Combining Forecasting Procedures: Some Theoretical Results
Author(s): Yuhong Yang
Source: *Econometric Theory*, Vol. 20, No. 1 (Feb., 2004), pp. 176-222
Published by: Cambridge University Press
Stable URL: http://www.jstor.org/stable/3533509
Accessed: 10-07-2016 22:11 UTC

Econometric Theory, **20**, 2004, 176–222. Printed in the United States of America.
DOI: 10.1017/S0266466604201086

COMBINING FORECASTING PROCEDURES: SOME THEORETICAL RESULTS

YUHONG YANG
Iowa State University

We study some methods of combining procedures for forecasting a continuous random variable. Statistical risk bounds under the square error loss are obtained under distributional assumptions on the future given the current outside information and the past observations. The risk bounds show that the combined forecast automatically achieves the best performance among the candidate procedures up to a constant factor and an additive penalty term. In terms of the rate of convergence, the combined forecast performs as well as if the best candidate forecasting procedure were known in advance.

Empirical studies suggest that combining procedures can sometimes improve forecasting accuracy over the original procedures. Risk bounds are derived to theoretically quantify the potential gain and price of linearly combining forecasts for improvement. The result supports the empirical finding that it is not automatically a good idea to combine forecasts. Indiscriminate combining can degrade performance dramatically as a result of the large variability in estimating the best combining weights. An automated combining method is shown in theory to achieve a balance between the potential gain and the complexity penalty (the price of combining), to take advantage (if any) of sparse combining, and to maintain the best performance (in rate) among the candidate forecasting procedures if linear or sparse combining does not help.

1. INTRODUCTION

Forecasting or prediction is of interest in many scientific fields. Depending on the nature of the process of constructing forecasts, model-based methods or more subjective opinion-based methods are used. Model-based methods include empirically based modeling, utilizing statistical tools and theory-based modeling, which depends mainly on theories from the subject field. Opinion-based methods require extensive experience in the subject field. Usually, experts' predictions are based on a difficult-to-quantify mixture of many different

This research was supported by U.S. National Security Agency Grant MDA9049910060 and U.S. National Science Foundation CAREER Grant DMS0094323. The author sincerely thanks three reviewers and Poti Giannakouros for their very valuable comments, which led to a substantial improvement of the paper. Address correspondence to: Yuhong Yang, Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011-1210, USA; e-mail: yyang@iastate.edu.

  0266-4666/04 $12.00

sources of information (qualitative and/or quantitative) and their subjective intuition or beliefs. Research in various fields has strongly suggested that the performance of prediction can be enhanced when (sometimes even in simple fashion) forecasts are combined. A survey of research results on combining forecasts is in Clemen (1989).

Although the idea of combining models/wisdom is attractive, the challenge, of course, is how to do it right. Although many approaches have been proposed for combining forecasts, with much empirical evidence of success and some theoretical justification, there is still a lack of general theoretical understanding. A difficulty is that the forecasts to be combined can be completely different in their derivations and assumptions, and therefore the commonly used methods for comparing statistical models such as model selection criteria or the traditional hypothesis testing techniques are not generally applicable. As pointed out by, e.g., Armstrong (1989) and others, the research in the area of combining forecasts so far provides few guidelines for applications.

In this paper, we propose some methods for combining point forecasts of a continuous variable and address several issues on combining forecasts. We focus on theoretical developments.

1.1. Setup of the Problem

Suppose that we are interested in forecasting or predicting a real-valued continuous random variable Y. Let $Y_1, Y_2, \ldots$ be the true values at time 1, 2,..., respectively. At each time $i \geq 1$ (possibly) related qualitative and/or quantitative information, denoted jointly by X_i (which thus could be multidimensional) is observed prior to the occurrence of Y_i. We need to predict Y_i based on X_i and the earlier data $Z^{i-1} = (X_l, Y_l)_{l=1}^{i-1}$. After Y_i is revealed, a possible discrepancy (loss) occurs between the forecast and the true value Y_i. The goal of forecasting is to make the discrepancy as small as possible.

Assume that for $i \geq 1$, the conditional distribution of Y_i given $X_i = x_i$ and $Z^{i-1} = z^{i-1}$ has (conditional) mean m_i and variance v_i. That is, $Y_i = m_i + e_i$, where e_i is the random error representing the conditional uncertainty in Y at time i. The conditional mean and variance m_i and v_i depend on x_i and z^{i-1} in general. Note that $E(e_i | X_i, Z^{i-1}) = 0$ almost surely for $i \geq 1$. Let $\hat{y}_i$ be a predicted value of Y_i. Then

$$E_i(Y_i - \hat{y}_i)^2 = E_i(m_i + e_i - \hat{y}_i)^2 = (m_i - \hat{y}_i)^2 + v_i^2,$$

where E_i denotes the conditional expectation given $Z^{i-1} = z^{i-1}$ and $X_i = x_i$. The decomposition indicates that given z^{i-1} and x_i, the expected total square error of $\hat{y}_i$ has two components: the squared difference (bias) between $\hat{y}_i$ and the unknown conditional mean m_i, and the conditional uncertainty v_i. The latter cannot be controlled and therefore is always present regardless of which method is used for prediction. The best we can hope for is that $\hat{y}_i$ equals (or is

very close to) m_i. Because v_i is the same for all forecasts, it will not be included in our measure of performance. Accordingly, we may consider the loss function

$$L(Y_i, \hat{y}_i) = (m_i - \hat{y}_i)^2.$$

Let δ be a forecasting procedure yielding forecasts $\hat{y}_1, \hat{y}_2, \ldots$ at times 1, 2, and so on. We consider the average risk in prediction up to the present time n, denoted by $R(\delta; n)$, as the performance measure of the forecasting method δ, i.e.,

$$R(\delta; n) = \frac{1}{n} \sum_{i=1}^{n} E(m_i - \hat{y}_i)^2,$$

where the expectation is taken with respect to the randomness of Z^{n-1} and X_n.

Now suppose we have a collection of forecasting procedures, denoted by $\Delta = \{\delta_1, \delta_2, \ldots\}$. The number of forecasting procedures in Δ can be large or even countably infinite. This, from a theoretical point of view, is appropriate to study because it captures the complexity when the number of candidate forecasting procedures is large relative to n. For each forecasting procedure $\delta_j \in \Delta$, at each time $i \geq 1$, the procedure gives a predicted value or forecast $\hat{y}_{j,i}$ based on z^{i-1} and x_i. The outside information x may be completely, partially, or not at all available to a procedure. In this work, we assume that the outside information is not available for the purpose of combining forecasting procedures.

1.2. Some Views on Combining Forecasts

Despite the empirical successes of combining forecasts, some controversy attended the topic early in its history (see, e.g., discussions in Newbold and Granger, 1974). One reason is the lack of confidence in the rather ad hoc, from a statistical point of view, nature of the majority of the proposed methods. Another more philosophical reason is the view that combining forecasts is fundamentally flawed as a technique. The argument is that all the information that is used by the individual forecasters should be pooled and used to construct a single best model based on which the optimal forecast can be found.

We regard the first reason as valid. Useful methods have been proposed to pursue "optimal" combining based on the idea of minimizing the variance of a linear combination of the assumed unbiased forecasts (e.g., Bates and Granger, 1969). However, results on the performance of such combining methods in terms of statistical risks are rare. Simulation and case studies have shown that complicated combining methods pursuing "optimal" behavior often lead to unstable weights and the combined forecast even performs significantly worse than the individual forecasts (see, e.g., Figlewski and Urich, 1983; Kang, 1986; Clemen and Winkler, 1986). A related question is how many forecasts should be combined. Should one just combine all the forecasts available? Theoretical

understanding of these questions will provide helpful guidelines for combining procedures in practice.

Regarding the objection to combining forecasts based on the argument that combining information and then creating a comprehensive model is preferable, we strongly disagree. As pointed out by many researchers who favor combining forecasts, the combining information approach may not be feasible if one does not have access to the information used by the individual forecasters. Furthermore, even if one can in principle obtain the information, practical considerations (e.g., cost, time constraint, and lack of expertise in processing some of the information) may well prohibit such an action. Although these are valid defenses of combining forecasts, it can be further argued that even if these difficulties with the super model approach are ignored, combining forecasts is still a legitimate and constructive way to combine information (without building a super model). In light of nonparametric statistical research, combining information to build a super model, in our view, has very limited potential for successful applications.

Let us consider a relatively simple regression case to illustrate the point. Assume that the variable being forecast, Y_n, is a function of the information set $X_n = (X_{n1}, X_{n2}, \ldots)$ plus some noise, i.e., $Y_n = f(X_n) + e_n$, where the errors are independent. The individual forecasters have access to some or all of the components of X_n. Of course, f is unknown to any forecaster. In many (if not most) realistic situations, the estimation of f is a nonparametric problem. A super model of infinite dimension is not very helpful.

Besides fully nonparametric methods, such as kernel and nearest neighbor regressions, an automated parametric approximation approach is useful. The idea is that although honestly admitting that the target f cannot be described by a model with a finite number of parameters, one hopes that it can be approximated reasonably well by a sequence of parametric models. A key here is to allow the dimensions of the approximating models to increase appropriately (based on data) so that the approximation error (bias) of the model and the estimation error (variance of the estimator of f) are well balanced to achieve the optimal performance. The fundamental difference between this approach and the traditional parametric approach is that the latter performs estimation and prediction based on the selected model, taking it as the true model. It is becoming commonly acknowledged that such parametric approaches usually lead to underestimation of uncertainty in the estimation/prediction procedure (see, e.g., Chatfield, 1995; though much earlier research, e.g., Leamer, 1978, has already pointed in this direction). When f is not finite-dimensional, the best model, which balances bias and variability, generally depends on the sample size, and its dimension increases as more data arrive. From this point of view, identifying the correct model for forecasting is problematic. The true model is never recovered, but the prediction accuracy gets better and better. Even if one is given the right model with a large number of unknown parameters, it is still better to use a much simpler model trading the bias for much less variability.

The preceding discussion indicates that model/procedure selection is a constructive technique for combining information. Model selection can be regarded as a special case of combining forecasts: the individual forecasts are based on different models, and all the weight is allocated to a single model. Other combining strategies such as convex or linear combining are also useful.

1.3. Target Classes of Combined Forecasts

Let $\Delta = \{\delta_1, \delta_2, \ldots \delta_M\}$ be a finite collection of original forecasting procedures. A forecasting procedure δ, with forecasts $\hat{y}_i$ for $i \geq 1$, is said to be a combined procedure based on Δ if $\hat{y}_i$ is a measurable function of $Y_1, \ldots, Y_{i-1}$ and $\hat{y}_{j,l}, 1 \leq l \leq i, 1 \leq j \leq M$ for $i \geq 1$. Note that here no outside information is used for combining.

DEFINITION 1. *Any collection of combined forecasting procedures is called a target class of combining.*

For a given target class Ψ, let $\delta^* = \delta^*(\Psi; n)$ be the minimizer of the forecasting risk $R(\delta; n)$ over all $\delta \in \Psi$. Let $R(\Psi; n) = R(\delta^*; n)$ denote the minimum prediction risk in the class Ψ.

Consider two examples of target class, which follow.

Example 1

$\Psi_0 = \Delta$. For this case, $\delta^*(\Psi; n)$ is simply the best individual original procedure, and $R(\Psi_0; n) = \inf_{\delta \in \Delta} R(\delta; n)$ is the ideal performance in Ψ_0.

Example 2

Ψ_L consists of (constrained) linear combinations of the procedures:

$$\hat{y}_i^{\theta} = \sum_{j=1}^{M} \theta_j \hat{y}_{j,i}, \qquad i \geq 1,$$

where $\theta = (\theta_1, \ldots, \theta_M)$ satisfies the constraint $\sum_{j=1}^{M} |\theta_j| \leq 1$. Let δ^{θ} denote the corresponding forecasting procedure and let $\Theta = \Theta_M$ denote the set of all such θ. Then $\Psi_L = \{\delta^{\theta} : \theta \in \Theta\}$. Let θ_n^* be the minimizer of $R(\delta^{\theta}; n)$ over $\theta \in \Theta$. Then $R(\Psi_L; n) = R(\delta^{\theta_n^*}; n)$ is the ideal performance in Ψ_L.

Obviously $R(\Psi_L; n) \leq R(\Psi_0; n)$. Generally $R(\Psi; n)$ is nonincreasing in Ψ; i.e., the larger Ψ is, the smaller $R(\Psi; n)$ is. However, because the best procedure in the class Ψ is unknown, one may not be able to achieve the corresponding performance. In their influential paper, Bates and Granger (1969) consider the class Ψ of convexly combined forecasts. The combining coefficients are estimated based on the previous data. This approach is certainly well motivated, but it is rather unclear how close the actual performance (with the estimated coefficients) is to the ideal one. Such an issue is important for comparing

different target classes when considering combining forecasts. As with the familiar trade-off between bias and variance, though the ideal performance of a larger target class is generally better (at least no worse), it is more difficult to search for the best procedure in a larger class; in some sense one needs to pay a higher price for searching for (or "estimating") the best procedure. Thus it is not necessarily a good idea to simply pursue the ideal performance of a large target class.

In Example 1, one may use or develop a criterion to choose a model (or a procedure) from the original list. However, model (or procedure) selections are often unstable in the sense that a slight change in the data causes the selection of a different model, which usually makes the estimator or forecast based on the selected model have unnecessarily large variance. Several lines of research follow from this problem. Strategies for stabilizing such estimators or forecasts such as bagging (bootstrap aggregating) (Breiman, 1996b) have been suggested. There is evidence suggesting that appropriate weighting of the models (or procedures) can perform very well (in related contexts, see, e.g., Buckland, Burnham, and Augustin, 1997; Yang, 2001). Promising results have also been obtained under the Bayesian framework. See a recent review article by Hoeting, Madigan, Raftery, and Volinsky (1999) for references. In the theoretical direction, in a simplified density estimation context, Yang (2003) shows that the risk based on any selection rule is at least 27% higher than that based on an appropriate combining.

Along this line of combining (weighting) instead of selecting, we will consider convexly combining the original procedures (the convex coefficients are allowed to change in time). This may look like the linear combining in the second example, but the purposes of combining are very different: in the first example, we want the convex combination to perform as well as the best individual procedure in Δ; in the second example, we target the best (constrained) linear combination of the original procedures. To make a distinction, we will call the first case *combining for adaptation* and the second *combining for improvement*.

Now for a given target class Ψ, suppose $\hat{\delta}$ is a constructed combining procedure targeting the performance of $\delta^*(\Psi;n)$. Then the difference (regret) $R(\hat{\delta};n) - R(\delta^*(\Psi;n);n)$ measures the performance of $\hat{\delta}$ relative to the ideally combined procedure in Ψ. Note that $\delta^*(\Psi;n)$ depends also on the true distribution of $Y_1, Y_2, \ldots, Y_n$. To objectively measure the performance of $\hat{\delta}$, the worst value of $R(\hat{\delta};n) - R(\delta^*(\Psi;n);n)$ over different possible conditional means and variances of $Y_1, \ldots, Y_n$ is of interest. The smallest possible worst-case regret when $\hat{\delta}$ is over all valid combined procedures captures the price (in terms of prediction accuracy) we have to pay for not knowing $\delta^*(\Psi;n)$ in Ψ. The price of linearly combining in Ψ_L will be examined in Section 3.

In addition to the preceding minimax type of measure of performance, one may also consider pointwise performance as studied in e.g., Rissanen (1986) and Barron, Rissanen, and Yu (1998). Recently, Ploberger and Phillips (1999)

have derived lower bounds on pointwise predictive performance in terms of relative likelihood for some time series parametric models, including certain nonstationary ones. Such results are useful for model comparison and evaluation of estimators.

The rest of the paper is organized as follows. We present results for combining for adaptation in Section 2 and study combining for improvement in Section 3. Concluding remarks are given in Section 4. Technical proofs of the main results are in the Appendix.

