loosely speaking, we will show that the AA is (c, η) -realizable if and only if there exists λ such that Q is a supermartingale. More precisely, the “only if” part holds for some class of games only. For arbitrary games, the equivalence holds if we relax slightly the supermartingale definition (see Theorem 22).

Let us begin with some notation. For any functions $f: \Omega \rightarrow \mathbb{R}$ and $\pi: \Omega \rightarrow \mathbb{R}$ denote

$$\mathbb{E}_\pi f := \sum_{\omega \in \Omega} \pi(\omega) f(\omega).$$

Actually, this is the scalar product of f and π in \mathbb{R}^Ω . We will mostly use this for $\pi \in \mathcal{P}(\Omega)$; in this case $\mathbb{E}_\pi f$ can be interpreted as the expectation of f over distribution π . For functions $g \in [0, \infty]^\Omega$ and $\pi \in \mathcal{P}(\Omega)$, let

$$\mathbb{E}_\pi g := \sum_{\omega \in \Omega, \pi(\omega) \neq 0} \pi(\omega) g(\omega).$$

Recall that the function Q defined by (7) is a supermartingale if

$$\mathbb{E}_\pi (Q(g_1, \pi_1, \omega_1, \dots, g_N, \pi, \cdot) - Q(g_1, \pi_1, \omega_1, \dots, g_{N-1}, \pi_{N-1}, \omega_{N-1})) \leq 0$$

for any $g_1, \pi_1, \omega_1, \dots, g_{N-1}, \pi_{N-1}, \omega_{N-1}, g_N$ and π . The formula (7) can be rewritten as $Q = \prod_{n=1}^N q_{g_n}(\pi_n, \omega_n)$, where the functions $q_g: \mathcal{P}(\Omega) \times \Omega \rightarrow [0, \infty]$ are defined by the formula

$$q_g(\pi, \omega) = \exp \left(\eta \left(\frac{\lambda(\pi, \omega)}{c} - g(\omega) \right) \right) \quad (13)$$

for any $g \in \Sigma_\Lambda$. Clearly, Q is a supermartingale if and only if $\mathbb{E}_\pi q_g(\pi, \cdot) \leq 1$ for all $\pi \in \mathcal{P}(\Omega)$ and for all $g \in \Sigma_\Lambda$.

Let us say that a function $q: \mathcal{P}(\Omega) \times \Omega \rightarrow \mathbb{R}$ has the *supermartingale property* if for any $\pi \in \mathcal{P}(\Omega)$

$$\mathbb{E}_\pi q(\pi, \cdot) \leq 1.$$

The function q is *forecast-continuous* if for every $\omega \in \Omega$ it is continuous as the function of π .

So, Q defined by (7) is a forecast-continuous supermartingale if and only if the functions q_g defined by (13) are forecast-continuous and have the supermartingale property for all $g \in \Sigma_\Lambda$. In the sequel, we will discuss the properties of q_g instead of Q . Let us begin with a variant of Lemma 4.

Lemma 8. *Let a function $q: \mathcal{P}(\Omega) \times \Omega \rightarrow \mathbb{R}$ be forecast-continuous. If for all $\pi \in \mathcal{P}(\Omega)$ it holds that*

$$\mathbb{E}_\pi q(\pi, \cdot) \leq C,$$

where $C \in \mathbb{R}$ is some constant, then

$$\exists \pi \in \mathcal{P}(\Omega) \forall \omega \in \Omega \quad q(\pi, \omega) \leq C.$$

The proof of the lemma is given in Appendix. Here let us illustrate the idea behind the proof. Consider the function $\phi(\pi', \pi) = \mathbb{E}_{\pi'} q(\pi, \cdot)$ and assume that it has the *minimax* property: $\min_\pi \max_{\pi'} \phi(\pi', \pi) = \max_{\pi'} \min_\pi \phi(\pi', \pi)$. Looking at the right-hand side, note that $\min_\pi \phi(\pi', \pi) \leq \phi(\pi', \pi') \leq C$. Let π minimize the left-hand side, then we get $\max_{\pi'} \mathbb{E}_{\pi'} q(\pi, \cdot) \leq C$, that is, $\mathbb{E}_{\pi'} q(\pi, \cdot) \leq C$ for any π' , which implies the statement of the lemma if we consider distributions π' concentrated at each ω .

Note that Lemma 4 is a simple corollary of Lemma 8 applied to $C = 0$ and

$$q(\pi, \omega) = S(e_1, \pi_1, \omega_1, \dots, e_N, \pi, \omega) - S(e_1, \pi_1, \omega_1, \dots, e_{N-1}, \pi_{N-1}, \omega_{N-1}).$$

Now let us prove that if q_g defined by (13) have the supermartingale property for all $g \in \Sigma_\Lambda$ (in other words, Q is a supermartingale) then the AA is realizable.

Theorem 9. *Let λ map $\mathcal{P}(\Omega)$ to Σ_Λ , and let $c \geq 1$ and $\eta > 0$ be reals such that*

$$q_g(\pi, \omega) := \exp\left(\eta\left(\frac{\lambda(\pi, \omega)}{c} - g(\omega)\right)\right)$$

are forecast-continuous and have the supermartingale property for all $g \in \Sigma_\Lambda$. Then the AA is (c, η) -realizable.

Proof. Recall that the (c, η) -realizability is equivalent to the inequality (3) for any finite $G \subseteq \Sigma_\Lambda$ and for any distribution ρ on G . Let us consider the following function:

$$q(\pi, \omega) = \sum_{g \in G} \rho(g) q_g(\pi, \omega).$$

The function q is forecast-continuous and has the supermartingale property as a non-negative weighted sum of forecast-continuous functions with the supermartingale property. By Lemma 8 applied to this q and $C = 1$, there exists $\pi \in \mathcal{P}(\Omega)$ such that $q(\pi, \omega) \leq 1$ for all ω , that is,

$$\sum_{g \in G} \rho(g) \exp\left(\eta\left(\frac{\lambda(\pi, \omega)}{c} - g(\omega)\right)\right) \leq 1.$$

After trivial transformations, we get the inequality (3) with $\gamma(\omega)$ replaced by $\lambda(\pi, \omega)$. It remains to note that $\lambda(\pi, \cdot) \in \Sigma_\Lambda$. \square

3.4 Proper Loss Functions

The functions q_g defined by (13) have a loss function λ as a parameter. In this subsection, we consider an important property of this loss function.

A function $\lambda: \mathcal{P}(\Omega) \times \Omega \rightarrow [0, \infty]$ is called a *proper* loss function if for all $\pi, \pi' \in \mathcal{P}(\Omega)$

$$\mathbb{E}_\pi \lambda(\pi, \cdot) \leq \mathbb{E}_\pi \lambda(\pi', \cdot),$$

and λ is *strictly proper* if for all $\pi \neq \pi'$ the inequality is strict.

The intuition behind this definition is the following. Assume that the outcome ω is generated according to some distribution π . Then the expected loss $\mathbb{E}_\pi \lambda(\pi', \cdot)$ is minimal, if the prediction π' equals the true distribution. Informally speaking, proper loss functions encourage a forecaster to announce the true subjective probabilities. In a sense, if the loss function is proper then the predictions have a real, not just notational, probabilistic meaning. The proper loss functions are well-known in the Bayesian context; see [7] and [12] (note that these authors consider gains, or scores, instead of losses, so their notation differs from ours by the sign).

We say that λ is *proper with respect to a set* $X \subseteq [0, \infty]^\Omega$ if for all $\pi \in \mathcal{P}(\Omega)$, it holds that $\lambda(\pi, \cdot) \in X$ and for all $g \in X$ it holds that

$$\mathbb{E}_\pi \lambda(\pi, \cdot) \leq \mathbb{E}_\pi g$$

(in other words, $\lambda(\pi, \cdot) \in \arg \min_{g \in X} \mathbb{E}_\pi g$). If the inequality holds for a fixed π and all $g \in X$, we will say that λ is proper at π . Clearly, if λ is proper with respect to X then λ is proper in the usual sense. The definition has a simple geometrical interpretation. The inequality means that the set X lies on one side of the hyperplane $\{x \in \mathbb{R}^\Omega \mid \sum_{\omega \in \Omega} \pi(\omega)x(\omega) = \mathbb{E}_\pi \lambda(\pi, \cdot)\}$, and X touches the hyperplane at $\lambda(\pi, \cdot) \in X$. That is, $\lambda(\pi, \cdot)$ is a point where X touches the supporting hyperplane with normal π .

Lemma 10. *Let λ map $\mathcal{P}(\Omega)$ to Σ_Λ and $\eta > 0$ be such that the functions*

$$q_g(\pi, \omega) := e^{\eta(\lambda(\pi, \omega) - g(\omega))}$$

are forecast-continuous and have the supermartingale property for all $g \in \Sigma_\Lambda$ (the functions q_g are just (13) with $c = 1$). Then λ is a continuous proper loss function with respect to Σ_Λ .

Proof. The continuity is obvious. Since $e^x \geq 1 + x$ for all $x \in \mathbb{R}$, we get

$$\mathbb{E}_\pi e^{\eta(\lambda(\pi, \cdot) - g)} \geq \mathbb{E}_\pi (1 + \eta(\lambda(\pi, \cdot) - g)) = 1 + \eta(\mathbb{E}_\pi \lambda(\pi, \cdot) - \mathbb{E}_\pi g),$$

and from the supermartingale property we have $\mathbb{E}_\pi \lambda(\pi, \cdot) \leq \mathbb{E}_\pi g$ for all $g \in \Sigma_\Lambda$ and all $\pi \in \mathcal{P}(\Omega)$, since $\eta > 0$. (Remark: we get the strict inequality $\mathbb{E}_\pi \lambda(\pi, \cdot) < \mathbb{E}_\pi g$, if $\lambda(\pi, \omega_0) \neq g(\omega_0)$ and $\pi(\omega_0) \neq 0$ for some $\omega_0 \in \Omega$.) \square

From Theorem 9 we know that the conditions of the last lemma imply also that the game (Ω, Λ) is η -mixable. Let us show that the converse statement holds, i. e. the properness of λ and mixability are sufficient for the supermartingale property.

Lemma 11. *Suppose that the game (Ω, Λ) is η -mixable and $\lambda: \mathcal{P}(\Omega) \rightarrow \Sigma_\Lambda$ is a proper loss function with respect to Σ_Λ . Then the functions*

$$q_g(\pi, \omega) = e^{\eta(\lambda(\pi, \omega) - g(\omega))}$$

have the supermartingale property for every $g \in \Sigma_\Lambda$. If λ is continuous then q_g are forecast-continuous.

Proof. The forecast-continuity is obvious. Assume that the supermartingale property does not hold, in other words, that $E_\pi e^{\eta(\lambda(\pi, \cdot) - g)} = 1 + \delta$ for some $\pi \in \mathcal{P}(\Omega)$, $g \in \Sigma_\Lambda$ and $\delta > 0$. For any $\epsilon > 0$ consider the point

$$g_\epsilon = -\frac{1}{\eta} \ln \left((1 - \epsilon) e^{-\eta \lambda(\pi, \cdot)} + \epsilon e^{-\eta g} \right).$$

The point g_ϵ belongs to Σ_Λ^η by the definition of Σ_Λ^η , and $\Sigma_\Lambda^\eta = \Sigma_\Lambda$ since the game is η -mixable, that is, $g_\epsilon \in \Sigma_\Lambda$ for any $\epsilon > 0$. When $\epsilon \rightarrow 0$, we have

$$\begin{aligned} g_\epsilon &= \lambda(\pi, \cdot) - \frac{1}{\eta} \ln \left(1 + \epsilon \left(e^{\eta(\lambda(\pi, \cdot) - g)} - 1 \right) \right) \\ &= \lambda(\pi, \cdot) - \frac{\epsilon}{\eta} \left(e^{\eta(\lambda(\pi, \cdot) - g)} - 1 \right) + O(\epsilon^2). \end{aligned}$$

Taking the expectation E_π , we get

$$E_\pi g_\epsilon = E_\pi \lambda(\pi, \cdot) - \frac{\epsilon}{\eta} E_\pi \left(e^{\eta(\lambda(\pi, \cdot) - g)} - 1 \right) + O(\epsilon^2) = E_\pi \lambda(\pi, \cdot) - \frac{\epsilon \delta}{\eta} + O(\epsilon^2),$$

where $\delta > 0$ by our assumption. If ϵ is sufficiently small then $(\delta/\eta)\epsilon > O(\epsilon^2)$ and $E_\pi g_\epsilon < E_\pi \lambda(\pi, \cdot)$, which is impossible since $\lambda(\pi, \cdot)$ is proper with respect to Σ_Λ . \square

An alternative, more geometrical proof of the last lemma for binary games the reader can find in [5, Lemma 3].

