Econometric Theory, **26**, 2010, 598–631.
doi:10.1017/S0266466609100105

RECURSIVE FORECAST COMBINATION FOR DEPENDENT HETEROGENEOUS DATA

ALESSIO SANCETTA
FIRST, BNP Paribas

This paper studies a procedure to combine individual forecasts that achieve theoretical optimal performance. The results apply to a wide variety of loss functions and only require a tail condition on the data sequences. The theoretical results show that the bounds are also valid in the case of time varying combination weights.

1. INTRODUCTION

This paper considers the problem of online combination of individual forecasts to improve the prediction error in terms of a variety of loss functions. The main contributions of the paper are as follows: (1) to weaken existing conditions used to derive upper bounds for the error incurred by some algorithms used in the machine learning literature and (2) to adapt these existing algorithms to the problem of forecast combination. In particular, point (1) is achieved by (a) allowing both individual forecasts and also the object to be forecasted only to satisfy a tail condition; (b) allowing for a general class of loss functions. Point (2) is achieved by allowing the learning rate used in the algorithm to be time varying and possibly to be automatically calibrated within the algorithm. This reduces the amount of previous knowledge/discretion required in running the algorithm.

The forecast combination problem has attracted much attention in the econometrics literature, and the review article of Timmermann (2006) discusses the most recent advances. Combination of individual forecasts may lead to improvements over single individual forecasts. Several studies have shown that combining forecasts can be a useful hedge against structural breaks, and forecast combination is often more stable than single forecasts (e.g., Hendry and Clements, 2004; Stock and Watson, 2004). Moreover, forecast combination might be used not only to improve on the single forecasts but also as an alternative to model selection when the forecasts are unstable (e.g., Zou and Yang, 2004; in a different but related context, see also Breiman, 1996) and to improve testing between candidate models (e.g., Chen and Yang, 2007).

I am grateful to two anonymous referees and a co-editor for comments that improved the paper in both content and presentation. Address correspondence to Alessio Sancetta, FIRST, BNP Paribas, London, NW1 6AA, UK; e-mail: asancetta@gmail.com.

 © Cambridge University Press, 2009 0266-4666/10 $15.00

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Although optimal forecast combination is often derived by minimization of the user's expected loss over all possible decisions (e.g., Elliott and Timmermann, 2004), the presence of structural breaks might invalidate empirical estimation when we replace unknown expectations with sample expectations. A similar remark applies when the noise level or persistence is quite high relative to the sample size. In this case, it is often suggested to shrink the weights toward the equal combination weights (e.g., Diebold and Pauly, 1990; Aiolfi and Timmermann, 2006).

Clearly, if forecast combination is used to hedge against nonstationarities, there is no reason to expect that the optimal combination weights are time invariant. Hence, this paper studies the problem of optimal forecast combination when the optimal combination weights are possibly time varying. The procedure discussed here does not require any stability of the system besides a tail condition, and it works in the presence of dependent observation. Moreover, to account for possible breaks or instability, a nonzero weight is retained for all individual forecasts, effectively performing some form of shrinkage as discussed in the preceding references. We show that this allows us to track the best time varying combination weights. Although the combination weights might be time varying, there is no need to estimate changes in regime (e.g., Deutsch, Granger, and Teräsvirta, 1994).

Here, the combination weights are derived using one of the recursive algorithms introduced in the machine learning literature (Kivinen and Warmuth, 1997; Cesa-Bianchi, 1999; Herbster and Warmuth, 2001; Cesa-Bianchi and Lugosi, 2006, Ch. 11) and differ from recent applications in Yang (2004) and Sancetta (2007), which are variations of the original aggregating algorithm (AA) of Vovk (1990) (for a detailed list of related algorithms and references, see also Cesa-Bianchi and Lugosi, 2006).

Yang (2004) derives various interesting results. His bounds hold in expectation and do not require bounded random variables, but the conditions used can be restrictive. Most of his bounds are in terms of a multiplicative constant of the optimal loss. The bound that avoids this multiplicative constant and relates more closely to the ones given here (his Theorem 5) requires the existence of the moment generating function of the forecast errors and is restricted to loss functions that are basically quadratic (or powers of a quadratic, but imposing additional restrictive tail conditions). Effectively, the bounds in Yang (2004) compare to the best individual forecast and not to the best forecast combination. However, his Theorem 6 gives a bound that compares to the best forecast combination, but this requires a suitable discretization of the space of combination weights and does not seem computationally efficient.

The present results hold for more general data series (e.g., the moment generating function does not need to exist), and a wide variety of loss functions are allowed. Moreover, the algorithm allows us to directly compare to the best optimal forecast combination, where the optimal combination weights are even allowed to be time varying. All this extra flexibility comes at a price: the bounds derived are much weaker than the ones in Yang (2004, e.g., Thms. 5 and 6). Given that

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

the algorithms also are different, the results given here are complementary and particularly useful when the conditions in Yang (2004) do not apply.

The results of this paper are also related to the results in Sancetta (2007). There, the problem of probability forecasts is considered, and worst case bounds are derived for possibly time varying optimal individual probability forecasts. Clearly, the present framework is different, as interest here lies in forecast combinations of sequences and the goal is to perform as well as the best optimal time varying combined forecast. This more challenging goal does not allow us to derive worst case bounds, but only allows us to derive bounds that hold with high probability to avoid bounded sequences, usually assumed in the machine learning literature.

The plan of the paper is as follows. Section 2 states the algorithm used and carries out a theoretical analysis of its properties. Remarks on the results and generalizations are discussed in Section 3. There we provide more details on how the present results compare to those in Yang (2004). Proofs can be found in the Appendix.

2. RECURSIVE FORECAST COMBINATION

Suppose $(Y_t, X_t)_{t\in\mathbb{N}}$ are a sequence of random variables with values in $\mathcal{Y}\times\mathcal{X}$, where $\mathcal{Y}\subseteq\mathbb{R}$ and $\mathcal{X}\subseteq\mathbb{R}^K$ $(K>1)$. We interpret Y_t to be a quantity to be forecasted and X_t to be individual forecasts of that quantity. We do not discuss the nature of these individual forecasts: they could be exogenous to the econometrician's decision rule. However, both Y_t and X_t need to satisfy the following condition.

Condition 1. There are strictly positive constants c_1, c_2, and c_3 such that

$$\Pr(|Y_t|>x)\le c_1\exp\{-c_2x^{c_3}\},$$
$$\Pr(|X_{tk}|>x)\le c_1\exp\{-c_2x^{c_3}\},\qquad k=1,\dots,K.$$

Although this rules out trends in the data, we shall show that under suitable choice of the loss function used to evaluate the forecasts, we can allow for Y_t and X_t to have common trends as long as $(Y_t - X_{tk})$ is stationary. Details on this extension are discussed in Section 3.2.

We use the forecasts X_t to construct the combined forecast $\hat{P}_t := \langle w_t, X_t\rangle \in \mathcal{P}$ to predict Y_t where $w_t\in\mathcal{S}^K$ and $\mathcal{S}^K$ is the K-dimensional unit simplex ($\langle\dots,\dots\rangle$ is the inner product, and $\mathcal{P}$ is the space of predictions). For $\hat{P}_t$ to be a valid predictor, the weights w_t should only depend on the past and be independent of the present and future; i.e., we may only use $(Y_s, X_s)_{s<t}$ to construct w_t. The quality of the prediction $\hat{P}_t$ is evaluated by a loss function $l:\mathcal{Y}\times\mathcal{P}\to\mathbb{R}$, which needs to satisfy the following condition.

Condition 2. $l(y,p)$ is convex in the second argument and has a derivative with respect to the second argument.

A discussion of this condition can be found in Section 3.1.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

The cumulative loss when we predict with $(\hat{P}_t)_{t\le T}$ is defined by $\hat{L}_T := \sum_{t=1}^{T} l(Y_t, \hat{P}_t)$. We shall write $l_t(w) = l(Y_t, \langle w, X_t\rangle)$ to stress dependence on $w \in \mathcal{S}^K$. We may use whatever information up to time $t-1$ to construct w_t, and we will compare the loss from this approach to the loss incurred by $\inf_{u\in\mathcal{S}^K} \sum_{t=1}^{T} l_t(u)$. More generally, we will compare our loss with the loss incurred by using arbitrary weights $u_1, \ldots, u_T \in \mathcal{S}^K$, i.e., $L_T(u_1, \ldots, u_T) := \sum_{t=1}^{T} l_t(u_t)$. Because $u_1, \ldots, u_T$ are arbitrary, we may choose them such that $u_t := \arg\inf_{v\in\mathcal{S}^K} l_t(v)$. Clearly, in an arbitrary framework, we cannot expect to find any algorithm that produces results nearly as good as this choice of combination weights, but we still can do fairly well if the $u_1, \ldots, u_T$ are not allowed to change too often. Throughout, $u_1, \ldots, u_T$ will denote arbitrary but fixed combination weights from time 1 to T. Note that the time frame, i.e., T, does not need to be known in advance and the analysis is carried out uniformly in T. After all, when we carry out online estimation, we may not know in advance for how long we will use the same individual forecasts.

2.1. Recursive Choice of Regression Coefficients

The choice of combination weights is given by Algorithm 1, which follows. There, the gradient of the loss function with respect to w is denoted by $\nabla l_t(w)$ and $\nabla_k l_t(w)$ is its kth element. Generalizations to continuous but nondifferentiable loss functions are discussed in Section 3. The algorithm depends on the so-called learning rate $\eta_t := \eta t^{-\alpha}$ that is a function of a coefficient $\eta > 0$ and time with exponent $\alpha \in (0, 1/2]$. In both cases, the algorithm uses a second update that shrinks ("projects") the parameters onto a constrained set where the parameters $u_1, \ldots, u_T$ are supposed to lie. This "projection" can be interpreted as shrinkage and allows us to control the loss in case of changes in the underlying parameters. As discussed in the Introduction, shrinkage is a commonly used technique in economic forecasting and statistics.

Algorithm 1

Set
$w_{11} = \cdots = w_{1K} = 1/K$;
$\eta > 0$;
$\alpha \in (0, 1/2]$;
$\gamma \in [0, 1]$;
$\hat{L}_0 := 0$.
For $t = 1, \ldots, T$,
$\hat{L}_t = \hat{L}_{t-1} + l_t(w_t)$;
$\eta_t = \eta t^{-\alpha}$;
$w'_{t,k} := w_{tk} \exp\{-\eta_t \nabla_k l_t(w_t)\} / \sum_{k=1}^{K} w_{tk} \exp\{-\eta_t \nabla_k l_t(w_t)\}$;
$\psi := \sum_{k=1}^{K} I_{\{w'_{t,k} < \gamma/K\}}$;
$\omega := \sum_{k=1}^{K} I_{\{w'_{t,k} < \gamma/K\}} w'_{tk}$;
$w_{t+1,k} = w'_{tk}[1 - (\psi\gamma)/K]/[1-\omega]$
if $w'_{tk} \ge \gamma/K$, $w_{t+1,k} = \gamma/K$ otherwise.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

The algorithm uses two updates that will be motivated in the Section 2.2. The first update produces the ex post combination weight $w'_{t,k}$. This is called ex post because the update depends on the gradient of the observed loss at time t. This ex post weight is then projected on a subset of $\mathcal{S}^K$ to make sure that no weight is less than the user's prespecified level γ/K. In particular, all weights less than γ/K are set equal to the lower bound level γ/K. To impose the constraint that the weights are in $\mathcal{S}^K$ the weights that are greater than γ/K are then shrunk toward γ/K by an amount $[1-(\psi\gamma)/K]/[1-\omega]$ to make sure that the constraint is satisfied. Note that setting $\gamma = 1$ produces the equally weighted forecast combination.