2. PROPOSED METHODS FOR COMBINING FORECASTS FOR ADAPTATION

2.1. Combining with Normality when the Conditional Variances Are Known

We first consider the case in which the conditional distribution of Y_i given $Z^{i-1} = z^{i-1}$ and $X_i = x_i$ is Gaussian for all $i \geq 1$ with the conditional variances v_i known. This of course is not realistic in most applications, but it provides a relatively simple case to begin with. To combine the forecasting procedures $\Delta = \{\delta_1, \delta_2, \ldots\}$ at each time n, we look at their past performances and assign weights accordingly as follows. Let $\underline{\pi} = \{\pi_j : j \geq 1\}$ be the prior weights on the procedures, where π_j are positive numbers summing to 1. Let $W_{j,1} = \pi_j$ and for $n \geq 2$, let

$$W_{j,n} = \frac{\pi_j \exp\left(-\frac{1}{2}\sum_{i=1}^{n-1} \frac{(Y_i - \hat{y}_{j,i})^2}{v_i}\right)}{\sum_{j'} \pi_{j'} \exp\left(-\frac{1}{2}\sum_{i=1}^{n-1} \frac{(Y_i - \hat{y}_{j',i})^2}{v_i}\right)}.$$

Note that $\sum_{j=1}^{\infty} W_{j,n} = 1$ for $n \geq 1$ and $W_{j,n}$ depends only on the past forecasts and the corresponding actual realizations of Y. The combined forecasting procedure δ^* is a convex combination of the original procedures:

$$\hat{y}_n^* = \sum_{j=1}^{\infty} W_{j,n} \hat{y}_{j,n}$$

for $n \geq 1$.

Note that in the determination of $W_{j,n}$, the previous prediction errors are weighted relative to the actual conditional variances v_i. When there is larger uncertainty in Y_i given z^{i-1} and x_i, prediction error is likely to be larger. The division by v_i accounts for the willingness to tolerate a bigger error in such a case.

Note also that

$$W_{j,n} = \frac{W_{j,n-1} \exp\left(-\dfrac{(Y_{n-1} - \hat{y}_{j,n-1})^2}{2v_{n-1}}\right)}{\sum_{j'} W_{j',n-1} \exp\left(-\dfrac{(Y_{n-1} - \hat{y}_{j',n-1})^2}{2v_{n-1}}\right)}. \tag{1}$$

Thus after each additional observation, the weights on the candidate forecasts are updated. We call the algorithm aggregated forecast through exponential reweighting (AFTER).

This weighting method, as will be seen, leads to a general adaptation risk bound. From (1), the method has a Bayesian interpretation. If we view the weights $W_{j,n-1}, j \geq 1$ as the prior probabilities on the procedures before observing Y_{n-1}, then $W_{j,n}$ is the posterior probability of δ_j after Y_{n-1} is seen.

For the choice of prior weights $\{\pi_j\}$, if the size of Δ is not large, one can naturally use uniform weights. When the size of Δ is very large, however, one needs to be more subjective, but its effect becomes weaker and weaker as n gets larger.

Related algorithms focusing on worst-case performance have been studied in machine learning (see, e.g., Vovk, 1990; Littlestone and Warmuth, 1994; Cesa-Bianchi, Freund, Haussler, Schapire, and Warmuth, 1997; Haussler, Kivinen, and Warmuth, 1998). Closely related ideas on Bayesian updating, universal coding, and prequential statistical analysis have been considered earlier and since in information theory and statistics (see, e.g., Cover, 1965; Dawid, 1984; Foster, 1991; and recent review articles on universal prediction, Merhav and Feder, 1998; and the minimum description length principle in estimation and prediction, Barron et al., 1998). Results in the machine learning literature focus on bounding the cumulative loss $\sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$ (or other loss functions) without probabilistic assumptions. Results are usually obtained under the strong condition that the observations Y_i are uniformly bounded. Methods and statistical risk bounds for combining estimation procedures in the context of density estimation, regression, and conditional probability estimation are in Yang (2000a, 2000b, 2000c, 2001) and Catoni (1999).

Condition 0. There exists a constant $\tau > 0$ such that for all $i \geq 1$, with probability one we have

$$\sup_{j \geq 1} \frac{|\hat{y}_{j,i} - m_i|}{\sqrt{v_i}} \leq \sqrt{\tau}. \tag{2}$$

This condition means that the individual forecasts are not too far away from the conditional mean relative to the conditional standard deviation. For regression-based forecasts with the regression function bounded below and above by known constants, if the error variance does not diminish, Condition 0 is

satisfied when the forecasts are restricted to the same range. Condition 0 is not directly satisfied by the usual time series models such as autoregressive integrated moving average (ARIMA) models. However, like the regression case, for a case where the conditional mean of Y is naturally bounded (e.g., between 0 and 1), restricting the forecasts accordingly can lead to the satisfaction of Condition 0. For instance, for a qualitative threshold autoregressive conditional heteroskedasticity (ARCH) model (Gouriéroux and Monfort, 1992), one may restrict the parameters properly to satisfy Condition 0. It seems quite possible to relax Condition 0 so that specific time series models are allowed without restriction, but we will not pursue that direction in this work. From a practical point of view, in implementing AFTER, we feel that one does not need to worry about this technical condition. Indeed, following the work presented in this paper, Zou and Yang (2003) empirically study the performance of AFTER for ARIMA models and show through simulations and data examples that AFTER has a clear advantage in forecasting compared with popular model selection criteria (used for choosing ARIMA orders) when the instability of model selection cannot be ignored (as is usually the case in most applications).

THEOREM 1. *Assume that Condition 0 is satisfied. When the errors are normally distributed, the mean average square risk (relative to the conditional variance) of the combined procedure satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right) \le \left(2 + \frac{9\tau}{2}\right)\inf_{j\ge 1}\left(\frac{2\log(1/\pi_j)}{n} + \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right)\right).$$

Remarks.

1. Bounds of the type

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{y}_i^*)^2 \le \frac{C\log M}{n} + \inf_{1\le j\le M}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{y}_{j,i})^2$$

for combining M procedures based on related methods are given in, e.g., Haussler et al. (1998). But the result is obtained under the very strong assumption that Y_i are uniformly bounded. Such a requirement is usually not appropriate for modeling uncertainty in Y.

2. An upper bound of the type

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{y}_i^*)^2 \le C\inf_{1\le j\le M}\left(\frac{\log(M)}{n} + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{y}_{j,i})^2\right)$$

with a constant $C > 1$, when taking expectation, is much weaker than the bound given in the theorem.

3. Condition 0 implies that each of the candidate forecasting procedures has scaled mean average square risk upper bounded by τ.
4. Though the risk bound is given under Condition 0, we expect AFTER to perform well more generally. For a good forecasting procedure, (2) is most likely satisfied with high probability, and for a bad procedure, as seen from (1), the procedure's weight drops exponentially fast (if a forecasting procedure is really bad, it will soon be terminated).

From Theorem 1, up to a constant factor and an additive penalty $\log(1/\pi_j)/n$, the combined procedure achieves the performance of the optimal forecasting procedure among the candidates. In particular, if there are M forecasting procedures then with the uniform prior weight $\pi_j = 1/M$, we have

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right) \le \left(2 + \frac{9\tau}{2}\right)\left(\frac{2\log M}{n} + \inf_{1\le j\le M}\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right)\right).$$

For the nonparametric prediction case, $(1/n)\sum_{i=1}^{n} E((m_i - \hat{y}_{j,i})^2/v_i)$ converges more slowly than order $1/n$, and therefore the additive penalty is asymptotically negligible. If additionally when the sample size n increases, the individual forecasts get closer and closer to the conditional means in the sense that

$$\frac{\sup_{1\le j\le M}|\hat{y}_{j,i} - m_i|}{\sqrt{v_i}} \lesssim c_i \quad \text{a.s.} \tag{3}$$

for some positive constants $c_i \to 0$ as $i \to \infty$, then one can show

$$\limsup_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right)}{\inf_{1\le j\le M}\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right)} \le 2.$$

Thus the performance of the combined forecast is eventually at least half as good as the best individual one (which is unknown). (Note that the condition in (3) is put on all the original procedures and is very restrictive. It is used mainly for technical tractability, and we feel that the upper bound requirement is not really needed for the bad procedures.) The factor of 2 comes from relating the Kullback–Leibler divergence and the Hellinger distance in the technical proof of the theorem.

In Theorem 1, the risks of the forecasts are measured relative to the variability of Y. Under the following conditions (which unfortunately rule out some interesting models such as generalized autoregressive conditional hetero-

skedasticity (GARCH)), Theorem 1 implies good performance of the combined forecast under the ordinary square loss.

Condition 1. The conditional variances v_n are uniformly upper bounded away from ∞ and lower bounded away from zero in that there exist positive constants A_1 and A_2 such that $A_1 \le v_n \le A_2$ a.s. for all $n \ge 1$.

Condition 2. $\sup_{j\ge 1}|\hat{y}_{j,i} - m_i|$ is uniformly bounded above by A_3 with probability one for some constant $A_3 > 0$.

COROLLARY 1. *Assume that Conditions 1 and 2 are satisfied. Under Gaussian errors, the mean average square risk of the combined procedure satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E(m_i - \hat{y}_i^*)^2 \le C \inf_{j\ge 1}\left(\frac{\log(1/\pi_j)}{n} + \frac{1}{n}\sum_{i=1}^{n} E(m_i - \hat{y}_{j,i})^2\right),$$

where C is a constant depending only on A_1, A_2, and A_3.

Remark. One does not need to know the constants A_1, A_2, or A_3 to use the AFTER algorithm to combine forecasts.

2.2. Combining with Normality when the Variances Are Estimated

For Theorem 1, the conditional variances v_i are assumed to be known. Similar results can be obtained when the variances are estimated by the individual forecasters. Assume that for each forecasting procedure δ_j, prior to forecasting Y_n, an estimate of v_n, say, $\hat{v}_{j,n}$, is obtained based on z^{n-1} and x_n. For stationary observations a number of variance estimation methods have been proposed for different scenarios. Note that the procedures do not have to use different variance estimators and if some procedures do not provide variance estimates, we can borrow from others.

Condition 3. Assume the variance estimators $\hat{v}_{j,n}$ are not too far away from the true value: there exist constant $0 < \xi_1 \le 1 \le \xi_2 < \infty$ such that

$$\xi_1 \le \frac{\hat{v}_{j,i}}{v_i} \le \xi_2$$

with probability one for all $j \ge 1$ and $i \ge 1$.

We modify the AFTER algorithm using the variance estimates. Let $W_{j,1} = \pi_j$ as before and for $n \ge 2$, let

$$W_{j,n} = \frac{\dfrac{\pi_j}{\prod_{i=1}^{n-1}\hat{v}_{j,i}^{1/2}} \exp\left(-\dfrac{1}{2}\displaystyle\sum_{i=1}^{n-1}\frac{(Y_i - \hat{y}_{j,i})^2}{\hat{v}_{j,i}}\right)}{\displaystyle\sum_{j'\ge 1}\frac{\pi_{j'}}{\prod_{i=1}^{n-1}\hat{v}_{j',i}^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n-1}\frac{(Y_i - \hat{y}_{j',i})^2}{\hat{v}_{j',i}}\right)}. \tag{4}$$

Combine the forecasts by $\hat{y}_n^* = \sum_{j=1}^{\infty} W_{j,n}\hat{y}_{j,n}$.

THEOREM 2. *Assume that the errors are Gaussian and that Conditions 0 and 3 are satisfied. Then the risk of the combined procedure satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right) \le \left(1 + \xi_2 + \frac{9\tau}{2}\right)\inf_{j\ge 1}\left(\frac{2\log(1/\pi_j)}{n} + \frac{1}{n\xi_1}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right) + \frac{C(\xi_1,\xi_2)}{n}\sum_{i=1}^{n} E\left(\frac{(\hat{v}_{j,i} - v_i)^2}{v_i^2}\right)\right),$$

where $C(\xi_1,\xi_2) = (1/\xi_2 - 1 + \log\xi_2)/\xi_1^2(1/\xi_2 - 1)^2$.

Regarding the constant $C(\xi_1,\xi_2)$, e.g., when $\xi_1 = 1/\xi_2 = \frac{1}{2}$, $C(\xi_1,\xi_2) \approx 3.1$. Note that risks of the variance estimators also appear in the preceding performance bound and thus for this combining method variance estimation is also important. However, we do not need good variance estimation for all the procedures as long as there is an accurate one for a good procedure δ_j. The weighting method and the risk bound suggest that procedures with bad forecasts and/or bad variance estimation receive rather small weights and accordingly have little effect on the performance of good procedures. To have good estimators of v_i, assumptions on the relationship between observations are necessary. For example, one may consider stationary errors that are independent or that have short-range or long-range dependence. Performance of variance estimators can be theoretically justified under reasonable conditions. For various situations, the part of risk in the upper bound due to variance estimation in Theorem 2 may be dominated by the risk of prediction. For example, for prediction based on nonparametric regression with independent or short-range dependent errors, under mild smoothness conditions, the quantity $1/n\sum_{i=1}^{n} E((m_i - \hat{y}_{j,i})^2/v_i)$ typically converges at a rate $n^{-\gamma}$ for some $0 < \gamma < 1$ but $1/n\sum_{i=1}^{n} E((\hat{v}_{j,i} - v_i)^2/v_i^2)$ typically converges at the parametric rate $1/n$. In such a case, if the number of individual forecasting procedures in Δ is finite, and if in addition the individual variance estimators converge in the sense that $1/\xi_n \le \hat{v}_{j,n}/v_n \le \xi_n$ holds almost surely with $\xi_n \to 1$, then we again have

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right)}{\inf_{j\ge 1}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right)} \le 2.$$

If no variance estimator is available from the individual forecasters, one can estimate variances based on the previous realizations of Y_i and the corresponding forecasts as follows.

1. *Individual variance estimation for the forecasting procedures.* Fix a forecasting procedure δ_j. For $i = 1$, let $\hat{v}_{j,i}$ be any initial guess. For $i > 1$, let

$$\hat{v}_{j,i} = \frac{1}{i-1} \sum_{l=1}^{i-1} (Y_l - \hat{y}_{j,l})^2. \tag{5}$$

   Under some conditions (see Proposition 3 in the Appendix), one can show that $\hat{v}_{j,i} - v_i$ converges to zero in probability. Rates of convergence could also be worked out under additional conditions.
2. *Estimating variance based on the combined forecasts.* The combined forecasting procedure can also be used to estimate variance. For $i = 1$, let $\hat{v}_i$ be any guess. Then use this value as the variance estimate for all the forecasting procedures and accordingly assign weights by (4). For $i > 1$, let

$$\hat{v}_i = \frac{1}{i-1} \sum_{l=1}^{i-1} (Y_l - \hat{y}_l^*)^2. \tag{6}$$

Other estimators (e.g., based on exponential smoothing) may be more efficient under additional assumptions.

2.3. Combining under Non-Gaussian Errors

When there is a clear indication that the normality assumption on the conditional distributions of Y_n does not hold, the combining weights should be modified. Let $h(t)$, $t \in R$ be a probability density function with mean zero and variance 1, which is thought to better capture the error distribution. Non-Gaussian choices of h allow different degrees of heavy-tailedness in the error distributions.

Condition 4. For each pair $0 < s_0 < 1$ and $T > 0$, there exists a constant $B = B_{s_0,T}$ (depending on s_0 and T) such that

$$\int h(y) \log \frac{h(y)}{\frac{1}{s} h\left(\frac{y-t}{s}\right)} dy \le B((1-s)^2 + t^2)$$

for all $s_0 \le s \le s_0^{-1}$ and $-T \le t \le T$.

Note that $\int h(y) \log(h(y))/(1/s)h((y-t)/s)dy$ is the Kullback–Leibler divergence between density h and a rescaled and relocated density $(1/s) \times h((y-t)/s)$. Under mild conditions, this divergence should behave locally like

a squared distance. The condition is satisfied by Gaussian, double-exponential, t (with degrees of freedom bigger than 2), and many other distributions.

