# Forecasting Distributions with Experts Advice

Alessio Sancetta

April 2005

CWPE 0517

# Forecasting Distributions with Experts Advice

Alessio Sancetta
Faculty of Economics, University of Cambridge, UK

April 27, 2005

**Abstract**

This paper considers forecasts of the distribution of data whose distribution function is possibly time varying. The forecast is achieved via time varying combinations of experts' forecasts. We derive theoretical worse case bounds for general algorithms based on multiplicative updates of the combination weights. The bounds are useful to study the properties of forecast combinations when data are nonstationary and there is no unique best model. An application with an empirical study is used to highlight the results in practice.

**Keywords:** Expert, Forecast Combination, Multiplicative Update, Non-asymptotic Bound, On-line Learning, Shifting.

**JEL:** C53, C14.

## 1 Introduction

This paper studies forecasts combination that achieve optimal theoretical properties for online forecasting of distributions (with possibly time varying parameters). We show that this also covers the case of point predictions for arbitrary loss functions.

The goal is to use sequential strategies (or algorithms) that would allow us to forecast the distribution of new observations (within a given reference class) almost as well as if we knew them before hand. To do this, we borrow ideas from the literature in game theory (e.g. see special issue in Games and Economic Behavior, Vol. 29, 1999) and computational learning theory (e.g. Vovk, 1990, Cesa-Bianchi et al., 1997).

Predictions using forecast combinations are often called predictions with experts. We are interested algorithms that lead to optimal error bounds for worse case scenarios. These bounds do not make any assumption on the sequence, we do not even need to assume that the data are realisations of some sequence of random variables.

Worse case bounds derived here specialise to the bounds derived by Herbster and Warmuth (1998) and owe a lot to their presentation and results. An advantage of the present study is that a set of conditions are established so that results can be derived in general form without the need of deriving them on a case by case basis. This also allows us to gain a better understanding of the terms that contribute to the total error of the algorithm. Consequently the algorithms can be modified accordingly to improve the theoretical bounds. In fact, we state an additional algorithm that produces combination weights that, unlike the algorithms in Herbster and Warmuth (1998), do not depend on unknown parameters. Moreover, the results are derived for forecasting of distributions and not sequences (i.e. point prediction). As we willll show that this framework is more general.

Probabilistic bounds, which are related to the worse case bounds of this paper, have been studied by Yang (2004) in the case of forecast combination of point prediction under the square loss. Both probabilistic bounds and worse case bounds are of interest, so the two studies are complementary.

The literature on combination of forecasts is broad and an excellent survey is Timmermann (2004, section 7 for probability forecasts). Several studies have shown that combination of forecasts can be useful to hedge against structural breaks and forecasts combinations are often more stable than single forecasts (e.g. Hendry and Clements, 2004, Stock and Watson, 2004).

A fundamental component of forecast combination is the choice of prediction function and the combination weights. In particular given a prediction function, it is customary to derive combination weights using moment estimators. The experts' combination is chosen to minimise the user's expected utility over all possible decisions. This requires some stability of the system and assumptions about the data generating process.

Worse case bounds avoid the use of moment estimators. The combination weights are based on sequential updating and the problem is cast in a game theoretic framework. The econometrician needs to pool the experts and do at least as well as the best expert or combination of experts no matter what data are sampled by nature. He minimises his loss given that nature's goal is to sample data to maximise this loss. In this case, the objective is to do as well as the experts in the reference

class. Hence, it is not an expected utility problem, but a minimax problem with respect to the observed cumulated loss. The use of observed cumulated loss has an appealing interpretation in terms of the falsiability principle as addressed by Popper and adapted to the statistical framework via the prequential principle of Dawid (e.g. Dawid, 1984, 1985, 1986). Mutatis mutandis, this approach can be seen as a variation of the $\epsilon$-robust decision rule of Chamberlain (2000), where the set of data generating processes is restricted to the empirical measure.

Our main motivation is optimal forecasting of distributions with time varying parameters, where the parameters are obtained using some linear filters. Linear filters can be used to define parametric (e.g. regression estimators), semiparametric and nonparametric estimators. In this case, the choice of filter and the parameters in the filter is crucial and related to the model selection problem. Ideally, we would like to combine models to do as well as the best model with hindsight.

We highlight the framework. There is an arbitrary sequence of variables that are revealed sequentially over time. For example, returns on stock prices. We are given a distribution indexed in some finite parameter space. At each point in time we need to issue a value for the parameter that needs to be used in the next period forecast. For example we may think of a Gaussian distribution with mean zero and unknown variance that changes over time. Looking at an initial number of observations, we may select a finite number of models to provide a variance forecast given past observations. Then we would like to study algorithms that would allow us to pool the information provided by each model to issue forecasts that are almost as good as forecasting with hindsight using the best model. We also consider the case when changes in the reference class are allowed, i.e. one model may perform better over some period, but being outperformed in other periods. In the case of miss-specified models this is of fundamental importance. For example, the best model might change over time, especially when data are nonstationary.

The plan for the paper is as follows. Section 2 introduces background material. Section 3 states algorithms based on extensions of the exponential update of Vovk (1990), which is also the algorithm used in Yang (2004), and derives general worse case bounds. Section 4 provides an illustrative application to distributions with time varying parameters together with a study of empirical performance. Section 5 shows how to cover the case of point prediction for arbitrary loss functions. Further remarks can be found in Section 6.

### 1.1 Notation

Unless specified otherwise, throughout the paper the following notation is used. For a set $A$, $B \subset\subset A$, means that $B$ is a closed set inside $A$. If $A$ is a set with countable elements, $\#A$ stands for the cardinality of $A$. $\mathcal{S}_n$ stands for the $n \in \mathbb{N}$ dimensional unit simplex. $\mathbb{N}_+ := \mathbb{N} \backslash \{0\}$ is the set of positive integers, i.e. $1, 2, 3, ....$ Suppose $\mathcal{I}$ is a set with a countable number of elements, then $a_{\mathcal{I}} := (a_i)_{i \in \mathcal{I}}$ is a $\#\mathcal{I}$ dimensional vector. For vectors $a$ and $b$ having same dimension, $\langle a, b \rangle$ is their innner product. Suppose $a_s^t$ is a scalar or vector, for legibility reasons we may write $a(s,t)$ (i.e. the subscript first, then the superscript).

Suppose $X$ is a random variable. For $\Pr(X \leq x)$ define $\partial_x \Pr(X \leq x)$ to be the density function or the mass function of $X$ or the density function plus the atom at $x$, depending on the Lebesgue decomposition of the measure corresponding to $X$. If to this measure there corresponds a distribution function $P$, then $P(x) = \Pr(X \leq x)$ and $p(x) := \partial_x \Pr(X \leq x)$. Finally, $\delta(x)$ is the Dirac delta function, i.e. $\delta(0) = 1$, 0 otherwise.

## 2 Background

We face the following sequential problem at time $t = 0, ..., T-1$. Suppose $(X_t)_{t \in \mathbb{N}}$ is a sequence of random variables with values in $\mathbb{R}^S$, $S \geq 1$ and define $\mathcal{F}_t$ to be the sigma algebra generated by $(X_s)_{s \leq t}$. The data generating process is unknown. We observe realizations of $(X_s)_{s \leq t}$, say $x_0, .., x_t$. (Actually, we do not need $x_0, .., x_t$ to be realizations of random variables, but for the sake of explanation it is convenient to treat them as such.) These could be stock market returns from time 0 to $t$. Then, we suppose there is a collection of models $\left\{P_{\theta(e)} : \theta_e \in \Theta_e \subset\subset \mathbb{R}^{d(e)}, d(e) \geq 1\right\}\ e \in \mathcal{E}$ where $\mathcal{E}$ is called the experts' set. At time $t-1$, we are given the experts' forecasts $\left(\hat{\theta}_e^t\right)_{e \in \mathcal{E}}$ to be used as parameters in the models $\left\{P_{\theta(e)} : \theta_e \in \Theta_e\right\}_{e \in \mathcal{E}}$ at time $t$. We consider these forecasts as exogenous to the econometrician's decisions. The econometrician needs to issue the probability forecast $P_{W,t}$. When $x_t$ is observed, the econometrician suffers a loss $\mathcal{R}(p_{W,t}) := -\ln p_{W,t}(x_t)$. In particular, the econometrician will use an algorithm, say $W$, that will produce a probability on $\mathcal{E}$ at each point in time, say $(w_{e,t})_{e \in \mathcal{E}}$. The forecast $p_{W,t}$ will be a function of $(w_{e,t})_{e \in \mathcal{E}}$, $\left(\hat{\theta}_e^t\right)_{e \in \mathcal{E}}$ and $\left\{P_{\theta(e)}\right\}_{e \in \mathcal{E}}$ only. The econometrician's forecast must satisfy the following condition, but then it is arbitrary.

**Condition 1** *For any experts forecasts* $\hat{\theta}_{\mathcal{E}} = \left(\hat{\theta}_e\right)_{e \in \mathcal{E}}$, *outcome* $x$ *and* $w_{\mathcal{E}} = (w_e)_{e \in \mathcal{E}} \in$

$\mathcal{S}_{\#\mathcal{E}}$, $\exists c < \infty, \eta > 0$ *such that*

$$\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)\right) \leq -c \ln \sum_{e\in\mathcal{E}} w_e \exp\left\{-\eta \mathcal{R}\left(p_{\hat{\theta}(e)}\left(x\right)\right)\right\},$$

*where* $p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)$ *is the probability forecast based on an arbitrary vector* $w_{\mathcal{E}}$ *in the unit simplex, experts forecasts* $\hat{\theta}_{\mathcal{E}}$*, and model* $\{p_\theta\}$.

**Remark 2** *In most cases, we can choose* $c = 1/\eta$*, implying in the result below that* $c\eta = 1$.

**Example 3** *The prediction function is a mixture of the experts' models* $\{p_\theta\}$ : $p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right) = \sum_{e\in\mathcal{E}} w_e p_{\hat{\theta}(e)}\left(x\right)$. *This prediction function is often called the linear opinion poll (e.g. Genest and Zidek, 1986). In this case, Condition 1 is satisfied with equality with* $c = 1/\eta = 1$.

**Example 4** *Suppose* $\Theta_e = \Theta$ *and* $\Theta$ *is convex. Then, the prediction function is* $p_\theta$ *with parameter* $\theta$ *being the mean of the experts' forecasts with respect to the measure* $w_{\mathcal{E}}$*, i.e.* $p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right) = p\left(x|\left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right)$*, where* $p\left(x|\theta\right) := p_\theta\left(x\right)$. *In this case, Condition 1 is satisfied for* $c = 1/\eta$ *if* $\exists \eta > 0$ *such that* $\exp\left\{-\eta \mathcal{R}\left(\ p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)\right)\right\}$ *is concave in* $\theta$ *for any* $x$ *in the range of the sample observations. Several special examples when this is true will be provided below.*

If Condition 1 is satisfied with $c = 1/\eta = 1$, some nice interpretations are also possible, and will be provided below.