The exponent α is related to the frequency of change in $u_1, \ldots, u_T$. For any K-dimensional vector v, define $|v|_p = \left(\sum_{k=1}^K |v_k|^p\right)^{1/p}$ (with obvious modification when $p = \infty$). When l is the square loss and $\sum_{t=1}^{T-1} |u_t - u_{t+1}|_1 = O(T^{1-\epsilon})$ $\epsilon \in (0, 1]$, α should satisfy $\alpha = \epsilon/2$ (see Corollary 3 in Section 2.3) for the error in the procedure to be asymptotically negligible. (For simplicity we only consider polynomial growth.) Subsequently we provide motivation for these updates.

2.2. Motivation via Penalized Minimization

The algorithm can be motivated as an approximation to penalized risk minimization. The approximation argument was first suggested in Kivinen and Warmuth (1997). Define $D(u, w) := \sum_{k=1}^K u_k \ln(u_k/w_k)$, which is the relative entropy and will be used as a penalty on changes of combination weights. The ex post weight w'_t is such that, under regularity conditions for w close to w_t,

$$\arg \inf_{w \in \mathcal{S}^K} [D(w, w_t) + \eta_t l(Y_t, \langle w, X_t \rangle)] \tag{1}$$

$$\simeq \arg \inf_{w \in \mathcal{S}^K} \{D(w, w_t) + \eta_t [l(Y_t, \langle w_t, X_t \rangle) + l'(Y_t, \langle w_t, X_t \rangle)(w - w_t)]\} \tag{2}$$

$$=: w'_t,$$

where $l'(y, p) := (\partial/\partial p) l(y, p)$ throughout. From (1) it is clear that the objective is to minimize the loss function subject to the penalty D, where η_t can be interpreted as the inverse of the smoothing parameter. We could use some distance/penalty function other than D, and this would produce different updates. In the computation of w'_t, the loss function l is replaced by its first-order Taylor series approximation to allow for simple closed form solution. The optimization problem has two terms that act in opposite directions: the penalty D requires the ex post weight w'_t to be close to the ex ante combination weight w_t; the loss l requires w'_t to produce the best ex post forecast, and with no penalty this would lead to overfitting. The coefficient $\eta_t > 0$ controls these two opposing aspects. When η_t is large, the update changes the weight toward the weight that would have been optimal. A small η_t implies smoother ex post updates (i.e., more dependence on past observations), hence a higher degree of smoothness. As previously mentioned, η_t will be referred to as the learning rate and can be thought

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

of as the reciprocal of a smoothness parameter in a penalized risk minimization problem. The combination weights at time t are obtained by a further update. We use a function $Q_{(\mathcal{G},D)} : int(\mathcal{S}^K) \to \mathcal{S}^K$ that depends on some closed convex subset $\mathcal{G} \subset \mathcal{S}^K$ and on the "distance" D ($int(\mathcal{S}^K)$ denotes the interior of $\mathcal{S}^K$). In particular, $Q_{(\mathcal{G},D)}(w_t') = \arg\min_{u \in \mathcal{G}} D(u, w_t')$, and $w_t = Q_{(\mathcal{G},D)}(w_t')$ is the projection of w_t' onto

$$\mathcal{G} := \left\{ u \in \mathcal{S}^K \cap [\gamma / K, 1]^K \right\} \tag{3}$$

in terms of D. To constrain the parameters in $\mathcal{S}^K$, we add the constraint $\sum_{k=1}^K w_k = 1$, and the update to w_t' is obtained by simple optimization of the Lagrangian obtained from (2). The second update to w_t is obtained from $Q_{(\mathcal{G},D)}$. It is easy to see that the solution lies in $\mathcal{G}$. A proof that $Q_{(\mathcal{G},D)}(w_t') = w_{t+1}$ as in Algorithm 1 can be found in Herbster and Warmuth (2001).

For the case of competing linear models, when the estimation is not recursive, Leung and Barron (2006) derive exponential combination weights that bear some similarity to the recursive ones discussed here. Because of the nonsequential nature of their estimator they require minimization of a loss plus a penalty related to model complexity. This is not to be confused with the penalty on the weights: $D(w, w_t)$ in (1). A penalty related to model complexity is used to avoid overfitting because no validation sample is used. In recursive estimation, a new observation is revealed each time, and it is used to evaluate the model. The properties of recursive model evaluation have been extensively studied in prequential statistics (e.g., Dawid, 1984, 1997), which is at the hearth of the real time econometric approach discussed in Pesaran and Timmermann (2005). Here we are only interested in recursive estimation; hence we shall not further remark on estimation with a penalty related to model complexity.

2.3. Theoretical Performance of the Algorithms

A theoretical justification for Algorithm 1 is given. In particular, it is shown that reasonable bounds hold uniformly in the time horizon. Moreover, the bounds avoid the usual condition of bounded random variables in favor of tail conditions. Because of this, the bound is weaker than the bounds usually derived in the machine learning literature. Remarks on the conditions and extensions are discussed in Section 3. The bounds to be shown are of the kind

$$\sum_{t=1}^{T} [l_t(w_t) - l_t(u_t)] \leq error$$

with some high probability, where *error* refers to a bound for Algorithm 1 for any arbitrary $u_1, \ldots, u_T \in \mathcal{G}$. Hence, note that the weights we compare to are also constrained to lie in (3). Besides this, they are arbitrary, and we could choose them at time T, with hindsight. This will not be mentioned again. We will also show that these bounds hold in expectation.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Note that bounds that hold with high probability are of particular interest, because they show how the algorithm will perform not just on average but for any data sequence at a given level of confidence.

THEOREM 1. *Define* $m_T(\tau) := \left[\frac{1}{c_2}\tau + \frac{1}{c_2}\ln(c_1(K+1)T)\right]^{1/c_3}$, *and for any* $m > 0$

$$\bar{l}'(m) := \sup_{\substack{y\in[-m,m]\cap\mathcal{Y}\\ p\in[-m,m]\cap\mathcal{P}}} |l'(y,p)|.$$

Under Conditions 1 and 2, with probability at least $1-e^{-\tau}$,

$$\sum_{t=1}^{T}[l_t(w_t)-l_t(u_t)] \le error,$$

where

$$error \le \frac{T^\alpha}{\eta}\left(6\ln(K/\gamma)+[\ln(K/\gamma)+K]\sum_{t=1}^{T-1}|u_t-u_{t+1}|_1\right) + \frac{\eta T^{1-\alpha}}{8(1-\alpha)}\left[\bar{l}'(m_T(\tau))m_T(\tau)\right]^2.$$

The bound shows how the choice of $u_1,\ldots,u_T$ affects the relative performance. Note that we did not impose any dependence conditions. For loss functions that grow polynomially (e.g., the square loss), the preceding bounds only grow polynomially in τ being violated with exponentially small probability $e^{-\tau}$. The constant 6 in the first term is larger than necessary but leads to a tidier bound. Note that the bound depends on γ, i.e., the set $\mathcal{G}$ in (3). Setting γ small would increase the set of allowed weights but also increase the error in the bound.

The bounds of Theorem 1 require some further comment to be fully appreciated and compared with existing results, in particular with the results in Yang (2004). To this end, we use a series of corollaries to show its applications. The first important case is when we compare to the best time invariant combination weights.

COROLLARY 1. *Suppose* $u_t = u_{t+1}\ \forall t$. *Using the conditions and notation of Theorem 1, with probability at least* $1-e^{-\tau}$,

$$\sum_{t=1}^{T} l_t(w_t) - \inf_{u\in\mathcal{G}}\sum_{t=1}^{T} l_t(u) \le error,$$

where

$$error \le T^{1/2}\left[6\ln(K/\gamma)+\frac{1}{4}[\bar{l}'(m_T(\tau))m_T(\tau)]^2\right],$$

when we use the learning rate $\eta_t = t^{-1/2}$.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

For time invariant combination weights, the error only grows logarithmically in the number of combination weights K. Moreover, $error/T \to 0$ if $[\bar{l}'(m_T(\tau))m_T(\tau)]^2 = o(T^{1/2})$. This depends on the loss function and on the tails of the random variables. To avoid convoluted technical conditions, the following discussion only gives two special simple examples. To ease notation, dependence on K (finite and fixed) will be suppressed.

COROLLARY 2. *In Corollary 1 we have*

(i) if $l(y, p) = |y - p|^2$,

$$\frac{error}{T} = O\left(T^{-1/2}[\tau + \ln T]^{4/c_3}\right);$$

(ii) if $l(y, p) = \exp\{\theta(y - p)\} - \theta(y - p) - 1$, *and* $c_3 > 1$,

$$\frac{error}{T} = o\left(e^{b\tau^{1/c_3}}\right)$$

for some finite constant b depending on θ, c_2, *and* c_3.

The algorithms discussed here allow hedging against changes in the combination weights. For simplicity, we only consider the square loss case.

COROLLARY 3. *Suppose* $\sum_{t=1}^{T-1} |u_t - u_{t+1}|_1 = O(T^{1-\epsilon})$ $\epsilon \in (0, 1]$ *and* $l(y, p) = |y - p|^2$. *Then, under Condition 1, with probability at least* $1 - e^{-\tau}$,

$$error = O\left(T^{1-(\epsilon/2)}[\tau + \ln T]^{4/c_3}\right)$$

when we choose $\eta_t = O(t^{-(\epsilon/2)})$.

The result shows that we can allow for changes in the reference combination weights $u_1, \ldots, u_T$ as frequent as $O(T^{1-\epsilon})$ with $\epsilon > 0$ (note that $(\ln T)^{4/c_3} = o(T^{\epsilon/2})$) and still obtain a reasonable performance, though only asymptotically. In practice, if we believe that changes should be somehow frequent (e.g., $\epsilon \to 0$), then we should choose α in Algorithm 1 to be smaller than $1/2$.

By a moment condition, the previous results directly allow us to derive bounds that hold in expectation.

COROLLARY 4. *Suppose Conditions 1 and 2 hold and that* $\mathrm{E}|l(Y_t, \langle w_t, X_t \rangle)|^2 \le A^2$. *Then, all the previous bounds hold taking expectation on the left-hand side, adding* $2AT^{1/2}$, *and setting* $\tau = 0$ *in* $m_T(\tau)$, *i.e.,*

$$\sum_{t=1}^{T} \mathrm{E}[l_t(w_t) - l_t(u_t)] \le error + 2AT^{1/2}.$$

Square integrability of the loss function is not required, but this allows us to directly relate our results to the previous probabilistic bounds.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

The results in this section can be summarized as follows: (1) recursive forecast combination may allow us to perform almost as well as if we knew the combination weights for the whole sample; (2) *error* only grows logarithmically in the number K of individual forecasts under suitable tail conditions when we compare to the best time invariant combination weights.

3. DISCUSSION

We discuss the conditions used and suggest extensions. In particular, some comparison will be drawn with the results in Yang (2004) because of some similarities. Theorem 1 finds a natural application to the problem of portfolio selection, and we shall give some new results for this application. We then remark on the advantages and disadvantages of the gradient algorithm studied here. An illustrative simulation example is used for this purpose.

3.1. Remarks on Condition 2

Condition 2 is satisfied by any loss function $l(y, p) = l(y - p)$ that is convex, with continuous derivatives. Common examples are the square loss, LinEx, whereas some important loss functions such as the absolute loss and the ϵ insensitive loss (Vapnik, 1998) are left out. However, the proofs only require that

$$l_t(u) \geq l_t(w) + \langle \nabla l_t(w), u - w \rangle \tag{4}$$

holds for some function $\nabla l_t(w)$, not necessarily the gradient (see expression (A.4) in the proof of Lemma A.2 in the Appendix). Hence, it is sufficient that for any $w \in \mathcal{S}^K$, we are able to find a line going through $l_t(w)$ and lying below $l_t(u)$ for any $u \in \mathcal{S}^K$, implying that l only needs to be continuous and convex in p. Hence, we are free to interpret the elements of $\nabla l_t(w)$ as either right or left derivatives. This way we can allow for more robust loss functions.