We assume that variance estimators are available for the forecasting procedures. Note that the variance estimators given in (5) and (6) in the previous section do not require the normality assumption on the errors; thus they can be used here also. Let $\hat{s}_{j,i} = \sqrt{\hat{v}_{j,i}}$ for $j \geq 1$ and $i \geq 1$. The earlier AFTER algorithm is now modified to compute the weight based on the density function h and the variance estimates. Let $W_{j,1} = \pi_j$ and for $n \geq 2$, let

$$W_{j,n} = \frac{\pi_j \Pi_{i=1}^{n-1} \left(h\left(\frac{Y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}} \right) / \hat{s}_{j,i} \right)}{\sum_{j' \geq 1} \pi_{j'} \Pi_{i=1}^{n-1} \left(h\left(\frac{Y_i - \hat{y}_{j',i}}{\hat{s}_{j',i}} \right) / \hat{s}_{j',i} \right)}. \tag{7}$$

Combine the forecasts by $\hat{y}_n^* = \sum_{j=1}^{\infty} W_{j,n} \hat{y}_{j,n}$.

THEOREM 3. *Assume that Conditions 0, 3, and 4 are satisfied. Suppose that h used in (7) indeed characterizes the distributions of the errors in the sense that for $i \geq 1$, Y_i has the conditional density $h((y - m_i)/s_i)/s_i$ given $Z^{i-1} = z^{i-1}$ and $X_i = x_i$, where $s_i = \sqrt{v_i}$ is the conditional standard deviation of Y_i. Then the risk of the combined procedure satisfies*

$$\frac{1}{n} \sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i} \right)$$

$$\leq (2(1 + \xi_2) + 9\tau) \inf_{j \geq 1} \left(\frac{\log(1/\pi_j)}{n} + \frac{B_{s_0,T}}{n} \sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i} \right) + \frac{B_{s_0,T}}{n} \sum_{i=1}^{n} E\left(\frac{(\sqrt{\hat{v}_{j,i}} - \sqrt{v_i})^2}{v_i} \right) \right),$$

where $s_0 = \min(\sqrt{\xi_1}, 1/\sqrt{\xi_2})$ and $T = \sqrt{\tau}$.

2.4. Combining under Density Forecasts

Suppose that at each time i, a forecaster δ_j comes up with an estimate, say, $\hat{p}_{j,i}$, of p_i, the conditional density of Y_i given $Z^{i-1} = z^{i-1}$ and $X_i = x_i$. The estimate reflects the forecaster's assessment of the uncertainty in Y_i prior to its occurrence. Let $\hat{y}_{j,i}$ be the mean of the distribution of $\hat{p}_{j,i}$. Various methods have been suggested to combine probability distribution estimates (for a review on this topic, see, e.g., Genest and Zidek, 1986). Here our focus is on combining the density estimates to form a composite forecast of Y.

Let $W_{j,1} = \pi_j$, for $n \geq 2$, let

$$W_{j,n} = \frac{\pi_j \prod_{i=1}^{n-1} \hat{p}_{j,i}(Y_i)}{\sum_{j' \geq 1} \pi_{j'} \prod_{i=1}^{n-1} \hat{p}_{j',i}(Y_i)}, \tag{8}$$

and combine the forecasts by $\hat{y}_n^* = \sum_{j=1}^{\infty} W_{j,n} \hat{y}_{j,n}$.

Some technical conditions (to be given subsequently) are needed for part of the next result when $\hat{p}_{j,i}$ is of the form $g_{j,i}((y_i - \hat{y}_{j,i})/\hat{s}_{j,i})/\hat{s}_{j,i}$, where $g_{j,i}$ is a (nonrandom) density function with mean zero and variance 1. One hopes that for at least one forecaster δ_j, $g_{j,i}$ gets closer and closer (in some sense) to p_i as $i \to \infty$.

Let $F = \{g_\theta : \theta \in \Theta\}$ be a class of density functions on the real line. Let f be a given density (f may or may not be in F). Let $D(f\|g)$ denote the Kullback–Leibler divergence between two densities f and g. We say that f has a projection in F if there exists $\theta^* \in \Theta$ such that for all $\theta \in \Theta$, $D(f\|g_\theta) = D(f\|g_{\theta^*}) + D(g_{\theta^*}\|g_\theta)$. The density g_{θ^*} is called the information projection of f in F. When F is an exponential family with $g_\theta(x) = \exp(\sum_{i=1}^{k} \theta_i \varphi_i(x) - \psi(\theta))$, where φ_i are linearly independent basis functions and $\psi(\theta)$ is the normalizing constant, the information projection g_{θ^*} of f exists and is characterized by $\int \varphi_i(x) f(x)\,dx = \int \varphi_i(x) g_{\theta^*}(x)\,dx$ for $1 \leq i \leq k$ (assuming that the integrals are finite) (see Csiszár, 1975).

Now let $F_{j,i} = \{g_{j,i}^{\mu,\sigma} = (1/\sigma) g_{j,i}((x - \mu)/\sigma) : -\infty < \mu < \infty, \sigma > 0\}$ be the location-scale family generated by $g_{j,i}$ for $j \geq 1$ and $i \geq 1$. We need the following regularity condition for at least one forecasting procedure δ_j.

Condition 5. Suppose that at time i, the true conditional distribution of Y_i has density $p_i = h_i((y_i - m_i)/s_i)/s_i$, where h_i is a probability density function with mean 0 and variance 1 and is assumed to have an information projection in $F_{j,i}$. Assume also that the densities $g_{j,i}$, $i \geq 1$ satisfy Condition 4 uniformly in the sense that the constants $B_{s_0,T}$ in the condition do not depend on i. Assume further that there exists a constant $0 < c_0 < 1$ such that $D(g_{j,i}^{\mu,\sigma}\|g_{j,i}) \geq c_0 \min(\mu^2 + (\sigma - 1)^2, 1)$ for all $i \geq 1$.

As mentioned earlier, the projection requirement in Condition 5 is satisfied if $F_{j,i}$ are exponential families. The lower bound condition on the Kullback–Leibler divergence within each family is mild and typically satisfied.

THEOREM 4. *Assume that Conditions 0 and 3 are satisfied.*

1. *In general, the risk of the combined procedure satisfies*

$$\frac{1}{n} \sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right) \leq (2(1 + \xi_2) + 9\tau) \inf_{j \geq 1}\left(\frac{\log(1/\pi_j)}{n} + \frac{1}{n} \sum_{i=1}^{n} ED(p_i\|\hat{p}_{j,i})\right).$$

2. *For any $j \geq 1$, if δ_j satisfies Condition 5 and $D(h_i \| g_{j,i}) \to 0$ as $i \to \infty$, then the risk of the combined procedure satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right)$$

$$\leq C \inf_{j \geq 1}\left(\frac{\log(1/\pi_j)}{n} + \frac{1}{n}\sum_{i=1}^{n} D(h_i \| g_{j,i}) + \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{v_i}\right)\right.$$

$$\left. + \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(\sqrt{v_i} - \sqrt{\hat{v}_{j,i}})^2}{v_i}\right)\right),$$

where C is a constant depending on $\sup_{i \geq 1} D(h_i \| g_{j,i})$, τ, ξ_1, ξ_2, c_0, and B.

Remarks.

1. Note that if a procedure δ_j eventually correctly specifies the distribution h_i, then the projection requirement in Condition 5 is automatically satisfied for that procedure.
2. One could also use the weights in (8) to combine the density estimates. Yang (2000a, Sec. 5) gives a corresponding risk bound in terms of Kullback–Leibler divergence without the need of Conditions 0, 3, and 5.

Note that because the forms of the conditional densities h_i are unknown, the combining method pays an additional price of order $(1/n)\sum_{i=1}^{n} D(h_i \| g_{j,i})$. If for the forecaster δ_j, $g_{j,i}$ gets closer and closer to h_i in the Kullback–Leibler divergence, then this penalty approaches zero. It may or may not affect the rate of convergence of the combined procedure, depending on how close $g_{j,i}$ and h_i are to each other.

In particular, from the result, if there is at least one forecasting procedure that eventually picks h_i with good estimates of the conditional means and standard deviations, the combined procedure based on the AFTER algorithm behaves well. Thus different distribution shapes and assumptions on the dependence relationship between the errors e_i and also on the relationship between Y and the outside information X can be tried out to increase the chance of capturing the true uncertainty in Y. The theorem shows that if one such choice approximates the reality well, the combined forecast will automatically perform well.

2.5. A Consistency Property of the Combining Methods

Recall that the prediction error has a decomposition $E_i(Y_i - \hat{y}_i)^2 = (\hat{y}_i - m_i)^2 + v_i$. Equivalently, we have

$$E_i \frac{(Y_i - \hat{y}_i)^2}{v_i} = \frac{(\hat{y}_i - m_i)^2}{v_i} + 1.$$

This quantity measures prediction accuracy conditioning on the information available at the time. Note that it is always lower bounded by 1 and should be close to 1 for a good forecast. This motivates the following definition.

DEFINITION 2. *A forecasting procedure δ with forecasts $\hat{y}_i$ at time $i = 1,2,\ldots$ is said to be consistent if*

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(Y_i - \hat{y}_i)^2}{v_i}\right) \to 1 \quad \text{as } n \to \infty.$$

Note that from the decomposition mentioned previously, δ is consistent if and only if

$$\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{(\hat{y}_i - m_i)^2}{v_i}\right) \to 0 \quad \text{as } n \to \infty,$$

which roughly means that the forecasts approach the conditional mean m_i (relative to the conditional standard deviation $\sqrt{v_i}$). This is a desirable property for a forecasting procedure.

PROPOSITION 1. *Assume that the conditions for Theorems 1–3 are satisfied. Suppose that there is at least one consistent forecasting procedure in the original list. For the combining methods used in Theorems 2 and 3, we further assume that for at least one consistent forecasting procedure δ_{j^*}, the variance estimator becomes more and more accurate in the sense that $(1/n)\sum_{i=1}^{n} E((v_i - \hat{v}_{j^*,i})^2/v_i^2) \to 0$ as $n \to \infty$. Then the combined forecasting procedure is consistent.*

The proposition basically indicates that if any original procedure works well, so does the combined procedure in the consistency sense, even if nonconsistent procedures are present. Note that the key characteristic here is that the combining method does not need to know which original procedure is consistent. The conclusion of the proposition follows easily from the results of Theorems 1–3. A similar result can be derived for the combining method in Section 2.4.

2.6. Combining Forecasts without Variance Estimation

The methods in the previous sections require knowledge of the variances or appropriate estimates for determining weights for the candidate forecasts. Sometimes, e.g., when the data are nonstationary, it may be difficult to estimate the conditional variance of Y_n reliably. We present a method in this section that does not require variance estimation. A key idea is borrowed from Catoni (1999).

Condition 6. Assume that the forecasts satisfy that $\sup_{j\geq 1, i\geq 1}|\hat{y}_{j,i} - m_i|$ is bounded above by a positive constant $A > 0$ with probability one.

Let ψ be a fixed nonnegative convex function with $\psi(0) = 0$. Here ψ is not necessarily symmetric about 0. When $\psi(Y_i - \hat{y}_i)$ is used as a loss function, asymmetry of ψ is of interest to reflect the reality that over- and underforecasting might have very different consequences (cf. Newbold and Granger, 1974). Because ψ is convex, for any a_0, there exists a line $y = \theta_{a_0}(a - a_0) + \psi(a_0)$ (called a supporting line) such that ψ is above the line. We further assume subsequently that ψ is quadratically off the line and ψ does not change too dramatically in some sense. Let ψ'_+ and ψ'_- denote the right and left derivatives of ψ, respectively.

Condition 7. There exist constants $\underline{c} > 0$ and $\bar{c} > 0$ and $\beta > 0$ such that for each $a_0 \in R$ and a supporting line $y = \theta_{a_0}(a - a_0) + \psi(a_0)$ at any a_0, we have

$$\psi(a) - (\theta_{a_0}(a - a_0) + \psi(a_0)) \geq \underline{c}(a - a_0)^2 \tag{9}$$

and

$$\max_{-T \leq a \leq T} |\psi'_+(a)| \leq \bar{c}(1 + T)^\beta, \qquad \max_{-T \leq a \leq T} |\psi'_-(a)| \leq \bar{c}(1 + T)^\beta \tag{10}$$

for each $T > 0$.

Condition 7 essentially restricts ψ to be basically quadratic. The slight generality is used so as to include asymmetric function

$$\psi(a) = \begin{cases} c_1 a^2 & \text{when } a > 0 \\ c_2 a^2 & \text{when } a \leq 0 \end{cases}$$

for some positive constants $c_1 \neq c_2$. Such a loss $\psi(Y_i - \hat{y}_i)$ penalizes prediction error differently when under- or overpredicting Y.

We also need a condition on the conditional distributions of the errors. We do not require knowledge of its shape as long as certain moment generating like functions are uniformly upper bounded.

Condition 8. There exist a constant $t_0 > 0$ and continuous functions $0 < M_1(t), M_2(t) < \infty$ on $(-t_0, t_0)$ such that for all $n \geq 1$ and $-t_0 \leq t \leq t_0$, with probability one,

$$E_n|e_n|^{2\beta} \exp(t|e_n|^\beta) \leq M_1(t), \qquad E_n \exp(t|e_n|^\beta) \leq M_2(t),$$

where $\beta > 0$ is the same constant as in Condition 7.

Let $W_{j,1} = \pi_j$ and for $n \geq 2$, let

$$W_{j,n} = \frac{\pi_j \exp\left(-\lambda \sum_{i=1}^{n-1} \psi(Y_i - \hat{y}_{j,i})\right)}{\sum_{j' \geq 1} \pi_{j'} \exp\left(-\lambda \sum_{i=1}^{n-1} \psi(Y_i - \hat{y}_{j',i})\right)},$$

where λ is a positive constant to be chosen later. Then the combined forecasting procedure δ^* is defined by $\hat{y}_n^* = \sum_{j=1}^{\infty} W_{j,n}\hat{y}_{j,n}$.

THEOREM 5. *Assume that Conditions 6–8 are satisfied. Then when λ is chosen small enough, say, $0 < \lambda \le \lambda_0$ for some constant λ_0 depending on $\underline{c}$, $\bar{c}$, A, M_1, and M_2, the mean average risk (under the loss ψ) of the combined procedure satisfies*

$$\frac{1}{n}\sum_{i=1}^{n} E\psi(Y_i - \hat{y}_i^*) \le \inf_j \left(\frac{\log(1/\pi_j)}{\lambda n} + \frac{1}{n}\sum_{i=1}^{n} E\psi(Y_i - \hat{y}_{j,i}) \right).$$

In particular, when ψ is the square loss, we have

$$\frac{1}{n}\sum_{i=1}^{n} E(m_i - \hat{y}_i^*)^2 \le \inf_j \left(\frac{\log(1/\pi_j)}{\lambda n} + \frac{1}{n}\sum_{i=1}^{n} E(m_i - \hat{y}_{j,i})^2 \right).$$

Note that unlike the previous results, no variance estimation is required here nor is the form of the conditional distribution of Y specified. In addition, the loss (of a constant factor) beyond the additive penalty term is avoided, but under the strong moment generating function conditions. A difficulty with this method is the choice of λ, which can have a dramatic effect on the weights for moderate n. In general, the quantities A, M_1, and M_2 are unknown, and therefore it is hard to determine λ_0. If one is willing to assume a particular form of distribution (e.g., normal), the quantities can be appropriately bounded, and then λ_0 can be determined (for the regression case with independent errors, cf. Catoni, 1999). In contrast, the earlier algorithms do not require such a tuning parameter. Some simulation results in the context of nonparametric regression with independent errors in Yang (2001) suggest that when the true error distribution is double-exponential but weight calculations are based on normal error, estimation accuracy is significantly decreased. On the other hand, when both types of error distributions are considered, the combined procedure behaves as if it knew which error distribution is the correct one. Thus we tend to recommend the use of the earlier methods when one has reasonably strong confidence that some members of the chosen error distribution set closely reflect the true error distribution.