3.5 The Realizability of the AA and Supermartingales

Theorem 9 shows that if the functions q_g defined by (13) are forecast-continuous and have the supermartingale property then the AA is realizable. We want to show the converse, that if the AA is realizable then one can find λ such that the functions q_g are forecast-continuous and have the supermartingale property. For mixable games, we know already that a proper loss function works (though we do not know yet whether a proper loss function exists). In this subsection we show that we can obtain λ in any game if we can construct continuous proper loss functions for mixable games. How to do the latter and when it is possible is discussed in the next subsection.

To state and prove the main result of this subsection, we need two standard assumptions (see [25]) about the game (Ω, Λ) and some additional notation.

Assumption 1. Λ is a compact subset of $[0, \infty]^\Omega$ (in the extended topology).

Assumption 2. There exists $g_{\text{fin}} \in \Lambda$ such that $g_{\text{fin}}(\omega) < \infty$ for all $\omega \in \Omega$.

Note that if Λ is compact then Σ_Λ is also compact, as well as Σ_Λ^η . A nice feature of compact prediction sets is that the properties of the game are determined by the boundary of the prediction set.

For any set $X \subseteq [0, \infty]^\Omega$, by $\mathcal{M}X$ denote the set of minimal elements of X : $g_0 \in \mathcal{M}X$ if and only if for any $g \in X$ the inequality $g_0 \geq g$ implies $g_0 = g$. For a compact set X , for every $g \in X$ there is an element $g_0 \in \mathcal{M}X$ such that $g_0 \leq g$; that is, $X \subseteq (\mathcal{M}X + [0, \infty]^\Omega)$. Notice that $\mathcal{M}X$ is contained in the boundary ∂X of X .

Since $\Sigma_\Lambda = \Sigma_\Lambda + [0, \infty]^\Omega = \Lambda + [0, \infty]^\Omega$, we have $\mathcal{M}\Sigma_\Lambda = \mathcal{M}\Lambda \subseteq \Lambda$. For compact Λ , we have $\Sigma_\Lambda = \mathcal{M}\Sigma_\Lambda + [0, \infty]^\Omega = \Sigma_{\mathcal{M}\Lambda} = \mathcal{M}\Lambda + [0, \infty]^\Omega$. Note also that a game is η -mixable if and only if $\mathcal{M}\Sigma_\Lambda^\eta \subseteq \Lambda$, since this is equivalent to $\Sigma_\Lambda^\eta = \Sigma_\Lambda$. A loss function is proper with respect to Σ_Λ^η if and only if it is proper with respect to $\mathcal{M}\Sigma_\Lambda^\eta$.

Lemma 12. *Suppose that the game (Ω, Λ) satisfies Assumptions 1 and 2 and the AA is (c, η) -realizable for this game. Then there is a continuous mapping $V: \Sigma_\Lambda^\eta \rightarrow \partial\Sigma_\Lambda$ such that $V(g) \leq cg$ for all $g \in \Sigma_\Lambda^\eta$.*

The proof is given in Appendix. The mapping V is actually the central projection from Σ_Λ^η into the superprediction set Σ_Λ (which contains $c\Sigma_\Lambda^\eta$ when the AA is (c, η) -realizable).

Theorem 13. *Let the game (Ω, Λ) satisfy Assumptions 1 and 2, the AA be (c, η) -realizable for this game, and $\lambda^\eta: \mathcal{P}(\Omega) \rightarrow \Sigma_\Lambda^\eta$ be a continuous proper loss function with respect to Σ_Λ^η . Then for any continuous $\lambda: \mathcal{P}(\Omega) \rightarrow \partial\Sigma_\Lambda$ such that $\lambda(\pi, \cdot) \leq c\lambda^\eta(\pi, \cdot)$ for all $\pi \in \mathcal{P}(\Omega)$, the functions q_g defined by (13) are forecast-continuous and have the supermartingale property for every $g \in \Sigma_\Lambda$; and there exists a continuous $\lambda: \mathcal{P}(\Omega) \rightarrow \partial\Sigma_\Lambda$ such that $\lambda(\pi, \cdot) \leq c\lambda^\eta(\pi, \cdot)$ for all $\pi \in \mathcal{P}(\Omega)$.*

Proof. The forecast-continuity is obvious. Let us check the supermartingale property, i. e., that

$$\mathbb{E}_\pi e^{\eta(\frac{\lambda(\pi, \cdot)}{c} - g)} \leq 1$$

for all $\pi \in \mathcal{P}(\Omega)$ and all $g \in \Sigma_\Lambda$. Since $\lambda(\pi, \cdot) \leq c\lambda^\eta(\pi, \cdot)$, it suffices that

$$\mathbb{E}_\pi e^{\eta(\lambda^\eta(\pi, \cdot) - g)} \leq 1,$$

which follows from Lemma 11 applied to the η -mixable game $(\Omega, \Sigma_\Lambda^\eta)$ and the proper function λ^η (note that $\Sigma_\Lambda \subseteq \Sigma_\Lambda^\eta$, hence the lemma works for all $g \in \Sigma_\Lambda$).

It remains to observe that $\lambda(\pi, \cdot) = V(\lambda^\eta(\pi, \cdot))$, where V is defined in Lemma 12, has the properties we need. \square

3.6 Construction of a Continuous Proper Loss Function

In this subsection, we fix a game (Ω, Λ) , fix $\eta > 0$, and consider proper loss functions with respect to Σ_Λ^η . They can be interpreted also as proper loss functions for the η -mixable game $(\Omega, \Sigma_\Lambda^\eta)$.

Lemma 14. *Let λ_1 and λ_2 be functions from $\mathcal{P}(\Omega)$ to Σ_Λ^η . Suppose that they are proper with respect to Σ_Λ^η at some point $\pi \in \mathcal{P}(\Omega)$, that is, $\mathbb{E}_\pi \lambda_i(\pi, \cdot) \leq \mathbb{E}_\pi g$, $i = 1, 2$, for all $g \in \Sigma_\Lambda^\eta$. Then for all $\omega \in \Omega$ we have*

$$\pi(\omega) \neq 0 \quad \Rightarrow \quad \lambda_1(\pi, \omega) = \lambda_2(\pi, \omega).$$

The proof of the lemma is given in Appendix.

Let $\mathcal{P}^\circ(\Omega)$ be the set of all non-degenerate distributions, i. e.

$$\mathcal{P}^\circ(\Omega) = \{\pi \in \mathcal{P}(\Omega) \mid \forall \omega \in \Omega \ \pi(\omega) > 0\}.$$

Lemma 14 implies that a proper loss function is uniquely defined on $\mathcal{P}^\circ(\Omega)$. The following lemma gives a more explicit specification of the values of a proper loss function on $\mathcal{P}^\circ(\Omega)$.

Lemma 15. *Let the game (Ω, Λ) satisfy Assumptions 1 and 2. Let us define function $H: \mathbb{R}^\Omega \rightarrow [-\infty, \infty)$ by the formula*

$$H(\pi) = \min_{g \in \Sigma_\Lambda^\eta} \mathbb{E}_\pi g. \quad (14)$$

Let \mathcal{H} be the domain where H is differentiable. Then $\mathcal{H} \supseteq \mathcal{P}^\circ(\Omega)$, and the components of the gradient of H at $\pi \in \mathcal{H} \cap \mathcal{P}(\Omega)$ constitute a continuous function $\lambda: \mathcal{H} \cap \mathcal{P}(\Omega) \rightarrow \Sigma_\Lambda^\eta$ such that $\mathbb{E}_\pi \lambda(\pi, \cdot) = H(\pi)$. Moreover, if $\pi \in \mathcal{P}^\circ(\Omega)$ then $\lambda(\pi, \cdot)$ is the unique point where the minimum in (14) is attained.

Remark 16. The function $H(\pi)$ for $\pi \in \mathcal{P}(\Omega)$ is known as the generalized entropy of the game (Ω, Λ) ; see [13]. For the logarithmic loss game, $H(\pi)$ becomes the Shannon entropy of π (cf. (16)). It is worth mentioning that one can reconstruct the superprediction set Σ_Λ from the generalized entropy of the game, and also from the predictive complexity of the game (see [18] for the definitions and proofs in the case of binary games).

The proof of the lemma is given in Appendix. The proof is based on the fact that the function $-H(\pi)$ is convex. Note that $\lambda(\pi, \cdot) \in \mathcal{M}\Sigma_\Lambda^\eta$ for any $\pi \in \mathcal{P}^\circ(\Omega)$. Indeed, if for some $\pi \in \mathcal{P}^\circ(\Omega)$ we have $\lambda(\pi, \cdot) \notin \mathcal{M}\Sigma_\Lambda^\eta$ then there exists $g \leq \lambda(\pi, \cdot)$, $g \in \mathcal{M}\Sigma_\Lambda^\eta$ and $g(\omega) < \lambda(\pi, \cdot)$ for at least one ω . Since $\pi(\omega) > 0$, we get $\mathbb{E}_\pi g < \mathbb{E}_\pi \lambda(\pi, \cdot) = H(\pi)$, which contradicts the definition of H .

Recall that if a loss function λ is proper with respect to Σ_Λ^η then $\mathbb{E}_\pi \lambda(\pi, \cdot) = H(\pi)$. Lemma 15 shows that on $\mathcal{P}^\circ(\Omega)$ a proper loss function λ exists and it is unique and continuous. Our next task is to extend λ continuously from $\mathcal{P}^\circ(\Omega)$ to $\mathcal{P}(\Omega)$. Unfortunately, this is sometimes impossible. Consider an example.

Let $\Omega = \{1, 2, 3\}$, and let the prediction set be

$$\Lambda = \{(-\ln p, -\ln(1-p), 1) \mid p \in [0, 1]\}.$$

Actually, this is the binary logarithmic loss game with an additional dummy outcome. This game is 1-mixable and $\Sigma_\Lambda^1 = \Sigma_\Lambda$. It is easy to check that the proper loss function with respect to Σ_Λ is given on $\mathcal{P}^\circ(\Omega)$ by the formulas $\lambda(\pi, i) = -\ln \frac{\pi(i)}{\pi(1)+\pi(2)}$, $i = 1, 2$, and $\lambda(\pi, 3) = 1$. This function can be extended continuously to all π such that $\pi(1) + \pi(2) \neq 0$, so we have $\lambda(\pi, \cdot) = (\infty, 0, 1)$ if $\pi(1) = 0$ and $\lambda(\pi, \cdot) = (0, \infty, 1)$ if $\pi(2) = 0$. However, these continuations are inconsistent at the point $\pi = (0, 0, 1)$. Therefore, there is no continuous function on $\mathcal{P}(\Omega)$ which is proper with respect to Σ_Λ for this game.

Now let us consider three examples of games where a continuous proper (and even strictly proper) loss function exists.

The first example is the Brier game (see [31]), which is a generalization of the square loss game:

$$\lambda^B(\pi, \omega) = \sum_{o \in \Omega} (\delta_\omega(o) - \pi(o))^2$$

where $\delta_\omega(o) = 1$ if $o = \omega$ and $\delta_\omega(o) = 0$ if $o \neq \omega$. For the binary game $\Omega = \{0, 1\}$, distribution $\pi \in \mathcal{P}(\Omega)$ is pair $(1 - p, p)$ where $p \in [0, 1]$, and hence $\lambda^B(\pi, \omega) = 2(p - \omega)^2$, which is twice the loss $\lambda^{\text{sq}}(p, \omega) = (p - \omega)^2$ in the binary square loss game as defined in Subsection 3.2.

The Brier game is 1-mixable, that is, $\Sigma_{\Lambda^B}^\eta = \Sigma_{\Lambda^B}$ for $\eta \leq 1$. Let us calculate $H(\pi)$ defined by (14) for $\pi \in \mathcal{P}(\Omega)$:

$$\begin{aligned} H^B(\pi) &= \min_{g \in \Sigma_{\Lambda^B}^\eta} \mathbb{E}_\pi g = \min_{g \in \Lambda^B} \mathbb{E}_\pi g = \min_{\pi' \in \mathcal{P}(\Omega)} \mathbb{E}_\pi \lambda^B(\pi', \cdot) \\ &= \min_{\pi' \in \mathcal{P}(\Omega)} \sum_{\omega \in \Omega} \pi(\omega) \sum_{o \in \Omega} (\delta_\omega(o) - \pi'(o))^2 \\ &= 1 - \sum_{\omega \in \Omega} \pi^2(\omega) + \min_{\pi' \in \mathcal{P}(\Omega)} \sum_{\omega \in \Omega} (\pi'(\omega) - \pi(\omega))^2 = 1 - \sum_{\omega \in \Omega} \pi^2(\omega). \end{aligned}$$

Clearly, $H^B(\pi)$ is differentiable on $\mathcal{P}(\Omega)$, hence a continuous proper loss function for the Brier game can be computed as the gradient of H^B by Lemma 15. However, it is easier to note that the minimum of $\mathbb{E}_\pi \lambda^B(\pi', \cdot)$ is attained at $\pi' = \pi$ only, and thus the standard form of the loss function λ^B is proper.