The goal is to find a sequential algorithm, say $W$, that allows us to find $\left(w_{e,t}\right)_{e\in\mathcal{E}}$, such that for any $a_{\mathcal{E},t} := \left(a_{e,t}\right)_{e\in\mathcal{E}} \in \mathcal{S}_{\#\mathcal{E}}$, and any data sequence $x_1, ..., x_T$,

$$\sum_{t=1}^{T} \mathcal{R}\left(p_W\left(x_t|w_{\mathcal{E},t}, \hat{\theta}^t_{\mathcal{E}}\right)\right) \leq \sum_{t=1}^{T}\left(\sum_{e\in\mathcal{E}} a_{e,t}\mathcal{R}\left(p\left(x_t|\hat{\theta}^t_e\right)\right)\right) + error, \qquad (1)$$

where *error* is usually small, hopefully $o\left(T\right)$. It may not be possible to achieve this for arbitrary $a_{\mathcal{E},t} \in \mathcal{S}_{\#\mathcal{E}}$. However, by suitable restrictions, we obtain bounds that are known in the literature. Suppose $a_{\mathcal{E},t} = a_{\mathcal{E}} \in \mathcal{S}_{\#\mathcal{E}}$ has all entries zero, but one of them, i.e. it is one of the edges of the simplex. Since $a_{\mathcal{E}}$ is an arbitrary edge of the simplex, the previous bound implies

$$\sum_{t=1}^{T} \mathcal{R}\left(p_W\left(x_t|w_{\mathcal{E},t}, \hat{\theta}^t_{\mathcal{E}}\right)\right) \leq \min_{e\in\mathcal{E}} \sum_{t=1}^{T} \mathcal{R}\left(p\left(x_t|\hat{\theta}^t_e\right)\right) + error,$$

and in this case we have $error = O(\ln(\#\mathcal{E}))$ (Theorem 9, below). This bound says that the sequential algorithm used by the econometrician will produce forecasts as good as the forecasted probability of the best expert, plus a term $O(\ln(\#\mathcal{E}))$, i.e. the sequential forecast and the forecast using the best expert with hindsight produce almost the same error. This last statement makes sense if $p_W(x|w_{\mathcal{E}}, \theta_{\mathcal{E}})$ and $p(x|\theta_e)$ can be nested (in one another or within a larger family of distributions). This is the case for Examples 3 and 4.

If we expect different models and experts to perform better over different subsets of $x_1, ..., x_T$, then we may consider that it is preferable to have $a_{e,t}$ being dependent on $t$. In this case, the bound is relative to the best partition of experts.

### 2.0.1 Prequential Interpretation

The function $-\sum_{t=1}^{T} \mathcal{R}(p_{W,t}(x_t))$ differs from the usual likelihood function, as the the loglikelihood per observation at time $t$ is constructed using $\mathcal{F}_{t-1}$ measurable parameters. This loglikelihood is called the prequential likelihood (Dawid, 1986) and according to the same literature, $\sum_{t=1}^{T} \mathcal{R}(p_{W,t})$ is a proper scoring rule; smaller values are preferred to larger. This approach of model evaluation is consistent with the Popperian view that the validity of the model should be tested on observables. There is no need of introducing the concept of probability in this context: we are not finding an estimator for the maximum of the expected log-likelihood. We are only trying to minimize the total loss: this is not a probability problem, but a pattern recognition one (though the two may be related at some level).

# 3 The Algorithm

This section introduces the multiplicative algorithms that will be used for issuing the probability forecasts of the econometrician.

We need to find an $\mathcal{F}_{t-1}$ measurable strategy that produces the weights $(w_e^t)_{e\in\mathcal{E}}$. This is achieved using multiplicative updating algorithms. These algorithms have been studied by several authors (e.g. Vovk, 1990, Cesa-Bianchi et al, 1997, Herbster and Warmuth, 1998, Bousquet and Warmuth, 2002). We need to define transition functions $\mathsf{u}_t(e, e') : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$. These functions are called the share update functions. The choice of $\mathsf{u}_t$ is a fundamental ingredient that determines the order of magnitude of *error* in the above displays. A precise definition will be given later. Unless specified othewise, in the remaining of this section, we write $\theta_e^t$ for $\hat{\theta}_e^t$, which is an expert's forecast. The algorithm $W$ is a as follows.

Set
$v_{e,1} = w_{e,1} := 1/\left(\#\mathcal{E}\right), \forall e;$
$p_W = p_W\left(x_1 | w_{e,1}, \theta_e^1\right);$
$\mathcal{R}\left(x\right) := -\ln x.$

For $t = 1, ..., T-1,$
$p_{\theta(e,t)} = p_{\theta(e,t)}\left(x_t\right),$
$v'_{e,t} = v_{e,t} \exp\left\{-\eta \mathcal{R}\left(p_{\theta(e,t)}\right)\right\},$
$v_{e,t+1} = \sum_{e' \in \mathcal{E}} v'_{e',t} \mathsf{u}_{t+1}\left(e, e'\right),$
$v_{,t+1} = \sum_{e \in \mathcal{E}} v_{e,t+1},$
$w_{e,t+1} = v_{e,t+1} / v_{,t+1},$
$p_W = p_W\left(x_{t+1} | w_{\mathcal{E},t+1}, \theta_{\mathcal{E}}^{t+1}\right),$
$\mathcal{R}_{1,...,t+1} = \mathcal{R}_{1,...,t} + \mathcal{R}\left(p_W\right).$

**Remark 5** *The parameter $\eta$ is called the learning rate and depends on the prediction function used. For predictions based on model averaging as in Example 3, $\eta = 1$.*

**Remark 6** *For $\mathsf{u}_{t+1}\left(e, e'\right) = \delta\left(e - e'\right)$ the weight update is the one originally proposed by Vovk (1990) and also considered in Yang (2004).*

### 3.0.2 Bayesian Interpretation

When $\eta = 1$, the algorithm has a Bayesian interpretation. Suppose that $\left(E_t\right)_{t \in \mathbb{N}}$ is a sequence of random variables with values in $\mathcal{E}$, which does not need to be $\mathcal{F}_t$ measurable. Using the notation from the Introduction,

$$\partial_{x(t)} \Pr\left(X_t \leq x_t | E_t = e, \mathcal{F}_{t-1}\right) := \exp\left\{-\mathcal{R}\left(p_{\theta(e,t)}\right)\right\} = p_{\theta(e,t)}\left(x_t\right),$$

and

$$\partial_{x(t)} \Pr\left(X_t \leq x_t | \mathcal{E}, \mathcal{F}_{t-1}\right) = \sum_{e \in \mathcal{E}} w_{e,t} p_{\theta(e,t)}\left(x_t\right).$$

Therefore, the algorithm implies that the distribution of $E_t$ is characterised by the following quantities

$$\Pr\left(E_t = e | \mathcal{F}_{t-1}\right) = w_{e,t},$$

$$\Pr\left(E_t = e | \mathcal{F}_t\right) \propto \Pr\left(E_t = e | \mathcal{F}_{t-1}\right) \partial_{x(t)} \Pr\left(X_t \leq x_t | E_t = e, \mathcal{F}_{t-1}\right) \propto v'_{e,t}$$

$$\Pr\left(E_{t+1} = e_{t+1} | E_t = e_t, \mathcal{F}_t\right) \propto \mathsf{u}_{t+1}\left(e_{t+1}, e_t\right),$$

though this last display would be valid only under specific restrictions on $\mathsf{u}_{t+1}$, and

$$\begin{aligned}\Pr\left(E_{t+1}=e|\mathcal{F}_t\right) &= \sum_{e\in\mathcal{E}}\Pr\left(E_t=e|\mathcal{F}_t\right)\Pr\left(E_{t+1}=e_{t+1}|E_t=e_t|\mathcal{F}_t\right)\\ &\propto v_{e,t+1}\propto w_{e,t+1}.\end{aligned}$$

This last step is the share update and it is independent of the prediction scheme chosen by the econometrician.

Several prediction functions have been considered in the literature on forecasts of distributions combining experts (e.g. Genest and Zidek, 1986, see also Timmermann, 2004). The prediction function needs to satisfy Condition 1. Hence, prediction functions often found to be preferable (e.g. the logarithmic opinion poll) may not be adequate in the present context. The prediction function from Example 3 is usually multimodal and dispersed, but it always satisfies Condition 1 with $c=1/\eta=1$, admitting the above Bayesian interpretation.

### 3.0.3 Differences from a Bayesian Prediction

Notice that in a Bayesian framework, $e$ would be usually associated to a model depending on an unknown parameter. We can notice that the Bayes predictor assuming $E_t=e$ is

$$\theta^t_{B(e)}:=\arg\min_{\theta}\mathbb{E}\left[\mathcal{R}\left(p\left(X_t|\theta\right)\right)|E_t=e,\mathcal{F}_{t-1}\right].$$

The experts' forecast $\hat{\theta}^t_e$ does not need to be equal to $\theta^t_{B(e)}$. Then,

$$\theta^t_B:=\min_{e\in\mathcal{E}}\sum_{e'\in\mathcal{E}}\Pr\left(E_t=e'|\mathcal{F}_{t-1}\right)\mathbb{E}\left[\mathcal{R}\left(p\left(X_t|\theta^t_{B(e)}\right)\right)|E_t=e',\mathcal{F}_{t-1}\right] \tag{2}$$

is the Bayes choice of $\theta^t$. Alternatively, we can average over the models and

$$\theta^t_{BA}:=\min_{\theta}\sum_{e\in\mathcal{E}}\Pr\left(E_t=e|\mathcal{F}_{t-1}\right)\mathbb{E}\left[\mathcal{R}\left(p\left(X_t,\theta\right)\right)|E_t=e,\mathcal{F}_{t-1}\right] \tag{3}$$

is the Bayes average choice of $\theta^t$. We notice the following two differences. First, (3) delivers a value for $\theta$, and not the whole model. However, we can identify the whole model as the mixture of densities using the optimal parameter. Second, the criterion function for (2) and (3) is derived using expectation of the risk in terms of the conditioning model. The criterion function of the sequential algorithm is the prequential log-likelihood and no expectation is taken.

## 3.1 Properties of the Algorithm

We introduce the following.