2.7. Potential Applications

2.7.1. Forecasting Based on Regression Here we use an example to demonstrate how AFTER can be used to obtain forecasts that are adaptive to multiple scenarios.

Suppose that the response variable Y_i and the outside information $X_i = (X_{i1}, \ldots, X_{id})$ have the regression relationship

$$Y_i = m(X_i) + e_i,$$

where m is the unknown regression function assumed to be uniformly bounded and where e_i are Gaussian errors with mean zero and finite variances. Suppose

that the explanatory variable X takes values in a bounded cube, which is then taken to be $[0,1]^d$ without loss of generality. In addition, we assume that the errors and X_i are independent of each other and that the unknown design density of X with respect to Lebesgue measure is bounded above and away from zero on $[0,1]^d$. In practical situations, especially when d is large, it is usually difficult to know the form of m or how the errors are related to each other. Accordingly one may consider different scenarios in the hope that some of them capture the characteristics of the data and result in good forecasts. We consider several models/procedures here. In the first five cases, the errors are assumed to be independent. For such a scenario, as is well-known, prediction of the next response is basically equivalent to estimation of m.

1. *Linear regression.* Here m is assumed to be of the form $m(x_1,\ldots,x_d) = \beta_0 + \sum_{i=1}^d \beta_i x_i$. One can use the least squares method to estimate the parameters. The squared L_2 risk for estimating m is of the parametric rate $1/n$ if the model holds. One may also consider other plausible parametric models (e.g., including quadratic or higher order and/or cross product terms of the variables).

2. *Wavelet regression.* For $1 \le \sigma, q \le \infty$ and $\alpha > 0$, let $B_{q,\sigma}^{\alpha,d}(C)$ be the collections of all functions $g \in L_q[0,1]^d$ such that the Besov norm satisfies $\|g\|_{B_{q,\sigma}^{\alpha,d}} \le C$ (see, e.g., Triebel, 1975; DeVore and Lorentz, 1993). Besov (and similarly defined Triebel's F) classes contain a rich collection of function classes, including Sobolev classes, fractional Sobolev classes, and Hardy classes. The richness of these classes provides a great deal of flexibility (e.g., spatial inhomogeneity) for statistical function estimation. The minimax rate of convergence for estimating a function $m \in B_{q,\sigma}^{\alpha,d}(C)$ under the squared L_2 loss is $n^{-2\alpha/(2\alpha+d)}$ (see, e.g., Donoho and Johnstone, 1998; Yang and Barron, 1999). For $d = 1$, Donoho and Johnstone (1998) show that wavelet thresholding techniques naturally lead to minimax-rate optimal estimators. Furthermore, the estimators are adaptive in the sense that the hyperparameters σ, q, and α need not be known in advance. For $d > 1$, one can consider tensor-product wavelet bases and expect similar minimax optimal results.

3. *Tensor-product wavelets with different interaction order.* Suppose d is large. When the smoothness parameter α is small or moderate, the convergence rate $n^{-2\alpha/(2\alpha+d)}$ for the Besov class $B_{q,\sigma}^{\alpha,d}(C)$ is slow: the phenomenon known as the "curse of dimensionality." To overcome the "curse," one may consider (somewhat enlarged) Besov classes of different interaction orders as follows:

$$S_{q,\sigma}^{\alpha,1}(C) = \left\{ \sum_{i=1}^d g_i(x_i) : g_i \in B_{q,\sigma}^{\alpha,1}(C),\ 1 \le i \le d \right\},$$

$$S_{q,\sigma}^{\alpha,2}(C) = \left\{ \sum_{1\le i<j\le d} g_{i,j}(x_i,x_j) : g_{i,j} \in B_{q,\sigma}^{\alpha,2}(C),\ 1 \le i < j \le d \right\},$$

$$\vdots$$

$$S_{q,\sigma}^{\alpha,d}(C) = B_{q,\sigma}^{\alpha,d}(C).$$

Note that the simplest function class $S_{q,\sigma}^{\alpha,1}(C)$ contains additive functions (no interaction). These classes have different effective input dimensions (between 1 and d). By results of Yang and Barron (1999), the minimax rate of convergence under squared L_2 loss for estimating m in $S_{q,\sigma}^{\alpha,r}(C)$ is $n^{-2\alpha/(2\alpha+r)}$ for $1 \le r \le d$, as suggested by the heuristic dimensionality reduction principle of Stone (1985). Note that when r is small relative to d, the convergence rate is much improved compared with $n^{-2\alpha/(2\alpha+d)}$. Corresponding to these classes of different interaction orders, one can consider tensor-product wavelets of different interaction orders. With an appropriate thresholding, one expects to obtain an estimator that converges optimally for every $S_{q,\sigma}^{\alpha,r}(C)$ without knowing the hyperparameters σ, q, α, and r.

4. *Neural nets.* Let $N(C)$ be the closure in $L_2[0,1]^d$ of the set of all functions $g: R^d \to R$ of the form $g(x) = c_0 + \sum_i c_i \sigma(v_i' x + b_i)$ (where the prime denotes the transpose), with $|c_0| + \sum_i |c_i| \le C$, and $\|v_i\| = 1$, where σ is the step function $\sigma(t) = 1$ for $t \ge 0$, and $\sigma(t) = 0$ for $t < 0$. The minimax rate for estimating $m \in N(C)$ under the squared L_2 loss is shown to be bounded between

$$n^{-(1+2/d)/(2+1/d)}(\log n)^{-(1+1/d)(1+2/d)/(2+1/d)} \quad \text{and} \quad (n/\log n)^{-(1+1/d)/(2+1/d)} \tag{11}$$

(see Yang and Barron, 1999). When d is large, the rate is slightly better than $n^{-1/2}$ (independent of d), which avoids the "curse of dimensionality." Estimators at rate $O(\log n/n^{1/2})$ using finite-dimensional neural network models are in, e.g., Barron (1994).

5. *Kernel regression.* With kernel methods, universally consistent estimators have been derived under L_q loss without any assumption on the joint distribution of (X,Y) other than the necessary existence of the corresponding moment of Y (see, e.g., Stone, 1977; Devroye and Wagner, 1980). Thus we have a consistent kernel estimator under the squared L_2 loss.

Now suppose that one suspects that the random errors may be correlated. For simplicity, assume that the errors are stationary and normally distributed with $\mathrm{Cov}(e_i, e_{i+j})$ of order $|j|^{-\gamma}$ for some $\gamma > 0$. When $0 < \gamma \le 1$, the errors are said to be long-range dependent.

6. *Prediction under long-range dependence.* Suppose that the errors are long-range dependent. Then estimation of the regression function can be severely hampered (see Hall and Hart, 1990; Wang, 1996; Johnstone and Silverman, 1997; Efromovich, 1999; Yang, 1997). However, as long as the dependence parameter is known, prediction is not harder at least in terms of rate of convergence (Yang, 1997). When the dependence is unknown, the prediction problem becomes more complicated. At this point, we are not aware of general results in that direction. We here assume that γ is known and the regression function m has the neural net representation as in case 4. Then one can construct a forecasting procedure converging at rate $O(n^{-1/2})$.

Now we combine the forecasts based on the preceding six cases using the method in Section 2.2. If the errors are independent, then when the true regres-

sion function is linear in the variables $x_i, 1 \le i \le d$, the prediction risk of the combined procedure converges at rate $1/n$; when m is in $B_{q,\sigma}^{\alpha,d}(C)$ with α relatively large compared to d, the risk converges at a good rate $n^{-2\alpha/(2\alpha+d)}$; when m is in $S_{q,\sigma}^{\alpha,r}(C)$ for some small r, then the risk converges at rate $n^{-2\alpha/(2\alpha+r)}$; when m is not in any of these cases, but has the neural net representation, then the risk also converges at a good rate $O(\log n/n^{1/2})$. Even if m is not smooth at all, the combined forecasting procedure is still consistent. If the errors are long-range dependent with the dependence parameter correctly specified in case 6, and if m has the neural net representation, the risk converges at rate $O(n^{-1/2})$.

The key point here is that by combining the different forecasts, the mixed forecasting procedure automatically adapts to different scenarios for a good rate of convergence.

2.7.2. Single series forecasting For this case, one is interested in predicting a variable based on its past values. Here no outside information is used to help prediction.

Let $Y_1, Y_2, \ldots$ be a univariate time series. Different tools, including Box–Jenkins methods using ARIMA models (Box and Jenkins, 1976), exponential smoothing (e.g., Holt, 1957; Winters, 1960), and models based on, e.g., econometric theories have been proposed and widely used in practice. Note that forecasting based on ARIMA modeling typically requires selection of appropriate autoregressive, moving average, and differencing orders. As pointed out in the introduction to this paper, in the regression (and classification) context(s), it is now known that selection often causes unnecessarily large variability and does not perform as well as combining the models appropriately. It seems reasonable to expect a similar phenomenon in the time series forecasting context. Thus the AFTER algorithms could be used to combine plausible ARIMA models of different orders. Another way of using the algorithms is to combine, e.g., a forecast based on the ARIMA technique and a forecast based on exponential smoothing.

Many authors on combining forecasts have commented that one might expect a greater improvement when the procedures are more distinct in nature (see, e.g., Newbold and Granger, 1974). It may also be worthwhile to combine procedures/models of similar nature to avoid a large variability due to selection. See Zou and Yang (2003) for empirical results in that direction.

3. COMBINING FORECASTS FOR A BETTER PERFORMANCE

The results given so far deal with combining forecasts for adaptation; i.e., the purpose of combining the individual forecasts is to achieve the best overall performance over time among all the competing forecasts. In applications, the more aggressive goal of combining procedures for improving the individual forecasts is certainly appealing. See Clemen, Murphy, and Winkler (1995) for a discussion of the difference between the two directions in combining forecasts.

In this section, we consider linear combinations of forecasts with a certain constraint. The treatment is mainly theoretical because the combining methods we study are hard to implement in applications. Our purpose is to address several important issues on combining forecasts for a better performance. The theoretical understanding can be very helpful when evaluating practically feasible combining methods. Related theoretical results on linearly combining nonparametric regression procedures with independent errors are in Juditsky and Nemirovski (2000) and Yang (in press). Some practical combining methods for regression have been proposed by Wolpert (1992), Breiman (1996a), and LeBlanc and Tibshirani (1996).

Let $\Delta = \{\delta_1, \delta_2, \ldots\}$ be the set of original forecasting procedures. One hopes that a certain linear combination of them will significantly outperform any of the individual procedures. There are two related issues that we want to address:

1. stability versus pursuing further improvement;
2. sparsity in combining.

In addition to higher computational cost, there must also be a price in terms of statistical performance for searching for the best linearly combined procedure. The price will be quantified in a minimax framework in terms of the number of procedures being combined. As intuition suggests, the more procedures involved in combining, the bigger the price one needs to pay as a result of the increased variability of the combined forecast. The first issue then amounts to the comparison of the potential gain and the price of combining more forecasting procedures. If the potential gain is larger than the price, then there is an advantage in combining more procedures (ignoring the computational cost). Of course, in applications, one does not know the optimal trade-off between the potential gain and the price. Thus it would be nice to have a combined procedure that automatically behaves the right way: that would be both conservative (which targets the best performance based on a small number of procedures to avoid a large variability) and aggressive (which targets the best performance based on a large number of procedures to increase the potential gain), whichever is better. Sparseness in combining is also an important issue. If many coefficients in a linear combination of the procedures are rather small, keeping the important ones only can lead to a much better performance because of the much reduced variability of the resulting forecast. Because one does not know which coefficients are important in advance, some sort of search over sparse combinations is needed.

Empirical evidence suggests that simple combining methods sometimes outperform more complicated versions. A possible explanation, as proposed by several researchers (e.g., Figlewski and Urich, 1983; Kang, 1986; Clemen and Winkler, 1986), is that more complicated combining methods have a much larger variability. Our theoretical results tend to support this view.

For simplicity in illustration, in this section we will assume normality of the conditional distributions of the future response given the past data and the present

outside information and will also assume that the variances v_i are known and bounded above and below by some positive constants as in Corollary 1. However, similar results can be obtained with little difficulty for the other cases considered in Section 2 with weaker assumptions.

3.1. What Is the Price of Linearly Combining for Improvement?

Let $\Delta = \{\delta_1, \ldots, \delta_M\}$ be a finite collection of forecasting procedures with δ_j producing forecast $\hat{y}_{j,i}$ at time i. Consider the linear combination of the procedures:

$$\hat{y}_i^{\theta} = \sum_{j=1}^{M} \theta_j \hat{y}_{j,i}, \qquad i \geq 1,$$

where $\theta = (\theta_1, \ldots, \theta_M)$ satisfies the constraint $\sum_{j=1}^{M} |\theta_j| \leq 1$. Let δ^{θ} denote the corresponding forecasting procedure and let $\Theta = \Theta_M$ denote the set of all such θ. Let θ^* be the minimizer of $(1/n)\sum_{i=1}^{n} E(m_i - \hat{y}_i^{\theta})^2$ over $\theta \in \Theta$ and let $R(M; n; \Delta)$ denote the minimized value. It captures the best performance among all the linearly combined procedures subject to the constraint. Targeting δ^{θ^*} (which is unknown), we combine the procedures δ^{θ} using the AFTER algorithm with θ restricted to a finite set $\Theta_D \subset \Theta$. An appropriate choice of Θ_D (see the proof of Theorem 6 in the Appendix) leads to the following result. Let δ^M denote the combined forecasting procedure. Recall that for a forecasting procedure δ with forecast $\hat{y}_i$ for $i \geq 1$, $R(\delta; n) = (1/n)\sum_{i=1}^{n} E(m_i - \hat{y}_i)^2$ is the mean average square risk.

THEOREM 6. *Assume that Conditions 1 and 2 are satisfied. The risk of the combined procedure δ^M can be bounded as follows:*

$$R(\delta^M; n) \leq C \begin{cases} R(M; n; \Delta) + \dfrac{M \log(1 + n/M)}{n} & \text{when } M < \sqrt{n} \\ R(M; n; \Delta) + \dfrac{\log M}{\sqrt{n \log n}} & \text{when } M \geq \sqrt{n} \end{cases},$$

where C is a constant depending on A_1, A_2, and A_3 (as in Conditions 1 and 2). In particular, if $M = M_n \leq C_0 n^{\tau}$ for some $\tau > 0$ and $C_0 > 0$, then

$$R(\delta^M; n) \leq C' \begin{cases} R(M; n; \Delta) + \dfrac{\log n}{n^{1-\tau}} & \text{when } 0 \leq \tau < \frac{1}{2} \\ R(M; n; \Delta) + \left(\dfrac{\tau \log n}{n}\right)^{1/2} & \text{when } \frac{1}{2} \leq \tau < \infty, \end{cases} \tag{12}$$

where the constant C' depends on A_1, A_2, A_3, and C_0.

Remark. A similar result in the regression context suitable when $M \geq \sqrt{n}$ is given in Juditsky and Nemirovski (2000).

The result characterizes the performance of the combined procedure in terms of the best linear combination (subject to the constraint) and a penalty. For nonparametric estimation or prediction, exact risks are usually too difficult to obtain without restrictive conditions, and accordingly convergence rates are used to compare procedures. The result implies that when combining $M = C_0 n^\tau$ procedures, the combined forecast has mean average risk converging at the rate of the best combination plus a penalty term or price, which is of order $(\tau \log n/n)^{1/2}$ when $\tau \geq \frac{1}{2}$ and is of order $n^{-(1-\tau)} \log n$ when $0 \leq \tau < \frac{1}{2}$.