Remark 17. Note that in the example above we computed the value of $H(\pi)$ assuming that $\pi \in \mathcal{P}(\Omega)$. If we want to compute $\lambda(\pi, \omega)$ as the partial derivatives of $H(\pi)$ with respect to $\pi(\omega)$, we must consider $H(\pi)$ as a function on \mathbb{R}^Ω (as stated in Lemma 15). To this end, just note that H is homogeneous:

$$H(\pi) = H\left(\frac{\pi}{\sum_{\omega \in \Omega} \pi(\omega)}\right) \sum_{\omega \in \Omega} \pi(\omega) \quad (15)$$

for $\pi \in \mathbb{R}^\Omega$. In the Brier game example we have

$$H^B(\pi) = \left(1 - \frac{\sum_{\omega \in \Omega} \pi^2(\omega)}{(\sum_{\omega \in \Omega} \pi(\omega))^2}\right) \sum_{\omega \in \Omega} \pi(\omega),$$

and the partial derivatives are

$$1 - 2\pi(\omega) + \sum_{o \in \Omega} \pi^2(o) = \lambda^B(\pi, \omega)$$

for any $\pi \in \mathcal{P}(\Omega)$. In general, if we have a function $\phi: \mathbb{R}^\Omega \rightarrow \mathbb{R}$ such that $\phi(\pi) = H(\pi)$ for all $\pi \in \mathcal{P}(\Omega)$, taking the derivatives of (15) we get that the proper loss function λ can be computed by the following formula for any $\pi \in \mathcal{P}(\Omega)$:

$$\lambda(\pi, \omega) = \phi(\pi) - \sum_{o \in \Omega} \pi(o) \phi'_o(\pi) + \phi'_\omega(\pi),$$

where ϕ'_ω is the partial derivative of ϕ with respect to $\pi(\omega)$. This formula is known from the Savage theorem [20] (see also [12, Theorem 3.2]; recall that they consider scores, or gains, $-\lambda$ instead of losses λ).

The second example is the Hellinger game:

$$\lambda^H(\pi, \omega) = \frac{1}{2} \sum_{o \in \Omega} \left(\sqrt{\delta_\omega(o)} - \sqrt{\pi(o)}\right)^2.$$

Similarly to the Brier game, we can find that

$$H^H(\pi) = \min_{\pi' \in \mathcal{P}(\Omega)} \sum_{\omega \in \Omega} \pi(\omega) \left(1 - \sqrt{\pi'(\omega)}\right) = \sum_{\omega \in \Omega} \pi(\omega) - \sqrt{\sum_{\omega \in \Omega} \pi^2(\omega)}.$$

Here the minimum is not attained at $\pi' = \pi$ and λ^H is not proper. Taking the derivatives, we find a proper loss function for the Hellinger game:

$$\lambda(\pi, \omega) = 1 - \frac{\pi(\omega)}{\sqrt{\sum_{\omega \in \Omega} \pi^2(\omega)}}.$$

This loss function is known as the spherical loss.

The spherical loss and the Hellinger loss specify the same game but under different parameterization. For binary games, this kind of “reparameterization” was considered in [14, Section 3.1], where a proper function $\lambda(\pi, \cdot)$ was called a Bayes-optimal prediction for bias π . More precisely, the paper [14] discusses binary games specified by a loss function $\lambda(\gamma, \omega)$, where ω is 0 or 1 and $\gamma \in [0, 1]$. Their Lemma 3.5 states conditions (on derivatives of λ as a function of γ) when there exists a unique γ_p that minimizes $(1-p)\lambda(\gamma, 0) + p\lambda(\gamma, 1)$ for each $p \in [0, 1]$. This γ_p can be obtained from Equation (3.8) in [14]:

$$(1-p) \frac{d}{d\gamma} \lambda(\gamma, 0) \Big|_{\gamma=\gamma_p} + p \frac{d}{d\gamma} \lambda(\gamma, 1) \Big|_{\gamma=\gamma_p} = 0.$$

Our Lemma 15 can be regarded as a generalization of this approach.

Our third example is the general logarithmic loss game defined by

$$\lambda^{\log}(\pi, \omega) = -\ln \pi(\omega).$$

Similarly to the Brier loss function, the logarithmic loss function is strictly proper. Indeed, let us calculate the entropy H^{\log} for $\pi \in \mathcal{P}(\Omega)$:

$$\begin{aligned} H^{\log}(\pi) &= \min_{\pi' \in \mathcal{P}(\Omega)} \sum_{\omega \in \Omega} \pi(\omega) (-\ln \pi'(\omega)) \\ &= -\sum_{\omega \in \Omega} \pi(\omega) \ln \pi(\omega) - \max_{\pi' \in \mathcal{P}(\Omega)} \sum_{\omega \in \Omega} \pi(\omega) \ln \frac{\pi'(\omega)}{\pi(\omega)} = -\sum_{\omega \in \Omega} \pi(\omega) \ln \pi(\omega). \end{aligned} \quad (16)$$

Here the partial derivatives are infinite at the bound of $\mathcal{P}(\Omega)$. Nevertheless, it is easy to check that the minimum in the definition of $H^{\log}(\pi)$ is always attained at one point $\pi' = \pi$ only. The last equality in (16) holds since logarithm is concave $\sum_{\omega \in \Omega} \pi(\omega) \ln \frac{\pi'(\omega)}{\pi(\omega)} \leq \ln \left(\sum_{\omega \in \Omega} \pi(\omega) \frac{\pi'(\omega)}{\pi(\omega)} \right) = 0$ and the inequality is strict unless $\frac{\pi'(\omega)}{\pi(\omega)}$ are equal for all $\omega \in \Omega$ or $\pi(\omega_0) = 1$ for some ω_0 . In the former case, $\pi = \pi'$, since $\pi, \pi' \in \mathcal{P}(\Omega)$. In the latter case, we get $\max_{\pi' \in \mathcal{P}(\Omega)} \ln \pi'(\omega_0)$, which is attained if $\pi'(\omega_0) = 1$, and hence $\pi = \pi'$ too.

Now we consider a general way to construct proper loss functions, even in the case when H is not differentiable on all $\mathcal{P}(\Omega)$. Note that the only way to extend λ continuously is to define it at $\mathcal{P}(\Omega) \setminus \mathcal{P}^\circ(\Omega)$ as a limit from $\mathcal{P}^\circ(\Omega)$, where $\lambda(\pi, \cdot)$ is defined as a point of minimum. The following lemma proved in Appendix states that a limit of such points is again a point of minimum.

Lemma 18. *Let $\pi_i \in \mathcal{P}(\Omega)$ and $\gamma_i \in \mathcal{M}\Sigma_\Lambda^\eta$ be such that $E_{\pi_i} \gamma_i = \min_{g \in \Sigma_\Lambda^\eta} E_{\pi_i} g$, $i = 1, 2, \dots$. Assume that $\pi_i \rightarrow \pi$ and $\gamma_i \rightarrow \gamma$ as $i \rightarrow \infty$. Then $\gamma \in \mathcal{M}\Sigma_\Lambda^\eta$ and $E_\pi \gamma = \min_{g \in \Sigma_\Lambda^\eta} E_\pi g$.*

In particular, the lemma implies that a continuous proper loss function exists in games where each minimum is attained in a unique point. Let us formulate this assumption explicitly and prove the existence theorem.

Assumption 3. For every $\pi \in \mathcal{P}(\Omega)$ such that $\pi(\omega_1) = 0$ and $\pi(\omega_2) = 0$ for some $\omega_1, \omega_2 \in \Omega$, $\omega_1 \neq \omega_2$, there exists only one point where the minimum of $E_\pi g$ over all $g \in \mathcal{M}\Sigma_\Lambda^\eta$ is attained.

Remark 19. Assumption 3 holds automatically for all binary games. The games with differentiable H , such as the general square loss game, satisfy Assumption 3 as well.

Theorem 20. *Suppose that the game (Ω, Λ) satisfies Assumptions 1 and 2, and Assumption 3 for certain $\eta > 0$. Then there exists a continuous loss function $\lambda: \mathcal{P}(\Omega) \rightarrow \mathcal{M}\Sigma_\Lambda^\eta$ that is proper, and even strictly proper, with respect to Σ_Λ^η .*

Proof. Let us show first that the minimum of $E_\pi g$ over all $g \in \mathcal{M}\Sigma_\Lambda^\eta$ is attained at one point only for all $\pi \in \mathcal{P}(\Omega)$. For $\pi \in \mathcal{P}^\circ(\Omega)$, it follows from Lemma 14. Let $\pi \in \mathcal{P}(\Omega)$ be such that $\pi(\omega_0) = 0$ for some $\omega_0 \in \Omega$ and $\pi(\omega) \neq 0$ for $\omega \neq \omega_0$. Let $g_1, g_2 \in \mathcal{M}\Sigma_\Lambda^\eta$ be any two points of minimum. Again by Lemma 14, $g_1(\omega) = g_2(\omega)$ for all $\omega \neq \omega_0$. Therefore $g_1 \leq g_2$ or $g_1 \geq g_2$ (since $g_1(\omega_0)$ and $g_2(\omega_0)$ are comparable, being two reals), and the greater of them cannot belong to $\mathcal{M}\Sigma_\Lambda^\eta$. Thus, $g_1 = g_2$. Assumption 3 works for all other $\pi \in \mathcal{P}(\Omega)$.

Let us take $\lambda(\pi, \cdot) = \arg \min_{g \in \mathcal{M}\Sigma_\Lambda^\eta} E_\pi g$ for all $\pi \in \mathcal{P}(\Omega)$. Clearly, λ is proper with respect to Σ_Λ^η (recall that every point in Σ_Λ^η is minorized by some point in $\mathcal{M}\Sigma_\Lambda^\eta$). Let us prove continuity. Take any converging sequence $\pi_i \in \mathcal{P}(\Omega)$, let π be its limit, and consider the corresponding $\lambda(\pi_i, \cdot)$. Lemma 18 implies that all accumulation points of the set $\{\lambda(\pi_i, \cdot)\}$ are points where $\min_{g \in \mathcal{M}\Sigma_\Lambda^\eta} E_\pi g$ is attained, therefore $\lambda(\pi, \cdot)$ is the only accumulation point and $\lambda(\pi_i, \cdot)$ converges to $\lambda(\pi, \cdot)$. \square

3.7 Defensive Forecasting Revisited

Let us review the results we obtained so far. Theorems 1 and 5 gives us the same loss bound for a game (Ω, Λ) , if the AA is realizable and if Q defined by (7) is a forecast-continuous supermartingale, respectively.

The algorithms are very close in their internal structure. We can say even more: with the same parameters and inputs, they give the same predictions, in some sense. More precisely, two sets coincide: the set of $\gamma_N \in \Lambda$ satisfying (1) and the set of $\gamma_N \in \Lambda$ such that γ_N minorizes $\lambda(\pi_N, \cdot)$ for π_N satisfying (9).

Both algorithms are applicable under almost the same conditions: Theorem 9 says that if Q is a forecast-continuous supermartingale then the AA is realizable; Theorems 13 and 20 show the converse for games satisfying Assumptions 1–3.

Whereas Assumptions 1 and 2 are standard and natural, and the AA is usually considered only for the games satisfying these assumptions, Assumption 3 is new and quite cumbersome. However, it turns out that with the help of a more

complicated version of the DFA we can get rid of Assumption 3 and get a perfect equivalence between the realizability of the AA and some supermartingale condition (under the standard Assumptions 1 and 2 only).

To begin with, let us slightly relax the definitions concerning supermartingales. We say that a function $q: \mathcal{P}^\circ(\Omega) \times \Omega \rightarrow \mathbb{R}$ has the *supermartingale property on $\mathcal{P}^\circ(\Omega)$* if for any $\pi \in \mathcal{P}^\circ(\Omega)$

$$\mathbb{E}_\pi q(\pi, \cdot) \leq 1.$$

The function q is *forecast-continuous on $\mathcal{P}^\circ(\Omega)$* if for every $\omega \in \Omega$ it is continuous as the function of π for all $\pi \in \mathcal{P}^\circ(\Omega)$.

Lemma 21. *Let a function $q: \mathcal{P}^\circ(\Omega) \times \Omega \rightarrow \mathbb{R}$ be non-negative and forecast-continuous on $\mathcal{P}^\circ(\Omega)$. Suppose that for all $\pi \in \mathcal{P}^\circ(\Omega)$ it holds that*

$$\mathbb{E}_\pi q(\pi, \cdot) \leq C,$$

where $C \in [0, \infty)$ is some constant. Then there exists a sequence $\{\pi^{(i)}\}_{i \in \mathbb{N}}$ such that $\pi^{(i)} \in \mathcal{P}^\circ(\Omega)$, the sequence $\pi^{(i)}$ converges in $\mathcal{P}(\Omega)$, the sequences $q(\pi^{(i)}, \omega)$ converge for every $\omega \in \Omega$, and

$$\forall \omega \in \Omega \quad \lim_{i \rightarrow \infty} q(\pi^{(i)}, \omega) \leq C.$$

The proof of the lemma is given in Appendix after the proof of Lemma 8.