Example 1

For the absolute loss function $l(y, p) = |y - p|$,

$$l(Y_t, \langle X_t, u \rangle) \geq l(Y_t, \langle X_t, w \rangle) + sign(Y_t - \langle X_t, w \rangle)\langle X_t, u - w \rangle.$$

3.2. Remarks on Condition 1

In Condition 1 the exponent c_3 can be arbitrarily small so that the moment generating function does not need to exist. Moreover, the condition is in terms of the target sequence $(Y_t)_{t \in \mathbb{N}}$ and the individual forecasts $(X_t)_{t \in \mathbb{N}}$ and not in terms of the forecast error, whose distribution might be more difficult to derive. For loss functions with polynomial growth, the results are also valid for polynomial tails but with bounds worse than the ones shown. For any random variable Z define $\|Z\|_r := (\mathrm{E}|Z|^r)^{1/r}$ (the L_r norm, $r > 0$). Then, using the square loss, it can be shown that $error/T$ in Corollary 1 is $o(T^{-(8-\gamma/2\gamma)})$ (hence $\gamma \geq 8$) as soon as

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

$\|X_{tk}\|_{\gamma+\epsilon}$ and $\|Y_t\|_{\gamma+\epsilon}$ are finite for $\epsilon > 0$, requiring the existence of an $8+\epsilon$ moment. For the sake of simplicity the polynomial case is not considered. These details are left to the interested reader, who will be able to discover that the dependence on K in Theorem 1 also deteriorates from logarithmic to polynomial growth (with exponent less than one).

Condition 8 in Yang (2004) requires $\mathrm{E}|R_{tk}|^{2\beta}\exp\{t|R_{tk}|^{\beta}\} < \infty$, where R_{tk} is the forecast error at time t for the kth forecast and $\beta > 0$ is as determined in his Condition 7. His Condition 7 requires $\beta \geq 1$; hence the forecast errors need to have a distribution with tails that decay at least as fast as an exponential.

Moreover, suppose m_t is the conditional mean of Y_t. Yang basically assumes that $|m_t - X_{tk}|$ is uniformly bounded in t with probability one for any k (his Condition 6, where uniformity is also in k; see also his Conditions 0 and 2). This is restrictive in some cases. Suppose Y_t is a real valued AR(1) process with conditional mean $m_t = aY_{t-1}$ and the kth individual forecast is $X_{tk} = (a+\epsilon_k)Y_{t-1}$ for some real ϵ_k. Then, $\sup_t |m_t - X_{tk}| = \sup_t |\epsilon_k Y_{t-1}|$, and unless $\epsilon_k = 0$ for any k, Yang's Condition 6 fails if Y_t does not take values in a compact subset of the reals.

There seems to be an advantage in working directly with the forecast error, as this allows us to consider trending data when the loss is a function of the forecast error. Suppose that the individual forecasts $X_t := (X_{t1}, \ldots, X_{tK})$ are such that the forecast error $R_{tk} := (Y_t - X_{tk})$ satisfies a tail bound as in Condition 1 but Y_t and X_{tk} do not. This occurs when there is a trend either deterministic or stochastic, say, μ_t, that is not bounded in probability. Then, for the forecast error to be bounded in probability, this means that Y_t and X_{tk} share the common component μ_t, so that when the loss is a function of R_{tk} only, we only need Condition 1 to hold for $(Y_t - \mu_t)$ and $(X_{tk} - \mu_t)$ rather than Y_t and X_{tk} directly. This complicates the notation but has no impact on the results.

Example 2

Suppose $l(Y_t, \langle X_t, w\rangle) = |Y_t - \langle X_t, w\rangle|^2$. This loss can be written as

$$|(Y_t - \mu_t) - \langle (X_t - \mu_t), w\rangle|^2,$$

where the subtraction in $X_t - \mu_t$ is meant elementwise.

Whether μ_t is equal to or different than zero, the algorithm requires no change because of cancellations (see Algorithm 1).

3.3. Remarks on the Results

Depending on the loss function, it is clear that the stated bounds do not allow $error/T$ to be smaller than $T^{-1/2}$ in order of magnitude. This is not optimal (e.g., Yang, 2004, Thms. 6 and 7); however, it is not clear how to find an optimal algorithm that achieves faster convergence in the general framework considered here.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

For the sake of comparison, note that Corollary 4 gives a convergence rate $T^{-1/2}\ln^2 T$ for Gaussian random variables. Theorem 6 in Yang (2004) leads to the following bound:

$$\mathrm{E}\sum_{t=1}^{T} l_t(w_t) \leq C\left(\inf_{u\in\mathcal{S}^K} \mathrm{E}\sum_{t=1}^{T} l_t(u) + \frac{\ln T}{T}\right),$$

where $C > 2$ under the square loss. Recall that the procedure that leads to this bound requires a finite approximation for the unit simplex and it can be computationally challenging. Actually, unlike here, his result does not require u to be restricted in $\mathcal{S}^K$. The proof given in Yang (2004) uses the fact that the forecast errors are Gaussian (i.e., it uses his Corollary 1), though this is not mentioned in the statement of Theorem 6. It seems possible to avoid the annoying multiplicative factor $C > 2$ (apply Yang's Theorem 5 rather than his Corollary 1 in the proof of Theorem 7 in Yang, 2004). Hence the result in Yang (2004) is stronger, but it requires (1) a square loss, (2) stringent tail conditions (see Section 3.2), and (3) a suitable discretization of the space of combination weights, which add to the computational burden, increasing considerably the number of combination weights. Moreover, the procedure does not account for time varying optimal combination weights. Hence, the present results can complement Yang's results in the following situations: (1) the loss function is not quadratic with possibly discontinuous derivatives (see Section 3.1); (2) the random variables do not possess a moment generating function, and the difference between the individual forecasts and the conditional mean of Y_t is not uniformly bounded (see Section 3.2); (3) we wish to perform as well as the best optimal combined forecast using a simple recursive algorithm; and (4) the optimal combination weights are time varying. These conditions are not necessarily crucial, and when possible one should rather use the algorithm given in Yang (2004) because of to its superior performance under the described conditions.

3.3.1. Tightness of the Bounds. The discussion about the bounds in Yang (2004) may induce the reader to think that the bounds considered here are too loose. Results in the machine learning literature suggest that this is not the case when we consider general loss functions and not necessarily the square loss—recall that no dynamic restrictions are imposed on the data. Let $\hat{P}(X_t)$ be any forecasting procedure based on X_t and $(Y_s, X_s)_{s<t}$ only. Define the minimax regret

$$V_T^{(K)} := \sup_{X_1}\inf_{\hat{P}(X_1)}\sup_{Y_1}\cdots\sup_{X_T}\inf_{\hat{P}(X_T)}\sup_{Y_T}\sum_{t=1}^{T}\left(l\left(Y_t, \hat{P}(X_t)\right) - \min_{k\in\{1,\ldots,K\}} l(Y_t, \langle u, X_{tk}\rangle)\right).$$

This is the worst case *error* we can expect in the bounds given in this paper when $\hat{P}(X_t) = \langle w_t, X_t\rangle$ and w_t is chosen according to Algorithm 1. For definiteness

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

consider the absolute loss $l(y, p) = |y - p|$ and suppose that $Y_t \in \{1, 0\}$ and $X_t \in [1, 0]^K$. By Theorem 3.7 in Cesa-Bianchi and Lugosi (2006) we have

$$\sup_{T>0} \frac{V_T^{(K)}}{\sqrt{(T/2)\ln K}} \geq 1.$$

This is a lower bound on the minimax regret showing that we cannot expect *error* to be smaller than $O(\sqrt{T \ln K})$ under the absolute loss. This is a very special example, and more general results of the same flavor can also be stated (e.g., Cesa-Bianchi and Lugosi, 2006, Thm. 3.6). When we compare to $u_T = \cdots = u_1 \in \mathcal{S}^K$, under the preceding conditions on Y and X, Theorem 1 gives

$$\sum_{t=1}^{T} \left[l_t(w_t) - l_t(u_T) \right] \leq T^{1/2} \left(\frac{6 \ln(K)}{\eta} + \frac{\eta}{4} \right)$$

choosing $\alpha = 1/2$, $\gamma = 1$ because $u_1 \in \mathcal{S}^K$ and noting that $\bar{l}'(m_T(\tau))m_T(\tau) \leq 1$ under the conditions on X and l (see the first two equalities in the proof of Theorem 1 for more details). Here, we are not comparing to the best individual forecast but to the best convex combination that will still achieve the lower bound in order of magnitude, setting $\eta = \sqrt{\ln K}$ in the previous display.

3.3.2. Changes in the Reference Combination Weights. Theorem 1 uses $\sum_{t=1}^{T-1} |u_t - u_{t+1}|_1$ to quantify the effect of changes in the combination weights. Corollary 3 restricts it to be $O(T^{1-\epsilon})$ for some $\epsilon > 0$. This implies some sort of negligibility in the number of changes relative to the sample size T. In the context of structural breaks, this is conceptually similar to Assumption A5 in Bai and Perron (1998). There, it is required that the durations within each break are increasing with the sample size, i.e., $T_i = \lfloor \lambda_i T \rfloor$, where T_i is the time of the ith break, $\lambda_i > 0$ ($\lfloor \ldots \rfloor$ is the integer part). This is clearly required for inference. However, for forecasting, it is enough to use the more general approach in terms of $\sum_{t=1}^{T-1} |u_t - u_{t+1}|_1$ with the aforementioned constraint. This constraint is clearly implied by Assumption 5A in Bai and Perron (1998); the other way around is not true. Note that negligible does not imply $|u_{t+1} - u_t|_1 \to 0$. This possibility is the object of study of asymptotic local power in the literature on structural breaks. In this case, note that changes such that $|u_{t+1} - u_t|_1 = o(T^{-1/2})$ usually cannot be detected (e.g., Andrews, 1993, p. 826 and Cor. 2). On the other hand a wide range of possibilities is allowed here, including $|u_{t+1} - u_t|_1$ proportional to some nonzero constant for some finite number of t values, as implied by Assumption A5 in Bai and Perron (1998).

We further discuss the meaning attached to the combination weights $u_1, \ldots, u_T$ and its implications. As mentioned in Section 2, $u_1, \ldots, u_T$ are arbitrary but fixed reference combination weights. Theorem 1 establishes a relation between the weights derived by the algorithm and the reference combination weights. The reference combination weights are not necessarily the "optimal" combination

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

weights. They represent the reference target against which we wish to compete, i.e., our goal. Hence, once we set our goal—e.g., perform as well as the best ex post time invariant combination weights—Theorem 1 can be used to find the error incurred by the algorithm in achieving this goal. A very simple example will clarify the point of this remark.