Define

$$\psi_n(M) = \begin{cases} \dfrac{M \log(1 + n/M)}{n} & 1 \leq M < \sqrt{n} \\ \dfrac{\log M}{\sqrt{n \log n}} & M \geq \sqrt{n}. \end{cases}$$

Note that $\psi_n(M)$ is the penalty term associated with the combined procedure for not knowing which linear coefficients work the best. The term $\psi_n(M)$ increases as M increases and thus can be viewed as a complexity penalty of combining M procedures (relative to the sample size). When n is small, the price of combining a large number of procedures is very high, but when n becomes large, more forecasts can be combined for a relatively low price.

One might wonder if the price we gave previously for combining forecasts is really necessary or rather is due to suboptimal combining. In fact, the upper bound given previously is almost optimal in the sense that it cannot be improved much in general as shown in the following result. For simplicity, assume that the conditional variance of Y_i given z^{i-1} and x_i is 1 for $i \geq 1$. For a real number a, let $\lfloor a \rfloor$ denote the largest integer no bigger than a.

THEOREM 7. *Consider $M_n = \lfloor n^\tau \rfloor$ for some $\tau > 0$. There exist M_n procedures $\Delta_{M_n} = \{\delta_j, 1 \leq j \leq M_n\}$ satisfying Condition 2 such that for any combined procedure δ based on Δ_{M_n}, one can find a case of true data generating process such that*

$$R(\delta; n) - R(M; n; \Delta_{M_n}) \geq C'' \begin{cases} n^{-(1-\tau)} & \text{when } 0 \leq \tau \leq \frac{1}{2} \\ \left(\dfrac{\log n}{n}\right)^{1/2} & \text{when } \frac{1}{2} < \tau < \infty, \end{cases}$$

where the constant C'' does not depend on n.

Thus no matter how carefully a combining procedure is constructed, it cannot be uniformly close to the ideal performance $R(M; n; \Delta_{M_n})$ within an order

smaller than the one given earlier. Note that the lower rate matches the upper rate in Theorem 6 when $\tau > \frac{1}{2}$ and the upper and lower rates differ only in logarithmic factors when $0 \leq \tau \leq \frac{1}{2}$.

Theorems 6 and 7 together with Corollary 1 provide some understanding of the question of whether one should combine for adaptation or combine for improvement. By Corollary 1, and taking the uniform prior weight on the M_n procedures, combining for adaptation yields risk of order

$$\inf_{1 \leq j \leq M_n} R(\delta_j; n) + \frac{\log M_n}{n}.$$

If M_n does not increase exponentially fast, $\log M_n/n$ is of order $\log n/n$, which is usually negligible for nonparametric estimation or prediction. Then combining for adaptation achieves the best performance (rate) among the M_n forecasting procedures. In terms of rate of convergence of the mean average risk, linearly combining n^τ procedures for better performance is guaranteed to do better than combining for adaptation only when both $R(M_n; n; \Delta_{M_n})$ and the penalty term $(\log n/n)^{1/2}$ for $\tau > \frac{1}{2}$ and $n^{-(1-\tau)}$ for $0 \leq \tau \leq \frac{1}{2}$ converge at a faster rate than $\inf_{1 \leq j \leq M_n} R(\delta_j; n) + (\log M_n)/n$. Otherwise, the linearly combined procedure can do worse or even much worse than combining for adaptation. Thus it is not automatically advantageous to combine forecasting procedures for improvement.

Theorems 6 and 7 can also help us to understand how many procedures should be combined. Note that $\inf_{1 \leq j \leq M_n} R(\delta_j; n)$ is nonincreasing as more forecasting procedures are included in Δ, but on the other hand the penalty $\psi_n(M_n)$ increases. Thus one needs a good trade-off between the potential gain and the complexity penalty. In the next section, a method will be given to automatically achieve the best trade-off.

3.2. Combining with Multiple Discretization Accuracies

In the construction of the combined forecast in the previous section, a discretization is used that is chosen according to n (among other things). This is undesirable because it requires knowledge of how many times one will forecast in the future. Theoretically speaking, the problem can be fixed by considering different discretization accuracies at the same time. In addition, we will seek automatic balance between the potential gain and the complexity penalty.

Let $\Delta = \{\delta_1, \delta_2, \ldots\}$ be a list of forecasting procedures. Let $\Delta_M = \{\delta_1, \delta_2, \ldots, \delta_M\}$. Fix M for the time being. For each $k \geq 1$, consider a best ϵ-net N_k (under the l_1^M distance) in Θ_M with 2^k points (which minimizes the maximum distance between a point in Θ_M and the closest point in the net). From Schütt (1984, Theorem 1), the entropy number under the l_1^M distance is upper bounded by $c\phi(M, k)$, where

$$\phi(M,k) = \begin{cases} 1 & \text{if } 1 \le k \le \log_2 M \\ \dfrac{\log_2(1+M/k)}{k} & \text{if } \log_2 M \le k \le M \\ 2^{-2k/M}M^{-1} & \text{if } k > M \end{cases}$$

and c does not depend on M or k. We combine the 2^k corresponding forecasting procedures (linear combinations of the original M procedures with θ in N_k) for adaptation using the uniform prior weight. Denote the procedure by $\delta(M,k)$. By Corollary 1,

$$R(\delta(M,k);n) \le C\left(\frac{k}{n} + R(M;n;\Delta_M) + \phi^2(M,k)\right),$$

where C is a constant depending on A_1, A_2, and A_3. For convenience, we will use the same symbol C to denote any constant that may depend on A_1, A_2, and A_3.

Define $\log^*$ by $\log^* x = \log(x+1) + 2\log\log(x+1)$. Now we combine the procedures $\{\delta(M,k): M \ge 1, k \ge 1\}$ for adaptation with the prior weight $ce^{-\log^* M - \log^* k}$, where the constant c is chosen to normalize the weights to sum to 1. Let δ^* denote this forecasting procedure. It follows again from Corollary 1:

$$R(\delta^*;n) \le C \inf_{M,k}\left(R(M;n;\Delta_M) + \frac{k}{n} + \phi^2(M,k) + \frac{\log M}{n} + \frac{\log k}{n}\right).$$

Compared with the other terms, $(\log M/n) + (\log k/n)$ are asymptotically negligible. Thus the final combined procedure behaves automatically as well as if it knew the optimal discretization accuracy that balances k/n and $\phi^2(M,k)$ and knew also the optimal number of procedures M_n that balances $R(M;n;\Delta_M)$ and $\inf_k((k/n) + \phi^2(M,k))$. In particular, for each n and M, by taking k as in the proof of Theorem 6, we have

$$R(\delta^*;n) \le C \inf_M (R(M;n;\Delta_M) + \psi_n(M)). \tag{13}$$

Note that the upper bound agrees in order with that in Theorem 6 but the combining procedure does not need to know n in advance. Up to a constant factor, the combined forecast automatically achieves the best trade-off between risk reduction and the complexity penalty for combining.

3.3. Sparsely Combining

In some cases, when combining many procedures, the best linear combination may concentrate on only a few of them. For such a case, combining only the important procedures leads to a much better performance because of the significantly reduced variability in the final combined procedure. Usually one does not know which of the original procedures are important individually, and even

if one does, some of the individually important procedures may not be needed in the best linear combination. Without such knowledge, we consider different sparse combinations of the candidate forecasts.

The idea is to combine each suitable subset of the candidate procedures for improvement and then combine the new ones for adaptation (over the different subsets). For each integer $M > 1$, $1 \le L < M$, and a subset S of $\{1,2,\ldots,M\}$ of size L, let $\delta(S)$ be the linearly combined (for improvement) procedure based on $\{\delta_j : j \in S\}$ using multiple discretization accuracies as in the construction of δ^* in the previous section. Then let $\delta^{M,L}$ be the combined (for adaptation) procedure based on all such $\delta(S)$ with the uniform prior weight $1/\binom{M}{L}$ (there are $\binom{M}{L}$ many such procedures). Then let $\delta^{(M)}$ be the combined (for adaptation) procedure based on $\delta^{M,1},\ldots,\delta^{M,M-1}$ using the uniform prior weight $1/(M-1)$. Let δ_F^* denote the combined (for adaptation) procedure based on $\delta^{(M)}$, $M \ge 2$ with the prior weight $c' \exp(-\log^* M)$, where the constant c' is chosen such that $\sum_{M=2}^{\infty} c' e^{-\log^* M} = 1$. Let $\Delta(S)$ denote the collection of procedures $\{\delta_j : j \in S\}$. Based on Corollary 1, we have

$$R(\delta_F^*;n) \le C \inf_{M\ge 2}\left(\inf_{1\le L\le M-1}\left(\inf_{|S|=L,\,S\subset\{1,2,\ldots,M\}} R(M;n;\Delta(S)) + \psi_n(L) + \frac{\log\binom{M}{L}}{n}\right)\right)$$
$$\le C \inf_{M\ge 2}\left(\inf_{1\le L\le M-1}\left(\inf_{|S|=L,\,S\subset\{1,2,\ldots,M\}} R(M;n;\Delta(S)) + \psi_n(L) + \frac{L\log M}{n}\right)\right). \tag{14}$$

Compared with upper bound (13), we see the potential advantage of sparse combining. Let M_n achieve the best trade-off between $R(M;n;\Delta_M)$ and $\psi_n(M)$. Suppose there exists $\Delta(S) \subset \{1,2,\ldots,M_n\}$ with $L = |\Delta(S)|$ much smaller than M_n and $R(M_n;n;\Delta(S))$ is only slightly larger than $R(M_n;n;\Delta_M)$. Then $\psi_n(L) + (L\log M_n)/n$ is much smaller than $\psi_n(M_n)$ if M_n does not grow exponentially fast. For such a case, sparse combining avoids unnecessarily large variability in combining all the first M_n procedures.

Sparse approximation/estimation has been effectively used for wavelet estimation (see, e.g., Donoho and Johnstone, 1994, 1998; Johnstone, 1999) and also has been applied to general model selection (see Yang and Barron, 1998, 1999; Barron, Birgé, and Massart, 1999).

3.4. Adaptively Combining Forecasts

The results in the previous sections provide helpful insight into the difference between combining for adaptation and combining for improvement; on potential gain and the price of combining forecasts; and on the advantage of sparse combining. In practice, however, one does not know which scenario one is in and therefore can not gauge beforehand the optimal combining strategy. If one indiscriminately combines a large number of procedures, one might end up with a much worse performance compared with combining for adaptation. In this

section, we give an adaptive combining method that overcomes these difficulties, at least in theory. Here we show that when combining the procedures properly, one can have the potential of obtaining a large gain in forecasting accuracy yet without losing much when there happens to be no advantage in considering sophisticated linear combinations for improvement.

Consider three different approaches for combining procedures in $\Delta = \{\delta_1, \delta_2, \ldots\}$.

The first approach is to combine the procedures for adaptation. Here one intends to capture the best performance (in terms of rate of convergence) among the candidate procedures. Let δ_A^* denote this combined procedure based on Δ using the AFTER algorithm as given in Section 2. When Δ is not a finite collection, one cannot use uniform prior weight. One choice of prior weight π_j is $ce^{-\log^* j}$, where the constant c is chosen to normalize the weights to sum to 1. Based on Corollary 1, we have

$$R(\delta_A^*; n) \le C \inf_j \left(\frac{\log(j+1)}{n} + R(\delta_j; n) \right) =: CR_1^*(n; \Delta). \tag{15}$$

If one procedure, say, δ_{j^*}, behaves the best asymptotically, then the penalty is of order $1/n$. If the best procedure changes according to n, then $\inf_j(\log(j+1)/n + R(n;\delta_j))$ is a trade-off between complexity and estimation accuracy.

The second approach targets the best performance among all the linear combinations of the original procedures up to different orders. This is the combining procedure described in Section 3.2. Let δ_B^* denote this procedure. Then

$$R(\delta_B^*; n) \le C \inf_M (R(M; n; \Delta_M) + \psi_n(M)) =: CR_2^*(n; \Delta).$$

The third approach is to combine the procedures through various sparse combinations as described in Section 3.3. Let δ_C^* denote this procedure. Then

$$R(\delta_C^*; n) \le C \inf_{M \ge 2} \left(\inf_{1 \le L \le M-1} \left(\inf_{|S|=L, S \subset \{1,2,\ldots,M\}} R(M; n; \Delta(S)) + \psi_n(L) + \frac{L \log M}{n} \right) \right)$$
$$=: CR_3^*(n; \Delta).$$

Now we combine these three procedures δ_A^*, δ_B^*, and δ_C^* with equal prior weight $\frac{1}{3}$ and let δ_F denote the final combined procedure. Note that this is still a linear combination of the original procedures. We have the following conclusion.

PROPOSITION 2. *Assume Conditions 1 and 2 are satisfied. Then the combined procedure satisfies*

$$R(\delta_F; n) \le C \min(R_1^*(n; \Delta), R_2^*(n; \Delta), R_3^*(n; \Delta)),$$

where C is a constant depending only on A_1, A_2, and A_3.

Thus in terms of rate of convergence, first of all, the combined forecasting procedure converges as fast as any original procedure. Second, when linear combinations of the first M_n procedures (for some $M_n > 1$) can improve estimation accuracy dramatically, one pays a price at most of order $\psi_n(M_n)$ for the better performance. When certain linear combinations of a small number of procedures perform well, the combined procedure can also take advantage of that. In summary, the combined procedure is automatically both aggressive in pursuing greater accuracy improvement and conservative in avoiding paying a price higher than the potential gain.

4. CONCLUDING REMARKS

Combining forecasts is a useful technique for sharing strengths of different forecasting procedures. There are two closely related directions in combining forecasts: combining for adaptation and combining for improvement.

An algorithm, AFTER, is proposed for combining forecasts for adaptation. The method assigns weights to the candidate procedures according to their performances thus far and relies on the specification of the forms (e.g., Gaussian or double-exponential) of the conditional distributions of Y given the past data and the current outside information. One may use a fixed specification for combining or use specifications from the individual forecasters. When the common specification for the first case or the individual specifications corresponding to the best forecasts capture the true uncertainty in Y well, the AFTER algorithm is shown to have a prediction risk automatically close to the best procedure. No assumption on the relationship between different forecasts is needed for the result. In applications, various model assumptions on the relationship between Y and the outside information X and on dependence structure among the errors (noise) can be considered in constructing individual predictions. The resulting forecasts and also additional forecasts based on subjective assessments can be combined. The algorithm AFTER is easy to implement in practice. Combining without specifying or estimating the conditional distributions of Y is also studied under assumptions on certain moment generating functions.

An alternative to combining procedures is selecting the best one. Model selection criteria can be used in that regard when combining model-based forecasts. However, model selection is often unstable, which causes a rather large variability in prediction. Combining with appropriate weighting as by AFTER gives more stable predictions and accordingly yields better performance.

In the other direction, we studied linearly combining (subject to a constraint) forecasts for improving the individual procedures. Because the "optimal" weights are estimated, one must pay a price in the sense that the risk of the combined forecast is larger in general than that of the best linear combination. The potential gain and the price in terms of prediction accuracy are theoretically quantified. The result supports the empirical finding that combining does not necessarily lead to performance improvement and indiscriminate combining of

a number of forecasts can have a much worse performance compared to the best individual procedure. In addition, we showed an advantage of sparse combining. A multipurpose combining method is shown to be able to automatically balance the potential gain and the complexity penalty; it takes advantage of sparse combining when only a few procedures have large weights for the best linear combination; and it always achieves (in rate) the best individual performance.

Our results on linearly combining forecasts for improvement are mainly theoretical. Computationally feasible methods need to be developed with similar properties. In addition to linear combining, nonlinear methods may also prove helpful for sharing strengths of different procedures. Another interesting and challenging research direction is combining prediction intervals. See Taylor and Bunn (1999) for some simulation results. Further theoretical research on these topics, in our opinion, will be a fruitful guide for combining forecasts in practice.

REFERENCES

Armstrong, J.S. (1989) Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting* 5, 585–588.

Barron, A.R. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* 39, 930–945.