Theorem 22. *Let the game (Ω, Λ) satisfy Assumptions 1 and 2. The AA is (c, η) -realizable for this game if and only if there exists λ such that the functions q_g defined by (13) are forecast-continuous on $\mathcal{P}^\circ(\Omega)$ and have the supermartingale property on $\mathcal{P}^\circ(\Omega)$ for all $g \in \Sigma_\Lambda$.*

Proof. The “only if” part easily follows from Lemma 15 combined with (the proof of) Theorem 13.

The “if” part is analogous to Theorem 9. We need to prove inequality (3) for any finite $G \subseteq \Sigma_\Lambda$ and for any distribution ρ on G . Let us consider the function

$$q(\pi, \omega) = \sum_{g \in G} \rho(g) q_g(\pi, \omega),$$

which is non-negative, forecast-continuous on $\mathcal{P}^\circ(\Omega)$, and has the supermartingale property on $\mathcal{P}^\circ(\Omega)$. By Lemma 21 applied to this q and $C = 1$, there exist $\pi^{(i)} \in \mathcal{P}^\circ(\Omega)$ such that

$$\forall \omega \in \Omega \quad \lim_{i \rightarrow \infty} \sum_{g \in G} \rho(g) \exp \left(\eta \left(\frac{\lambda(\pi^{(i)}, \omega)}{c} - g(\omega) \right) \right) \leq 1.$$

Let $\gamma^{(i)} = \lambda(\pi^{(i)}, \cdot) \in \Sigma_\Lambda$. Since Σ_Λ is compact (by Assumption 1), the sequence $\gamma^{(i)}$ contains a convergent subsequence; let $\gamma \in \Sigma_\Lambda$ be its limit. Then $\sum_{g \in G} \rho(g) \exp(\eta(\gamma/c - g))$ is a limit of the corresponding convergent subsequence of the sequence $\sum_{g \in G} \rho(g) \exp(\eta(\gamma^{(i)}/c - g))$, and for every $\omega \in \Omega$ we get inequality (3):

$$\sum_{g \in G} \rho(g) \exp \left(\eta \left(\frac{\gamma(\omega)}{c} - g(\omega) \right) \right) \leq 1.$$

□

Let us state also the algorithm DFA*, a variant of the DFA suitable for supermartingales on $\mathcal{P}^\circ(\Omega)$. At step N , the DFA* defines the function

$$q(\pi, \omega) = \sum_{\theta \in \Theta} P_0(\theta) \exp \left(\eta \sum_{n=1}^{N-1} \left(\frac{\gamma_n(\omega_n)}{c} - \gamma_n^\theta(\omega_n) \right) \right) \\ \times \exp \left(\eta \left(\frac{\lambda(\pi, \omega)}{c} - \gamma_N^\theta(\omega) \right) \right)$$

and chooses any sequence of $\pi^{(i)} \in \mathcal{P}^\circ(\Omega)$ such that

$$\forall \omega \in \Omega \quad \lim_{i \rightarrow \infty} q(\pi^{(i)}, \omega) \leq 1.$$

Then the algorithm chooses as γ the limit of any convergent subsequence of the sequence $\lambda(\pi^{(i)}, \cdot)$, and announces $\gamma_N = \sigma(\gamma)$ as Learner's prediction. It is clear that the DFA* guarantees the same loss bound as Theorem 5.

It is important for applications that the AA is rather efficient computationally (though it is more complicated than some other algorithms). The DFA* is designed to obtain a nice theory, and it makes little sense to discuss its efficiency. The DFA is much more practical than the DFA*. Unfortunately, the DFA seems to be less practical than the AA. Its main step hidden in the proof of Lemma 8 requires finding a fixed point (or a minimax), which is generally a hard task (PPAD-complete). For binary games, however, the fixed points can be found by bisection method, which gives us a not so inefficient implementation of the DFA. Some tricks can also help for games with three outcomes.

Remark 23. After this paper had been finished, the authors have discovered another way to deal with games that do not satisfy Assumption 3. The idea is to consider a multivalued loss function: to every π it assigns all points where the minimum of $E_\pi g$ is attained. The definition of supermartingale should be modified accordingly, and a variant of Lemma 4 can be proved for such multivalued supermartingales. The details will be added later or published elsewhere.

3.8 On Continuous Outcomes

We assumed so far that the space of outcomes, Ω , is finite. However, it is often natural to consider a continuous space of outcomes. For example, for the square loss function $\lambda^{\text{sq}}(p, \omega) = (p - \omega)^2$, one can take $\omega \in [0, 1]$ instead of $\omega \in \{0, 1\}$.

In this subsection we consider one important case of continuous outcome spaces: a finite-dimensional simplex. We will consider a simplex as the space $\mathcal{P}(\Omega)$ of distributions on some finite Ω . A game of prediction is a pair $(\mathcal{P}(\Omega), \Lambda)$, where $\Lambda \subseteq [0, \infty]^{\mathcal{P}(\Omega)}$; predictions are functions $\gamma: \mathcal{P}(\Omega) \rightarrow [0, \infty]$; the protocol is the same. Each game of prediction with the outcomes from a simplex $\mathcal{P}(\Omega)$ can be restricted to a game on Ω : we identify each $\omega \in \Omega$ with the distribution δ_ω concentrated on this ω . Thus we may assume $\mathcal{P}(\Omega) \supset \Omega$. Denote by $\Lambda_\Omega \subseteq [0, \infty]^\Omega$ the set of functions from Λ restricted to Ω .

We will show how the supermartingale technique works for games having some regularity property. (A similar extension for the AA is discussed in [14, Section 4.1].)

To motivate this kind of property, let us start from the other side and assume that we have a prediction (recall that our prediction is a vector of our losses for every possible outcome) γ defined on Ω and want to extend it to $\mathcal{P}(\Omega)$. The most natural way to do this is to say that an element of $\mathcal{P}(\Omega)$ is just a probability distribution on the outcomes, and consider the expected loss with respect to this distribution, that is, $\gamma(p) := \mathbb{E}_p \gamma$ for every $p \in \mathcal{P}(\Omega)$. It is also natural to expect that having this property one should be able to transfer a regret bound from the game on Ω to the respective game on $\mathcal{P}(\Omega)$. However, the equality $\gamma(p) = \mathbb{E}_p \gamma$ is too restrictive. For example, it does not hold for the square loss. At the same time, what does hold for the square loss (and will be checked later) is an equality concerning the difference of two predictions: $\gamma_1(p) - \gamma_2(p) = \mathbb{E}_p(\gamma_1 - \gamma_2)$. This is quite natural in our context, since the difference is a regret, loosely speaking, and a regret is the value we are optimizing. This leads to the following requirement (formally weaker than the condition for the square loss).

We say that $\Lambda \subseteq [0, \infty]^{\mathcal{P}(\Omega)}$ has the *relative exp-convexity property* for certain c and η if for all $\gamma_1, \gamma_2 \in \Lambda$ and for all $p \in \mathcal{P}(\Omega)$ it holds that

$$\exp\left(\eta\left(\frac{\gamma_1(p)}{c} - \gamma_2(p)\right)\right) \leq \sum_{\omega \in \Omega} p(\omega) \exp\left(\eta\left(\frac{\gamma_1(\omega)}{c} - \gamma_2(\omega)\right)\right).$$

Remark 24. The relative exp-convexity property for any $c > 0$ and η follows from

$$\forall \gamma \in \Lambda \quad \forall p \in \mathcal{P}(\Omega) \quad \gamma(p) = \sum_{\omega \in \Omega} p(\omega) \gamma(\omega)$$

due to convexity of the exponent function. For $c = 1$ and any η , it follows also from

$$\forall \gamma_1, \gamma_2 \in \Lambda \quad \forall p \in \mathcal{P}(\Omega) \quad \gamma_1(p) - \gamma_2(p) = \sum_{\omega \in \Omega} p(\omega) (\gamma_1(\omega) - \gamma_2(\omega)).$$

Let $\sigma_\Omega: \Lambda_\Omega \rightarrow \Lambda$ be any mapping inverse to the restriction from Λ to Λ_Ω , that is, for any $\gamma \in \Lambda_\Omega$, the function $\sigma_\Omega(\gamma) \in \Lambda$ restricted to Ω is γ . Such a mapping exists since every element of Λ_Ω is a restriction of some element of Λ .

Theorem 25. *For a game $(\mathcal{P}(\Omega), \Lambda)$, suppose that Λ has the relative exp-convexity property for some $c \geq 1$ and $\eta > 0$. For the restricted game (Ω, Λ_Ω) , suppose that for some $\lambda: \mathcal{P}(\Omega) \rightarrow \Sigma_{\Lambda_\Omega}$, the functions q_g defined by (13) are forecast-continuous and have the supermartingale property for all $g \in \Sigma_{\Lambda_\Omega}$. Let $\sigma: \Sigma_{\Lambda_\Omega} \rightarrow \Lambda_\Omega$ be a substitution function (that is, $\sigma(g) \leq g$ for all $g \in \Sigma_{\Lambda_\Omega}$). Then for the game $(\mathcal{P}(\Omega), \Lambda)$ there is Learner's strategy (in fact, a variant of the DFA) with parameters $c, \eta, \lambda, P_0, \sigma$, and σ_Ω guaranteeing that, at each step N and for all experts θ , it holds*

$$L_N \leq cL_N^\theta + \frac{c}{\eta} \ln \frac{1}{P_0(\theta)}.$$

Proof. Assume that we are at step N and need to announce the next prediction. Let $\gamma_n^\theta \in \Lambda$, $n = 1, \dots, N$ be the experts' prediction up to step N , γ_n , $n = 1, \dots, N-1$ be the Learner's previous predictions, and p_n , $n = 1, \dots, N-1$ be the previous outcomes. Define the function Q_{N-1} from Θ to \mathbb{R}

$$Q_{N-1}(\theta) = \prod_{n=1}^{N-1} \exp\left(\eta\left(\frac{\gamma_n(p_n)}{c} - \gamma_n^\theta(p_n)\right)\right)$$

and consider the following function on $\mathcal{P}(\Omega) \times \Omega$:

$$q_N(\pi, \omega) = \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta) \times \exp \left(\eta \left(\frac{\lambda(\pi, \omega)}{c} - \gamma_N^\theta(\omega) \right) \right).$$

Due to the assumptions about the last multiplier, q_N is forecast-continuous and $E_\pi q_N(\pi, \cdot) \leq \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta)$. By Lemma 8, we can find $\pi_N \in \mathcal{P}(\Omega)$ such that for all $\omega \in \Omega$

$$q_N(\pi_N, \omega) \leq \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta).$$

The prediction of the strategy is $\gamma_N = \sigma_\Omega(\sigma(\lambda(\pi_N, \cdot))) \in \Lambda$.

Let $p_N \in \mathcal{P}(\Omega)$ be the outcome at step N . The relative exp-convexity property implies that

$$\exp \left(\eta \left(\frac{\gamma_N(p_N)}{c} - \gamma_N^\theta(p_N) \right) \right) \leq \sum_{\omega \in \Omega} p_N(\omega) \exp \left(\eta \left(\frac{\gamma_N(\omega)}{c} - \gamma_N^\theta(\omega) \right) \right).$$

We have $\gamma_N(\omega) = \sigma(\lambda(\pi_N, \cdot))(\omega)$ by the definition of σ_Ω , hence we have $\gamma_N(\omega) \leq \lambda(\pi_N, \omega)$ by definition of σ . Thus,

$$\begin{aligned} \sum_{\theta \in \Theta} P_0(\theta) Q_N(\theta) &= \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta) \times \exp \left(\eta \left(\frac{\gamma_N(p_N)}{c} - \gamma_N^\theta(p_N) \right) \right) \\ &\leq \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta) \times \sum_{\omega \in \Omega} p_N(\omega) \exp \left(\eta \left(\frac{\lambda(\pi_N, \omega)}{c} - \gamma_N^\theta(\omega) \right) \right) \\ &= \sum_{\omega \in \Omega} p_N(\omega) q_N(\pi_N, \omega) \leq \sum_{\theta \in \Theta} P_0(\theta) Q_{N-1}(\theta), \end{aligned}$$

and the loss bound follows as usual. \square

As an example, let us again consider the Brier game (the general square loss function), now with distributions as outcomes: Ω is a finite non-empty set, outcomes p are from $\mathcal{P}(\Omega)$, and the loss of decision $\pi \in \mathcal{P}(\Omega)$ for outcome p is

$$\lambda^B(\pi, p) = \sum_{\omega \in \Omega} (p(\omega) - \pi(\omega))^2.$$

It is easy to check that this game has the relative exp-convexity property for $c = 1$ and any η due to Remark 24:

$$\begin{aligned} \sum_{\omega \in \Omega} p(\omega) (\lambda^B(\pi_1, \omega) - \lambda^B(\pi_2, \omega)) \\ &= \sum_{\omega \in \Omega} (\pi_1^2(\omega) - \pi_2^2(\omega)) + 2 \sum_{\omega \in \Omega} p(\omega) (\pi_2(\omega) - \pi_1(\omega)) \\ &= \lambda^B(\pi_1, p) - \lambda^B(\pi_2, p). \end{aligned}$$

Another important example is the Kullback-Leibler game (its restricted version is the logarithmic loss game):

$$\lambda^{KL}(\pi, p) = \sum_{\omega \in \Omega} p(\omega) \ln \frac{p(\omega)}{\pi(\omega)}.$$

This game also has the relative exp-convexity property for $c = 1$ and any η : $\lambda^{KL}(\pi_1, p) - \lambda^{KL}(\pi_2, p) = \sum_{\omega \in \Omega} p(\omega) (\lambda^{KL}(\pi_1, \omega) - \lambda^{KL}(\pi_2, \omega))$.