Suppose that $Y_t = A_t U_{t-1} + Z_t$ where $(U_t)_{t\in\mathbb{N}}$ and $(Z_t)_{t\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) mean zero with finite moments, independent of each other and independent of the Markov chain $(A_t)_{t\in\mathbb{N}}$ with values in $\{0, 0.5, 1\}$ and transition matrix

$$\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}. \tag{5}$$

Hence, the (i, j)th entry of (5) gives the probability of A_{t+1} being equal to the jth entry in $\{0, 0.5, 1\}$ given that A_t is equal to the ith entry in $\{0, 0.5, 1\}$. It is simple to see that the stationary invariant distribution gives equal probability to each individual entry in $\{0, 0.5, 1\}$. Suppose that at time t we use $K = 3$ individual forecasts: $a_k U_{t-1}$ where $a_1 = 0$, $a_2 = 0.5$, and $a_3 = 1$. Hence, every time one of these forecasts is the best mean square error predictor of Y_t, conditioning on the regressor U_{t-1} and the stochastic regression coefficient A_t. Clearly, the best individual forecast changes every time period, and tracking these changes implies $u_t \neq u_{t+1}$ for each t. Recalling that the reference weights are constrained to be in $\mathcal{G}$ (with $\mathcal{G}$ as in (3)), we shall have $|u_t - u_{t+1}|_1 = 2(1-\gamma)$ for each t. This is because ideally we would like to give weight one to the best individual forecast and zero to the remaining two; however the constraint $\mathcal{G}$ only allows us to give weight $1 - 2\gamma/3$ to the best and $\gamma/3$ to each of the remaining two. Hence in Theorem 1, $\sum_{t=1}^{T-1} |u_t - u_{t+1}|_1 = 2T(1-\gamma)$ because the best individual forecast changes every time. Therefore, the goal is not achievable with negligible asymptotic error. However, we may consider the best forecast combination—in a mean square error sense—conditioning on U_{t-1} only. Accordingly, the true conditional model can be written as

$$Y_t = U_{t-1}(\mathrm{E}A_t) + Z'_t,$$

where, conditioning on U_{t-1}, $Z'_t = Z_t + U_{t-1}[(1 - \mathrm{E})A_t]$ is mean zero. From the invariant distribution implied by (5), we have $\mathrm{E}A_t = 1/2$. From our choice of individual forecasts, it is clear that $u_t = (1/3, 1/3, 1/3)$ is the mean square error optimal combination weight (conditioning on U_{t-1} only). Hence although we cannot track very frequent changes in the optimal individual forecasts, it is sensible to look at the best time invariant optimal combination weights as our reference class. The results of the paper also tell us the error incurred in trying to achieve this more modest goal. In this context—and related ones—using time invariant combination weights is equivalent to integrating out the uncertainty due to

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

a mechanism that cannot be learned. Similar remarks are often made in the literature on structural breaks. In the case of a linear model, changes in the regression coefficient due to the omission of some stationary variables can be interpreted as leading to many structural breaks. However, the true relationship, conditioning only on the nonomitted variables (i.e., integrating out the omitted variables) leads to a constant regression coefficient that can then be consistently estimated by ordinary least squares (e.g., Andrews, 1993, p. 826). Hence, in some situations, it may be appropriate to consider changes in the reference combination weights that only have a persistent effect on the values of the combination weights, e.g., in the preceding example if $(\mathrm{E}A_t)$ were slowly changing in time. The practical implication for the algorithm is that we should avoid choosing α close to zero only because we envisage that there are many changes in the "optimal" combination weights. Indeed, a very small value of α considerably inflates the last term in *error* in Theorem 1. Given that α controls the learning rate, the crucial related aspect is the selection of the learning rate. This issue is further discussed in Section 3.6. We now turn to a related application of Theorem 1.

3.4. A Bound in Expectation when Condition 1 Fails

The main interest in this paper is to derive bounds that hold with high probability and try to reduce the effect of the number K of individual forecasts on the bounds. This is achieved using Condition 1. However, this condition may fail. We give a simple alternative that is applicable in this situation.

THEOREM 2. *Suppose that for any* $k = 1, \ldots, K$,

$$\sup_{t\in\mathbb{N}} \max_{k\le K} \mathrm{E} \sup_{w\in\mathcal{G}} |l'(Y_t, \langle w, X_t\rangle) X_{tk}|^{2p} < \infty \tag{6}$$

with $\mathcal{G}$ *as in (3). Then, under Condition 2,*

$$\sum_{t=1}^{T} \mathrm{E}[l_t(w_t) - l_t(u)] \le \mathit{error},$$

where

$$\mathit{error} \le 6\frac{T^\alpha}{\eta}[1+\ln(K/\gamma)] + [\ln(K/\gamma)+K]\frac{T^\alpha}{\eta}\sum_{t=1}^{T}|u_t - u_{t+1}|_1$$

$$+ \eta T^{1-\alpha}\frac{K^{1/p}}{8(1-\alpha)} \sup_{t\in\mathbb{N}} \max_{k\le K} \left(\mathrm{E}\sup_{w\in\mathcal{G}} |l'(Y_t, \langle w, X_t\rangle) X_{tk}|^{2p}\right)^{1/p}.$$

This bound is similar to the previous ones; the only difference is that the last term grows as a power of K. Note that (6) can be quite a mild condition in some circumstances as shown in the following result.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

COROLLARY 5. *Suppose* $l(y,p)=|y-p|$. *If* $|X_{tk}|^2$ *is uniformly integrable in* k *and* t, *then, in Theorem 2,*

$$error \leq 6\frac{T^{\alpha}}{\eta}[1+\ln(K/\gamma)]+[\ln(K/\gamma)+K]\frac{T^{\alpha}}{\eta}\sum_{t=1}^{T}|u_t-u_{t+1}|_1 + \eta T^{1-\alpha}\frac{K}{8(1-\alpha)}\sup_{t\in\mathbb{N}}\max_{k\leq K}\mathrm{E}|X_{tk}|^2.$$

As discussed in Example 2, in the presence of a common trend μ_t, we can clearly replace X_{tk} with $(X_{tk}-\mu_t)$ both in the conditions and in the bound.

3.5. Related Application: Universal Portfolios

A nice application of the forecast combination problem is the case of unsupervised learning, i.e., when the target is unobservable. The most notable example is the portfolio choice problem, where there is no target sequence and the goal is to maximize wealth. Algorithms that allow us to construct portfolios that have performance comparable to the best constantly rebalanced portfolio with hindsight have attracted the attention of the mathematical finance literature (e.g., Cover, 1991; Helmbold, Schapire, Singer, and Warmuth, 1998; Cross and Barron, 2003; Györfi, Lugosi, and Udina, 2006). The results of this paper allow us to derive theoretical bounds for this problem. For the sake of concreteness, let $X_t:=(X_1,\ldots,X_K)$ be a vector of relative returns (e.g., closing price divided by opening price at time t) and let $-\ln(p):\mathcal{P}\to\mathbb{R}$ be the loss function. Minimization of the cumulated negative log loss is equivalent to maximization of relative wealth (Breiman, 1961; Cover, 1991; for a critique of this criterion, see Samuelson, 1979). Clearly, our framework works just as well with other loss functions (i.e., any change sign of a utility function), but to better relate to the existing literature in this area, $\ln(p)$ will be used, and the problem is turned into a maximization problem. To this end, we state the following result.

THEOREM 3. *For any sequence of relative returns* $(X_t)_{t\in\mathbb{N}}$, *and* $u_1,\ldots,u_T\in\mathcal{G}$ ($\mathcal{G}$ *as in (3)) using Algorithm 1,*

$$\sum_{t=1}^{T}\ln(\langle w_t,X_t\rangle)\geq\sum_{t=1}^{T}\ln(\langle u_t,X_t\rangle)-error,$$

where

$$error\leq\frac{T^{\alpha}}{\eta}\left(6\ln(K/\gamma)+[\ln(K/\gamma)+K]\sum_{t=1}^{T-1}|u_t-u_{t+1}|_1\right)+\frac{\eta T^{1-\alpha}}{8(1-\alpha)}\left(\frac{K}{\gamma}\right)^2$$

and for the constantly rebalanced portfolio, i.e., $u_t=u_{t+1}$, $\forall t$,

$$error<2\frac{K}{\gamma}(T\ln(K/\gamma))^{1/2},$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

choosing $\eta = \gamma / K(24\ln(K/\gamma))^{1/2}$, $\alpha = 1/2$. *Hence, the portfolio constructed from Algorithm 1 is universal for any portfolio weights in* $\mathcal{G}$.

Note that this result is quite strong: no probabilistic assumption is necessary. This result also allows for nonstationary portfolios (i.e., time varying rebalanced portfolios) and extends results in the literature on universal portfolios improving the bounds for the best constantly rebalanced case (e.g., Helmbold et al., 1998, give a bound $O(T^{3/4})$ as compared to the $O(T^{1/2})$ derived here, though Cover, 1991, gives a bound $O(\ln T)$, using a different algorithm, which is difficult to implement in practice). However, we require the portfolio weights to be in $\mathcal{G}$, and this can be restrictive. The constants in the bound of Theorem 3 are not optimal, but they allow for a tidier bound. For to this reason, the derived η might not be the best choice. Because of the very low signal to noise ratio in relative returns, the choice of learning rate seems to be crucial. Further study seems to be required to make these algorithms usable in real trading. Empirical results reported in the literature are very encouraging (e.g., Cover, 1991; Helmbold et al., 1998; Györfi et al., 2006) but also very dependent on the data set used (e.g., Nikandrova, 2005). Empirical results carried out by the author, but not reported here, also suggest that more work is required in the direction of practical implementation, casting doubt on the finite-sample performance. Nevertheless, this remains an interesting application that might deserve further study.

3.6. Choosing the Learning Rate

As remarked previously, in some applications, the choice of learning rate η_t can have a substantial impact. From the author's experience, $\eta \in (0, 1]$ usually works well. For the algorithm used in Yang (2004), under certain conditions, there is a theoretical justification for choosing the learning rate η_t equal to one (e.g., Sancetta, 2007). However, for the gradient algorithm used here, such a claim is more difficult to establish. For bounded random variables, the learning rate should not much affect the results. However, when the random variables are unbounded, the thicker the tails, the smaller η should be chosen. Hence, it might be convenient to use an automatic procedure to do so. We discuss two possible alternatives.

Suppose that $\hat{P}_t^{(j)}$ is the combined prediction based on Algorithm 1 using $\eta = \eta_j$, $j = 1, \ldots, J$. In doubt about which η_j should be used, we combine using combination weights $(v_{tj})_{j\in\{1,\ldots,J\}}$ in the unit simplex and use the combined prediction $\hat{P}_t^{(v)} := \sum_{j=1}^{J} v_{tj}\hat{P}_t^{(j)}$. Basically we construct a combination of the forecast combinations, where each $\hat{P}_t^{(j)}$ corresponds to the combined forecast at time t using learning rate $\eta_t = \eta_j t^{-\alpha}$ (with slight abuse of notation). The combination weights $(v_{tj})_{j\in\{1,\ldots,J\}}$ are then chosen using a modified version of the aggregating algorithm of Vovk (1990), which is closely related to the algorithms studied in Yang (2004) and Sancetta (2007).

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Algorithm 2

Set

$v_{11} = \cdots = v_{1K} = 1/J;$

$L_0^{(v)} := 0.$

For $t = 1, \ldots, T,$

$L_t^{(v)} = L_{t-1}^{(v)} + l_t^{(v)},$

$v_{t+1,j} := v_{tj} \exp\left\{-t^{-1/2} l_t^{(j)}\right\} \Big/ \sum_{j=1}^{J} v_{tj} \exp\left\{-t^{-1/2} l_t^{(j)}\right\},$

where

$l_t^{(v)} = l\left(Y_t, \hat{P}_t^{(v)}\right)$

and

$l_t^{(j)} = l\left(Y_t, \hat{P}_t^{(j)}\right).$

Writing $w_t^j = (w_{t1}^j, \ldots, w_{tK}^j)$ for the combination weights in Algorithm 1 using $\eta = \eta_j$, $\hat{P}_t^{(v)} = \langle w_t^v, X_t \rangle$ where $w_t^v = (w_{t1}^v, \ldots, w_{tK}^v)$ and $w_{tk}^v = \sum_{j=1}^{J} v_{tj} w_{tk}^j$. We then state the following result, which is deduced from Theorem 2.3 in Cesa-Bianchi and Lugosi (2006) using Theorem 1.

THEOREM 4. *Define* $m_T(\tau) := \left[\frac{1}{c_2}\tau + \frac{1}{c_2}\ln(c_1(K+1)T)\right]^{1/c_3}$ *and for any* $m > 0$,

$$\bar{l}(m) := \sup_{\substack{y \in [-m,m] \cap \mathcal{Y} \\ p \in [-m,m] \cap \mathcal{P}}} |l(y,p)|.$$

Under Conditions 1 and 2, with probability at least $1 - e^{-\tau}$, *combining Algorithms 1 and 2,*

$$\sum_{t=1}^{T} [l_t(w_t^v) - l_t(u_t)] \leq error,$$

where

$$error \leq T^{1/2} \frac{|\bar{l}(m_T(\tau))|^2}{4} + 2T^{1/2} \ln J + \min_{j \in \{1,\ldots,J\}} error_j$$

and

$$error_j \leq \frac{T^\alpha}{\eta_j} \left(6\ln(K/\gamma) + [\ln(K/\gamma) + K] \sum_{t=1}^{T-1} |u_t - u_{t+1}|_1\right) + \frac{\eta_j T^{1-\alpha}}{8} \left[\bar{l}'(m_T(\tau)) m_T(\tau)\right]^2.$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Theorem 4 shows that we can increase the error in the bound but this will lead to an automatic selection of the best learning rate. Theoretically, this bound can be disappointing, as under special circumstances, the additional error term can be of the same order of magnitude as $error_j$. However, in practical situations, it is valuable to select the best learning rate.