**Condition 7** $\sum_{e\in\mathcal{E}} v'_{e,t} \geq \sum_{e\in\mathcal{E}} v_{e,t+1}$.

**Remark 8** *Condition 7 put a restriction on* $\mathsf{u}_{t+1}(e_{t+1}, e_t)$ *saying that*

$$\sum_{e\in\mathcal{E}} v_{e,t+1} = \sum_{e\in\mathcal{E}} \left( \sum_{e'\in\mathcal{E}} v'_{e',t} \mathsf{u}_{t+1}(e, e') \right) \leq \sum_{e\in\mathcal{E}} v'_{e,t}.$$

*The most simple example is given by* $\mathsf{u}_{t+1}(e, e') = 1/(\#\mathcal{E})$, *or* $\mathsf{u}_{t+1}(e, e') = \delta(e - e')$, *where* $\delta$ *is the Dirac delta function (i.e.* $\delta(0) = 1$, *zero elsewhere). This condition is satisfied with equality if* $(\mathsf{u}_{t+1}(e, e'))_{(e,e')\in\mathcal{E}\times\mathcal{E}}$ *is a doubly stochastic matrix (i.e. a Markov transition matrix).*

**Theorem 9** *Under Conditions 1 and 7,*

$$\mathcal{R}_{1,...,T}(p_W) \leq c\eta \mathcal{R}_{1,...,T}\left(p_{\theta(e)}\right) - c \ln \left( \prod_{t=1}^{T} \mathsf{u}_{t+1}(e, e) \right) - c \ln v_{e,1}$$

**Remark 10** *The bound shows that we need*

$$-\ln \left( \prod_{t=1}^{T} \mathsf{u}_{t+1}(e, e) \right) - \ln v_{e,1}$$

*to be as small as possible. This can be achieved by choosing* $v_{e,1}$ *as in the algorithm, i.e.* $v_{e,1} = 1/(\#\mathcal{E})$, *and* $\mathsf{u}_{t+1}(e, e') = \delta(e - e')$. *If we restrict* $a_{\mathcal{E},t}$ *in (1) to be on one of the edges of the simplex, then* $v_{e,1} = 1/(\#\mathcal{E})$, *and* $\mathsf{u}_{t+1}(e, e') = \delta(e - e')$ *are optimal choices.*

To state the next result we introduce some extra notation.

**Notation 11** *We divide the segment* $\mathcal{I}_T = (1, ..., T)$ *into K+1 subsegments,* $\mathcal{I}_{T(k)} = (t_k, ..., t_{k+1} - 1)$ *that are mutually exclusive and exhaustive,* $\mathcal{I}_T = \bigcup_{k=0}^{K} \mathcal{I}_{T(k)}$. *According to this notation,* $t_0 = 1$, *and* $\#\mathcal{I}_{T(k)} = t_{k+1} - t_k$. *Define* $e_k \in \mathcal{E}$.

**Theorem 12** *Under Conditions 1 and 7,*

$$\begin{aligned} \mathcal{R}_{1,...,t}(p_W) \quad \leq \quad & c\eta \sum_{k=0}^{K} \mathcal{R}_{t(k),...,t(k+1)-1}\left(p_{\theta(e(k))}\right) + c \ln(\#\mathcal{E}) \\ & -c \sum_{k=1}^{K} \ln \mathsf{u}_{t(k)}(e_k, e_{k-1}) - c \sum_{k=0}^{K} \sum_{s=t(k)}^{t(k+1)-2} \ln\left(\mathsf{u}_{s+1}(e_k, e_k)\right). \end{aligned}$$

**Remark 13** *The additional terms*

$$-\sum_{k=1}^{K}\ln \mathsf{u}_{t(k)}\left(e_k,e_{k-1}\right)-\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln\left(\mathsf{u}_{s+1}\left(e_k,e_k\right)\right)$$

*account for the arbitrary $K$ partition. For the sake of explanation, suppose $\mathsf{u}_{t+1}\left(e,e'\right)$ were the transition probability of going from $e'$ to $e$ at time $t+1$. Comparing Theorems 9 and 12, we have*

$$-\sum_{t=1}^{T}\ln\left(\mathsf{u}_{t+1}\left(e,e\right)\right)\geq-\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln\left(\mathsf{u}_{s+1}\left(e_k,e_k\right)\right)$$

*if the probability of keeping the same expert is the same across experts. However, in Theorem 12 we also have the extra term*

$$-\sum_{k=1}^{K}\ln \mathsf{u}_{t(k)}\left(e_k,e_{k-1}\right),$$

*which accounts for shifting from expert $e_{k-1}$ to expert $e_k$ at time $t_k$. To minimise the bound we need to redistribute transition probabilities. It is clear that we should put little weight to transition from and to experts that are likely to provide bad performance. Unfortunately, this cannot be done without knowledge of who are the best performing experts. Herbster and Warmuth, 1998, provide two specific choices of $\mathsf{u}_{t+1}\left(e,e'\right)$ depending on a parameter that depends on $K$, the number of shifts only (the Fixed Share and Variable Share algorithms). Bousquet, 2003, provides a Bayesian algorithm that put some prior in the parameter of the Fixed Share algorithm in order to update this parameter sequentially.*

## 3.2 Algorithm to Learn the Share Update

Suppose that $\mathsf{u}_{t+1}\left(e,e'|\lambda\right),\ \lambda\in\Lambda$ is a class of share updates. Suppose we choose finite number of these updates functions with parameter $\lambda_l\ l\in\mathcal{L}$. We can extend the previous algorithm to the case where we want to find the best $\lambda_l$. For simplicity, but with abuse of notation, $\mathsf{u}_{t+1}\left(e,e'|l\right):=\mathsf{u}_{t+1}\left(e,e'|\lambda_l\right)$.

The following algorithm, $WL$, depends on a constant $\kappa>0$ which will be defined later. The algorithm is a as follows.

Set

$v_{e,1}=w_{e,1}:=1/\left(\#\mathcal{E}\right),\ \forall e;$

$v_{l,1}=\omega_{l,1}:=1/\left(\#\mathcal{L}\right),\ \forall l;$

$p_{\theta(e,t)}=p_{\theta(e,t)}\left(x_t\right);$

$p_W = p_W\left(x_1 | w_{e,1}\theta_e^1\right);$
$\mathcal{R}(x) := -\ln x.$

For $t = 1, ..., T-1$,
$p_{\theta(e,t)} = p_{\theta(e,t)}(x_t),$
$v'_{e,l,t} = v_{e,l,t} \exp\left\{-\eta \mathcal{R}\left(p_{\theta(e,t)}\right)\right\},$
$v_{e,l,t+1} = \sum_{e' \in \mathcal{E}} v'_{e',l,t} \mathsf{u}_{t+1}(e, e'|l),$
$v_{,l,t+1} = \sum_{e \in \mathcal{E}} v_{e,l,t+1},$
$w_{e,l,t+1} = v_{e,l,t+1} / v_{,l,t+1},$
$p_{W(l)} = p_W\left(x_{t+1} | w_{\mathcal{E},l,t+1}\theta_{\mathcal{E}}^{t+1}\right),$
$v_{l,t+1} = v_{l,t} \exp\left\{-\kappa \mathcal{R}\left(p_{W(l)}\right)\right\},$
$v_{,t+1} = \sum_{e \in \mathcal{E}} v_{e,l,t+1},$
$\omega_{l,t+1} = v_{l,t+1} / v_{,t+1},$
$p_{WL} = p_{WL}\left(x_{t+1} | \omega_{\mathcal{L},t+1}, w_{\mathcal{E},\mathcal{L},t+1}, \theta_{\mathcal{E}}^{t+1}\right),$
$\mathcal{R}_{1,...,t+1} = \mathcal{R}_{1,...,t} + \mathcal{R}(p_{WL}).$

We need to extend Condition 1.

**Condition 14** *For any experts' forecast* $\hat{\theta}_{\mathcal{E}} = \left(\hat{\theta}_e\right)_{e \in \mathcal{E}}$, *outcome* $x$, *and weights* $w_{\mathcal{E}} = (w_e)_{e \in \mathcal{E}} \in \mathcal{S}_{\#\mathcal{E}}$, *and* $\omega_{\mathcal{L}} = (\omega_l)_{l \in \mathcal{L}} \in \mathcal{S}_{\#\mathcal{L}}$, $\exists b < \infty, \kappa > 0$ *such that*

$$\mathcal{R}\left(p_{WL}\left(x|, \omega_{\mathcal{L}}, w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)\right) \leq -b \ln \sum_{l \in \mathcal{L}} \omega_l \exp\left\{-\kappa \mathcal{R}\left(p_{W(l)}\left(x | w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)\right)\right\},$$

*where* $p_{W(l)}\left(x | w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)$ *is the econometrician's forecast using algorithm* $W$ *and share update* $\mathsf{u}_{t+1}(e, e'|l)$.

**Remark 15** *Notice that if the forecast is through mixtures, then,*

$$p_{WL} = \sum_{l \in \mathcal{L}} v_{l,t} p_{W(l)}$$

*and Condition 14 holds automatically with* $b = 1/\kappa = 1$.

**Theorem 16** *Under Conditions 1, 7, and 14,* $\forall e, l, K$

$$\begin{aligned}\mathcal{R}_{1,...,t}(p_{WL}) \quad \leq \quad & (b\kappa c\eta) \sum_{k=0}^{K} \mathcal{R}_{t(k),...,t(k+1)-1}\left(p_{\theta(e(k))}\right) + bc \ln(\#\mathcal{E}) + b \ln(\#\mathcal{L}) \\ & -bc \sum_{k=1}^{K} \ln \mathsf{u}_{t(k)}(e_k, e_{k-1}|l) - bc \sum_{k=0}^{K} \sum_{s=t(k)}^{t(k+1)-2} \ln \mathsf{u}_{s+1}(e_k, e_k|l).\end{aligned}$$

**Corollary 17** *Under Conditions 1, 7, and 14,* $\forall K$

$$\begin{aligned}\mathcal{R}_{1,...,t}\left(p_W\right) &\leq (b\kappa c\eta)\sum_{k=0}^{K}\mathcal{R}_{t(k),...,t(k+1)-1}\left(p_{\theta(e(k))}\right)+bc\ln\left(\#\mathcal{E}\right)+b\ln\left(\#\mathcal{L}\right)\\ &\quad+(bc)\min_{\substack{e\in\mathcal{E}\\ l\in\mathcal{L}}}\left(-\sum_{k=1}^{K}\ln \mathsf{u}_{t(k)}\left(e_k,e_{k-1}|l\right)-\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln \mathsf{u}_{s+1}\left(e_k,e_k|l\right)\right).\end{aligned}$$

**Remark 18** *Theorem 16 says that increasing the bound by* $b\ln\left(\#\mathcal{L}\right)$ *we can learn the minimising* $\lambda_l$ $l\in\mathcal{L}$*. In this case, to bound*

$$-\sum_{k=1}^{K}\ln \mathsf{u}_{t(k)}\left(e_k,e_{k-1}|l\right)-\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln \mathsf{u}_{s+1}\left(e_k,e_k|l\right),$$

*we need to choose a specific family* $\mathsf{u}_{t+1}\left(e,e'|\lambda\right)$ $\lambda\in\Lambda$ *and a finite collection of* $(\lambda_l)_{l\in\mathcal{L}}$ *over which to minimise.*

**Remark 19** *As for Theorem 12, we can choose a prediction function (e.g. the one of Example 3) such that* $\kappa=\eta=1$ *and* $b=c=1$*. More generally, the prediction functions considered in the examples of this paper are such that* $\kappa b=1$ *and* $\eta c=1$*. However, the results cover possibly more general cases.*

## 3.3 Some Choices for the Share Update

The bound of Theorem 16 says that the algorithm $WL$ leads to an efficient choice of expert and parameter $\lambda_l$ for $\mathsf{u}_{t+1}\left(e,e'|\lambda\right)$. However, to find an explicit bound we need to spicery the class of functions. There are several choices, and here three are presented.