Barron, A.R. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning* 14, 115–133.

Barron, A.R., L. Birgé, & P. Massart (1999) Risk bounds for model selection via penalization. *Probability Theory and Related Fields* 113, 301–413.

Barron, A.R., J. Rissanen, & B. Yu (1998) The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory* 44, 2743–2760.

Bates, J.M., & C.W.J. Granger (1969) The combination of forecasts. *Operational Research Quarterly* 20, 451–468.

Box, G.E.P. & G.M. Jenkins (1976) *Time Series Analysis: Forecasting and Control*, 2nd ed. Holden-Day.

Breiman, L. (1996a) Stacked regressions. *Machine Learning* 24, 49–64.

Breiman, L. (1996b) Bagging predictors. *Machine Learning* 24, 123–140.

Buckland, S.T., K.P. Burnham, & N.H. Augustin (1997) Model selection: An integral part of inference. *Biometrics* 53, 603–618.

Catoni, O. (1999) "Universal" Aggregation Rules with Exact Bias Bounds. Preprint no. 510, Laboratoire de Probabilites et Modeles Aleatoires, Université Paris VI & Université Paris VII.

Cesa-Bianchi, N., Y. Freund, D.P. Haussler, R. Schapire, & M.K. Warmuth (1997) How to use expert advice? *Journal of the Association for Computing Machinery* 44, 427–485.

Chatfield, C. (1995) Model uncertainty, data mining, and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A* 158, 419–466.

Clemen, R.T. (1989) Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5, 559–583.

Clemen, R.T., A.H. Murphy, & R.L. Winkler (1995) Screening probability forecasts: Contrasts between choosing and combining. *International Journal of Forecasting* 11, 133–145.

Clemen, R.T. & R.L. Winkler (1986) Combining economic forecasts. *Journal of Business and Economic Statistics* 4, 39–46.

Cover, T.M. (1965) Behavior of sequential predictors of binary sequences. In *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes,* pp. 263–271. Publishing House of the Czechoslovak Academy of Sciences.

Csiszár, I. (1975) I-Divergence geometry of probability distributions and minimization problems. *Annals of Probability* 3, 146–158.

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

DeVore, R.A. & G.G. Lorentz (1993) *Constructive Approximation.* Springer.

Devroye, L.P. & T.J. Wagner (1980) Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics* 8, 231–239.

Donoho, D.L. & I.M. Johnstone (1994) Ideal denoising in an orthonormal basis chosen from a library of bases. *C.R. Acad. Sci. Paris* 319, 1317–1322.

Donoho, D.L. & I.M. Johnstone (1998) Minimax estimation via wavelet shrinkage. *Annals of Statistics* 26, 879–921.

Efromovich, S. (1999) How to overcome curse of long-memory? *IEEE Transactions on Information Theory* 45, 1735–1741.

Figlewski, S. & T. Urich (1983) Optimal aggregation of money supply forecasts: Accuracy, profitability, and market efficiency. *Journal of Finance* 28, 695–710.

Foster, D.P. (1991) Prediction in the worst case. *Annals of Statistics* 19, 1084–1090.

Genest, C. & J.V. Zidek (1986) Combining probability distributions: A critique and an annotated bibliography. *Statistical Science* 1, 114–148.

Gouriéroux, C. & A. Monfort (1992) Qualitative threshold ARCH models. *Journal of Econometrics* 52, 159–199.

Hall, P. & J.D. Hart (1990) Nonparametric regression with long-range dependence. *Stochastic Processes and Their Applications* 36, 339–351.

Haussler, D., J. Kivinen, & M.K. Warmuth (1998) Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory* 44, 1906–1925.

Hoeting, J.A., D. Madigan, A.E. Raftery, & C.T. Volinsky (1999) Bayesian model averaging: A tutorial (with discussion). *Statistical Science* 14, 382–401.

Holt, C.C. (1957) Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. Carnegie Institute of Technology. ONR Research Memorandum 52.

Johnstone, I. (1999) Function Estimation in Gaussian Noise: Sequence Models. Manuscript.

Johnstone, I. & B.W. Silverman (1997) Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Association, Series B* 59, 319–351.

Juditsky, A. & A. Nemirovski (2000) Functional aggregation for nonparametric estimation. *Annals of Statistics* 28, 681–712.

Kang, H. (1986) Unstable weights in the combination of forecasts. *Management of Science* 32, 683–695.

Leamer, E.E. (1978) *Specification Searches: Ad hoc Inference with Nonexperimental Data.* Wiley.

LeBlanc, M. & R. Tibshirani (1996) Combining estimates in regression and classification. *Journal of the American Statistical Association* 91, 1641–1650.

Littlestone, N. & M.K. Warmuth (1994) The weighted majority algorithm. *Information and Computation* 108, 212–261.

Merhav, N. & M. Feder (1998) Universal prediction. *IEEE Transactions on Information Theory* 44, 2124–2147.

Newbold P. & C.W.J. Granger (1974) Experience with forecasting univariate times series and the combination of forecasts. *Journal of the Royal Statistical Society, Series A* 137, 131–165 (with discussion).

Ploberger, W. & P.C.B. Phillips (1999) Empirical Limits for Time Series Econometric Models. Cowles Foundation Discussion paper 1220, Yale University.

Rissanen, J. (1986) Stochastic complexity and modeling. *Annals of Statistics* 14, 1080–1100.

Schütt, C. (1984) Entropy numbers of diagonal operators between symmetric Banach spaces. *Journal of Approximation Theory* 40, 121–128.

Stone, C.J. (1977) Consistent nonparametric regression. *Annals of Statistics* 5, 595–620.

Stone, C.J. (1985) Additive regression and other nonparametric models. *Annals of Statistics* 13, 689–705.

Taylor, J.W. & D.W. Bunn (1999) Investigating improvements in the accuracy of prediction intervals for combinations of forecasts: A simulation study. *International Journal of Forecasting* 15, 325–339.

Triebel, H. (1975) Interpolation properties of ϵ-entropy and diameters. Geometric characteristics of embedding for function spaces of Sobolev-Besov type. *Mat. Sbornik* 98, 27–41; English trans. in *Math. USSR Sb.* 27, 23–37, 1977.

Wang, Y. (1996) Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics* 24, 466–484.

Vovk, V.G. (1990) Aggregating strategies. In M. Fulk & J. Case (eds.), *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pp. 372–383.

Winters, P.R. (1960) Forecasting sales by exponentially weighted moving averages. *Management of Science* 6, 324–342.

Wolpert, D. (1992) Stacked generalization. *Neural Networks* 5, 241–259.

Yang, Y. (1997) Nonparametric Regression and Prediction with Dependent Errors. Technical Report 29, Department of Statistics, Iowa State University. A shorter version appeared in *Bernoulli* 7, 633–655, 2001.

Yang, Y. (in press) Aggregating regression procedures for a better performance. *Bernoulli*, forthcoming.

Yang, Y. (2000a) Mixing strategies for density estimation. *Annals of Statistics* 28, 75–87.

Yang, Y. (2000b) Combining different procedures for adaptive regression. *Journal of Multivariate Analysis* 74, 135–161.

Yang, Y. (2000c) Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica* 10, 1069–1089.

Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–588.

Yang, Y. (2003) Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica* 13, 783–809.

Yang, Y. & A.R. Barron (1998) An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* 44, 95–116.

Yang, Y. & A.R. Barron (1999) Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 27, 1564–1599.

Zou, H. & Y. Yang (2003) Combining time series models for forecasting. *International Journal of Forecasting*, forthcoming.

APPENDIX: PROOFS OF THE RESULTS

Proof of Theorem 1. Let

$$f^n = \Pi_{i=1}^n \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(y_i - m_i)^2\right)$$

$$= \frac{1}{(2\pi)^{n/2}\Pi_{i=1}^n v_i^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(y_i - m_i)^2}{v_i}\right)$$

and

$$q^n = \sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^n \frac{1}{\sqrt{2\pi v_i}} \exp\left(-\frac{1}{2v_i}(y_i - \hat{y}_{j,i})^2\right)$$

$$= \frac{1}{(2\pi)^{n/2}\Pi_{i=1}^n v_i^{1/2}} \sum_{j=1}^{\infty} \pi_j \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(y_i - \hat{y}_{j,i})^2}{v_i}\right).$$

Consider $\log(f^n/q^n)$. By monotonicity of the log function, for each fixed $j^* \geq 1$, we have

$$\log(f^n/q^n) \leq \log\left(\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(y_i - m_i)^2}{v_i}\right)}{\pi_{j^*}\exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(y_i - \hat{y}_{j^*,i})^2}{v_i}\right)}\right)$$

$$= \log(1/\pi_{j^*}) + \frac{1}{2}\sum_{i=1}^n \frac{(y_i - \hat{y}_{j^*,i})^2 - (y_i - m_i)^2}{v_i}. \tag{A.1}$$

On the other hand, observe

$$q^n = \sum_{j=1}^{\infty} \frac{\pi_j}{\sqrt{2\pi v_1}} \exp\left(-\frac{1}{2v_1}(y_1 - \hat{y}_{j,1})^2\right)$$

$$\times \frac{\sum_{j=1}^{\infty} \frac{\pi_j}{\sqrt{4\pi^2 v_1 v_2}} \exp\left(-\frac{1}{2v_1}(y_1 - \hat{y}_{j,1})^2 - \frac{1}{2v_2}(y_2 - \hat{y}_{j,2})^2\right)}{\sum_{j=1}^{\infty} \frac{\pi_j}{\sqrt{2\pi v_1}} \exp\left(-\frac{1}{2v_1}(y_1 - \hat{y}_{j,1})^2\right)}$$

$$\times \cdots \times \frac{\sum_{j=1}^{\infty} \frac{\pi_j}{\Pi_{i=1}^n \sqrt{2\pi v_i}} \exp\left(-\sum_{i=1}^n \frac{1}{2v_i}(y_i - \hat{y}_{j,i})^2\right)}{\sum_{j=1}^{\infty} \frac{\pi_j}{\Pi_{i=1}^{n-1} \sqrt{2\pi v_i}} \exp\left(-\sum_{i=1}^{n-1} \frac{1}{2v_i}(y_i - \hat{y}_{j,i})^2\right)}.$$

Let $p_i = (1/\sqrt{2\pi v_i})\exp(-(y_i - m_i)^2/2v_i)$ and $g_i = 1/\sqrt{2\pi v_i}\sum_{j=1}^{\infty} W_{j,i}\exp(-[(y_i - \hat{y}_{j,i})^2/2v_i])$. It follows by definition of $W_{j,i}$ that $\log(f^n/q^n) = \sum_{i=1}^n \log(p_i/g_i)$. Together with (A.1), we have

$$\sum_{i=1}^n \log\left(\frac{p_i}{g_i}\right) \leq \log(1/\pi_{j^*}) + \frac{1}{2}\sum_{i=1}^n \frac{(y_i - \hat{y}_{j^*,i})^2 - (y_i - m_i)^2}{v_i}. \tag{A.2}$$

Taking expectation on both sides of (A.2) under the conditional distribution of Y_n given $Z^{n-1} = z^{n-1}$ and x_n, we have

$$\sum_{i=1}^{n-1} \log\left(\frac{p_i}{g_i}\right) + E_n \log\left(\frac{p_n}{g_n}\right)$$

$$\le \log(1/\pi_{j^*}) + \frac{1}{2}\sum_{i=1}^{n-1} \frac{(y_i - \hat{y}_{j^*,i})^2 - (y_i - m_i)^2}{v_i}$$

$$+ E_n\left(\frac{(y_n - \hat{y}_{j^*,n})^2 - (y_n - m_n)^2}{2v_n}\right). \tag{A.3}$$

Because $y_n = m_n + e_n$, where $E_n e_n = 0$, we have

$$(y_n - \hat{y}_{j^*,n})^2 - (y_n - m_n)^2 = (m_n - \hat{y}_{j^*,n})^2 - 2e_n(\hat{y}_{j^*,n} - m_n)$$

and

$$E_n[e_n(\hat{y}_{j^*,n} - m_n)] = (\hat{y}_{j^*,n} - m_n)E_n e_n = 0.$$

It follows that

$$E_n\left(\frac{(y_n - \hat{y}_{j^*,n})^2 - (y_n - m_n)^2}{2v_n}\right) = \frac{(m_n - \hat{y}_{j^*,n})^2}{2v_n}.$$

Observe that under the normality assumption on the errors,

$$E_n \log\left(\frac{p_n}{g_n}\right) = \int p_n \log\frac{p_n}{g_n}\, dy_n \ge \int (\sqrt{p_n} - \sqrt{g_n})^2\, dy_n,$$

where the inequality is the familiar relationship between the Kullback–Leibler divergence and the squared Hellinger distance. Let $\mu_{p_n}(= m_n), v_{p_n}(= v_n)$, and $\mu_{g_n}(= \hat{y}_n^*), v_{g_n}$ denote the means and variances of p_n and g_n, respectively. Under Condition 0, it can be easily verified that $v_{g_n} \le v_n + (2\sup_{j\ge 1}|\hat{y}_{j,n} - m_n|)^2 \le v_n(1 + 4\tau)$. Observing that $|\mu_{p_n} - \mu_{g_n}| = |m_n - \hat{y}_n^*| \le \sup_{j\ge 1}|\hat{y}_{j,n} - m_n| \le \sqrt{v_n \tau}$, together with Lemma 1, which follows, we have that

$$\int (\sqrt{p_n} - \sqrt{g_n})^2\, dy_n \ge \frac{(m_n - \hat{y}_n^*)^2}{2v_n + 2v_{g_n} + |\mu_{p_n} - \mu_{g_n}|^2} \ge \frac{(m_n - \hat{y}_n^*)^2}{2v_n + 2v_n(1 + 4\tau) + v_n\tau}.$$

It follows that

$$E_n \log\left(\frac{p_n}{g_n}\right) \ge \int (\sqrt{p_n} - \sqrt{g_n})^2\, dy_n \ge \frac{(m_n - \hat{y}_n^*)^2}{v_n(4 + 9\tau)}.$$

Together with (A.3), we have

$$\sum_{i=1}^{n-1} \log\left(\frac{p_i}{g_i}\right) + \frac{(m_n - \hat{y}_n^*)^2}{v_n(4+9\tau)}$$
$$\leq \log(1/\pi_{j^*}) + \frac{1}{2}\sum_{i=1}^{n-1} \frac{(y_i - \hat{y}_{j^*,i})^2 - (y_i - m_i)^2}{v_i} + \frac{(m_n - \hat{y}_{j^*,n})^2}{2v_n}.$$

Thus

$$E\left(\sum_{i=1}^{n-1} \log\left(\frac{p_i}{g_i}\right) - \frac{1}{2}\sum_{i=1}^{n-1} \frac{(y_i - \hat{y}_{j^*,i})^2 - (y_i - m_i)^2}{v_i}\right)$$
$$\leq \log(1/\pi_{j^*}) - E\left(\frac{(m_n - \hat{y}_n^*)^2}{v_n(4+9\tau)}\right) + E\left(\frac{(m_n - \hat{y}_{j^*,n})^2}{2v_n}\right).$$

Handle $\log(p_{n-1}/g_{n-1}) - [(y_{n-1} - \hat{y}_{j^*,n-1})^2 - (y_{n-1} - m_{n-1})^2]/v_{n-1}$ the same way as for $i = n$, and similarly for other $1 \leq i < n-1$, we have

$$\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i(4+9\tau)}\right) \leq \log(1/\pi_{j^*}) + \sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j^*,i})^2}{2v_i}\right).$$

Because the preceding analysis holds for every j^*, it follows that

$$\sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i(4+9\tau)}\right) \leq \inf_{j\geq 1}\left(\log(1/\pi_j) + \sum_{i=1}^{n} E\left(\frac{(m_i - \hat{y}_{j,i})^2}{2v_i}\right)\right).$$

The conclusion of Theorem 1 then follows. This completes the proof of Theorem 1. ■