If we find the combination of the two algorithms complicated, a heuristic approach that delivers reasonable results is as follows. From the proof of Theorem 2 we might be induced to choose $\eta \simeq (\mathrm{E}\max_{k\leq K} |l'(Y_t, \langle w_t, X_t\rangle)X_{tk}|^2)^{-1/2}$ so that under ergodicity this implies

$$\eta^{-1/2} \simeq \left((t-1)^{-1}\sum_{s<t}\max_{k\leq K}|l'(Y_s, \langle w_s, X_s\rangle)X_{sk}|^2\right)^{-1/2} \tag{7}$$

at time t. Unfortunately, the finite-sample bounds proved here require that η be fixed. Hence, the bounds would be valid only for large T, under ergodicity conditions.

3.7. Further Remarks

Algorithm 1 uses the gradient of the loss function, whereas the algorithms considered in Yang (2004) and Sancetta (2007), among others, are based on the loss function. Based on both simulation and empirical work carried out by the author, the difference between a gradient-based algorithm and one based on the loss function can be marginal. However, the choice of learning rate can be important in some situations. Based on the author's experience, the performance of a gradient algorithm is more sensitive to the choice of learning rate than the performance of algorithms based on the loss function. Section 3.6 provided some suggestions to mitigate the problem.

The advantage of the gradient algorithm discussed here is that by a simple recursive update, it allows for time varying optimal combination weights. In this respect, bounds with time varying optimal combination weights are unknown for algorithms based on the loss function, unless we are willing to consider time varying optimal weights that lie on the edges of the simplex (e.g., Sancetta, 2007). As previously remarked, a loss-based algorithm (e.g., Yang, 2004) might be used to track the best time invariant optimal forecast combination only by some ad hoc extension, which adds to the computational burden.

Here, we can allow for time varying optimal combination weights shrinking the weights (i.e., $\gamma > 0$ in Algorithm 1) or, more formally, by projecting the weights onto $\mathcal{G}$ as in (3). The advantage of time varying optimal combination weights depends on the context, and it is not always the case that it will outperform algorithms that do not account for this. In general we make the following observations.

1. If there is one single forecast that clearly outperforms the others, shrinkage leads to suboptimal results. This is because the optimal combination weights

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

will lie on an edge of the simplex (i.e., it is optimal to choose the best forecast).
2. If the optimal combination weights change slowly over time, shrinkage may deliver results that are similar to using no shrinkage and a smaller learning rate (e.g., small η).
3. If the optimal combination weights change abruptly resulting in a large loss, then shrinkage does usually help.
4. In some situations, shrinkage may not lead to an improvement over the average loss but may decrease the degree of variability in the individual losses and avoid extreme losses. This was also noticed in related empirical evidence provided in Sancetta (2007).

In general Algorithm 1 allows us to incorporate learning in forecast combination (as in the algorithms studied by Yang, 2004) and, setting $\gamma > 0$, to reduce variability (as done when we choose equally weighted forecast combinations).

We conclude by providing a simulation example, whose purpose is not to show that one method is superior to another (this clearly depends on the problem we choose) but to provide some illustration on when one method is likely to perform better than the others. For this reason, the model chosen does not pretend to be realistic but is simple enough to allow discrimination between methods with special focus on the algorithms in Yang (2004) and those considered here.

Illustration: AR(1) with Time Varying Coefficient. Suppose

$$Y_t = A_t Y_{t-1} + Z_t,$$

where

$$A_t = A_{t-1} I_t + U_t (1 - I_t) \tag{8}$$

and $(I_t)_{t\in\mathbb{N}}$ is i.i.d. Bernoulli (p), $(U_t)_{t\in\mathbb{N}}$ is i.i.d. $(0, 1)$ uniform, and $(Z_t)_{t\in\mathbb{N}}$ is i.i.d. chi-square one. We set $A_1 = U_1$ and $Y_0 = 0$, so that when $p = 1$, $(Y_t)_{t\in\mathbb{N}}$ collapses to the usual stationary AR(1) model because $A_t = U_1$ when $p = 1$, by the initial condition. When $p < 1$, the time series dynamics of the data may change abruptly from one period to another. The optimal conditional mean square error predictor is $\mathrm{E}_{t-1} Y_t = (\mathrm{E}_{t-1} A_t) Y_{t-1} + 1$ and $(\mathrm{E}_{t-1} A_t) \in (0, 1)$. To verify the range of A_t, we just use the recursion (8) with $A_1 \in (0, 1)$ implying $A_t \in (0, 1)$. Hence, the data satisfy the conditions of the paper. For simplicity suppose we use the following individual forecast of Y_t: $a_k Y_{t-1} + 1, k = 1, 2$, where $a_1 = 0, a_2 = 1$. A convex combination of the two forecasts gives $\hat{P}_t = (1 - w_t) Y_{t-1} + 1$, which can equal $\mathrm{E}_{t-1} Y_t$ for suitable choice of w_t (the combination weight on $a_1 Y_{t-1} + 1 = 1$). The two forecasts are combined using Algorithm 1. We consider $\alpha = 1/2$, $\eta = 0.01, 0.1, 1$ and $\gamma = 0.05$, and we combine the combined forecasts based on different η's using Algorithm 2. We refer to the resulting forecast combination as Alg.L, whereas Alg.B will denote the results for Algorithm 1 choosing the best $\eta \in \{0.01, 0.1, 1\}$ ex post. We also consider η chosen according to (7) and refer to

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

it as Alg.E. Combination weights are also produced using the algorithm studied in Yang (2004). One version of this algorithm requires an estimate of the variance of the forecast error based on the past loss for the combined forecast (i.e., Yang's eqns. (4) and (6)), and we call this Yang1, whereas the second version considered is based on his Theorem 5 (the most relevant to the present study) with a learning rate that here was chosen equal to one (for details on that algorithm and the choice of learning rate, see also Sancetta, 2007, Sect. 5) and is referred to as Yang2. We also compute the equally weighted forecast (EW), the unconstrained least squares forecast with hindsight (LS), and the least squares forecast with hindsight with weights constrained to lie in (3) (LSC). Finally, we also consider the best individual forecast (Best).

We simulate 100 samples of 400 observations for various choices of p, and we start predicting from the second observation of each sample. For each sample, the average square loss is recorded. Figure 1 shows the box plots for the 100 average losses (one for each sample) when $p = 1, 0.95, 0.85, 0.35$. Throughout, LS and LSC have similar performance, and so we only report LS. EW has a very mediocre performance for large p but starts to perform well when $p \leq 0.65$. For $p = 0.35$ EW performs similarly to LS. To avoid visual distortions in the scale of the box plots we will not report it.

The mediocre performance of EW is expected: we combine forecasts that are potentially very biased, and their equal combination does not reduce the bias unless $p = 0$ implying $\mathrm{E}_{t-1} A_t = 1/2$ so that the EW forecast leads to the optimal conditional mean square error forecast. Indeed EW forecasts lead to excellent results when either all the forecasts are not too biased (reducing variability) or when the forecasts are biased, but EW reduces this bias. As mentioned previously, the goal is to provide some further insights for the comparison of the algorithms in Yang (2004) and those discussed here (e.g., for empirical evidence and comparison of EW and the algorithms of the kind that are used here, see Sancetta, 2007).

When $p = 1$, Alg.L and Alg.E perform similarly to the best optimal combined forecast with hindsight (same story when we choose any other $\eta \in \{0.01, 0.1, 1\}$). Yang1 and Yang 2 perform similarly to the best individual forecast. These results are in line with the theory. When $p = 0.95$ we have a different picture. The gradient-based algorithms can be quite unstable, and this is shown by the big outlier in Alg.L and Alg.E. Choosing the best learning rate would have avoided the outlier (see Alg.B). As remarked by Yang (2004), Yang2 can be unstable in some practical situations, though in theory a stronger bound holds for Yang2 than for Yang1. Nevertheless when compared to the gradient algorithms, Yang1 and Yang2 are clearly more stable. To convey an idea about the problem, Figure 2 plots the sequence that led to the big outlier; this has quite erratic behavior. In these situations, the algorithms of Yang (2004) might be more robust. Although one should guard against the problem of data snooping in making these claims about instability, the author's experience with other data sets confirms the possibility of this instability, which might be less of a problem with algorithms based on the loss functions rather than on the gradient.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

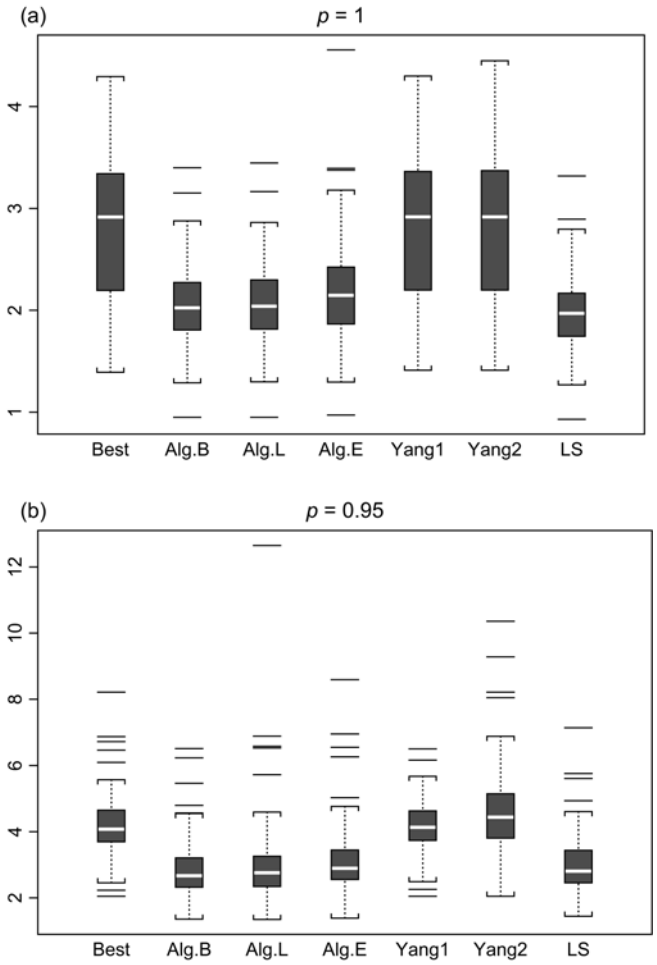


FIGURE 1. Simulation results.

When $p = 0.85$ there are a sufficient number of changes that lead Yang1 to perform as well as Best, but worse than LS—the performance of Yang2 is worse, confirming Yang's suggestion to use Yang1 unless we are sure about the value of the learning rate. Alg.L and Alg.E perform almost as well as Alg.B, whose performance is close to LS. A large number of shifts (i.e., $p = 0.85$) avoids a persistent trending behavior as shown in Figure 2, which can be a cause of instability in the gradient algorithms.