4 Second-Guessing Experts

In this section, we apply the supermartingale technique and the DFA to a new variant of the prediction with expert advice setting. Protocol 2 is an extension of Protocol 1, where the game is specified by the same elements (Ω, Λ) as before, but the Experts have a new power.

Protocol 2 Prediction with Second-Guessing Expert Advice

$L_0 := 0$.
 $L_0^\theta := 0$, for all $\theta \in \Theta$.
for $n = 1, 2, \dots$ **do**
 All Experts $\theta \in \Theta$ announce $\Gamma_n^\theta: \Lambda \rightarrow \Lambda$.
 Learner announces $\gamma_n \in \Lambda$.
 Reality announces $\omega_n \in \Omega$.
 $L_n := L_{n-1} + \gamma_n(\omega_n)$.
 $L_n^\theta := L_{n-1}^\theta + \Gamma_n^\theta(\gamma_n, \omega_n)$, for all $\theta \in \Theta$.
end for

The new protocol contains only one substantial change. Every Expert θ announces a function Γ^θ from Λ to Λ instead of an element of Λ (to simplify notation, we consider Γ also as a function from $\Lambda \times \Omega$ to $[0, \infty]$, as we did with the proper loss functions λ). Informally speaking, now an expert's opinion is not a prediction, but a conditional statement that specifies the actual prediction depending on Learner's next step. Therefore, the loss of each expert is determined by the Learner's prediction as well as by the outcome chosen by Reality. We will call the experts in Protocol 2 *second-guessing experts*. Second-guessing experts are a generalization of experts in the standard Protocol 1: a standard expert can be interpreted in Protocol 2 as a constant function.

The phenomenon of "second-guessing experts" occurs, for example, in real-world finance. In particular, commercial banks serve as "second-guessing experts" for the central bank when they use variable interest rates (that is, the interest rate for the next period is announced not as a fixed value but as an explicit function of the central bank base rate).

In game theory, the notion of internal regret [9, 3, 22, 23] is somewhat related to the idea of second-guessing experts. The internal regret appears in the framework where for each prediction, which is called action in that context, there is an expert that consistently recommends this action, and Learner follows one of the experts at each step. The internal regret for a pair of experts (i, j) shows by how much Learner could have decreased his loss if he had followed expert j each time he followed expert i . This can be modeled by a second-guessing expert that "adjusts" Learner's predictions: agrees with Learner if Learner does not follow i , and recommends following j when the Learner follows i .

The internal regret is usually studied in randomized prediction protocols. In the case of deterministic Learner's predictions, one cannot hope to get any interesting loss bound without additional assumptions. Indeed, Experts can always suggest exactly the "opposite" to the Learner's prediction (for example, in the log loss game, they predict 1 if Learner predicts p_n ("the probability of 1") less than 0.5 and they predict 0 otherwise), and Reality can "agree" with them (choosing the outcome equal to Experts' prediction); then the Experts' losses remain zero, but the Learner's loss grows linearly in the number of steps.

A non-trivial bound is possible if Learner is allowed to give predictions in the form of a distribution on Experts. This can be formalized as the Freund-Schapire game [25, Example 7]. Then the second-guessing expert modeling an internal regret is a continuous transformation of the distribution given by Learner. The results of [3] and others are bounds of the form $L_N \leq L_N^\theta + O(\sqrt{N})$ for the Freund-Schapire game, which is non-mixable. A discussion of bounds of this form achievable by the defensive forecasting method will be published elsewhere: in this paper we consider another kind of bounds. However, here we will also make the assumption that second-guessing experts modify the prediction of Learner *continuously*.

4.1 The DFA for Second-Guessing Experts

First consider the case when Γ_n^θ are continuous mappings from Λ to Λ . The DFA requires virtually no modifications for this task and gives the same loss bounds as in Theorem 5.

Theorem 26. *Suppose that for some c, η , and some continuous $\lambda: \mathcal{P}(\Omega) \rightarrow \Lambda$ the functions q_g defined by (13) are forecast-continuous and have the supermartingale property for all $g \in \Lambda$. Then for the game following the protocol of prediction with second-guessing expert advice where all experts θ at all steps n announce continuous functions $\Gamma_n^\theta: \Lambda \rightarrow \Lambda$, there is Learner's strategy (in fact, the DFA applied to Q^{P_0} defined by (17)) with parameters c, η, λ, P_0 , (where P_0 is a distribution on Θ) guaranteeing that, at each step N and for all experts θ , it holds*

$$L_N \leq cL_N^\theta + \frac{c}{\eta} \ln \frac{1}{P_0(\theta)}.$$

Proof. For any continuous $\Gamma: \Lambda \rightarrow \Lambda$ consider the function

$$\tilde{q}_\Gamma(\pi, \omega) = \exp \left(\eta \left(\frac{\lambda(\pi, \omega)}{c} - \Gamma(\lambda(\pi, \cdot), \omega) \right) \right).$$

It is forecast-continuous as a composition of continuous functions, and has the supermartingale property since for any $\pi \in \mathcal{P}(\Omega)$, taking $g = \Gamma(\lambda(\pi, \cdot))$ we have $E_\pi \tilde{q}_\Gamma = E_\pi q_g \leq 1$. Similarly to (8), define Q^{P_0} on $((C(\Lambda \rightarrow \Lambda))^\Theta \times \mathcal{P}(\Omega) \times \Omega)^*$, where $C(\Lambda \rightarrow \Lambda)$ is the set of continuous functions on Λ , by the formula

$$Q^{P_0}(\{\Gamma_1^\theta\}_{\theta \in \Theta}, \pi_1, \omega_1, \dots, \{\Gamma_N^\theta\}_{\theta \in \Theta}, \pi_N, \omega_N) = \sum_{\theta \in \Theta} P_0(\theta) \prod_{n=1}^N \exp \left(\eta \left(\frac{\lambda(\pi_n, \omega_n)}{c} - \Gamma_n^\theta(\lambda(\pi_n, \cdot), \omega_n) \right) \right). \quad (17)$$

As in Theorem 5, Q^{P_0} is a forecast-continuous supermartingale.

At step N , the strategy chooses any π_N satisfying (9) and announces $\gamma_N = \lambda(\pi_N, \cdot)$ as Learner's prediction (we do not need a substitution function here since the range of λ is in Λ by the theorem assumption). The loss bound

follows, since

$$\begin{aligned} \exp\left(\eta \sum_{n=1}^N \left(\frac{\lambda(\pi_n, \omega_n)}{c} - \Gamma_n^\theta(\gamma_n, \omega_n)\right)\right) &= \\ \exp\left(\eta \sum_{n=1}^N \left(\frac{\lambda(\pi_n, \omega_n)}{c} - \Gamma_n^\theta(\lambda(\pi_n, \cdot), \omega_n)\right)\right) &\leq \frac{1}{P_0(\theta)}. \end{aligned}$$

□

Recall that Theorem 20 provides us (under Assumptions 1–3) with a continuous proper loss function $\lambda: \mathcal{P}(\Omega) \rightarrow \mathcal{M}\Sigma_\Lambda^\eta$. For any η -mixable game, we have $\mathcal{M}\Sigma_\Lambda^\eta \subseteq \Lambda$, and due to Theorem 13 we can take this λ and get forecast-continuous q_g with the supermartingale property.

For non-mixable games there is no guarantee that such λ exists. Theorem 13 gives a function λ ranging over $\partial\Sigma_\Lambda$ (the boundary of the superpredictions set Σ_Λ), which is not necessarily contained in Λ . Moreover, it may happen that even for continuous experts $\Gamma_n^\theta: \Lambda \rightarrow \Lambda$ it is impossible to get any interesting loss bound, for any strategy. Indeed, consider a game where Λ is not connected (e.g., the simple prediction game [25, Example 1] with $\Lambda = \{(0, 1), (1, 0)\}$). Then the example with “opposite” predictions works: the experts just need to map Learner’s predictions into another connected component.

By this reason, let us consider a modification of Protocol 2 that changes the sets of predictions allowed for Learner and for Experts. Namely, for the game (Ω, Λ) , Experts $\theta \in \Theta$ announce $\Gamma_n^\theta: \partial\Sigma_\Lambda \rightarrow \Sigma_\Lambda$, and Learner announces $\gamma_n \in \partial\Sigma_\Lambda$ (the rest of Protocol 2 does not change). We will assume that the game satisfies Assumptions 1 and 2 (for non-compact Λ the boundary $\partial\Sigma_\Lambda$ may be empty). Then the modified protocol usually gives more freedom to Learner: since $\mathcal{M}\Lambda \subseteq \partial\Sigma_\Lambda$, the predictions in $\Lambda \setminus \partial\Sigma_\Lambda$ are minorized by some better predictions in $\mathcal{M}\Lambda$. The Experts are allowed to give predictions (which are $\Gamma_n^\theta(\gamma_n)$) in a larger set Σ_Λ , however, they need to cope with Learner predictions from a larger set too.

For the modified protocol, Theorem 26 holds with minimal changes: λ is allowed to range over Σ_Λ instead of Λ , the functions q_g have the supermartingale property for all $g \in \Sigma_\Lambda$ (instead of $g \in \Lambda$ only), Γ_n^θ are continuous functions from $\partial\Sigma_\Lambda$ to Σ_Λ ; the proof does not change. Theorem 13 provides us with λ such that q_g have the required properties.

4.2 The AA for Second-Guessing Experts

In contrast to the DFA, the AA cannot be applied to the second-guessing protocol in a straightforward way. However, the AA can be modified for this case. Recall that the AA is based on the inequality (1), which is already solved for γ_N . In the second-guessing protocol, both sides of this inequality will contain γ_N :

$$\gamma_N(\omega_N) \leq -\frac{c}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta \in \Theta} P_{N-1}(\theta)} \exp(-\eta \Gamma_N^\theta(\gamma_N, \omega_N)) \right).$$

The DFA implicitly solves this inequality in (the proof of) Lemma 4, using a kind of fixed point theorem. We will present a modification of the AA which uses a fixed point theorem explicitly.

A topological space X has the *fixed point property* if every continuous function $f: X \rightarrow X$ has a fixed point, that is, $\exists x \in X f(x) = x$.

Let us show that if the game (Λ, Ω) satisfies Assumptions 1 and 2 then the set $\mathcal{M}\Sigma_\Lambda^\eta$ (the set of minimal points of Σ_Λ^η) has the fixed point property for any $\eta > 0$. First consider the homeomorphism from $[0, \infty]^\Omega$ to $[0, 1]^\Omega$ that maps $g \mapsto \exp(-\eta g)$. As mentioned in Section 2, the set $\exp(-\eta\Sigma_\Lambda^\eta)$ is convex. It is non-empty due to Assumption 2 and compact due to Assumption 1. Thus, $\exp(-\eta\Sigma_\Lambda^\eta)$ has the fixed point property by [1, Theorem 4.10], and Σ_Λ^η has the property as its homeomorphic image [1, Theorem 4.1]. Now we need the following technical lemma proved in Appendix.

Lemma 27. *There is a continuous mapping $F: \Sigma_\Lambda^\eta \rightarrow \mathcal{M}\Sigma_\Lambda^\eta$ such that $F(g) \leq g$ for any $g \in \Sigma_\Lambda^\eta$.*

Remark 28. Essentially, the main contents of Lemma 27 is a construction of a continuous substitution function. In many natural games, the standard substitution functions are continuous without additional efforts.