### 3.3.1 Fixed Share

Following Remark 13, we can choose the transition share update for keeping the same expert to be independent of the expert and obtain the Fixed Share update of Herbster and Warmuth (1998),

$$\mathsf{u}_{t+1}\left(e,e'|\lambda\right)=(1-\lambda)\,\delta\left(e-e'\right)+\frac{\lambda}{\#\mathcal{E}-1}\left[1-\delta\left(e-e'\right)\right], \tag{4}$$

where $\lambda\in\Lambda=[0,1]$. From Remark 8, we see that (4) satisfies Condition 7. We can set $\lambda_l=l/L$, $l=0,...,L$, $L\in\mathbb{N}$, so that $\#\mathcal{L}=L$. To get a bound for this update, we need the following.

**Lemma 20**

$$\sum_{k=1}^{K} \ln \mathsf{u}_{t(k)}\left(e_k, e_{k-1}|\lambda\right) = K \ln\left(\frac{\lambda}{\#\mathcal{E}-1}\right)$$

$$\sum_{k=0}^{K} \sum_{s=t(k)}^{t(k+1)-2} \ln \mathsf{u}_{s+1}\left(e_k, e_k|\lambda\right) = (T-K-1)\ln\left(1-\lambda\right).$$

Applying Theorem 16, together with this lemma, we have the following.

**Corollary 21** *Under the Conditions of Theorem 16, using the* Fixed Share *update (4), and* $\lambda_l = l/L$, $l = 0, ..., L$, $L \in \mathbb{N}$,

$$\begin{aligned}\mathcal{R}_{1,...,t}\left(p_W\right) \leq\ & (b\kappa c\eta)\sum_{k=0}^{K}\mathcal{R}_{t(k),...,t(k+1)-1}\left(p_{\theta(e(k))}\right) + bc\ln\left(\#\mathcal{E}\right) + b\ln\left(\#\mathcal{L}\right) \\ & + (bc)\min_{l\in\mathcal{L}}\left(-K\ln\left(\frac{l/L}{\#\mathcal{E}-1}\right) - (T-K-1)\ln\left(1-l/L\right)\right).\end{aligned}$$

### 3.3.2 Variable Share

Alternatively, we can choose the transition share update for keeping the same expert to depend on the expert ( e.g. the Variable Share update of Herbster and Warmuth, 1998). Here we propose the following new update,

$$\mathsf{u}_{t+1}\left(e, e'|\lambda\right) = \left[1-\lambda\left(1-\beta\left(e', t\right)\right)\right]\delta\left(e-e'\right) + \lambda\beta\left(e', t\right)\left[1-\delta\left(e-e'\right)\right], \qquad (5)$$

where $\beta\left(e,t\right) \in \mathcal{S}_{\#\mathcal{E}}$ and $\lambda \in [0,1]$.

**Example 22** *Suppose* $\beta\left(e,t\right) := p_{\theta(e,t)} / \sum_{e\in\mathcal{E}} p_{\theta(e,t)}$. *Then, the probability of switching expert is affected by the performance in the last trial.*

**Example 23** *Suppose* $r_h\left(e,t\right)$ *is the ranking of expert* $e$ *in the interval* $[t-h,t]$ *relative to the other experts over the same time span. The ranking can be based on the median of* $\left(\mathcal{R}\left(p_{\theta(e,s)}\right)\right)_{s\in\{t-h,...,t\}}$. *Then,*

$$\beta\left(e,t\right) := \left[r_h\left(e,t\right)\right]^{-1} / \sum_{e\in\mathcal{E}}\left[r_h\left(e,t\right)\right]^{-1}$$

*(e.g. Timmermann, 2004, and references therein). In this case, the switching probability depends on the performance over* $[t-h,t]$ *and not just on the last trial. Moreover, ranking is less sensitive against outliers, hence this could be a robust rule to use.*

Since $\sum_{e\in\mathcal{E}}\beta(e,t)=1$ it is easy to see that (5) satisfies Condition 7. The interpretation of (5) in terms of transition probabilities helps our intuition. In this case, the probability of changing state from $e_t$ to $e_{t+1}$ is a function of $\beta(e_t,t)$, hence it depends on the original state $e_t$. For (4), this probability is independent of the original state.

To get a bound for this update, we need the following.

**Lemma 24** *Suppose* $\mathsf{u}_{t+1}(e,e'|\lambda)$ *is as in (5). Then,*

$$\sum_{k=1}^{K}\ln\mathsf{u}_{t(k)}(e_k,e_{k-1}|\lambda)=K\ln\lambda+\sum_{k=1}^{K}\ln\left(1-\beta(e_{k-1},t_k-1)\right),$$

*and*

$$\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln\mathsf{u}_{s+1}(e_k,e_k|\lambda)\geq(T-K-1)\ln(1-\lambda)\vee\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln\left(\beta(e(k),s)\right),$$

*with strict inequality if* $(\lambda,\beta(e,s))\in(0,1)^2\ \forall e,s$.

**Corollary 25** *If* $\beta(e,t)<(1-1/\#\mathcal{E})\ \forall t,e$ *then*

$$-\sum_{k=1}^{K}\ln\mathsf{u}_{t(k)}(e_k,e_{k-1}|l)$$

*is strictly smaller for (5) than for (4); if* $\beta(e,s)<(1-\lambda)$ *then*

$$-\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln\mathsf{u}_{s+1}(e_k,e_k|\lambda)$$

*is smaller for (5) than for (4), otherwise they are equal.*

**Remark 26** *By Corollary 25, if we restrict* $\beta(e,t)<(1-1/\#\mathcal{E})$, *there is an improvement in the bound of Theorem 16 if we use (5) instead of (4). The second part of Corollary 25 requires knowledge of* $\lambda$. *Fortunately, even without knowledge of* $\lambda$, *only restricting* $\beta(e,t)<(1-1/\#\mathcal{E})$, *the bound in Theorem 16 cannot be worse than the one obtained by using (4).*

**Mixing Past** Another choice for $\mathsf{u}_{t+1}(e,e'|\lambda)$ is given by taking averages of past weights over the same expert. In this case,

$$\mathsf{u}_{t+1}(e,e'|\lambda)=[\lambda_t-\beta_t]\,\delta(e-e'),$$

where $\lambda = (\lambda_0, ..., \lambda_t) \in \mathcal{S}_t$, $\beta_t = \left(v'_{e,t}\right)^{-1} \sum_{s=0}^{t-1} \lambda_s v'_{e,s}$ and $v'_{e,s}$ is the intermediate weight for expert $e$ at time $s$. This essentially leads to the algorithm in Bousquet and Warmuth (2002). The bound for this updating scheme can be obtained from our previous results with no extra effort. For the sake of brevity, details are left to the reader.

# 4 Illustration: Choosing the Right Linear Filter

Consider the family of distributions $\left\{P_\theta, \theta \in \Theta \subset\subset \mathbb{R}^d, d \geq 1\right\}$. For the sake of simple explanation restrict $d = 1$. Suppose there is some function $g : \mathbb{R} \to \mathbb{R}$ such that $\theta^t = \mathbb{E}\left(g\left(X_t\right)|\mathcal{F}_{t-1}\right)$. This is often the case. For example, the exponential family model satisfies this (e.g. Normal distribution, Poisson, Bernoulli). Then, suppose $g\left(X_t\right)$ admits the semimartingale representation

$$g\left(X_t\right) = f_t + v_t \varepsilon_t, \tag{6}$$

where $f_t$ and $v_t$ are $\mathcal{F}_{t-1}$ measurable and $\mathbb{E}\varepsilon_t = 0$, $\mathbb{E}\varepsilon_t^2 = 1$. Hence, $\theta^t = f_t$ (by reparametrisation of the marginal distributions, this covers the case $g'\left(\theta^t\right) = f_t$, for some function $g'$). The parameter estimation is equivalent to estimation of the $\mathcal{F}_{t-1}$ measurable trend in $g\left(X_t\right)$. We can estimate or at least approximate $\theta^t$ by $\hat{\theta}^t = \sum_{s<t} w\left(s,t\right) g\left(x_s\right)$, where $\left(w\left(s,t\right)\right)_{0\leq s<t\in\mathbb{N}_+}$ is a linear filter possibly depending on $\left(X_s\right)_{s<t}$ so that $\left(w\left(s,t\right)\right)_{0\leq s<t\in\mathbb{N}_+}$ is $\mathcal{F}_{t-1}$ measurable. This framework encompasses many different methods like averages, moving averages, exponential smoothing, kernel smoothing and linear projections. In the case of linear projections, the whole filter is given by the projection matrix.

The case where we suppose that there is a function $g : \mathbb{R} \times \Theta \to \mathbb{R}$ such that $\mathbb{E}\left(g\left(X_t, \theta_t\right)|\mathcal{F}_{t-1}\right) = 0$, can be covered similarly via linear approximation (e.g. Sancetta and Nikandrova, 2005, and Polzehl and Spokoiny, 2004, for an application to GARCH). The $K > 1$ dimensional case is dealt similarly either by defining a vector of estimating equations or by direct solution if the parameter defining equation admits an explicit solution as a function of $X_t$. Reparametrisation may be used to simplify the estimation and to imply different dynamics.

In these cases, the crucial step is the choice of $\left(w\left(s,t\right)\right)_{0\leq s<t\in\mathbb{N}_+}$. Different experts that use different filters can be combined in order to obtain a probability forecast. The following empirical example pursue this route.

## 4.1 Empirical Example

We use the Gaussian distribution with time varying mean and variance as an empirical illustration of forecasts' combination using experts for the log-returns of the S&P500 index. The period chosen is 02/Jan/1970-04/Mar/2005, which will lead to 8600 predictions considering a start up period of about 200 observations.

The Gaussian distribution may not be a good choice to model assets returns. However, this is irrelevant to the purpose of illustrating the theoretical results of the previous section. Given this choice, the parameters to be estimated are the mean $\mu_t = \mathbb{E}(X|\mathcal{F}_{t-1})$ and the variance $\sigma_t^2 = \mathbb{E}(X^2|\mathcal{F}_{t-1})$. We suppose that the experts give us forecasts for these parameters.