As p becomes smaller, the picture is similar to the previous cases. An increase in the number of shifts finds an improvement in the equally weighted combined forecast (for $p = 0.35$, the performance of EW and LS is similar). This appears

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

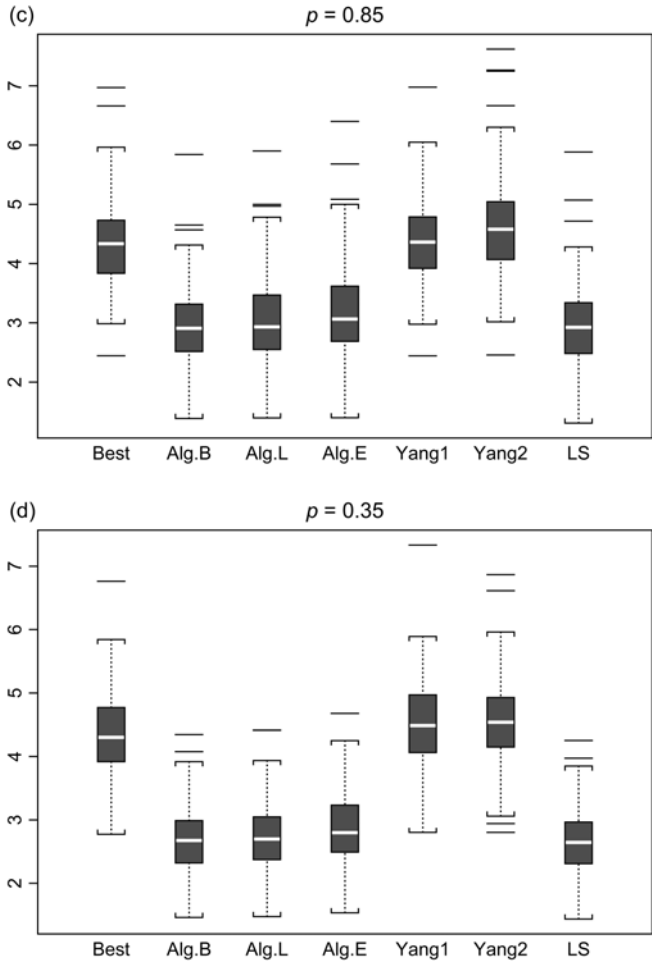


FIGURE 1. *continued.*

to reduce the performance of Yang1 and Yang2 relative to the gradient algorithms that are specifically devised to perform in situations when the optimal combined forecast does not lie at one of the edges of the simplex.

In the previous example, the individual forecasts could be much improved by forecast combination. We briefly consider the opposite case when the best individual forecast cannot be improved. We include the true optimal conditional mean square error forecast $\mathrm{E}_{t-1}Y_t = (\mathrm{E}_{t-1}A_t)Y_{t-1} + 1$, $\mathrm{E}_{t-1}A_t = pA_{t-1} + (1-p)/2$. Hence, we use the following individual forecasts, $a_{kt}Y_{t-1} + 1$, $k = 1, 2, 3$, where $a_{1t} = 0$, $a_{2t} = 1$, and $a_{3t} = \mathrm{E}_{t-1}A_t$. Clearly, this is quite idealistic, but our goal is just to illustrate the relative loss of the present algorithms in situations that are

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

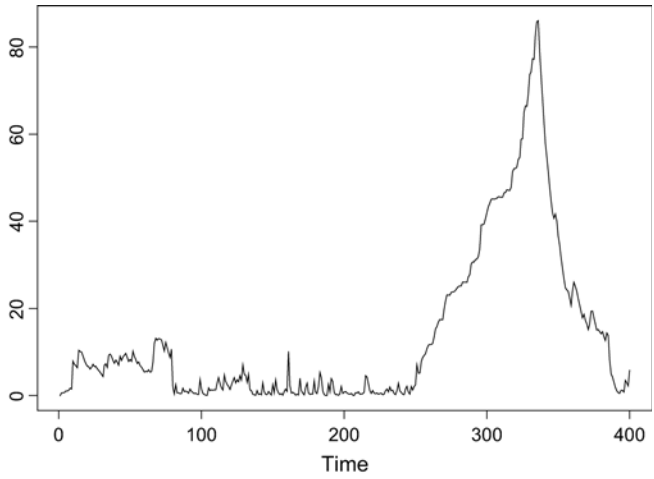


FIGURE 2. Simulated sequence causing outliers in Alg.L and Alg.E.

more favorable to the procedures studied in Yang (2004). It is enough to restrict $p = 1$, because $p < 1$ is not of great interest when we include the third forecast $\mathrm{E}_{t-1} Y_t$. In fact, now the optimal weights in a conditional mean square error sense are time invariant, i.e., $u_{1t} = u_{t2} = 0$ and $u_{t3} = 1$, favoring the third (unfeasible) forecast $\mathrm{E}_{t-1} Y_t$, no matter what p is. Figure 3 compares the different algorithms. When one of the forecasts is very good, we incur an additional loss when we use the algorithms considered in this paper. On average, the increase in the loss relative to Yang1 and Yang2 appears marginal. This example shows the possibility

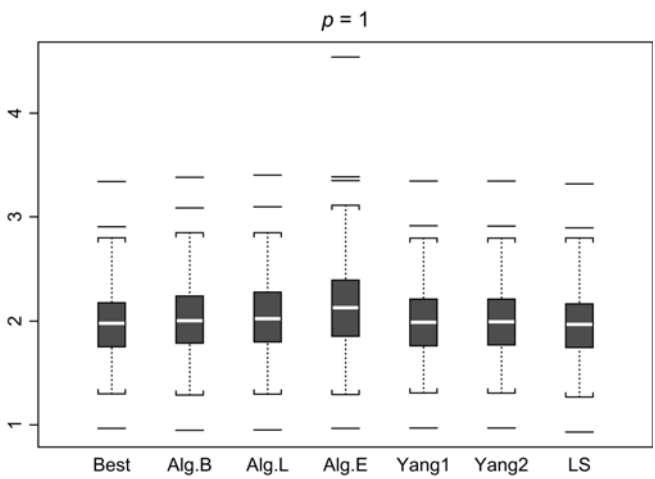


FIGURE 3. Simulation when one individual forecast is optimal.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

of instability in the results when a gradient-based algorithm is used. This confirms the suggestion that we should use a loss-based algorithm if the more restrictive conditions in Yang (2004) are satisfied.

REFERENCES

Aiolfi, M. & A. Timmermann (2006) Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics* 135, 31–53.

Andrews, D. (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–856.

Bai, J. & P. Perron (1998) Estimating and testing linear models with multiple structural changes. *Econometrica* 66, 47–78.

Breiman, L. (1961) Optimal gambling systems for favorable games. In J. Neyman (ed.), *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* 1, 65–78. University of California Press.

Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Annals of Statistics* 24, 2350–2383.

Cesa-Bianchi, N. (1999) Analysis of two gradient-based algorithms for online regression. *Journal of Computer and System Sciences* 59, 392–411.

Cesa-Bianchi, N. & G. Lugosi (2006) *Prediction, Learning , and Games.* Cambridge University Press.

Chen, Z. & Y. Yang (2007) Time series models for forecasting: Testing or combining? *Studies in Nonlinear Dynamics & Econometrics* 11, no. 1, article 3. Downloadable: http://www.bepress.com/snde/vol11/iss1/art3.

Cover, T. (1991) Universal portfolios. *Mathematical Finance* 1, 1–29.

Cross, J.E. & A.R. Barron (2003) Efficient universal portfolios for past dependent target classes. *Mathematical Finance* 13, 245–276.

Dawid, A.P. (1984) Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A* 147, 278–292.

Dawid, A.P. (1997) Prequential analysis. In S. Kotz, C.B. Read, & D.L. Banks (eds.), *Encyclopedia of Statistical Sciences*, vol. 1, pp. 464–470. Wiley.

Deutsch, M., C.W.J. Granger, & T. Teräsvirta (1994) The combination of forecasts using changing weights. *International Journal of Forecasting* 10, 47–57.

Devroye, L., L. Györfi, & G. Lugosi (1996) *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag.

Diebold, F.X. & P. Pauly (1990) The use of prior information in forecast combination. *International Journal of Forecasting* 6, 503–508.

Elliott, G. & A. Timmermann (2004) Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics* 122, 47–79.

Györfi, L., G. Lugosi, & F. Udina (2006) Nonparametric kernel-based sequential investment strategies. *Mathematical Finance* 16, 337–358.

Helmbold, D.P., R.E. Schapire, Y. Singer, & M.K. Warmuth (1998) Online portfolio selection using multiplicative updates. *Mathematical Finance* 8, 325–347.

Hendry, D.F. & M.P. Clements (2004) Pooling of forecasts. *Econometrics Journal* 7, 1–31.

Herbster, M. & M.K. Warmuth (2001) Tracking the best linear predictor. *Journal of Machine Learning Research* 1, 281–309.

Kivinen, J. & M.K. Warmuth (1997) Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation* 132, 1–63.

Leung, G. & A.R. Barron (2006) Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52, 3396–3410.

Nikandrova, A. (2005) Universal portfolios selection from a practical perspective. Masters Dissertation, Faculty of Economics, University of Cambridge.

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Pesaran, M.H. & A. Timmermann (2005) Real time econometrics. *Econometric Theory* 21, 212–231.
Samuelson, P.A. (1979) Why we should not make mean log of wealth big though years to act are long. *Journal of Banking and Finance* 3, 305–307.
Sancetta, A. (2007) Online forecast combinations of distributions: Worst case bounds. *Journal of Econometrics* 141, 621–651.
Stock, J.H. & M.W. Watson (2004) Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* 23, 405–430.
Timmermann, A. (2006) Forecast combinations. In G. Elliott, C.W.J. Granger, & A. Timmermann (eds.), *Handbook of Economic Forecasting* 1, pp. 135–196. Elsevier.
Vapnik, V. (1998) *Statistical Learning Theory*. Wiley.
Vovk, V. (1990) Aggregating strategies. In M.A. Fulk (ed.), *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pp. 371–386. Morgan Kaufmann. Downloadable: http://www.vovk.net/aa/01.zip.
Yang, Y. (2004) Combining forecasting procedures: Some theoretical results. *Econometric Theory* 20, 176–222.
Zou, H. & Y. Yang (2004) Combining time series models for forecasting. *International Journal of Forecasting* 20, 69–84.

APPENDIX: Proofs

A.1. Proofs of Theorems and Corollaries. The proof of Theorem 1 will use a few lemmas stated and proved in Section A.2.

Proof of Theorem 1. Define

$$g_t := \bar{l}'(m_t). \tag{A.1}$$

With probability at least $1-\epsilon$, where $\epsilon := e^{-\tau}$,

$$\sum_{t=1}^{T}[l_t(w_t)-l_t(u_t)] = \sum_{t=1}^{T}[l_t(w_t)I_{M_t}-l_t(u_t)]$$

[by Lemma A.1]

$$\leq \sum_{t=1}^{T}\frac{D(u_t,w_t)-D(u_t,w_t')}{\eta_t}I_{M_t}+\frac{\eta}{8}\sum_{t=1}^{T}\frac{|g_tX_tI_{\{|X_t|\leq m_t\}}|_\infty^2}{t^\alpha}$$

[by Lemma A.2]

$$\leq \sum_{t=1}^{T}\frac{D(u_t,w_t)-D(u_t,w_t')}{\eta_t}I_{M_t}+\frac{\eta}{8}\sum_{t=1}^{T}\frac{\left[\bar{l}'(m_t(\tau))m_t(\tau)\right]^2}{t^\alpha}$$

[by (A.1) and using the constraint]

$$\leq \sum_{t=1}^{T}\frac{D(u_t,w_t)-D(u_t,w_{t+1})}{\eta_t}I_{M_t}+\frac{\eta}{8}\sum_{t=1}^{T}\frac{\left[\bar{l}'(m_t(\tau))m_t(\tau)\right]^2}{t^\alpha}$$

[by Lemma A.3]

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

$$\leq 6\frac{T^\alpha}{\eta}(1+\ln(K/\gamma))+[\ln(K/\gamma)+K]\frac{T^\alpha}{\eta}\sum_{t=1}^{T}|u_t-u_{t+1}|_1$$
$$+\frac{\eta}{8}\sum_{t=1}^{T}\frac{\left[\bar{l}'(m_t)m_t\right]^2}{t^\alpha}$$

[by Lemma A.4]

$$\leq \frac{T^\alpha}{\eta}\left(6(1+\ln(K/\gamma))+\sum_{t=1}^{T}[\ln(\gamma/K)+K]|u_t-u_{t+1}|_1\right)$$
$$+\frac{\eta}{8}\left[\bar{l}'(m_T(\tau))m_T(\tau)\right]^2\sum_{t=1}^{T}t^{-\alpha}$$

$$\leq \frac{T^\alpha}{\eta}\left(6(1+\ln(K/\gamma))+\sum_{t=1}^{T-1}[\ln(\gamma/K)+K]|u_t-u_{t+1}|_1\right)$$
$$+\frac{\eta T^{1-\alpha}}{8(1-\alpha)}\left[\bar{l}'(m_t(\tau))m_t(\tau)\right]^2$$

summing over t and noting that there is no prediction at time $T+1$; hence we can choose $u_{T+1}=u_T$, and the last term in the summation (inside the parentheses) drops. ■

The proofs of the corollaries are next.