The definition of $\mathcal{M}\Sigma_\Lambda^\eta$ implies that if $F(g) \leq g$ then $F(g) = g$ for any $g \in \mathcal{M}\Sigma_\Lambda^\eta$, and hence F defined in the lemma is a *retraction* (by definition, a continuous mapping from a topological space into its subset that does not move elements of the subset). Due to [1, Theorem 4.2], since Σ_Λ^η has the fixed point property, its retract $\mathcal{M}\Sigma_\Lambda^\eta$ has the fixed point property too.

Theorem 29. *Suppose that the game (Ω, Λ) satisfies Assumptions 1 and 2 and is η -mixable. Then for the prediction with second-guessing expert advice protocol, there exists Learner's strategy (a modification of the AA) with parameters η and P_0 guaranteeing that, at each step N and for all experts θ , it holds*

$$L_N \leq L_N^\theta + \frac{1}{\eta} \ln \frac{1}{P_0(\theta)}.$$

Proof. At step N , the modified AA announces as Learner's prediction γ_N any solution of the following equation with respect to $\gamma \in \mathcal{M}\Sigma_\Lambda^\eta$:

$$\gamma = F \left(-\frac{1}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta \in \Theta} P_{N-1}(\theta)} \exp(-\eta \Gamma_N^\theta(\gamma)) \right) \right), \quad (18)$$

where Γ_N^θ are announced by the experts, the weights P_{N-1} are defined in the usual way with the help of the previous losses:

$$P_{N-1}(\theta) = P_0(\theta) \prod_{n=1}^{N-1} \exp(-\eta \Gamma_n^\theta(\gamma_n, \omega_n)),$$

and F is the continuous mapping from Lemma 27.

Since for an η -mixable game we have $\Sigma_\Lambda^\eta = \Sigma_\Lambda$, and since $\mathcal{M}\Sigma_\Lambda \subseteq \Lambda$, the functions Γ_N^θ are defined on γ . By the definition of Σ_Λ^η , the argument of F in equation (18) belongs to Σ_Λ^η , and F maps it to $\mathcal{M}\Sigma_\Lambda^\eta$. The mapping is continuous as the composition of continuous mappings. Therefore, since $\mathcal{M}\Sigma_\Lambda^\eta$ has the fixed point property, equation (18) has a solution.

The property $F(g) \leq g$ implies that

$$\gamma_N \leq -\frac{1}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta \in \Theta} P_{N-1}(\theta)} \exp(-\eta \Gamma_N^\theta(\gamma_N)) \right),$$

and the usual analysis of the AA gives us the bound. \square

Let us outline briefly how the construction of Theorem 29 can be applied to non-mixable games under the modified second-guessing protocol (where experts are defined on $\partial\Sigma_\Lambda$). Let the AA be (c, η) -realizable. Now we are looking for $\gamma \in \partial\Sigma_\Lambda$ satisfying the following equation:

$$\gamma = V \left(F \left(-\frac{1}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta \in \Theta} P_{N-1}(\theta)} \exp(-\eta \Gamma_N^\theta(\gamma)) \right) \right) \right), \quad (19)$$

where after F we apply V , the mapping defined in the proof of Lemma 12. Since V is continuous and maps Σ_Λ^η to $\partial\Sigma_\Lambda$, we get a continuous mapping of $V(\Sigma_\Lambda^\eta) \subseteq \partial\Sigma_\Lambda$ into itself. It remains to show that $V(\Sigma_\Lambda^\eta)$ has the fixed point property. Similarly to the proof of Lemma 12, consider the set $Z = \{g \in \Sigma_\Lambda^\eta \mid \forall r \in [0, 1) rg \notin \Sigma_\Lambda^\eta\}$. For any $g \in \Sigma_\Lambda^\eta$, there exists a unique r such that $rg \in Z$, and the continuity of this mapping $g \rightarrow rg$ follows in the same way as in the proof of Lemma 12; thus Z is a retract of Σ_Λ^η and has the fixed point property. Since $V(g) = V(rg)$ for any non-negative real r such that g and rg belong to Σ_Λ^η , we have $V(\Sigma_\Lambda^\eta) = V(Z)$. The definition of Z implies that V is bijective on Z , and again as in the proof of Lemma 12 one can show that the inverse mapping $V^{-1}: V(Z) \rightarrow Z$ is continuous. Therefore $V(Z)$ has the fixed point property as the homeomorphic image of Z .

Let $\gamma_N \in V(\Sigma_\Lambda^\eta)$ be any solution of the equation (19). By the properties of F and V , we have

$$\gamma_N \leq -\frac{c}{\eta} \ln \left(\sum_{\theta \in \Theta} \frac{P_{N-1}(\theta)}{\sum_{\theta \in \Theta} P_{N-1}(\theta)} \exp(-\eta \Gamma_N^\theta(\gamma_N)) \right),$$

and the usual AA bound follows.

5 Predictions with Respect to Several Loss Functions

In this section, we illustrate the use of the supermartingale technique for another extension of Protocol 1: a game with several loss functions (for a more detailed discussion of this setting see [5]). In contrast to the case of second-guessing experts, it is not clear yet whether the AA can help in this case.

Up to now a game was (Ω, Λ) where Λ was the set of admissible predictions, common for Learner and Experts. Here we return to the game specification by a loss function on the decision space $\mathcal{P}(\Omega)$. However, now each Expert θ has its own loss function λ^θ . So, the game is specified by $(\Omega, \mathcal{P}(\Omega), \{\lambda^\theta\}_{\theta \in \Theta})$, where $\lambda^\theta: \mathcal{P}(\Omega) \times \Omega \rightarrow [0, \infty]$ are proper loss functions. The sets of predictions $\Lambda(\theta)$ and superpredictions $\Sigma_{\Lambda(\theta)}$ may be different for different experts θ . The game follows Protocol 3.

Protocol 3 Prediction with Expert Evaluators' Advice

$L_0^{(\theta)} := 0$, for all $\theta \in \Theta$.
 $L_0^\theta := 0$, for all $\theta \in \Theta$.
for $n = 1, 2, \dots$ **do**
 All Experts $\theta \in \Theta$ announce $\pi_n^\theta \in \mathcal{P}(\Omega)$.
 Learner announces $\pi_n \in \mathcal{P}(\Omega)$.
 Reality announces $\omega_n \in \Omega$.
 $L_n^{(\theta)} := L_{n-1}^{(\theta)} + \lambda^\theta(\pi_n, \omega_n)$, for all $\theta \in \Theta$.
 $L_n^\theta := L_{n-1}^\theta + \lambda^\theta(\pi_n^\theta, \omega_n)$, for all $\theta \in \Theta$.
end for

There are two changes in Protocol 3 compared to Protocol 1. The accumulated loss L^θ of each Expert θ is calculated according to his own loss function λ^θ . Learner does not have one accumulated loss anymore, but the losses $L^{(\theta)}$ of Learner are calculated separately for comparisons with each Expert θ and according to the loss function of this Expert.

Now it does not make much sense to speak about the best expert: their performance is evaluated by different loss functions and thus the losses may have different scale. What remains meaningful are bounds for every expert θ of the form

$$L_N^{(\theta)} \leq c^\theta L_N^\theta + a^\theta,$$

where c^θ and a^θ may be different for different experts $\theta \in \Theta$.

Informally speaking, Protocol 3 describes the following situation. We have some practical task and a number of prediction algorithms (they will be our Experts). Each of them minimizes some loss, maybe different for different algorithms. We do not know which algorithms fits our task best. As usual in practice, we do not have a loss that measures the quality of predictions for our task; we only know that predictions must be close to the real outcomes. A safe option in this case would be to predict in such a way that our prediction are not bad compared to predictions of any of the algorithms even if the quality is evaluated by the loss function ascribed to this algorithm.

The DFA can be adapted to Protocol 3 straightforwardly.

Theorem 30. *Suppose that for each $\theta \in \Theta$, there exist reals $c^\theta \geq 1$ and $\eta^\theta > 0$ such that the functions*

$$\exp\left(\eta^\theta\left(\frac{\lambda^\theta(\pi, \omega)}{c^\theta} - g(\omega)\right)\right)$$

(they are direct analogs of q_g defined by (13)) are forecast-continuous and have the supermartingale property for all $g \in \Sigma_{\Lambda^\theta}$. Then for any initial distribution $P_0 \in \mathcal{P}(\Theta)$ there is Learner's strategy (in fact, the DFA applied to Q^{P_0} defined by (20)) guaranteeing that, at each step N and for all experts θ , it holds

$$L_N^{(\theta)} \leq c^\theta L_N^\theta + \frac{c^\theta}{\eta^\theta} \ln \frac{1}{P_0(\theta)}.$$

Proof. Similarly to the proofs of Theorems 5 and 26, we can construct the

supermartingale Q^{P_0} :

$$Q^{P_0}(\{\pi_1^\theta\}_{\theta \in \Theta}, \pi_1, \omega_1, \dots, \{\pi_N^\theta\}_{\theta \in \Theta}, \pi_N, \omega_N) = \sum_{\theta \in \Theta} P_0(\theta) \prod_{n=1}^N \exp\left(\eta^\theta \left(\frac{\lambda^\theta(\pi_n, \omega_n)}{c^\theta} - \lambda^\theta(\pi_n^\theta, \omega_n)\right)\right) \quad (20)$$

and choose π_N satisfying (9) with the help of Lemma 4. The loss bound follows in the same way as in Theorem 5. \square

Protocol 3 can handle also the following task. We have several experts and several candidates for the loss function, and a priori some experts may perform well for two or more of the loss functions. In this case, it is natural to require that Learner's loss is small with respect to every expert and with respect to every loss function. A simple trick reduces the task to Protocol 3: for each original expert (supplying us with a prediction), we consider several new experts who announce the same prediction but use different loss functions. If our predictions are good in the game with these new experts then our predictions are good in the original game with respect to any of the loss functions.

For example, assume that we want to compete with K experts according to the logarithmic loss function and square loss function in the game with outcomes $\{0, 1\}$. Lemmas 6 and 7 imply that the following function is a forecast-continuous supermartingale:

$$\frac{1}{2K} \sum_{k=1}^K \exp\left(\sum_{n=1}^N (-\ln \pi_n(\omega_n) + \ln \pi_n^k(\omega_n))\right) + \frac{1}{2K} \sum_{k=1}^K \exp\left(2 \sum_{n=1}^N ((\omega_n - \pi_n(1))^2 - (\omega_n - \pi_n^k(1))^2)\right),$$

where π_n^k is the prediction of Expert k and π_n is the prediction of Learner. Choosing π_n according to Lemma 4, we can achieve that the regret term with respect to the logarithmic loss function is bounded by $\ln(2K) < \ln K + 0.7$, and the regret with respect to the square loss function is bounded by $0.5 \ln(2K) < 0.5 \ln K + 0.4$ — practically the same as the regrets against K experts that are achievable when we compete with respect to only one of the loss functions.

Acknowledgements

This work was partly supported by EPSRC grant EP/F002998/1. Discussions with Alex Gammernan, Glenn Shafer, and Alexander Shen, and detailed comments of the anonymous referees for the conference version [4] and for a journal submission have helped us improve the paper.

References

- [1] R. Agarwal, M. Meehan, D. O'Regan. *Fixed Point Theory and Applications*, volume 141 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, England, 2001.

- [2] D. Blackwell, M. A. Girschik. *Theory of Games and Statistical Decisions*, New York: Wiley, 1954.
- [3] A. Blum, Y. Mansour. From External to Internal Regret. *J. Mach. Learn. Res.*, 8:1307–1324, 2007.
- [4] A. Chernov, Y. Kalnishkan, F. Zhdanov, V. Vovk. Supermartingales in Prediction with Expert Advice. In: Y. Freund, L. Györfi, G. Turán, T. Zeugmann (eds.) *ALT 2008 Proceedings*, LNCS(LNAI) vol. 5254, pp. 199–213. Springer, 2008.
- [5] A. Chernov, V. Vovk. Prediction with Expert Evaluators’ Advice. In: R. Gavaldà, G. Lugosi, T. Zeugmann, S. Zilles (eds.) *ALT 2009 Proceedings*, LNCS(LNAI) vol. 5809, pp. 8–22. Springer, 2009. (See also Technical Report arXiv:0902.4127v1 [cs.LG].)
- [6] N. Cesa-Bianchi, G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, England, 2006.
- [7] A. P. Dawid. The Geometry of Proper Scoring Rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.
- [8] D. Foster, R. Vohra. Asymptotic Calibration. *Biometrika*, 85:379–390, 1998.
- [9] D. Foster, R. Vohra. Regret in the Online Decision Problem. *Games Econ. Behav.*, 29:104–130, 1999.
- [10] P. Gács. Uniform Test of Algorithmic Randomness over a General Space. *Theoretical Computer Science*, 341:91–137, 2005.
- [11] P. Gács. *Lecture Notes on Descriptive Complexity and Randomness*, Unpublished, available online at <http://www.cs.bu.edu/faculty/gacs/papers/ait-notes.pdf>
- [12] T. Gneiting, A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. American Statistical Association*, 102:359–378, 2007.
- [13] P. D. Grunwald, A. P. Dawid. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Annals of Statistics*, 32(4), pp. 1367–1433, 2004.
- [14] D. Haussler, J. Kivinen, M. Warmuth. Sequential Prediction of Individual Sequences under General Loss Functions. *IEEE Trans. on Information Theory*, 44(5):1906–1925, 1998.
- [15] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *J. American Statistical Association*, 58:13–30, 1963.
- [16] L. Levin. Uniform Tests of Randomness. *Soviet Mathematics Doklady*, 17:337–340, 1976. The Russian original: *Doklady AN SSSR*, 227(1), 1976.
- [17] M. Li, P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd edition, New York: Springer, 1997.