The experts estimate the parameters as follows. Let $\lfloor x \rfloor$ be the integer part of $x \in \mathbb{R}$, and $G_n : \mathbb{R} \to \mathbb{Z}$ be a function such that

$$G_n(x) := \begin{cases} 2^{-n} \lfloor x 2^n \rfloor, & \text{if } |x| < n \\ n(x/|x|), & \text{if } |x| \geq n \end{cases}.$$

Define $x_{t-l}^{t-1} := x_{t-l}, ..., x_{t-1}$. With abuse of notation

$$G_n\left(x_{t-l}^{t-1}\right) := \left[G_n(x_{t-l}), ..., G_n(x_{t-1})\right].$$

We define the following linear filters

$$w(s,t) = \begin{cases} \frac{I\left\{G_n\left(x_{t-l}^{t-1}\right)=G_n\left(x_{s-l}^{s-1}\right)\right\}}{\left(\#\left\{0\leq s<t: G_n\left(x_{t-l}^{t-1}\right)=G_n\left(x_{s-l}^{s-1}\right)\right\}\right)}, & \text{if } \#\left\{0 \leq s < t : G_n\left(x_{t-l}^{t-1}\right) = G_n\left(x_{s-l}^{s-1}\right)\right\} > 0 \\ 1/t & \text{otherwise} \end{cases}$$

and

$$w'(s,t) = (1-h) h^{t-s} I(s < t).$$

The first filter leads to the regression on $l$ past values binned using $G_n$, while the second leads to exponential smoothing. The first filter has been used by Yakowitz et al. (1999) to construct strongly consistent forecasts of stationary and ergodic time series and by Györfi and Lugosi (2002) in the context of experts' forecasting. The second is commonly employed for trend estimation of time series data and may be justified as optimal filter in a mean square error sense for random walk plus noise dynamics.

Using the arguments from the previous subsection, with $g(x) = x$, $(x - \hat{\mu}_t)^2$ and $\hat{f}_t = \hat{\mu}_t$, $\hat{\sigma}_t^2$, $\hat{\mu}_t'$, $\hat{\sigma}_t^{2\prime}$,

$$\begin{aligned} \hat{\mu}_t(h_1) &= \sum_{s<t} (1-h_1) h_1^{t-s} x_s \\ \hat{\sigma}_t^2(h_2, h_1) &= \sum_{s<t} (1-h_2) h_2^{t-s} (x_s - \hat{\mu}_t(h_1))^2 \end{aligned}$$

and

$$\begin{aligned}\hat{\mu}'_t(l,n) &= \frac{\sum_{\left\{s<t:G_n\left(x_{t-l}^{t-1}\right)=G_n\left(x_{s-l}^{s-1}\right)\right\}} x_s}{\#\left\{0\leq s<t:G_n\left(x_{t-l}^{t-1}\right)=G_n\left(x_{s-l}^{s-1}\right)\right\}} \\ \hat{\sigma}^{2\prime}_t(l,n) &= \frac{\sum_{\left\{s<t:G_n\left(x_{t-l}^{t-1}\right)=G_n\left(x_{s-l}^{s-1}\right)\right\}} \left(x_s-\hat{\mu}_t(n,l)\right)^2}{\#\left\{0\leq s<t:G_n\left(x_{t-l}^{t-1}\right)=G_n\left(x_{s-l}^{s-1}\right)\right\}}.\end{aligned}$$

We use two sets of experts in terms of the forecasts $\left(\hat{\mu}_t(h_1), \hat{\sigma}^2_t(h_2,h_1)\right)$ with $h_i = 5,10,20,40,80,160$ $(i=1,2)$ and $\left(\hat{\mu}'_t(l,n), \hat{\sigma}^{2\prime}_t(l,n)\right)$ $l = 1,2,4,8,16,32$, $n = 0,1,2,3,4$. This means that for the exponential smoothing estimators we take all the possible combinations of mean and variance estimators based on $h$ in the given grid of values. For the regression estimators of mean and variance we keep the same order of autoregression but use different binwidth. This leads to a total number of 66 experts some of which are redundant (e.g. for $n=0$, $\hat{\mu}_t(l,n) = \hat{\mu}_t(l',n)$, $\forall l,l'$).

We compute the following algorithms. Algorithm $W$ is computed for the no share update (i.e. $\mathsf{u}_{t+1}(e,e') = \delta(e-e')$) and the fixed share and the variable share with $\beta$ as in Example 22 and different values of $\lambda$. Algorithm $WL$ is also computed for the fixed and variable share update. In particular we choose $\lambda = 1/8, 2/8, ..., 7/8$. We also computed the variable share update with $\beta(e,t) < (1-1/\#\mathcal{E})$ as discussed in Remark 26, but the constraint was never binding.

The prediction function used is the mixture of distributions, as in Example 3, so that $\eta = 1/c = 1$. Five of the 66 experts incurred an infinite loss at some point. Table I gives summary statistics for the prequential loglikelihood of the worse expert with finite loss function (expert 57: $\hat{\mu}_t(4,3)$, $\hat{\sigma}^2_t(4,3)$), the best expert (expert 24: $\hat{\mu}_t(160)$, $\hat{\sigma}^2_t(40,160)$), the best experts' partition, the $W$ algorithm using the no share update, and the $WL$ algorithm using the fixed share update and the variable share update. Clearly, the best expert partition cannot be achieved. However, algorithm $W$ does achieve the best expert bound.

Algorithm $WL$ improves on the best expert bound without assuming a specific value of $\lambda$. This algorithm also reduces the variability of the loss almost to the level of variability achieved by the best experts' partition. A close look at the predictions can reveal that the performance of algorithm $WL$ started to improve on algorithm $W$ from the big crash of October 1987. As shown in Figure I, before then, the difference was marginal. Hence, algorithm $WL$ may help to hedge against nonstationary behaviour as the crash of October 1987.

Table II reports some details about the last weights for algorithm $W$ using no share update and the share updates for different values of $\lambda$ together with the

weights assigned by algorithm $WL$ to the different values of $\lambda$. Algorithm $W$ with no share update learnt that expert 24 was the best one. Algorithm $WL$ algorithms keep a positive weight for all experts, as no experts in both updates receives less than 0.3% of the weight, but no expert receives more than 3.3% of the weight. This is expected, as Algorithm $WL$ allows for the best expert to change overtime, and the worse expert could be the next best performing expert. Moreover, both updates favour a infrequent change in the best expert.

Figure III plots the cumulative loss over the last 600 observations for the best expert and the no share update, and the fixed share and variable share updates using algorithm $WL$. Figure IV shows the results for the last 600 observations using algorithm $W$ with fixed share update and different values of $\lambda$.

Table I. Summary of Experts' Performance.

| | Min. | 1st Qu. | Median | Mean | St.Dev. | 3rd Qu. | Max. | Tot. Loss |
|---|---|---|---|---|---|---|---|---|
| Expert 57 | -1.1632 | 0.8363 | 1.0264 | 1.9741 | 11.5071 | 1.5437 | 706.4390 | 16977.10 |
| Expert 24 | -0.0585 | 0.7192 | 1.0616 | 1.3054 | 1.5363 | 1.5336 | 86.8573 | 11226.73 |
| Best Experts' Partition | -2.1915 | 0.3907 | 0.7182 | 0.8112 | 0.8048 | 1.1247 | 23.8878 | 6976.02 |
| No Share Update | -0.1008 | 0.7192 | 1.0644 | 1.3059 | 1.5368 | 1.5347 | 86.8573 | 11230.92 |
| Fixed Share Update | -0.1950 | 0.7920 | 1.0497 | 1.2938 | 0.9455 | 1.5030 | 27.2893 | 11126.50 |
| Variable Share Update | -0.1451 | 0.7916 | 1.0497 | 1.2937 | 0.9471 | 1.4999 | 27.5577 | 11125.45 |

Figure I. Total Loglikelihood for Algorithms Comparison.

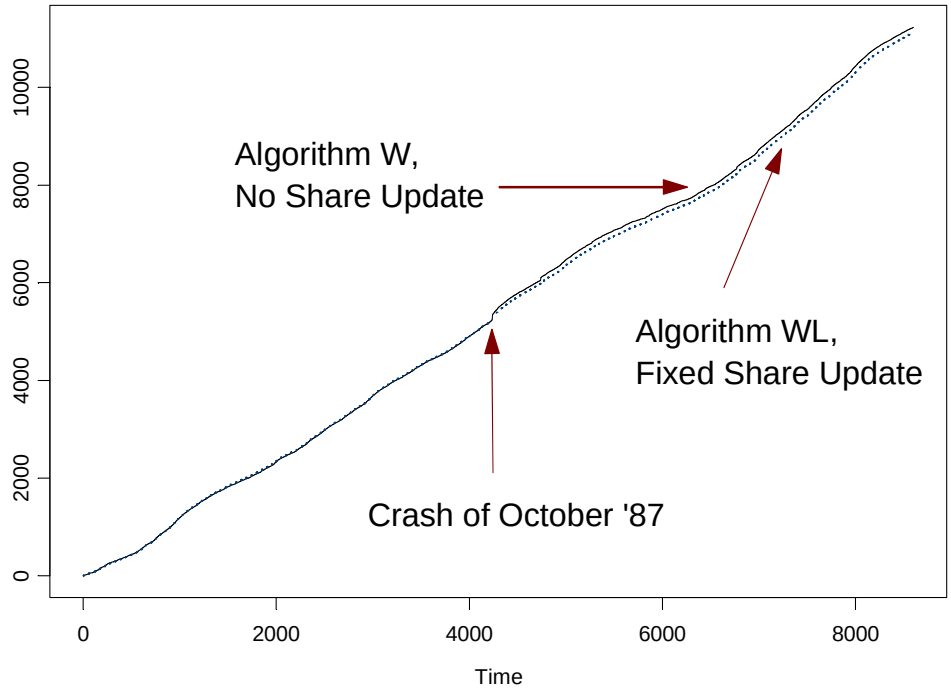



Table II. Experts' Weights.