Proof of Corollary 2. By direct computation,

$$\left|\frac{\partial l(y,p)}{\partial p}\right| = 2|y-p| \leq 2(|y|\vee|p|),$$

and Condition 2 applies with $\bar{l}'(m)\leq 2m$, so that

$$\left[\bar{l}'(m_t(\tau))m_t(\tau)\right]^2 = 4m_T(\tau)^4 = O\left([\tau+\ln T]^{4/c_3}\right)$$

and we obtain the first result by application of Corollary 1. For LinEx loss,

$$\left|\frac{\partial l(y,p)}{\partial p}\right| \leq 2|\theta|\exp\{|\theta||y-p|\} \leq 2|\theta|\exp\{2|\theta|(|y|\vee|p|)\}$$

so that, by Condition 2,

$$\left[\bar{l}'(m_t(\tau))m_t(\tau)\right]^2 \leq 4|\theta|^2\exp\{2|\theta|m_T(\tau)\}m_T(\tau)^2 \leq 4|\theta|^2\exp\{4|\theta|m_T(\tau)\}$$
$$= O\left(\exp\left\{4|\theta|\left[\left(\frac{\tau}{c_2}\right)^{1/c_3}+\left(\frac{\ln T}{c_2}\right)^{1/c_3}\right]\right\}\right)$$
$$= o\left(T\exp\left\{4|\theta|\left(\frac{\tau}{c_2}\right)^{1/c_3}\right\}\right),$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

using the fact that $4|\theta|(\ln T/c_2)^{1/c_3} = o(\ln T)$ for $c_3 > 1$. The result follows by application of Corollary 1. ■

Proof of Corollary 3. From the proof of Corollary 2,

$$\left[\bar{l}'(m_t(\tau))m_t(\tau)\right]^2 = O\left([\tau + \ln T]^{4/c_3}\right).$$

Hence, solving $T^{\alpha+(1-\epsilon)} = T^{1-\alpha}$ with respect to α and substituting back in Theorem 1 gives the result. ■

Proof of Corollary 4. Take expectation on both sides of (A.2) in the proof of Lemma A.1 in Section A.1 The first term on the right-hand side is bounded by Theorem 1, and we shall only bound the second term using Hölder inequality:

$$\sum_{t=0}^{T} \mathrm{E} l\left(Y_t, w_t X_t\right) I_{M_t^c} \leq A \sum_{t=0}^{T} \Pr\left(M_t^c\right)^{1/2}.$$

To bound $\Pr(M_t^c)$, note that the event $\{\max_{1\leq k\leq K} |X_{tk}| \leq m_t\}$ implies the event $\{|\hat{P}_t| \leq m_t\}$ (see the proof of Lemma A.1). Hence, by the arguments in the proof of Lemma A.1,

$$\begin{aligned}\Pr(M_t^c) &\leq \Pr\left(|Y_t| > m_t\right) + \Pr\left(\max_{1\leq k\leq K} |X_{tk}| > m_t\right) \\ &\leq (K+1)c_1 \exp\{-c_2 m_t^{c_3}\} \\ &= t^{-1}\end{aligned}$$

choosing

$$m_t := \left[\frac{1}{c_2}\ln(c_1(K+1)t)\right]^{1/c_3}.$$

Hence,

$$\mathrm{E} l(Y_t, \langle w_t, X_t\rangle) I_{M_t^c} \leq A \Pr(M_t^c)^{1/2} \leq A t^{-1/2}.$$

Summing over t gives the result. ■

Proof of Theorem 2. For any $w_t, u_t \in \mathcal{S}^K$, conditioning on Y_t and X_t, by Lemma A.2

$$\begin{aligned}\mathrm{E} l(Y_t, \langle w_t, X_t\rangle) - l(Y_t, \langle u_t, X_t\rangle) \leq \mathrm{E}\frac{D(u_t, w_t) - D(u_t, w_t')}{\eta_t} \\ + \eta_t \frac{\mathrm{E}\max_k |l'(Y_t, \langle w_t, X_t\rangle) X_{tk}|^2}{8}\end{aligned}$$

and summing over t, by Lemma A.4

$$\begin{aligned}&\sum_{t=1}^{T} \mathrm{E}[l(Y_t, \langle w_t, X_t\rangle) - l(Y_t, \langle u_t, X_t\rangle)] \\ &\quad \leq 6\frac{T^\alpha}{\eta}[1 + \ln(K/\gamma)] + [\ln(K/\gamma) + K]\frac{T^\alpha}{\eta}\sum_{t=1}^{T} |u_t - u_{t+1}|_1\end{aligned}$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

$$+\eta\sum_{t=1}^{T}\frac{\mathrm{E}\max_k|l'(Y_t,\langle w_t,X_t\rangle)X_{tk}|^2}{8t^\alpha}$$

$$\le 6\frac{T^\alpha}{\eta}[1+\ln(K/\gamma)]+[\ln(K/\gamma)+K]\frac{T^\alpha}{\eta}\sum_{t=1}^{T}|u_t-u_{t+1}|_1$$

$$+\eta T^{1-\alpha}\frac{\sup_t \mathrm{E}\max_k \sup_{w_t\in\mathcal{G}}|l'(Y_t,\langle w_t,X_t\rangle)X_{tk}|^2}{8(1-\alpha)}$$

summing over t and using standard uniform upper bounds. Now note that for any K real random variables $Z_1,\ldots,Z_K$ and $p\ge 1$,

$$\left(\mathrm{E}\max_k|Z_k|\right)^p \le \mathrm{E}\left(\max_k|Z_k|\right)^p$$

[by Jensen inequality]

$$\le \sum_{k=1}^{K}\mathrm{E}|Z_k|^p$$

$$\le K\max_k \mathrm{E}|Z_k|^p,$$

which implies

$$\mathrm{E}\max_k|Z_k| \le K^{1/p}\max_k(\mathrm{E}|Z_k|^p)^{1/p}.$$

Setting $Z_k=\sup_{w_t\in\mathcal{G}}|l'(Y_t,\langle w_t,X_t\rangle)X_{tk}|^2$, an application of the previous display gives the result. ■

Proof of Theorem 3. By a sign change we can consider $-\ln(p)$, as for the previous results. Hence, follow the proof of Theorem 1 in the following order: apply Lemma A.2 with no need to apply Lemma A.1 first. Then, Lemma A.2 gives a bound in terms of $|l'(y,\langle w,x\rangle)x|_\infty^2$, which, for the specific choice of loss function, corresponds to

$$\max_k\left|\frac{x_{tk}}{\langle w_t,x_t\rangle}\right|^2 \le \max_k\left|\frac{x_{tk}}{(\gamma/K)x_{tk}}\right|^2 \le \left|\frac{K}{\gamma}\right|^2$$

using the constraint. Then, the proof proceeds as for Theorem 1 and the corollaries. ■

Proof of Theorem 4. Following the same arguments in the proof of Theorem 1, deduce that with probability at least $1-\epsilon$, where $\epsilon:=e^{-\tau}$,

$$\sum_{t=1}^{T}[l_t(w_t^v)-l_t(u_t)] = \sum_{t=1}^{T}\left[l_t\left(w_t^v\right)I_{M_t}-l_t(u_t)\right]$$

$$\le \min_{j\in\{1,\ldots,J\}}\sum_{t=1}^{T}\left[l_t\left(w_t^j\right)I_{M_t}-l_t(u_t)\right]+T^{1/2}\frac{|l(m_T(\tau))|^2}{4}$$

$$+2T^{1/2}\ln J.$$

An application of Theorem 1 gives the result. ■

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

A.2. Technical Lemmas. These lemmas are used in the proof of Theorem 1.

LEMMA A.1. *Define* $m_t = \left[\frac{1}{c_2}\ln(\frac{(K+1)c_1 t}{\epsilon})\right]^{1/c_3}$ *and the set*

$M_t := \{|Y_t| \le m_t, |\hat{P}_t| \le m_t\}$.

Under Conditions 1 and 2, with probability at least $1-\epsilon$,

$$\sum_{t=1}^{T} l(Y_t, \hat{P}_t) = \sum_{t=1}^{T} l\left(Y_t, \hat{P}_t\right) I_{M_t}.$$

Proof. Let M_t^c be the complement of M_t. Consider the following identity:

$$\sum_{t=1}^{T} l(Y_t, \hat{P}_t) = \sum_{t=1}^{T} l\left(Y_t, \hat{P}_t\right) I_{M_t} + \sum_{t=1}^{T} l\left(Y_t, \hat{P}_t\right) I_{M_t^c} \tag{A.2}$$

$$= \mathrm{I} + \mathrm{II}.$$

Note that

$$\Pr\left(\left|\hat{P}_t\right| > m_t\right) \le \Pr\left(|w_t|_1 \max_{1\le k\le K} |X_{tk}| > m_t\right)$$

$$= \Pr\left(\max_{1\le k\le K} |X_{tk}| > m_t\right) \tag{A.3}$$

[because $w_t \in \mathcal{S}^K$]

$$\le \sum_{k=1}^{K} \Pr\left(\left|X_{tk}\right| > m_t\right)$$

by the union bound, so that the event $\{\max_{1\le k\le K} |X_{tk}| \le m_t\}$ implies the event $\left\{|\hat{P}_t| \le m_t\right\}$. Hence,

$$\Pr\left(\sum_{t=1}^{T} l_t\left(Y_t, \hat{P}_t\right) I_{M_t^c} > 0\right) \le \Pr\left(\max_{1\le t\le T} |Y_t| > m_t\right) + \Pr\left(\max_{1\le t\le T}\left|\hat{P}_t\right| > m_t\right)$$

$$\le \sum_{t=1}^{T}\left[\Pr(|Y_t| > m_t) + \sum_{k=1}^{K} \Pr(|X_{tk}| > m_t)\right]$$

[by (A.3) and the union bound]

$$\le (K+1)c_1 \sum_{t=1}^{T} \exp\{-c_2 m_t^{c_3}\}$$

[by Condition 1]

$$= \epsilon \quad \text{choosing } m_t = \left[\frac{1}{c_2}\ln\left(\frac{(K+1)c_1 t}{\epsilon}\right)\right]^{1/c_3}.$$

Hence with this choice of m_t, the result holds with probability at least $1-\epsilon$. ■

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

The next result follows from Lemma 2 in Cesa-Bianchi (1999). A sketch of the proof is provided for the sake of completeness and also because it is a key ingredient.