- [18] Y. Kalnishkan, V. Vovk, M. V. Vyugin. Loss Functions, Complexities, and the Legendre Transformation. *Theoretical Computer Science*, 313(2):195–207, 2004.
- [19] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [20] L. J. Savage, Elicitation of Personal Probabilities and Expectations. *J. American Statistical Association*, 66:783–801, 1971.
- [21] G. Shafer, V. Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.
- [22] G. Stoltz, G. Lugosi. Internal Regret in On-Line Portfolio Selection. *Machine Learning*, 59:125–159, 2005.
- [23] G. Stoltz, G. Lugosi. Learning Correlated Equilibria in Games with Compact Sets of Strategies. *Games and Economic Behavior*, 59:187–209, 2007.
- [24] V. Vovk. Aggregating Strategies. In M. Fulk, J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pp. 371–383, San Mateo, CA, Morgan Kaufmann, 1990.
- [25] V. Vovk. A Game of Prediction with Expert Advice. *Journal of Computer and System Sciences*, 56:153–173, 1998
- [26] V. Vovk. Defensive Prediction with Expert Advice. In: S. Jain, H. Simon, E. Tomita (eds.) *ALT 2005 Proceedings*, LNCS(LNAI) vol. 3734, pp. 444–458. Springer, 2005. (See also: Competitive On-line Learning with a Convex Loss Function. Technical Report [arXiv:cs/0506041v3](#) [cs.LG], [arXiv.org](#) e-Print archive, September 2005.)
- [27] V. Vovk. On-line Regression Competitive with Reproducing Kernel Hilbert Spaces. In: J. Cai, S. Barry Cooper, A. Li (eds.) *TAMC 2006 Proceedings*, LNCS(LNAI) vol. 3959, pp. 452–463. Springer, 2006. (See also Technical Report [arXiv:cs/0511058v2](#) [cs.LG], [arXiv.org](#) e-Print archive, January 2006.)
- [28] V. Vovk. Metric Entropy in Competitive On-line Prediction. Technical Report [arXiv:cs/0609045v1](#) [cs.LG], [arXiv.org](#) e-Print archive, September 2006.
- [29] V. Vovk. Defensive Forecasting for Optimal Prediction with Expert Advice. Technical Report [arXiv:0708.1503](#) [cs.LG], [arXiv.org](#) e-Print archive, August 2007.
- [30] V. Vovk. Continuous and Randomized Defensive Forecasting: Unified View. Technical Report [arXiv:0708.2353v2](#) [cs.LG], [arXiv.org](#) e-Print archive, August 2007.
- [31] V. Vovk, F. Zhdanov. Prediction with Expert Advice for the Brier Game. *ICML '08: Proceedings of the 25th International Conference on Machine learning*, pp. 1104–1111, 2008.

Appendix

Proof of Lemma 8. Given the function q , let us define the following function ϕ on $\mathcal{P}(\Omega) \times \mathcal{P}(\Omega)$:

$$\phi(\pi', \pi) = \mathbb{E}_{\pi'} q(\pi, \cdot).$$

For each fixed π' , the function $\phi(\pi', \cdot)$ is continuous, since q is continuous. For each fixed π , the function $\phi(\cdot, \pi)$ is linear, and thus concave. Note also that $\mathcal{P}(\Omega)$ is a convex compact set. Therefore, ϕ satisfies the conditions of Ky Fan's minimax theorem (see e.g. [1, Theorem 11.4]), and thus there exists $\tilde{\pi} \in \mathcal{P}(\Omega)$ such that for any $\pi' \in \mathcal{P}(\Omega)$ it holds that

$$\mathbb{E}_{\pi'} q(\tilde{\pi}, \cdot) = \phi(\pi', \tilde{\pi}) \leq \sup_{\pi \in \mathcal{P}(\Omega)} \phi(\pi, \pi) = \sup_{\pi \in \mathcal{P}(\Omega)} \mathbb{E}_{\pi} q(\pi, \cdot) \leq C. \quad (21)$$

It is easy to see that $\tilde{\pi}$ has the property that the lemma must guarantee: $q(\tilde{\pi}, \omega) \leq C$ for all $\omega \in \Omega$. Indeed, if we substitute the distribution δ_{ω} (which is concentrated on ω) for π' in (21), the left-hand side will be just $q(\tilde{\pi}, \omega)$. \square

Lemma 8 is a very important statement in our supermartingale framework, so let us outline an alternative proof for it (for details see [10, Theorem 6], [11, Theorem 16.1] or [30, Theorem 1]). Consider the sets $F_{\omega} = \{\pi \mid q(\pi, \omega) \leq C\}$. These sets are closed and for any $\Omega_0 \subseteq \Omega$ the union $\cup_{\omega \in \Omega_0} F_{\omega}$ contains all the measures concentrated on Ω_0 . Then all F_{ω} has a non-empty intersection by Sperner's lemma.

Lemma 31. *Let a function $q: \mathcal{P}^{\circ}(\Omega) \times \Omega \rightarrow \mathbb{R}$ be non-negative and forecast-continuous on $\mathcal{P}^{\circ}(\Omega)$. Suppose that for any $\pi \in \mathcal{P}^{\circ}(\Omega)$ it holds that*

$$\mathbb{E}_{\pi} q(\pi, \cdot) \leq C,$$

where $C \in [0, \infty)$ is some constant. Then it holds that

$$\exists \pi \in \mathcal{P}^{\circ}(\Omega) \forall \omega \in \Omega \quad q(\pi, \omega) \leq (1 + \epsilon)C.$$

Proof. Let $\delta \in (0, 1)$ be a constant to be chosen later.

Let $\mathcal{P}^{\delta}(\Omega) = \{\pi \in \mathcal{P}(\Omega) \mid \forall \omega \in \Omega \pi(\omega) \geq \delta\}$. This set is a non-empty convex compact subset of $\mathcal{P}^{\circ}(\Omega)$. Repeating the construction from the proof of Lemma 8 and applying Ky Fan's theorem for the function on $\mathcal{P}^{\delta}(\Omega)$, we get that there exists $\tilde{\pi} \in \mathcal{P}^{\delta}(\Omega)$ such that for any $\pi' \in \mathcal{P}^{\delta}(\Omega)$ it holds that $\mathbb{E}_{\pi'} q(\tilde{\pi}, \cdot) \leq C$.

For each ω_0 , consider the distribution π_{δ, ω_0} such that $\pi_{\delta, \omega_0}(\omega) = \delta$ for $\omega \neq \omega_0$ and $\pi_{\delta, \omega_0}(\omega_0) = 1 - \delta(|\Omega| - 1)$. Substituting π_{δ, ω_0} for π' , we get

$$(1 - \delta(|\Omega| - 1))q(\tilde{\pi}, \omega_0) + \delta \sum_{\omega \neq \omega_0} q(\tilde{\pi}, \omega) \leq C.$$

Since $q(\tilde{\pi}, \omega) \geq 0$ (the supermartingale S is non-negative), the last inequality implies that $(1 - \delta(|\Omega| - 1))q(\tilde{\pi}, \omega_0) \leq C$. It remains to note that we can choose δ so small that $1/(1 - \delta(|\Omega| - 1)) \leq 1 + \epsilon$. \square

Proof of Lemma 21. According to Lemma 31, we can find $\pi_k \in \mathcal{P}^{\circ}(\Omega)$ such that

$$\forall \omega \in \Omega \quad q(\pi_k, \omega) \leq \left(1 + \frac{1}{k}\right)C.$$

Since $\mathcal{P}(\Omega)$ is compact, there exists a strictly increasing index sequence $k(j)$, $j \in \mathbb{N}$, such that the sequence $\pi_{k(j)}$ converges to some $\pi \in \mathcal{P}(\Omega)$.

The points $g_j = q(\pi_{k(j)}, \cdot)$ belong to a compact set $[0, 2C]^\Omega$. Hence there exists a strictly increasing index sequence $j(i)$, $i \in \mathbb{N}$, such that the sequence $g_{j(i)}$ converges to some g_0 . For every $\omega \in \Omega$, we have $g_j(\omega) = q(\pi_{k(j)}, \omega) \leq (1 + 1/k(j))C$, therefore

$$g_0(\omega) = \lim_i g_{j(i)}(\omega) \leq \lim_i \left(1 + \frac{1}{k(j(i))}\right) C = C.$$

It remains to set $\pi^{(i)} = \pi_{k(j(i))}$ and note that $q(\pi^{(i)}, \omega) = g_{j(i)}$. \square

Proof of Lemma 12. Let $\mathbf{0}: \Omega \rightarrow [0, \infty]$ be the constant zero function (that is, $\mathbf{0}(\omega) = 0$ for all $\omega \in \Omega$). If $\mathbf{0} \in \Sigma_\Lambda$ then $\mathbf{0} \in \partial\Sigma_\Lambda$ and we can let $V(g) = \mathbf{0}$ for any g .

Assume that $\mathbf{0} \notin \Sigma_\Lambda$. Let $V(g) = R(g)g$, where $R: \Sigma_\Lambda^\eta \rightarrow (0, c]$ is defined by the following rule: $R(g) = \min\{r \in (0, c] \mid rg \in \Sigma_\Lambda\}$ for any $g \in \Sigma_\Lambda^\eta$.

Since the AA is (c, η) -realizable, it holds that $c\Sigma_\Lambda^\eta \subseteq \Sigma_\Lambda$, that is, $cg \in \Sigma_\Lambda$ for any $g \in \Sigma_\Lambda^\eta$. The minimum is attained since Σ_Λ is compact (by Assumption 1). Thus $R(g)$ is well defined. It is obvious from the definition that $V(g) = R(g)g$ belongs to the boundary $\partial\Sigma_\Lambda$ of Σ_Λ for all $g \in \Sigma_\Lambda^\eta$.

It remains to check that $V(g) = R(g)g$ is continuous in g . We prove that R is continuous, namely, we take any $g_i \rightarrow g_0$ and for any infinite subsequence $\{g_{i_k}\}$, we show that if $R(g_{i_k})$ converges then $\lim_k R(g_{i_k}) = R(g_0)$. If $R(g_{i_k})$ converges then $R(g_{i_k})g_{i_k}$ converges, and $\lim_k R(g_{i_k})g_{i_k} = (\lim_k R(g_{i_k}))g_0 \in \Sigma_\Lambda$ since Σ_Λ is compact. Therefore $R(g_0) \leq \lim_k R(g_{i_k})$. For the other inequality, consider $R(g'|g) = \min\{r \in (0, \infty) \mid rg' \geq V(g)\}$ for $g, g' \in \Sigma_\Lambda^\eta$ such that if $g(\omega) \neq 0$ for some $\omega \in \Omega$ then $g'(\omega) \neq 0$ too. Clearly, the function $R(g'|g)$ is continuous in g' for any fixed g (note that $R(g'|g) = \max_{\omega: g(\omega) \neq 0} V(g)(\omega)/g'(\omega)$) and $R(g|g) = R(g)$. Since $rg' \geq V(g) \in \Sigma_\Lambda$ implies $rg' \in \Sigma_\Lambda$, we have $R(g'|g) \geq R(g')$. In particular, $R(g_{i_k}|g_0) \geq R(g_{i_k})$ (assuming k large enough so that $g(\omega) \neq 0$ implies $g_{i_k}(\omega) \neq 0$) and $R(g_0) = \lim_k R(g_{i_k}|g_0) \geq \lim_k R(g_{i_k})$. \square

Proof of Lemma 14. Assume that $\lambda_1(\pi, \omega_0) \neq \lambda_2(\pi, \omega_0)$ and $\pi(\omega_0) > 0$ for some $\omega_0 \in \Omega$.

Since $\lambda_1(\pi, \cdot)$ and $\lambda_2(\pi, \cdot)$ belong to Σ_Λ^η , the point

$$g = -\frac{1}{\eta} \ln \frac{e^{-\eta\lambda_1(\pi, \cdot)} + e^{-\eta\lambda_2(\pi, \cdot)}}{2}$$

also belongs to Σ_Λ^η by the definition of Σ_Λ^η .