Proof of Theorem 2. We only need to modify the proof of Theorem 1. Define f^n as before but modify q^n to be

$$q^n = \sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^n \frac{1}{\sqrt{2\pi\hat{v}_{j,i}}} \exp\left(-\frac{1}{2\hat{v}_{j,i}}(y_i - \hat{y}_{j,i})^2\right)$$
$$= \sum_{j=1}^{\infty} \pi_j \frac{1}{(2\pi)^{n/2}\Pi_{i=1}^n \hat{v}_{j,i}^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{j,i})^2}{\hat{v}_{j,i}}\right).$$

As before, for each fixed $j^* \geq 1$, we have

$$\log(f^n/q^n) \leq \log\left(\frac{(\Pi_{i=1}^n (2\pi v_i)^{-1/2})\exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - m_i)^2}{v_i}\right)}{\pi_{j^*}(\Pi_{i=1}^n (2\pi\hat{v}_{j^*,i})^{-1/2})\exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{j^*,i})^2}{\hat{v}_{j^*,i}}\right)}\right)$$
$$= \log(1/\pi_{j^*}) + \frac{1}{2}\sum_{i=1}^{n}\left(\log\frac{\hat{v}_{j^*,i}}{v_i} + \frac{(y_i - \hat{y}_{j^*,i})^2}{\hat{v}_{j^*,i}} - \frac{(y_i - m_i)^2}{v_i}\right). \quad \textbf{(A.4)}$$

Taking expectation conditioned on z^{i-1} and x_i, we have

$$E_i\left(\log\frac{\hat{v}_{j^*,i}}{v_i}+\frac{(y_i-\hat{y}_{j^*,i})^2}{\hat{v}_{j^*,i}}-\frac{(y_i-m_i)^2}{v_i}\right)=\frac{(m_i-\hat{y}_{j^*,i})^2}{\hat{v}_{j^*,i}}+\frac{v_i}{\hat{v}_{j^*,i}}-1-\log\frac{v_i}{\hat{v}_{j^*,i}}. \tag{A.5}$$

Similarly as before, $\log(f^n/q^n)=\sum_{i=1}^n\log(p_i/g_i)$, where g_i is modified to be

$$g_i=\sum_{j=1}^{\infty}W_{j,i}\frac{1}{\sqrt{2\pi\hat{v}_{j,i}}}\exp\left(-\frac{(y_i-\hat{y}_{j,i})^2}{2\hat{v}_{j,i}}\right).$$

Together with (A.4) and (A.5), we have

$$\sum_{i=1}^n E\log\left(\frac{p_i}{g_i}\right)\le\log(1/\pi_{j^*})+\frac{1}{2}\sum_{i=1}^n E\left(\frac{(m_i-\hat{y}_{j^*,i})^2}{\hat{v}_{j^*,i}}+\frac{v_i}{\hat{v}_{j^*,i}}-1-\log\frac{v_i}{\hat{v}_{j^*,i}}\right). \tag{A.6}$$

Notice that as before $E_i\log(p_i/g_i)\ge\int(\sqrt{p_i}-\sqrt{g_i})^2\,dy_i$, and using the fact that $v_{g_i}\le\xi_2 v_i+4v_i\tau$ and applying Lemma 1, we have

$$E_i\log\left(\frac{p_i}{g_i}\right)\ge\frac{(m_i-\hat{y}_i^*)^2}{v_i(2(1+\xi_2)+9\tau)}.$$

Together with (A.6), we have

$$\sum_{i=1}^n E\left(\frac{(m_i-\hat{y}_i^*)^2}{v_i(2(1+\xi_2)+9\tau)}\right)\le\log(1/\pi_{j^*})+\frac{1}{2}\sum_{i=1}^n E\left(\frac{(m_i-\hat{y}_{j^*,i})^2}{\hat{v}_{j^*,i}}+\frac{v_i}{\hat{v}_{j^*,i}}-1-\log\frac{v_i}{\hat{v}_{j^*,i}}\right).$$

It is straightforward to verify that if $x\ge x_0>0$, $x-1-\log x\le c_{x_0}(x-1)^2$ for a constant $c_{x_0}=(x_0-1-\log x_0)/(x_0-1)^2$. Together with the fact that the preceding inequality holds for every j^*, under Condition 3, it follows that

$$\sum_{i=1}^n E\left(\frac{(m_i-\hat{y}_i^*)^2}{v_i(2(1+\xi_2)+9\tau)}\right)\le\inf_{j\ge1}\left(\log(1/\pi_j)+\sum_{i=1}^n E\left(\frac{(m_i-\hat{y}_{j,i})^2}{2\xi_1 v_i}\right)+C(\xi_1,\xi_2)E\left(\frac{(\hat{v}_{j,i}-v_i)^2}{2v_i^2}\right)\right),$$

where $C(\xi_1,\xi_2)=(1/\xi_2-1+\log\xi_2)/\xi_1^2(1/\xi_2-1)^2$. The conclusion then follows. This completes the proof of Theorem 2. ■

Proof of Theorem 3. Redefine

$$f^n = \Pi_{i=1}^n \frac{1}{s_i} h\left(\frac{y_i - m_i}{s_i}\right)$$

and

$$q^n = \sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^n \frac{1}{\hat{s}_{j,i}} h\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right).$$

Similarly as before, we have that for every j^*,

$$\sum_{i=1}^n E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i(2(1+\xi_2)+9\tau)}\right) \le \log(1/\pi_{j^*}) + \sum_{i=1}^n E \log \frac{\frac{1}{s_i} h\left(\frac{y_i - m_i}{s_i}\right)}{\frac{1}{\hat{s}_{j,i}} h\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right)}.$$

By a simple linear transformation,

$$E_i \log \frac{\frac{1}{s_i} h\left(\frac{y_i - m_i}{s_i}\right)}{\frac{1}{\hat{s}_{j,i}} h\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right)} = \int \frac{1}{s_i} h\left(\frac{y_i - m_i}{s_i}\right) \log \frac{\frac{1}{s_i} h\left(\frac{y_i - m_i}{s_i}\right)}{\frac{1}{\hat{s}_{j,i}} h\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right)} dy_i$$

$$= \int h(z) \log \frac{h(z)}{\frac{1}{\hat{s}_{j,i}/s_i} h\left(\frac{z - (\hat{y}_{j,i} - m_i)/s_i}{\hat{s}_{j,i}/s_i}\right)} dz$$

$$\le B_{s_0,T}\left((1 - \hat{s}_{j,i}/s_i)^2 + \frac{(\hat{y}_{j,i} - m_i)^2}{s_i^2}\right),$$

where for the last inequality, we use Condition 4 with $s_0 = \min(\xi_1^{1/2}, \xi_2^{-1/2})$ and $T = \tau^{1/2}$. It follows that

$$E \log \frac{\frac{1}{s_i} h\left(\frac{y_i - m_i}{s_i}\right)}{\frac{1}{\hat{s}_{j,i}} h\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right)} \le B_{s_0,T}\left(E\frac{(\hat{s}_{j,i} - s_i)^2}{s_i^2} + E\frac{(\hat{y}_{j,i} - m_i)^2}{s_i^2}\right).$$

The rest of the proof then follows easily. This completes the proof of Theorem 3. ■

Proof of Theorem 4. We prove the first part of the result first. Let $\Pi_{i=1}^n p_i$ be the true joint distribution of $Y_1, \ldots, Y_n$ and $\Pi_{i=1}^n \hat{p}_{j,i}$ be the estimated version of forecaster j for $j \ge 1$. Let $q^n = \sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^n \hat{p}_{j,i}$. Then as before, for each $j^* \ge 1$,

$$\log\left(\frac{\Pi_{i=1}^n p_i}{q^n}\right) \le \log(1/\pi_{j^*}) + \sum_{i=1}^n \log \frac{p_i}{\hat{p}_{j^*,i}}.$$

Let $g_i(y_i) = \sum_{j=1}^{\infty} W_{j,i}\hat{p}_{j,i}(y_i)$ for $i \geq 1$. By definition of $W_{j,i}$, $q^n(y_1,\ldots,y_n) = \Pi_{i=1}^n g_i(y_i)$. Thus

$$E\left(\sum_{i=1}^n E_i \log\frac{p_i}{g_i}\right) \leq \log(1/\pi_{j^*}) + \sum_{i=1}^n ED(p_i\|\hat{p}_{j^*,i}).$$

Then lower bounding the Kullback–Leibler divergence by the squared Hellinger distance and applying Lemma 1 as in the proof of Theorem 1, we have

$$\frac{1}{n}\sum_{i=1}^n E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i}\right) \leq (2(1+\xi_2)+9\tau)\inf_{j\geq 1}\left(\frac{\log(1/\pi_j)}{n} + \frac{1}{n}\sum_{i=1}^n ED(p_i\|\hat{p}_{j,i})\right).$$

For the second result of the theorem, let

$$f^n = \Pi_{i=1}^n \frac{1}{s_i}h_i\left(\frac{y_i - m_i}{s_i}\right) \quad \text{and} \quad q^n = \sum_{j=1}^{\infty}\pi_j\Pi_{i=1}^n \frac{1}{\hat{s}_{j,i}}g_{j,i}\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right).$$

For δ_j satisfying Condition 5, taking expectation conditioned on z^{i-1} and x_i, by a simple linear transformation and then applying Lemma 2 (which follows), we have

$$E_i \log\left(\frac{\frac{1}{s_i}h_i\left(\frac{y_i - m_i}{s_i}\right)}{\frac{1}{\hat{s}_{j,i}}g_{j,i}\left(\frac{y_i - \hat{y}_{j,i}}{\hat{s}_{j,i}}\right)}\right) = \int h_i(t)\log\left(\frac{h_i(t)}{\frac{1}{\frac{\hat{s}_{j,i}}{s_i}}g_{j,i}\left(\frac{t - \frac{(\hat{y}_{j,i} - m_i)}{s_i}}{\frac{\hat{s}_{j,i}}{s_i}}\right)}\right)dt$$

$$\leq C\left(D(h_i\|g_{j,i}) + \frac{(\hat{s}_{j,i} - s_i)^2}{s_i^2} + \frac{(\hat{y}_{j,i} - m_i)^2}{s_i^2}\right), \quad \textbf{(A.7)}$$

where C is a constant depending on c_0 and $B_{s_0,T}$ with s_0 and T being constants depending on ξ_1, ξ_2, and τ. Then as before, we have

$$\sum_{i=1}^n E\left(\frac{(m_i - \hat{y}_i^*)^2}{v_i(2(1+\xi_2)+9\tau)}\right)$$

$$\leq \log(1/\pi_j) + C\left(\sum_{i=1}^n D(h_i\|g_{j,i}) + \sum_{i=1}^n E\frac{(\hat{s}_{j,i} - s_i)^2}{s_i^2} + \sum_{i=1}^n E\frac{(\hat{y}_{j,i} - m_i)^2}{s_i^2}\right).$$

The conclusion then follows. This completes the proof of Theorem 4. ■

Proof of Theorem 5. Define $h(x) = \exp(-\lambda\psi(x))$ (it is not necessarily a probability density function). Define

$$q^n = \sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^n h(y_i - \hat{y}_{j,i}).$$

For each fixed $j^* \geq 1$, we have

$$\log(1/q^n) \leq \log(1/\pi_{j^*}) + \lambda \sum_{i=1}^{n} \psi(y_i - \hat{y}_{j^*,i}). \tag{A.8}$$

As before,

$$q^n = \left(\sum_{j=1}^{\infty} \pi_j h(y_1 - \hat{y}_{j,1}) \right) \times \frac{\sum_{j=1}^{\infty} \pi_j h(y_1 - \hat{y}_{j,1}) h(y_2 - \hat{y}_{j,2})}{\sum_{j=1}^{\infty} \pi_j h(y_1 - \hat{y}_{j,1})}$$

$$\times \cdots \times \frac{\sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^n h(y_i - \hat{y}_{j,i})}{\sum_{j=1}^{\infty} \pi_j \Pi_{i=1}^{n-1} h(y_i - \hat{y}_{j,i})}.$$

It follows that

$$\log(1/q^n) = -\sum_{i=1}^{n} \log\left(\sum_{j=1}^{\infty} W_{j,i} h(y_i - \hat{y}_{j,i}) \right). \tag{A.9}$$

We now bound $\log(\sum_{j=1}^{\infty} W_{j,i} h(y_i - \hat{y}_{j,i})) = \log(E^J \exp\{-\lambda\psi(y_i - \hat{y}_{J,i})\})$, where E^J (depending on i) denotes expectation with respect to J under the probability mass function $P(J = j) = W_{j,i}$ (for a fixed i). By Lemma 10.1 of Catoni (1999), under (10) of Condition 7, we have

$$\log(E^J \exp(-\lambda\psi(y_i - \hat{y}_{J,i}))) \leq -\lambda E^J \psi(y_i - \hat{y}_{J,i}) + I, \tag{A.10}$$

where I denotes

$$\frac{\lambda^2}{2} E^J(\psi(y_i - \hat{y}_{J,i}) - E^J\psi(y_i - \hat{y}_{J,i}))^2$$

$$\times \exp\left(\bar{c}\lambda 2^{\beta} \left(|y_i - m_i|^{\beta} + \left(1 + \sup_{j \geq 1} |\hat{y}_{j,i} - m_i| \right)^{\beta} \right) \right).$$

For handling the term I, note that by (10) of Condition 7,

$$\begin{aligned} &E^J(\psi(y_i - \hat{y}_{J,i}) - E^J\psi(y_i - \hat{y}_{J,i}))^2 \\ &\quad \leq E^J(\psi(y_i - \hat{y}_{J,i}) - \psi(y_i - E^J\hat{y}_{J,i}))^2 \\ &\quad \leq \bar{c}^2 \left(|y_i - m_i| + \left(1 + \sup_{j \geq 1} |\hat{y}_{j,i} - m_i| \right) \right)^{2\beta} E^J(\hat{y}_{J,i} - E^J\hat{y}_{J,i})^2 \\ &\quad \leq \bar{c}^2 2^{2\beta - 1} \left(|y_i - m_i|^{2\beta} + \left(1 + \sup_{j \geq 1} |\hat{y}_{j,i} - m_i| \right)^{2\beta} \right) E^J(\hat{y}_{J,i} - E^J\hat{y}_{J,i})^2. \end{aligned} \tag{A.11}$$

Under (9) of Condition 7, let $a_0 = y_i - \hat{y}_i^* = y_i - E^J\hat{y}_{J,i}$; we have

$$\psi(y_i - \hat{y}_{j,i}) - (\theta_{a_0}(y_i - \hat{y}_{j,i} - (y_i - \hat{y}_i^*)) + \psi(y_i - \hat{y}_i^*))$$
$$\geq \underline{c}|y_i - \hat{y}_{j,i} - y_i + \hat{y}_i^*|^2 = \underline{c}(\hat{y}_{j,i} - \hat{y}_i^*)^2. \tag{A.12}$$

Taking expectation under E^J, we have

$$E^J\psi(y_i - \hat{y}_{J,i}) - (\theta_{a_0}E^J(\hat{y}_i^* - \hat{y}_{J,i}) + \psi(y_i - \hat{y}_i^*)) \geq \underline{c}E^J(\hat{y}_{J,i} - \hat{y}_i^*)^2.$$

Observing that $E^J(\hat{y}_i^* - \hat{y}_{J,i}) = 0$ by definition of $\hat{y}_i^*$, we have

$$E^J\psi(y_i - \hat{y}_{J,i}) - \psi(y_i - \hat{y}_i^*) \geq \underline{c}E^J(\hat{y}_{J,i} - \hat{y}_i^*)^2. \tag{A.13}$$