LEMMA A.2. *For any* $y \in \mathbb{R}$, $x \in \mathbb{R}^K$, *and* $\eta > 0$, $u \in \mathcal{S}^K$, $w \in int(\mathcal{S}^K)$, *and*

$$w'_k := \frac{w_k \exp\{-\eta l'(y, \langle w, x\rangle)x_k\}}{\sum_{k=1}^K w_k \exp\{-\eta l'(y, \langle w, x\rangle)x_k\}},$$

the following result holds:

$$l(y, \langle w, x\rangle) - l(y, \langle u, x\rangle) \le \frac{D(u, w) - D(u, w')}{\eta} + \eta \frac{|l'(y, \langle w, x\rangle)x|_\infty^2}{8}.$$

Proof. Define $l'(y, p) := (\partial/\partial p) l(y, p)$ and $l''(y, p)$ for the second-order derivative with respect to p. Hence, by Taylor expansion, for some w^* in the convex hull of $\{u, w\}$,

$$l(y, \langle w, x\rangle) - l(y, \langle u, x\rangle) = \langle w - u, x\rangle l'(y, \langle w, x\rangle) - \frac{\langle w - u, x\rangle^2 l''(y, \langle w^*, x\rangle)}{2}$$

$$\le \langle w - u, x\rangle l'(y, \langle w, x\rangle) \tag{A.4}$$

by convexity of $l(y, p)$ with respect to p. Recall that $D(w, u) := \sum_k w_k \ln(w_k/u_k)$ and define $z_k := -\eta l'(y, \langle w, x\rangle)x_k$. Then,

$$\begin{aligned} D(u, w) - D(u, w') &= \sum_{k=1}^K u_k \ln(w'_k/w_k) \\ &= \sum_{k=1}^K u_k \left[\ln \frac{w_k \exp\{z_k\}}{\sum_{l=1}^K w_l \exp\{z_l\}} - \ln w_k\right] \\ &\quad \text{[by definition of } w'_k\text{]} \\ &= \sum_{k=1}^K u_k z_k - \ln\left(\sum_{l=1}^K w_l \exp\{z_l\}\right) \\ &= \sum_{k=1}^K (u_k - w_k) z_k - \ln\left(\sum_{l=1}^K w_l \exp\left\{z_l - \sum_{k=1}^K w_k z_k\right\}\right) \\ &\ge \sum_{k=1}^K (u_k - w_k) z_k - \frac{|z|_\infty^2}{8}, \end{aligned}$$

using Hoeffding's inequality (e.g., Devroye, Györfi, and Lugosi, 1996) for the moment generating function of bounded random variables in the last step. Substituting in (A.4), we obtain

$$\langle w - u, x\rangle l'(y, \langle w, x\rangle) \le \frac{D(u, w) - D(u, w')}{\eta} + \eta \frac{|l'(y, \langle w, x\rangle)x|_\infty^2}{8},$$

implying the last result. ■

The next lemma is Corollary 3 in Herbster and Warmuth (2001). A concise proof for general "distance" functions can be found in Lemma 11.3 of Cesa-Bianchi and Lugosi (2006).

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

LEMMA A.3. *For any vector $u_t \in \mathcal{G}$ ($\mathcal{G}$ as in (3)), $w_{t+1} = \arg\min_{v\in\mathcal{G}} D(v, w'_t)$ (i.e., w_{t+1} as in Algorithm 1) and $w'_t \in int(\mathcal{S}^K)$,*

$$D(u_t, w'_t) \geq D(u_t, w_{t+1}).$$

For $u \in \mathcal{S}^K$ and $w \in int(\mathcal{S}^K)$, define

$$D_F(u, w) = F(u) - F(w) - \langle (u - w), \nabla F(w)\rangle.$$

We note that when $F(u) := \sum_{k=1}^K u_k(\ln u_k - 1)$, $D_F(u, w) = D(u, w)$. To remind us of the preceding decomposition, we will use $D_F(u, w)$ instead of $D(u, w)$ in the last two lemmas. Different choices of F would lead to different "distance" functions and different updates via (2) (for details, see, e.g., Cesa-Bianchi and Lugosi, 2006, Ch. 11).

LEMMA A.4. *For $u_t \in \mathcal{S}^K$, η_t and w_t as in Algorithm 1 ($t = 1, \ldots, T$),*

$$\sum_{t=1}^T \frac{D_F(u_t, w_t) - D_F(u_t, w_{t+1})}{\eta_t} \leq 6\frac{T^\alpha}{\eta}[1 + \ln(K/\gamma)] + [\ln(K/\gamma) + K]\frac{T^\alpha}{\eta} \times \sum_{t=1}^T |u_t - u_{t+1}|_1.$$

Proof. To prove the lemma we need to find a telescoping sum. To this end

$$\begin{aligned}\frac{D_F(u_t, w_t) - D_F(u_t, w_{t+1})}{\eta_t} &= \frac{D_F(u_t, w_t) - D_F(u_{t+1}, w_{t+1})}{\eta_t} \\ &\quad - \frac{D_F(u_t, w_{t+1}) - D_F(u_{t+1}, w_{t+1})}{\eta_t} \\ &\quad \text{[adding and subtracting } D_F(u_{t+1}, w_{t+1})] \\ &= \mathrm{I} + \mathrm{II}.\end{aligned}$$

We shall deal with the sum over t of each term separately.

Sum over I.

$$\begin{aligned}\sum_{t=1}^T \mathrm{I} &= \frac{1}{\eta_1} D_F(u_1, w_1) - \frac{1}{\eta_T} D_F(u_{T+1}, w_{T+1}) - \sum_{t=1}^{T-1} D_F(u_{t+1}, w_{t+1})\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right) \\ &= \frac{1}{\eta} D_F(u_1, w_1) - \frac{T^\alpha}{\eta} D_F(u_{T+1}, w_{T+1}) - \sum_{t=1}^{T-1} D_F(u_{t+1}, w_{t+1})\frac{t^\alpha - (t+1)^\alpha}{\eta} \\ &\quad \text{[by definition of } \eta_t] \\ &\leq \frac{1}{\eta} D_F(u_1, w_1) + \max_{1\leq t\leq T-1} D_F(u_{t+1}, w_{t+1}) \sum_{t=1}^{T-1} \frac{(t+1)^\alpha - t^\alpha}{\eta} \\ &\quad \text{[using the fact that } D_F \text{ is nonnegative]}\end{aligned}$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

$$= \frac{1}{\eta} D_F(u_1, w_1) + \max_{1 \le t \le T-1} D_F(u_{t+1}, w_{t+1}) \frac{T^\alpha}{\eta}$$

[because the sum telescopes]

$$\le 2 \max_{1 \le t \le T} D_F(u_t, w_t) \frac{T^\alpha}{\eta}.$$

Sum over II. By definition of D_F,

$$-D_F(u_t, w_{t+1}) + D_F(u_{t+1}, w_{t+1}) = F(u_{t+1}) - F(u_t) + \langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle.$$

Hence,

$$\text{II} = \frac{F(u_{t+1}) - F(u_t)}{\eta_t} + \frac{\langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle}{\eta_t}$$
$$= \text{IV} + \text{V}.$$

Note that $F(u) \in [0, -1]$ for $u \in [0, 1]$. Hence, summing over IV,

$$\sum_{t=1}^{T} \text{IV} = -\sum_{t=1}^{T} \frac{F(u_t) - F(u_{t+1})}{\eta_t}$$

$$= -\frac{1}{\eta} F(u_1) + \frac{1}{\eta_T} F(u_{T+1}) + \sum_{t=1}^{T-1} F(u_{t+1}) \frac{t^\alpha - (t+1)^\alpha}{\eta}$$

$$\le -\frac{1}{\eta} F(u_1) + \max_{1 \le t \le T-1} [-F(u_{t+1})] \sum_{t=1}^{T-1} \frac{(t+1)^\alpha - t^\alpha}{\eta}$$

[because $F(u)$ is nonpositive]

$$\le -\frac{1}{\eta} F(u_1) + \max_{1 \le t \le T-1} [-F(u_{t+1})] \frac{T^\alpha}{\eta}$$

[because the sum telescopes]

$$\le 2 \max_{1 \le t \le T} [-F(u_t)] \frac{T^\alpha}{\eta}.$$

Moreover,

$$\sum_{t=1}^{T} \text{V} \le \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |\langle (u_t - u_{t+1}), \nabla F(w_{t+1}) \rangle|$$
$$\le \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1 |\nabla F(w_{t+1})|_\infty.$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Hence

$$\sum_{t=1}^{T} \frac{D_F(u_t, w_t) - D_F(u_t, w_{t+1})}{\eta_t}$$

$$\leq 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} D_F(u_t, w_t) + 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} [-F(u_t)] + \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |\langle (u_t - u_{t+1}),$$

$$\nabla F(w_{t+1})\rangle|$$

$$= 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} (F(u_t) - F(w_t) - \langle (u_t - w_t), \nabla F(w_t)\rangle)$$

$$+ 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} [-F(u_t)] + \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1 |\nabla F(w_t)|_\infty$$

[by definition of $D_F(u_t, w_t)$]

$$\leq 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} \langle (w_t - u_t), \nabla F(w_t)\rangle + 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} [-F(u_t)]$$

$$+ 2\frac{T^\alpha}{\eta} \max_{1\leq t\leq T} [-F(w_t)] + \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1 |\nabla F(w_t)|_\infty$$

$$=: \text{VI}.$$

Using Lemma A.5 (stated next) the preceding display is bounded by

$$\text{VI} \leq 2\frac{T^\alpha}{\eta} \ln(K/\gamma) + 4\frac{T^\alpha}{\eta} [\ln(K/\gamma) + 1] + [\ln(K/\gamma) + K] \frac{T^\alpha}{\eta} \sum_{t=1}^{T} |u_t - u_{t+1}|_1;$$

hence the result. ■

Using the constraint $\mathcal{G}$, we can derive the following bounds used previously.

LEMMA A.5. *For* $F(u) := \sum_{k=1}^{K} u_k(\ln u_k - 1)$, $u \in \mathcal{S}^K$, *and* $w \in \mathcal{G}$

$$\langle (w_t - u_t), \nabla F(w_t)\rangle \leq \ln(K/\gamma),$$

$$|u_t - u_{t+1}|_1 |\nabla F(w_{t+1})|_\infty \leq [\ln(K/\gamma) + K] |u_t - u_{t+1}|_1,$$

and

$$F(u_t) \leq \ln(\gamma/K) - 1.$$

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press

Proof. By direct calculation,

$$\begin{aligned}\langle (w_t - u_t), \nabla F(w_t)\rangle &= \sum_{k=1}^{K} (w_{tk} - u_{tk})[\ln w_{tk} + (K-1)] \\ &= \sum_{k=1}^{K} (w_{tk} - u_{tk}) \ln w_{tk} \\ &\quad \text{[because the coefficients add up to one]} \\ &\leq -\sum_{k=1}^{K} u_{tk} \ln w_{tk} \\ &\quad \text{[because the first term is negative]} \\ &\leq -\ln(\gamma / K),\end{aligned}$$

using the constraint. Finally,

$$|u_t - u_{t+1}|_1 |\nabla F(w_{t+1})|_\infty \leq [-\ln(\gamma / K) + K - 1]|u_t - u_{t+1}|_1,$$

using the constraint. ■

The following result is directly deduced from Theorem 2.3 in Cesa-Bianchi and Lugosi (2006), assuming that the absolute value of the loss function $l(y, p)$ is bounded by a constant M.

LEMMA A.6. *Suppose* $|l(y, p)| \leq M$ *for any* y *and* p. *Then, using Algorithm 2,*

$$L_t^{(v)} \leq \min_{j \in \{1,\ldots,J\}} L_T^{(j)} + T^{1/2} \frac{M^2}{4} + 2T^{1/2} \ln J.$$

Proof. From the proof of Theorem 2.3 in Cesa-Bianchi and Lugosi (2006), accounting for the loss to be bounded by M rather than 1 (as in their proof), and using $t^{-1/2}$ as learning rate in their proof, deduce

$$\begin{aligned}L_T^{(v)} &\leq \min_{j \in \{1,\ldots,J\}} L_T^{(j)} + \frac{M^2}{8} \sum_{t=1}^{T} t^{-1/2} + 2 \sum_{t=1}^{T} \left[(t+1)^{1/2} - t^{1/2}\right] \ln J + \ln J \\ &\leq \min_{j \in \{1,\ldots,J\}} L_T^{(j)} + 2T^{1/2} \frac{M^2}{8} + 2\left[(T+1)^{1/2} - 1\right] \ln J + \ln J \\ &\quad \text{[bounding the first sum by } 2T^{1/2}\text{, whereas the second sum telescopes]} \\ &= \min_{j \in \{1,\ldots,J\}} L_T^{(j)} + T^{1/2} \frac{M^2}{4} + 2T^{1/2} \ln J.\end{aligned}$$

■

https://doi.org/10.1017/S0266466609100105 Published online by Cambridge University Press