For any reals x, y , we have $(e^x + e^y)/2 \geq e^{(x+y)/2}$, and the inequality is strict if $x \neq y$. Therefore, $g(\omega) \leq (\lambda_1(\pi, \omega) + \lambda_2(\pi, \omega))/2$ for all $\omega \in \Omega$ and $g(\omega_0) < (\lambda_1(\pi, \omega_0) + \lambda_2(\pi, \omega_0))/2$. Multiplying these inequalities by $\pi(\omega)$ and summing over all $\omega \in \Omega$, we get

$$\mathbb{E}_\pi g < \frac{1}{2} (\mathbb{E}_\pi \lambda_1(\pi, \cdot) + \mathbb{E}_\pi \lambda_2(\pi, \cdot))$$

(recall that $\pi(\omega_0) > 0$). Since λ_1 and λ_2 are proper with respect to Σ_Λ^η , we have $\mathbb{E}_\pi \lambda_1(\pi, \cdot) \leq \mathbb{E}_\pi g$ and $\mathbb{E}_\pi \lambda_2(\pi, \cdot) \leq \mathbb{E}_\pi g$. Hence we get a contradiction $\mathbb{E}_\pi g < \mathbb{E}_\pi g$. \square

For a convex function $U: \mathbb{R}^\Omega \rightarrow [-\infty, \infty]$, a *subgradient* at point $x \in \mathbb{R}^\Omega$ is a point $x^* \in \mathbb{R}^\Omega$ such that

$$\forall z \in \mathbb{R}^\Omega \quad U(z) \geq U(x) + \langle x^*, z - x \rangle.$$

Lemma 32. *Suppose that Y is a non-empty closed convex subset of $[0, \infty)^\Omega$. Let $U: \mathbb{R}^\Omega \rightarrow (-\infty, \infty]$ be the function*

$$U(x) = - \inf_{y \in Y} \langle x, y \rangle,$$

where $\langle x, y \rangle = \sum_{\omega \in \Omega} x(\omega)y(\omega)$ is the scalar product in \mathbb{R}^Ω . Then $U(x)$ is a convex function, and for any $\pi \in \mathcal{P}(\Omega)$, it holds that $U(\pi) < \infty$, and π^* is a subgradient of U at the point π if and only if $-\pi^* \in Y$ and $\langle \pi, -\pi^* \rangle = -U(\pi)$.

Proof. Since Y is not empty, the infimum is finite, and therefore $U(x) > -\infty$ for all x .

For any $\alpha \in [0, 1]$ and any $x_1, x_2 \in \mathbb{R}^\Omega$, we have $U(\alpha x_1 + (1 - \alpha)x_2) = -\inf_{y \in Y} (\alpha \langle x_1, y \rangle + (1 - \alpha) \langle x_2, y \rangle) \leq -\inf_{y \in Y} \alpha \langle x_1, y \rangle - \inf_{y \in Y} (1 - \alpha) \langle x_2, y \rangle = \alpha U(x_1) + (1 - \alpha)U(x_2)$, thus U is convex.

Let us fix some $\pi \in \mathcal{P}(\Omega)$. Then $\langle \pi, y \rangle \geq 0$ for all $y \in Y$, and $U(\pi) \leq 0 < \infty$.

Let $-\pi^* \in Y$ and $\langle \pi, -\pi^* \rangle = -U(\pi)$. Then $U(\pi) + \langle \pi^*, z - \pi \rangle = -\langle -\pi^*, z \rangle \leq -\inf_{y \in Y} \langle z, y \rangle = U(z)$ for any z , thus π^* is a subgradient of U at π .

Let π^* be any subgradient of U at π . Assume that $-\pi^* \notin Y$. Then $-\pi^*$ and Y can be strongly separated by Corollary 11.4.2 in [19], and Theorem 11.1(c) there implies that there exists $z \in \mathbb{R}^\Omega$ such that $\inf_{y \in Y} \langle y, z \rangle > \langle -\pi^*, z \rangle$. Let us choose $\delta > 0$ such that

$$\inf_{y \in Y} \langle y, z \rangle > \delta + \langle -\pi^*, z \rangle,$$

and then choose $y_0 \in Y$ such that

$$\langle \pi + z, y_0 \rangle < \inf_{y \in Y} \langle \pi + z, y \rangle + \delta.$$

From the definition of the subgradient, we get $U(\pi + z) \geq U(\pi) + \langle \pi^*, z \rangle$, and thus

$$\langle \pi + z, y_0 \rangle - \delta < \inf_{y \in Y} \langle \pi + z, y \rangle \leq \inf_{y \in Y} \langle \pi, y \rangle + \langle -\pi^*, z \rangle \leq \langle \pi, y_0 \rangle + \langle -\pi^*, z \rangle.$$

So, $\langle z, y_0 \rangle < \delta + \langle -\pi^*, z \rangle$, which contradicts the choice of δ . This means that $-\pi^* \in Y$.

It remains to note that the definition of the subgradient implies $U(0) \geq U(\pi) + \langle \pi^*, 0 - \pi \rangle$, and since $U(0) = 0$, we get $\inf_{y \in Y} \langle y, \pi \rangle = -U(\pi) \geq \langle -\pi^*, \pi \rangle$. \square

Proof of Lemma 15. By Assumption 2, there exists a finite point g_{fin} in $\Sigma_\Lambda^\eta \cap [0, \infty)^\Omega$, where $E_\pi g_{\text{fin}}$ is finite for any π . By Assumption 1, Σ_Λ^η is compact, and therefore the minimum is attained for all $\pi \in \mathbb{R}^\Omega$. Thus H is well defined. Note also that $H(\pi) \geq 0$ for $\pi \in \mathcal{P}(\Omega)$ and $H(\pi) = -\infty$ if $\pi(\omega) < 0$ for some $\omega \in \Omega$.

Now let us show that

$$H(\pi) = \inf_{g \in \Sigma_\Lambda^\eta \cap [0, \infty)^\Omega} E_\pi g.$$

Again by Assumption 2, the infimum is taken over a non-empty set. If $\pi(\omega) < 0$ for some $\omega \in \Omega$ then $H(\pi) = -\infty$ and the infimum is equal to $-\infty$ as well. Thus we need to consider only the case when $\pi(\omega) \geq 0$ for all $\omega \in \Omega$ and the minimum in the definition of H is attained at a point g such that $g(\omega) = \infty$ for some values of ω . Note that for these ω we have $\pi(\omega) = 0$, since $H(\pi) < \infty$. Choose a sequence $g_n \in \Sigma_\Lambda^\eta \cap [0, \infty)^\Omega$ that converges to g (for example, consider the segment between the points $e^{-\eta g}$ and $e^{-\eta g_{\text{fin}}}$, and take a sequence $e^{-\eta g_n}$ along this segment). Since $g_n(\omega)$ and $g(\omega)$ are finite for non-zero $\pi(\omega)$, we get $E_\pi g_n \rightarrow E_\pi g = H(\pi)$, and thus the infimum is not greater than $H(\pi)$.

Now we can apply Lemma 32 with $Y = \Sigma_\Lambda^\eta \cap [0, \infty)^\Omega$ and $U(\pi) = -H(\pi)$. It implies that for any $\pi \in \mathcal{P}(\Omega)$, the set of subgradients of U at π is the set of points where the infimum of $E_\pi g$ over $g \in \Sigma_\Lambda^\eta \cap [0, \infty)^\Omega$ is attained. If $\pi \in \mathcal{P}^\circ(\Omega)$, the infimum is attained indeed, and it is unique by Lemma 14. By Theorem 25.1 in [19], the function H is differentiable at π , and the point $\lambda(\pi, \cdot) = \arg \min_{g \in \Sigma_\Lambda^\eta} E_\pi g$ is the gradient of H . Thus $\mathcal{H} \supseteq \mathcal{P}^\circ(\Omega)$.

On the other hand, if H is differentiable, the set of subgradients consists of one element only, the gradient. Theorem 25.5 in [19] implies that the gradient mapping $\pi \mapsto \lambda(\pi, \cdot)$ is continuous on \mathcal{H} . \square

Proof of Lemma 18. Due to Assumption 1, $\mathcal{M}\Sigma_\Lambda^\eta$ is compact and therefore contains all its limit points, that is, $\gamma \in \mathcal{M}\Sigma_\Lambda^\eta$.

Let $E_{\pi'}^+ g$ be a shorthand for $\sum_{\omega \in \Omega, \pi(\omega) \neq 0} \pi'(\omega) g(\omega)$ for any $g \in [0, \infty)^\Omega$ and $\pi' \in \mathcal{P}(\Omega)$. By definition, $E_\pi \gamma = E_\pi^+ \gamma$.

Note first that $E_{\pi_i} g$ converges to $E_\pi g$ for any finite $g \in [0, \infty)^\Omega$.

Note also that $E_{\pi_i}^+ \gamma_i$ converges to $E_\pi^+ \gamma$. Indeed, $E_{\pi_i}^+ \gamma_i \leq E_{\pi_i} \gamma_i \leq E_{\pi_i} g_{\text{fin}} \leq \sum_{\omega \in \Omega} g_{\text{fin}}(\omega) < \infty$, where $g_{\text{fin}} \in \Sigma_\Lambda^\eta \cap [0, \infty)^\Omega$ exists by Assumption 2. If $\pi(\omega) \neq 0$ then $\pi_i(\omega)$ is separated from 0 for sufficiently large i , therefore $\gamma_i(\omega)$ are bounded, and their limit $\gamma(\omega)$ is finite. And for finite limits γ and π , the convergence is trivial.

Fix any $g_0 \in [0, \infty)^\Omega$ and any $\epsilon > 0$. For sufficiently large i , we have $E_{\pi_i}^+ \gamma_i \geq E_\pi^+ \gamma - \epsilon$ and $E_{\pi_i} g_0 \leq E_\pi g_0 + \epsilon$. Taking into account that $E_{\pi_i}^+ \gamma_i \leq E_{\pi_i} \gamma_i$ and $E_{\pi_i} \gamma_i = \min_{g \in \Sigma_\Lambda^\eta} E_{\pi_i} g \leq E_{\pi_i} g_0$, we get $E_\pi \gamma \leq E_\pi g_0 + 2\epsilon$. Since ϵ and g_0 are arbitrary, we have

$$E_\pi \gamma \leq \inf_{g \in \Sigma_\Lambda^\eta \cap [0, \infty)^\Omega} E_\pi g,$$

and the last infimum can be replaced by $\min_{g \in \Sigma_\Lambda^\eta}$ as shown in the proof of Lemma 15. \square

Proof of Lemma 27. We construct a continuous mapping $F: \Sigma_\Lambda^\eta \rightarrow \mathcal{M}\Sigma_\Lambda^\eta$ as a composition of mappings F_ω for all $\omega \in \Omega$. Each F_ω when applied to $g \in \Sigma_\Lambda^\eta$ preserves the values of $g(o)$ for $o \neq \omega$ and decreases as far as possible the value $g(\omega)$ so that the result is still in Σ_Λ^η . Formally, $F_\omega(g) = g'$ such that $g'(o) = g(o)$ for $o \neq \omega$ and $g'(\omega) = \min\{\gamma(\omega) \mid \gamma \in \Sigma_\Lambda^\eta, \forall o \neq \omega \gamma(o) = g(o)\}$.

Let us show that each F_ω is continuous. It suffices to show that $F_\omega(g)(\omega)$ depends continuously on g , since the other coordinates do not change. We will show that $F_\omega(g)(\omega)$ is convex in g , continuity follows (see, e.g. [19]). Indeed, take any $t \in [0, 1]$, and $g_1, g_2 \in \Sigma_\Lambda^\eta$. Since Σ_Λ^η is convex, then $tg_1 + (1-t)g_2 \in \Sigma_\Lambda^\eta$ and $tF_\omega(g_1) + (1-t)F_\omega(g_2) \in \Sigma_\Lambda^\eta$. The latter point has all the coordinates $o \neq \omega$ the same as the former. Thus, by definition of F_ω , we get $F_\omega(tg_1 + (1-t)g_2)(\omega) \leq$

$(tF_\omega(g_1) + (1-t)F_\omega(g_2))(\omega) = tF_\omega(g_1)(\omega) + (1-t)F_\omega(g_2)(\omega)$, which was to be shown.

All F_ω do not increase the coordinates. Since the set Σ_Λ^η contains any point g with all its majorants, $F_\omega(g_1) = g_1$ implies that $F_\omega(g_2) = g_2$ for any g_2 obtained from g_1 by applying any $F_{\omega'}$. Therefore, the image of a composition of F_ω over all $\omega \in \Omega$ is included in $\mathcal{M}\Sigma_\Lambda^\eta$. \square