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | No Share Update | | | |
| Weight | 1 for Expert 24, 0 for all Others | | | | | |
| | | | Fixed Share | | | |
| Lambda | 0.143 | 0.286 | 0.429 | 0.571 | 0.714 | 0.857 |
| Min Expert Weight | 0.003 | 0.006 | 0.008 | 0.010 | 0.012 | 0.013 |
| Max Expert Weight | 0.034 | 0.022 | 0.019 | 0.017 | 0.017 | 0.016 |
| Lambda Weight | 0.000 | 0.000 | 0.000 | 0.000 | 0.012 | 0.988 |
| | | | Variable Share | | | |
| Lambda | 0.143 | 0.286 | 0.429 | 0.571 | 0.714 | 0.857 |
| Min Expert Weight | 0.003 | 0.006 | 0.008 | 0.010 | 0.012 | 0.013 |
| Max Expert Weight | 0.033 | 0.022 | 0.019 | 0.017 | 0.017 | 0.016 |
| Lambda Weight | 0.000 | 0.000 | 0.000 | 0.000 | 0.036 | 0.963 |

Figure II. Total Loglikelihood for Best Expert, No Share,
Fixed Share and Variable Share over the Last 600 Observations

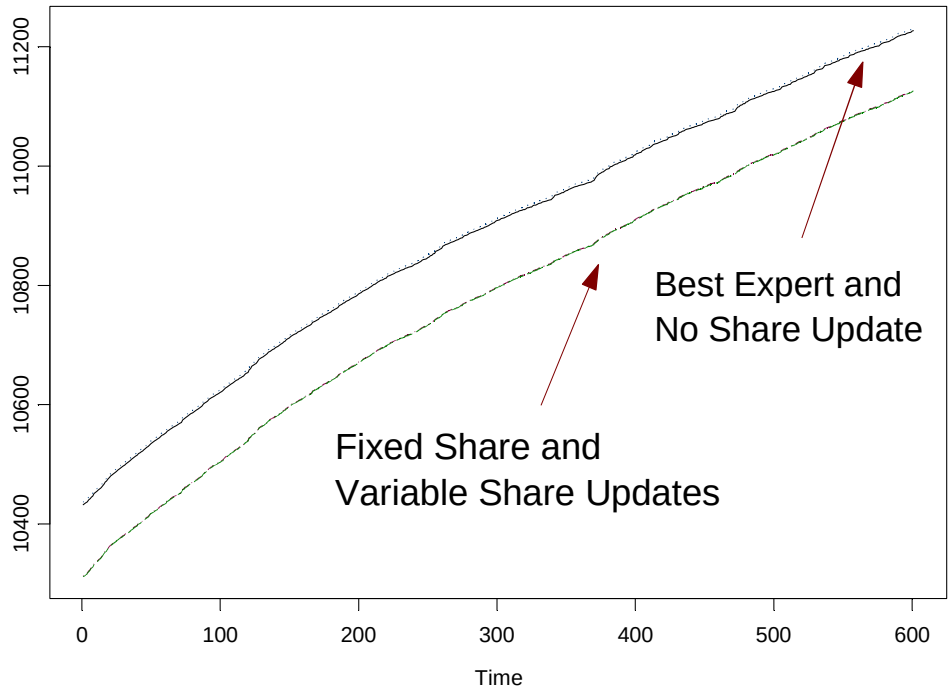
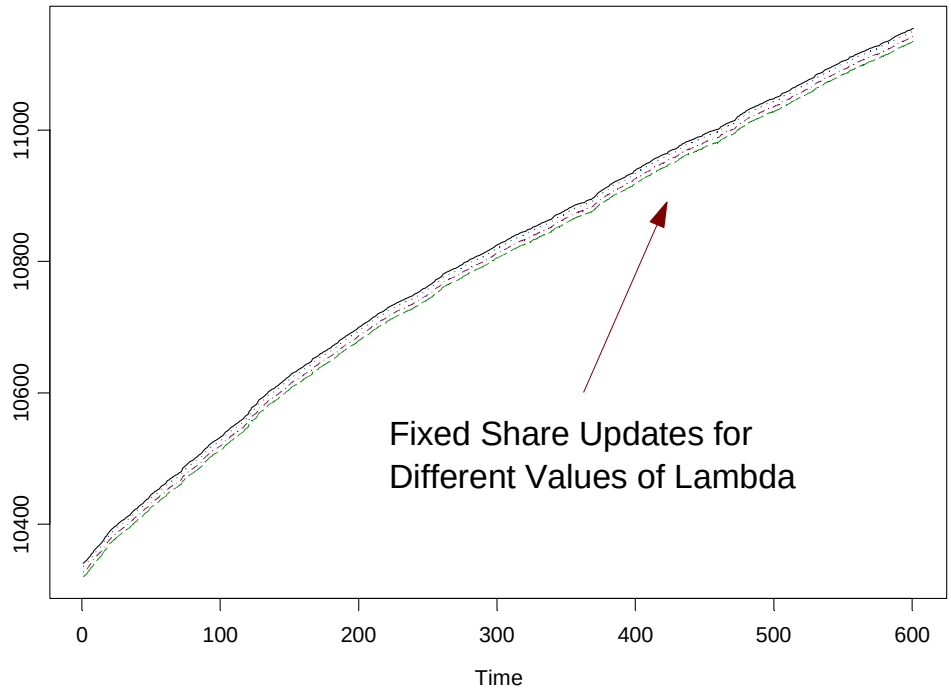


Figure III. Loglikelihood for Fixed Share for Different Values of $\lambda$.




# 5 Prediction of Individual Sequences

By a suitable choice of the family of distributions, forecast of distributions allows for forecast of individual sequences using the prediction function of Example 4.

**Example 27** *Define*

$$p_W\left(x|w_\mathcal{E},\hat{\theta}_\mathcal{E}\right)=\frac{1}{\sqrt{\pi}}\exp\left\{-\left|x-\left\langle w_\mathcal{E},\hat{\theta}_\mathcal{E}\right\rangle\right|^2\right\},$$

*which is the Gaussian density with mean* $\left\langle w_\mathcal{E},\hat{\theta}_\mathcal{E}\right\rangle$ *and variance* $1/2$. *Then the loss function is*

$$\mathcal{R}\left(p_W\left(x|w_\mathcal{E},\hat{\theta}_\mathcal{E}\right)\right)=\left|x-\left\langle w_\mathcal{E},\hat{\theta}_\mathcal{E}\right\rangle\right|^2+(1/2)\ln\pi.$$

*Since our results do not require* $p_W\left(x|w_\mathcal{E},\hat{\theta}_\mathcal{E}\right)$ *to integrate to one, the term* $(1/2)\ln\pi$ *can be dropped and* $\mathcal{R}\left(p_W\left(x|w_\mathcal{E},\hat{\theta}_\mathcal{E}\right)\right)$ *is exactly the square loss.*

**Example 28** *Define*

$$p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right)=a\exp\left\{-\left[\exp\left\{a\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\}-a\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right)\right]\right\},$$

*which is a scale location change of the Gumbel density. Then,*

$$\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right)\right)=\exp\left\{a\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\}-a\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right)+\ln\left(a\right)$$

*and replacing the irrelevant additive constant* $\ln(a)$ *with* $-1$, $\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right)\right)$ *becomes LinEx loss function with parameter* $a$.

**Example 29** *Define*

$$p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right)=\exp\left\{-\left|x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right|\right\},$$

*which is the double exponential density. Then*

$$\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right)\right)=\left|x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right|,$$

*which is the absolute loss function.*

For a loss function $\varphi(y)$, we do not need $\exp\{-\varphi(y)\}$ to be a density, what is required is that Conditions 1 and 14 are satisfied. If the predictions are obtained by parameters' averaging it is enough to check that Condition 1 is satisfied for some $c$ and $\eta$. In this case, all the bounds derived above apply to the prediction of individual sequences with $\kappa=\eta$ and $b=c$.

**Lemma 30** *Set* $c=1/\eta$ *and suppose* $p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right):=\exp\left\{-\varphi\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\}$, *where* $\varphi(y)$ *is a loss function. Suppose the sample of observations and their predictions are bounded. Then, Condition 1 is satisfied if for any finite absolute constant* $B$ *we can find an* $\eta\in(0,\infty)$ *such that* $\exp\{-\eta\varphi(y)\}$ *is concave for* $|y|\leq B$.

**Remark 31** *The condition that the sample observations are bounded implies that over the sample period we can find a constant large enough such that all the observations will be smaller in absolute values. It is rare to find applications were we observe data taking values equal to infinity. This is true for financial returns, as the exchange rules define a priori limits on the maximum and minimum price changes within a day. As discussed in Györfi and Lugosi (2002), if* $B$ *is unknown, we can fix* $B$ *to a large value, and if one of the observations happens to be larger, we just reset* $B$ *according to the new maximum value. Rerunning the algorithm with the new corresponding choice of* $\eta$ *makes sure that the bounds hold. Clearly, we could impose tail assumptions and truncate. But in the above bounds we are not even assuming that the segment* $x_1,...,x_t$ *is a realisation of some sequence of random variables, so this would not appear natural.*

**Example 32** *Suppose* $\varphi(y) = y^2$. *Then, Condition 1 is satisfied with* $\eta = 1/(2B^2)$. *To see this, differentiate* $\exp\{-\eta\varphi(y)\}$ *twice with respect to* $y$, *equate to zero to find the inflection points* $\pm 1/\sqrt{\eta 2}$. *Using the fact that* $y \in [-B, B]$ *we get the required value for* $\eta$.

**Example 33** *Suppose* $\varphi(y) = \exp\{ay\} - ay - 1$. *Then, Condition 1 is satisfied for* $\eta = \exp\{aB\}/(\exp\{aB\} - 1)^2$ *for* $y \in [-B, B]$.

The absolute norm $\varphi(y) = |y|$ does not satisfy the condition of Lemma 30. However, in this case, we use the following more general result that applies to all convex loss functions and sample sequences that only take bounded values.

**Lemma 34** *Define*

$$p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right) = \exp\left\{-\varphi\left(x - \left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\},$$

*and*

$$p_{\hat{\theta}(e)}(x) = \exp\left\{-\varphi\left(x - \hat{\theta}_e\right)\right\},$$

*where* $\varphi(y)$ *is a convex function. Then, for* $\left|x - \left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right| \leq B < \infty$

$$\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)\right) \leq -\eta^{-1}\ln\sum_{e\in\mathcal{E}} w_e \exp\left\{-\eta\mathcal{R}\left(p_{\hat{\theta}(e)}(x)\right)\right\} + \eta\varphi(2B)^2/8.$$

**Remark 35** *The extra term* $\eta\varphi(2B)^2/8$ *will result in an additional error equal to* $T\eta\varphi(2B)^2/8$ *in the bounds of the Theorems. By choice of* $\eta = O\left(T^{-1/2}\right)$ *the loss reduces to* $O\left(T^{1/2}\right)$.

# 6 Final Remarks

The bounds of this paper can be partially adapted to the case of an uncountable number of experts if the class of experts satisfies suitable entropy conditions (Cesa-Bianchi and Lugosi, 1999, for details). The uncountable case covers situations in which we average using a continuous mixing distribution instead of a finite number of weights for the forecast combination.

The algorithms considered enjoy some optimal theoretical properties. However, there could be other algorithms that lead to equivalent theoretical results or improve on the present ones. This will be the subject of future research.

We did not discuss how to choose our experts. We could clearly use a large number of them without any preliminary analysis. This was the case of our illustrative example, where few of them were known to be redundant. As shown in the Theorems, the error only grows logarithmically in the number of experts. Nevertheless, it is preferable to choose them carefully following some reasonable criterion. The role of sufficiency in forecast may play at the initial stage a fundamental role (see Timmermann, 2004).

In forecast combination, empirical evidence seems to suggest that it can be advantageous to trim the weights, setting very low weights equal to zero. Algorithm W with no share update does effectively set the weights of the worse performing experts equal to zero is run over long enough series. In general, if we dastard a fixed percentage of the worse models reducing the number of experts we track, then there is a gain if we are sure that the discarded models will never perform well. If we want to be able to resume these models, i.e. we still track these experts, then trimming can be carried out at the prediction stage leaving the weight updates unchanged. In this case, we need to check that Condition 1 is satisfied, which is the case if the reduced weights are not redistributed to the remaining weights. This may lead to problems as the weights that are kept would not add to one. If the weights are redistributed, we cannot be certainty that the worse performing expert suddenly becomes the best. The empirical example showed that this might be the case. To avoid such cases in deriving theoretical bounds, probabilistic assumptions need to be made and worse case bounds substituted by probabilistic bounds as in Yang (2004).

# A Proofs

## A.1 Theorems 9 and 12

The proof is based on the following Lemmata.

**Lemma 36** *Under Condition 1,*

$$\mathcal{R}\left(p_{W,t}\right) \leq -c \ln\left(\frac{\sum_{e\in\mathcal{E}} v'_{e,t}}{v_{,t}}\right).$$

**Proof.** By Condition 1,

$$\begin{aligned}\mathcal{R}\left(p_{W,t}\right) &\leq -c\ln\left(\sum_{e\in\mathcal{E}} w_{e,t}\exp\left\{-\eta\mathcal{R}\left(p_{\hat{\theta}(e,t)}\right)\right\}\right)\\ &= -c\ln\left(\frac{\sum_{e\in\mathcal{E}} v_{e,t}\exp\left\{-\eta\mathcal{R}\left(p_{\hat{\theta}(e,t)}\right)\right\}}{v_{,t}}\right)\\ &= -c\ln\left(\frac{\sum_{e\in\mathcal{E}} v'_{e,t}}{v_{,t}}\right).\end{aligned}$$

■

**Lemma 37** *Under Condition 7,*

$$\sum_{t=1}^{T}\ln\left(\frac{\sum_{e\in\mathcal{E}} v'_{e,t}}{v_{,t}}\right)\geq \ln v_{e,T+1}$$

*for any $e\in\mathcal{E}$. If also Condition 1 holds, this implies*

$$\mathcal{R}_{1,\ldots,t}\left(p_W\right)\leq -c\ln v_{e,t+1}$$

**Proof.** Using Condition 7,

$$\begin{aligned}\sum_{t=1}^{T}\ln\frac{\sum_{e\in\mathcal{E}} v'_{e,t}}{v_{,t}} &\geq \sum_{t=1}^{T}\ln\frac{v_{,t+1}}{v_{,t}}\\ &= \ln\frac{v_{,T+1}}{v_{,1}} = \ln v_{,T+1}\geq \ln v_{e,T+1},\end{aligned}$$

by definition of $v_{,1}$ in the penultimate step and because for non-negative scalars $a$ and $b$, $a+b\geq a\vee b$ in the last step. Using this inequality in Lemma 36, the second inequality follows. ■

**Lemma 38** *Under Condition 7,*

$$\begin{aligned}v_{e,t+1} &\geq \mathsf{u}_{t+1}\left(e,e\right)v_{e,t}\exp\left\{-\eta\mathcal{R}\left(p_{\theta(e,t)}\right)\right\},\\ v_{e,t+1} &\geq v'_{e',t}\mathsf{u}_{t+1}\left(e,e'\right),\end{aligned}$$

*and $\forall t'\leq t$*

$$\begin{aligned}v_{e,t+1} &\geq \left(\prod_{s=t'}^{t}\mathsf{u}_{s+1}\left(e,e\right)\right)\exp\left\{-\eta\mathcal{R}_{t',\ldots,t}\left(p_{\theta(e)}\right)\right\}v_{e,t'},\\ v'_{e,t+1} &\geq \left(\prod_{s=t'}^{t}\mathsf{u}_{s+1}\left(e,e\right)\right)\exp\left\{-\eta\mathcal{R}_{t',\ldots,t+1}\left(p_{\theta(e)}\right)\right\}v_{e,t'}.\end{aligned}$$

**Proof.** By definition of the algorithm,

$$\begin{aligned} v_{e,t+1} &= \sum_{e' \in \mathcal{E}} v'_{e',t} \mathsf{u}_{t+1}(e,e') \\ &\geq v'_{e,t} \mathsf{u}_{t+1}(e,e) = \mathsf{u}_{t+1}(e,e) v_{e,t} \exp\left\{-\eta \mathcal{R}\left(p_{\theta(e,t)}\right)\right\}, \end{aligned}$$

which proves the first inequality of the Lemma. The second inequality of the Lemma follows similarly from the first equality in the above display. Using the first inequality of the Lemma iteratively, gives the third inequality of the Lemma,

$$v_{e,t+1} \geq \left(\prod_{s=t'}^{t} \mathsf{u}_{s+1}(e,e) \exp\left\{-\eta \mathcal{R}\left(p_{\theta(e,s)}\right)\right\}\right) v_{e,t'}$$

and noting that

$$\exp\left\{\eta \mathcal{R}\left(p_{\theta(e,t+1)}\right)\right\} v'_{e,t+1} = v_{e,t+1},$$

the fourth inequality of the Lemma follows. ■

**Proof of Theorem 9.** Lemmata 37 and 38 imply that

$$\begin{aligned} \mathcal{R}_{1,\ldots,T}(p_W) &\leq -c \ln v_{e,T+1} \leq -c \ln \left(\left(\prod_{t=1}^{T} \mathsf{u}_{t+1}(e,e)\right) v_{e,1} \exp\left\{-\eta \mathcal{R}_{1,\ldots,T}\left(p_{\theta(e)}\right)\right\}\right) \\ &\leq c\eta \mathcal{R}_{1,\ldots,T}\left(p_{\theta(e)}\right) - c \ln \left(\prod_{t=1}^{T} \mathsf{u}_{t+1}(e,e)\right) - c \ln v_{e,1} \end{aligned}$$

■

**Proof of Theorem 12.** Consider the following telescoping product

$$v_{e(K),T+1} = v_{e(0),t(0)} \frac{v'_{e(0),t(1)-1}}{v_{e(0),t(0)}} \prod_{k=1}^{K} \left(\frac{v_{e(k),t(k)}}{v'_{e(k-1),t(k)-1}} \frac{v'_{e(k),t(k+1)-1}}{v_{e(k),t(k)}}\right) \frac{v_{e(K),T+1}}{v'_{e(K),t(K+1)-1}}. \quad (7)$$

From Lemma 38,

$$\frac{v_{e,t+1}}{v_{e,t}} \geq \mathsf{u}_{t+1}(e,e) \exp\left\{-\eta \mathcal{R}\left(p_{\theta(e,t)}\right)\right\},$$

$$\frac{v_{e,t+1}}{v'_{e',t}} \geq \mathsf{u}_{t+1}(e,e'),$$

and $\forall t' \leq t$

$$\frac{v'_{e,t+1}}{v_{e,t'}} \geq \left(\prod_{s=t'}^{t} \mathsf{u}_{s+1}(e,e)\right) \exp\left\{-\eta \mathcal{R}_{t',\ldots,t+1}\left(p_{\theta(e)}\right)\right\}.$$

Now by definition,

$$v_{e(0),t(0)} = 1/(\#\mathcal{E}),$$

from Lemma 38,

$$\frac{v_{e(k),t(k)}}{v'_{e(k-1),t(k)-1}} \geq \mathsf{u}_{t(k)}\left(e_k, e_{k-1}\right),$$

$$\frac{v'_{e(k),t(k+1)-1}}{v_{e(k),t(k)}} \geq \left(\prod_{s=t(k)}^{t(k+1)-2} \mathsf{u}_{s+1}\left(e_k, e_k\right)\right) \exp\left\{-\eta \mathcal{R}_{t(k),\ldots,t(k+1)-1}\left(p_{\theta(e(k))}\right)\right\},$$

and since there is no share update on the final trial

$$\frac{v_{e(K),T+1}}{v'_{e(K),t(K+1)-1}} = 1.$$

Substituting everything in (7),

$$\begin{aligned}
& v_{e(K),T+1} \\
\geq\ & (\#\mathcal{E})^{-1} \left(\prod_{s=t(0)}^{t(1)-2} \mathsf{u}_{s+1}\left(e_0, e_0\right)\right) \exp\left\{-\eta \mathcal{R}_{t(0),\ldots,t(1)-1}\left(p_{\theta(e(0))}\right)\right\} \\
& \times \prod_{k=1}^{K} \left[\mathsf{u}_{t(k)}\left(e_k, e_{k-1}\right) \left(\prod_{s=t(k)}^{t(k+1)-2} \mathsf{u}_{s+1}\left(e_k, e_k\right)\right) \exp\left\{-\eta \mathcal{R}_{t(k),\ldots,t(k+1)-1}\left(p_{\theta(e(k))}\right)\right\}\right] \\
=\ & (\#\mathcal{E})^{-1} \left(\prod_{s=t(0)}^{t(1)-2} \mathsf{u}_{s+1}\left(e_0, e_0\right)\right) \prod_{k=1}^{K} \left[\mathsf{u}_{t(k)}\left(e_k, e_{k-1}\right) \left(\prod_{s=t(k)}^{t(k+1)-2} \mathsf{u}_{s+1}\left(e_k, e_k\right)\right)\right] \\
& \times \prod_{k=0}^{K} \exp\left\{-\eta \mathcal{R}_{t(k),\ldots,t(k+1)-1}\left(p_{\theta(e(k))}\right)\right\}.
\end{aligned}$$

Taking natural log,

$$\begin{aligned}
& \ln v_{e(K),T+1} \\
\geq\ & -\ln\left(\#\mathcal{E}\right) + \ln\left(\prod_{s=t(0)}^{t(1)-2} \mathsf{u}_{s+1}\left(e_0, e_0\right)\right) \\
& + \ln \prod_{k=1}^{K} \left[\mathsf{u}_{t(k)}\left(e_k, e_{k-1}\right) \left(\prod_{s=t(k)}^{t(k+1)-2} \mathsf{u}_{s+1}\left(e_k, e_k\right)\right)\right] - \eta \sum_{k=0}^{K} \mathcal{R}_{t(k),\ldots,t(k+1)-1}\left(p_{\theta(e(k))}\right) \\
=\ & -\ln\left(\#\mathcal{E}\right) + \sum_{s=t(0)}^{t(1)-2} \ln\left(\mathsf{u}_{s+1}\left(e_0, e_0\right)\right) \\
& + \sum_{k=1}^{K} \left[\ln \mathsf{u}_{t(k)}\left(e_k, e_{k-1}\right) + \sum_{s=t(k)}^{t(k+1)-2} \ln\left(\mathsf{u}_{s+1}\left(e_k, e_k\right)\right)\right] - \eta \sum_{k=0}^{K} \mathcal{R}_{t(k),\ldots,t(k+1)-1}\left(p_{\theta(e(k))}\right)
\end{aligned}$$

$$= -\ln(\#\mathcal{E}) + \sum_{k=1}^{K} \ln \mathsf{u}_{t(k)}(e_k, e_{k-1}) + \sum_{k=0}^{K} \sum_{s=t(k)}^{t(k+1)-2} \ln(\mathsf{u}_{s+1}(e_k, e_k))$$
$$-\eta \sum_{k=0}^{K} \mathcal{R}_{t(k),\ldots,t(k+1)-1}\left(p_{\theta(e(k))}\right).$$

Using Lemmata 36 and 37 the result follows. ■

## A.2 Theorem 16

The proof is based on the following Lemmata.

**Lemma 39** *Under Condition 14,*

$$\mathcal{R}(p_{WL,t}) \leq -b \ln\left(\frac{\upsilon_{,t+1}}{\upsilon_{,t}}\right).$$

**Proof.** By Condition 14,

$$\begin{aligned}
\mathcal{R}(p_{WL,t}) &\leq -b \ln\left(\sum_{l\in\mathcal{L}} \omega_{l,t} \exp\left\{-\kappa \mathcal{R}\left(p_{W(l)}\right)\right\}\right) \\
&= -b \ln\left(\frac{\sum_{l\in\mathcal{L}} \upsilon_{l,t} \exp\left\{-\kappa \mathcal{R}\left(p_{W(l)}\right)\right\}}{\sum_{l\in\mathcal{L}} \upsilon_{l,t}}\right) \\
&\leq -b \ln\left(\frac{\upsilon_{,t+1}}{\upsilon_{,t}}\right).
\end{aligned}$$

■

**Lemma 40** *If Condition 14 holds*

$$\mathcal{R}_{1,\ldots,T}(p_{WL}) \leq -b \ln(\upsilon_{l,T+1}).$$

**Proof.** Use Lemma 39, sum over $t$, the sum telescopes and $\upsilon_{,1} = 1$. ■

**Lemma 41**

$$\upsilon_{l,t+1} = \upsilon_{l,1} \exp\left\{-\kappa \sum_{s=1}^{t} \mathcal{R}\left(p_{W(l,s)}\right)\right\}$$

**Proof.** By iteration of

$$\upsilon_{l,t+1} = \upsilon_{l,t} \exp\left\{-\kappa \mathcal{R}\left(p_{W(l,t)}\right)\right\}.$$

■

**Lemma 42** *Under Condition 14,*

$$\mathcal{R}_{1,\ldots,T}(p_{WL}) \leq -b \ln(\upsilon_{l,1}) + b\kappa \mathcal{R}_{1,\ldots,T}\left(p_{W(l)}\right).$$

**Proof.** By Lemmata 40 and 41. ■

**Proof of Theorem 16.** Use Lemma 42 and apply Theorem 12. ■

## A.3 Lemmata 20, 24, 30 and 34

**Proof of Lemma 20.** The first equality is immediate. The second follows noting that there are $T$ observations, hence $T-1$ share updates. Since $K$ of them are breaks, $T-K-1$ must be the remaining, and the second equality follows. ■

**Proof of Lemma 24.** By direct calculation, and the fact that $(\beta(e,t))_{e\in\mathcal{E}} \in \mathcal{S}_{\#\mathcal{E}}\ \forall t$,

$$\begin{aligned}\sum_{k=1}^{K} \ln \mathsf{u}_{t(k)}(e_k, e_{k-1}|\lambda) &= \sum_{k=1}^{K} \ln\left(\sum_{e\neq e_{k-1}} \lambda\beta(e, t_k-1)\right) = \sum_{k=1}^{K} \ln\lambda\left(1-\beta(e_{k-1}, t_k-1)\right)\\ &= K\ln\lambda + \sum_{k=1}^{K} \ln\left(1-\beta(e_{k-1}, t_k-1)\right).\end{aligned}$$

Notice that for $(a,b)\in[0,1]^2$,

$$\ln(1-ab) \geq \ln(1-a\wedge b) = \ln(1-a)\vee\ln(1-b),$$

with strict inequality if $(a,b)\in(0,1)^2$. Therefore,

$$\begin{aligned}\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2} \ln \mathsf{u}_{s+1}(e_k, e_k|\lambda) &= \sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2} \ln\left(1-\lambda\left(1-\beta(e(k), s)\right)\right)\\ &\geq (T-K-1)\ln(1-\lambda)\vee\sum_{k=0}^{K}\sum_{s=t(k)}^{t(k+1)-2}\ln\left(\beta(e(k), s)\right).\end{aligned}$$

■

**Proof of Lemma 30.** We need to check that

$$\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right)\right) \leq -\eta^{-1}\ln\sum_{e\in\mathcal{E}} w_e \exp\left\{-\eta\varphi\left(x-\left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\}$$

holds. The segment of observations $x_1, ..., x_T$ and their forecasts take finite values, hence set $\left|x-\left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right| \leq B < \infty$. By the conditions of the Lemma, we can choose $\eta$ such that

$$\exp\left\{-\eta\varphi\left(x-\left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\} \geq \sum_{e\in\mathcal{E}} \exp\left\{-\eta\varphi\left(x-\hat{\theta}_e\right)\right\}$$

for $\left|x-\left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right| \leq B$. Taking natural log and multiplying by $-\eta^{-1}$,

$$\varphi\left(x-\left\langle w_{\mathcal{E}}, \hat{\theta}_{\mathcal{E}}\right\rangle\right) \leq -\eta^{-1}\ln\sum_{e\in\mathcal{E}} \exp\left\{-\eta\varphi\left(x-\hat{\theta}_e\right)\right\}$$

and the result follows. ■

**Proof of Lemma 34.** From Hoeffding bound for the moment generating function of bounded random variables (Hoeffding, 1963, eq. 4.16) and convexity of $\varphi$,

$$\begin{aligned}\sum_{e\in\mathcal{E}} w_e \exp\left\{-\eta\varphi\left(x-\hat{\theta}_e\right)\right\} &\leq \exp\left\{-\eta\sum_{e\in\mathcal{E}} w_e\varphi\left(x-\hat{\theta}_e\right)\right\} + \exp\left\{\eta^2\varphi\left(2B\right)^2/8\right\} \\ &\leq \exp\left\{-\eta\varphi\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right)\right\} + \exp\left\{\eta^2\varphi\left(2B\right)^2/8\right\},\end{aligned}$$

which implies

$$\begin{aligned}&\mathcal{R}\left(p_W\left(x|w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right)\right) \\ =\ &\varphi\left(x-\left\langle w_{\mathcal{E}},\hat{\theta}_{\mathcal{E}}\right\rangle\right) \leq -\eta^{-1}\ln\sum_{e\in\mathcal{E}} w_e\exp\left\{-\eta\varphi\left(x-\hat{\theta}_e\right)\right\} + \eta\varphi\left(2B\right)^2/8.\end{aligned}$$

■

# References


[1] Bousquet, O. (2003) A Note on Parameter Tuning for On-Line Shifting Algorithms. Preprint. Downloadable: http://www.kyb.mpg.de/publications/pdfs/pdf2294.pdf.

[2] Bousquet, O. and M.K. Warmuth (2002) Tracking a Small Set of Experts by Mixing Past Posteriors Journal of Machine Learning Research 3, 363–396.

[3] Cesa-Bianchi, N. Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, M.K. Warmuth (1997) How to Use Expert Advice. Journal of the ACM 44, 427–485.

[4] Cesa-Bianchi, N. and G. Lugosi (1999) On Prediction of Individual Sequences. The Annals of Statistics 27, 1865-1895.

[5] Chamberlain, G. (2000) Econometrics and Decision Theory. Journal of Econometrics 95, 255-283.

[6] Dawid, A.P. (1984) Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. Journal of the Royal Statistical Society Ser. A 147, 278-292.

[7] Dawid, A.P. (1985) Calibration-Based Empirical Probability. The Annals of Statistics 13, 1251-1274.

[8] Dawid, A.P. (1986) Probability Forecasting. In S. Kotz, N.L. Johnson and C.B. Read (eds.), Encyclopedia of Statistical Sciences Vol. 7, 210-218. Wiley.

[9] Genest C. and J.V. Zidek (1986) Combining Probability Distributions: A Critique and an Annotated Bibliography. Statistical Science 1, 114-148.

[10] Györfi, L. and G. Lugosi (2002) Strategies for Sequential Prediction of Stationary Time Series. in M. Dror, P. L'Ecuyer and F. Szidarovszky (eds.), Modeling Uncertainty: An examination of its Theory, Methods, and Applications, 225-249. Kluwer Academic Publishers. Downloadable: http://www.econ.upf.es/~lugosi/autoreg.ps.

[11] Hansen, B.E. (1994) Autoregressive Conditional Density Estimation. International Economic Review 35, 705-730.

[12] Hendry, D.F. and M.P. Clements (2004) Pooling of Forecasts. Econometrics Journal 7, 1-31.

[13] Mark Herbster M. and M.K. Warmuth (1998) Tracking the Best Expert. Machine Learning 32, 151-178.

[14] Polzehl, J. and V. Spokoiny (2004) Varying Coefficient GARCH Versus Local Constant Volatility Modeling. Comparison of the Predictive Power. Preprint No. 977, Weierstrass Institute for Applied Analysis and Stochastics.

[15] Sancetta, A. and A. Nikandrova (2005) Forecasting Using Meta-Elliptical Distributions with a Study of Commodity Prices. Preprint

[16] James H. Stock, J.H. and M.W. Watson (2004) Combination Forecasts of Output growth in a Seven-Country Data Set. Journal of Forecasting 23, 405-430.

[17] Timmermann, A. (2004) Forecast Combinations. Forthcoming in G. Elliott, C.W.J Granger and A. Timmermann (eds.) Handbook of Economic Forecasting. North Holland.

[18] Vovk, V. (1990) Aggregating Strategies. Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT 1990), 371–383.

[19] Yakowitz, S., L. Györfi, J. Kieffer and G. Morvai (1999) Strongly Consistent Nonparametric Forecasting and Regression for Stationary Ergodic Sequences. Journal of Multivariate Analysis 71, 24-41.

[20] Yang, Y. (2004) Combining Forecasting Procedures: Some Theoretical Results. Econometric Theory 20, 176-222.