It follows that the term I is bounded by $E^J(\hat{y}_{J,i} - \hat{y}_i^*)^2$ multiplied by

$$\frac{\lambda^2}{2}\bar{c}^2 2^{2\beta-1}\left(\left(1 + \sup_{j\geq 1}|\hat{y}_{j,i} - m_i|\right)^{2\beta} + |e_i|^{2\beta}\right)$$
$$\times \exp\left(\bar{c}\lambda 2^\beta\left(|e_i|^\beta + \left(1 + \sup_{j\geq 1}|\hat{y}_{j,i} - m_i|\right)^\beta\right)\right)$$
$$\leq \lambda^2\bar{c}^2 2^{2\beta-2}((A+1)^{2\beta} + |e_i|^{2\beta})\exp(\bar{c}\lambda 2^\beta(|e_i|^\beta + (A+1)^\beta)).$$

Under Condition 8, by taking conditional expectation given z^{i-1} and x_i, we have that when $\bar{c}\lambda 2^\beta \leq t_0$,

$$E_i(I) \leq E^J((\hat{y}_{J,i} - E^J\hat{y}_{J,i})^2)\cdot\lambda^2\bar{c}^2 2^{2\beta-2}e^{\bar{c}\lambda 2^\beta(A+1)^\beta}((A+1)^{2\beta}M_2(\bar{c}\lambda 2^\beta) + M_1(\bar{c}\lambda 2^\beta)).$$

Take λ small enough, say, $0 < \lambda \leq \lambda_0$, so that $\lambda^2\bar{c}^2 2^{2\beta-2}e^{\bar{c}\lambda 2^\beta(A+1)^\beta}((A+1)^{2\beta} \times M_2(\bar{c}\lambda 2^\beta) + M_1(\bar{c}\lambda 2^\beta)) \leq \lambda\underline{c}/2$ and $\bar{c}\lambda 2^\beta \leq t_0$. Together with (A.13),

$$E_i(I) \leq \lambda\underline{c}/2E^J(\hat{y}_{J,i} - \hat{y}_i^*)^2 \leq \lambda/2E_i\{E^J\psi(Y_i - \hat{y}_{J,i}) - \psi(Y_i - \hat{y}_i^*)\}.$$

It follows that

$$E_i[\log E^J\exp(-\lambda\psi(Y_i - \hat{y}_{J,i}))] \leq -\lambda E_i\psi(Y_i - \hat{y}_i^*) + \lambda E_i[\psi(Y_i - \hat{y}_i^*) - E^J\psi(Y_i - \hat{y}_i^*)]$$
$$+ \lambda/2E_i[E^J\psi(Y_i - \hat{y}_i^*) - \psi(Y_i - \hat{y}_i^*)]$$
$$\leq -\lambda E_i\psi(Y_i - \hat{y}_i^*),$$

where the last step follows from $E^J\psi(Y_i - \hat{y}_i^*) \geq \psi(Y_i - \hat{y}_i^*)$ by Jensen's inequality based on convexity of ψ. Together with (A.9) and (A.8), we have

$$-\lambda E\sum_{i=1}^n \psi(Y_i - \hat{y}_i^*) \geq -E\log(1/q^n) \geq -\log(1/\pi_{j^*}) - \lambda\sum_{i=1}^n E\psi(Y_i - \hat{y}_{j^*,i}).$$

Because j^* is arbitrary, we have

$$\sum_{i=1}^n E\psi(Y_i - \hat{y}_i^*) \leq \inf_{j\geq 1}\left(\frac{\log(1/\pi_j)}{\lambda} + \sum_{i=1}^n E\psi(Y_i - \hat{y}_{j,i})\right).$$

This completes the proof of Theorem 5. ■

Proof of Theorem 6. We consider first the case when $M < \sqrt{n}$. Recall $\Theta_M = \{\theta = (\theta_1, \ldots, \theta_M) : \sum_{j=1}^M |\theta_j| \le 1\}$. Let N_ϵ be an ϵ-net in Θ_M under the l_1^M distance; i.e., for each $\theta \in \Theta_M$, there exists $\theta' \in N_\epsilon$ such that $\|\theta - \theta'\|_1^M = \sqrt{\sum_{j=1}^M |\theta_j - \theta_j'|} \le \epsilon$. An ϵ-net N_ϵ in Θ_M yields a suitable net in the set $\mathbf{F} = \{\delta^\theta : \theta \in \Theta_M\}$ of the linear combinations of the original forecasts. Let $\hat{y}_1, \ldots, \hat{y}_M$ denote the original forecasts by the M procedures at time i. Let F_ϵ be the set of the linear combinations of the forecasts $\hat{y}_1, \ldots, \hat{y}_M$ with coefficients in N_ϵ. Then for any $\hat{y} = \sum_{j=1}^M \theta_j \hat{y}_j$ with $\theta \in \Theta_M$, there exists $\theta' \in N_\epsilon$ such that

$$|\hat{y} - \sum_{j=1}^M \theta_j' \hat{y}_j| = |\sum_{j=1}^M (\theta_j - \theta_j')\hat{y}_j| \le A_3 \|\theta - \theta'\|_1^M \le A_3 \epsilon, \tag{A.14}$$

where A_3 is an upper bound on the magnitude of the original forecasts. Now we combine all the procedures in F_ϵ using the AFTER algorithm given in Section 2 for Theorem 1 with the uniform prior weight $1/|N_\epsilon|$. Let $\hat{y}_i^*$, $i \ge 1$ denote the combined forecasts at time i. By Corollary 1, we have

$$\frac{1}{n}\sum_{i=1}^n E(m_i - \hat{y}_i^*)^2 \le \frac{C \log(|N_\epsilon|)}{n} + C \inf_{\theta \in N_\epsilon} \frac{1}{n}\sum_{i=1}^n E(m_i - \hat{y}_i^\theta)^2,$$

where C depends only on A_1, A_2, and A_3. Because F_ϵ is an $(A_3\epsilon)$-net in $\mathbf{F}$, by the triangle inequality, for any set of m_i, we have $\inf_{\theta \in N_\epsilon} E(\sum_{i=1}^n (m_i - \hat{y}_i^\theta)^2) \le 2 \inf_{\theta \in \Theta} E(\sum_{i=1}^n (m_i - \hat{y}_i^\theta)^2) + 2nA_3^2\epsilon^2$. It follows that

$$\frac{1}{n}\sum_{i=1}^n E(m_i - \hat{y}_i^*)^2 \le \frac{C \log(|N_\epsilon|)}{n} + 2C \inf_{\theta \in \Theta_M} \frac{1}{n}\sum_{i=1}^n E(m_i - \hat{y}_i^\theta)^2 + 2A_3^2 C\epsilon^2. \tag{A.15}$$

To get the best upper bound (in order), we need to minimize $\log(|N_\epsilon|) + 2A_3^2\epsilon^2 n$ when discretizing Θ_M. We apply a result on entropy number, i.e., the worst case approximation error with the best net of size of 2^k points. Let ϵ_k denote the entropy number of Θ_M under the l_1^M distance. From Shütt (1984, Theorem 1), when $k \ge M$, $\epsilon_k \le c2^{-k/M}$ for some constant c independent of k and M. Take

$$k = \frac{M(\log(n/M) + 2\log 2)}{2 \log 2}$$

(note that $k \ge M$). (Strictly speaking, we need to round up to make k an integer.) Then

$$\log(|N_\epsilon|) + 2A_3^2\epsilon^2 n \le \frac{M(\log(n/M) + 2\log 2)}{2\log 2} + \frac{(A_3 c)^2 M}{2} \le c' M \log(1 + n/M),$$

where c' depends only on A_3 and c. The upper bound in Theorem 6 for $M < \sqrt{n}$ then follows.

Now consider the other case: $M \ge \sqrt{n}$. The preceding argument leads to a suboptimal rate $M \log(1 + n/M)$ for $M \le n$. For this case, as a result of the l_1 constraint, the number of large coefficients is small relative to M when $M \gg \sqrt{n}$. Working with only the large coefficients can result in the optimal rate of convergence.

Note that for $\|\theta\|_1^M \le 1$, $|\sum_{j=1}^M \theta_j \hat{y}_j| \le A_3$. Then by a sampling argument (see, e.g., Barron, 1993, Lemma 1) for each m, there exist a subset $I \subset \{1, \ldots, M\}$ of size m and $\theta_I' = (\theta_i', i \in I)$ such that $|\sum_{j=1}^M \theta_j \hat{y}_j - \sum_{j \in I} \theta_j' \hat{y}_j| \le A_3/\sqrt{m}$. Taking $m^* = \sqrt{n/\log n}$,

we have $|\sum_{i=1}^{M} \theta_i \hat{y}_i - \sum_{i\in I} \theta_i' \hat{y}_i| \le A_3(\log n/n)^{1/4}$. Consider an ϵ-net in $B_I = \{\theta_I: \sum_{j\in I}|\theta_j| \le 1\}$ under the $l_1^{m^*}$ distance. Again by Schütt (1984, Theorem 1), taking $k = [m^*(\log(n/m^*) + 2\log 2)]/2\log 2$, the best ϵ-net has approximation accuracy $\epsilon \le c/2\sqrt{m^*/n}$. Then as in (A.14), we know that there exists θ_I'' in this ϵ-net such that $|\sum_{i\in I} \theta_i' \hat{y}_i - \sum_{i\in I} \theta_i'' \hat{y}_i| \le A_3 c/2\sqrt{m^*/n}$. Thus for each $\delta^\theta \in \mathbf{F}$, there exist $I^* \subset \{1,\dots,M\}$ of size m^* and θ_I'' such that

$$\left|\sum_{j=1}^{M} \theta_j \hat{y}_j - \sum_{j\in I^*} \theta_j'' \hat{y}_j\right| \le \frac{A_3(\log n)^{1/4}}{n^{1/4}} + \frac{A_3 c}{2n^{1/4}(\log n)^{1/4}} \le \frac{c''(\log n)^{1/4}}{n^{1/4}},$$

where c'' depends only on A_3 and c.

Now for each fixed subset $I \subset \{1,\dots,M\}$ of size m^*, discretize the linear coefficients as described earlier. Then (with the uniform prior weight) combine the corresponding linear combinations of the forecasting procedures in Δ. Then combine these (combined) procedures over all possible choices of I (there are $\binom{M}{m^*}$ many such I altogether) with the uniform prior weight. Let δ^M denote this final procedure and let $\Delta_I = \{\delta_i, i \in I\}$. Applying Corollary 1 twice, we have that

$$R(\delta^M;n) \le C\left(\inf_{\theta\in\Theta_M} \frac{1}{n}\sum_{i=1}^{n} E(m_i - \hat{y}_i^\theta)^2 + \frac{(n\log n)^{1/2}}{n} + \frac{m^*\log(n/m^*)}{n} + \frac{\log\binom{M}{m^*}}{n}\right)$$

$$\le C'\left(\inf_{\theta\in\Theta_M} \frac{1}{n} E\left(\sum_{i=1}^{n}(m_i - \hat{y}_i^\theta)^2\right) + \frac{\log M}{\sqrt{n\log n}}\right),$$

where the constants C and C' depend on A_1, A_2, and A_3. This completes the proof of Theorem 6. ■

Proof of Theorem 7. Let $X_i = (X_{i1},\dots,X_{iM})$, $i \ge 1$ be independent and identically distributed (i.i.d.) where $X_{i1},\dots,X_{iM}$ are independent and uniformly distributed on $[0,1]$. Assume $Y_i = f(X_i) + e_i$, where the errors e_i are independent and normally distributed with mean zero and variance 1. Let $\varphi_1(x), \dots, \varphi_M(x)$ be uniformly bounded orthonormal basis functions (e.g., trigonometric basis). Take δ_i, $i \ge 1$ to be the procedure that forecasts Y_i by $\varphi_i(X_i)$. For each $M = n^\tau$, consider the class of regression functions $\mathsf{F} = \{f_\theta(x_1,x_2,\dots,x_M) = \theta_1\varphi_1(x_1) + \cdots + \theta_M\varphi_M(x_M): \|\theta\|_1^M \le 1\}$. It is clear that $R(M_n;n;\Delta_{M_n}) = E(\sum_{i=1}^{n}(m_i - \hat{y}_i^{\theta^*})^2) = 0$ for $f \in \mathsf{F}$. For this case, because of the independence of the errors, prediction and regression are essentially identical. The conclusion of Theorem 7 then follows from Theorem 2 of Yang (in press). This completes the proof of Theorem 7. ■

The following lemma relates the Hellinger distance between two densities to their means and variances.

LEMMA 1. *Let p and g be two probability densities on the real line with respect to a measure ν, with means μ_p and μ_g, variances $0 < \sigma_p^2 < \infty$ and $0 < \sigma_g^2 < \infty$, respectively. Then*

$$d_H^2(p,g) \ge \frac{(\mu_p - \mu_g)^2}{2(\sigma_p^2 + \sigma_g^2) + (\mu_p - \mu_g)^2}.$$

When $d_H^2(p,g) \le \frac{1}{2}$, *we have*

$$(\mu_p - \mu_g)^2 \le 4(\sigma_p^2 + \sigma_g^2)\, d_H^2(p,g).$$

Proof. Let a be any constant. Note that

$$\begin{aligned}(\mu_p - \mu_g)^2 &= \left(\int (x-a)p\,d\nu - \int (x-a)g\,d\nu\right)^2 \\ &= \left(\int (x-a)(p-g)\,d\nu\right)^2 \\ &\le \left(\int (|x-a||\sqrt{p}+\sqrt{g}|)|\sqrt{p}-\sqrt{g}|\,d\nu\right)^2 \\ &\le \left(\int (x-a)^2(\sqrt{p}+\sqrt{g})^2\,d\nu\right)\int (\sqrt{p}-\sqrt{g})^2\,d\nu,\end{aligned}$$

where the last step follows from the Cauchy–Schwarz inequality. Observe that

$$\begin{aligned}&\int (x-a)^2(\sqrt{p}+\sqrt{g})^2\,d\nu \\ &\quad \le 2\int (x-a)^2(p+g)\,d\nu \\ &\quad = 4\int (x-a)^2\,\frac{(p+g)}{2}\,d\nu.\end{aligned}$$

By taking $a = (\mu_p + \mu_g)/2$ (the mean of $(p_n + g_n)/2$), we have that $\int (x-a)^2((p+g)/2)d\nu$ is the variance of the distribution $(p+g)/2$, which is denoted by $\sigma^2_{(p+g)/2}$. Then with some elementary calculations, we have that

$$\sigma^2_{(p+g)/2} = \frac{1}{2}(\sigma_p^2 + \sigma_g^2) + \frac{1}{4}(\mu_p - \mu_g)^2.$$

From the preceding analysis, we have

$$\frac{(\mu_p - \mu_g)^2}{2(\sigma_p^2 + \sigma_g^2) + (\mu_p - \mu_g)^2} \le d_H^2(p,g).$$

The second conclusion follows easily. This completes the proof of Lemma 1. ■

LEMMA 2. *Let* f *and* g *be two probability densities on the real line (with respect to a measure* ν*), both with mean* 0 *and variance* 1. *Suppose* f *has an information projection in the location and scale family* $F = \{(1/\sigma)g((x-\mu)/\sigma) : -\infty < \mu < \infty, \sigma > 0\}$. *Suppose that the density* g *satisfies Condition 4 and assume further that there exists a constant* $0 < c < 1$ *such that* $D(g_{\mu,\sigma}\|g) \ge c\min(\mu^2 + (\sigma-1)^2, 1)$ *holds for all* $-\infty < \mu < \infty$ *and* $\sigma > 0$, *where* $g_{\mu,\sigma} = (1/\sigma)g((x-\mu)/\sigma)$. *Then when* $D(f\|g) \le c/2$, *for* $|\mu| \le a$ *and* $1/b \le \sigma \le b$ *for some constants* $a > 0$ *and* $b > 1$, *we have*

$$D(f\|g_{\mu,\sigma}) \le C(D(f\|g) + (\sigma-1)^2 + \mu^2)$$

for some constant $C > 0$ depending only on c and $B_{s_0,T}$ with $s_0 = (1 - 1/\sqrt{2})/b$ and $T = (a + 1/\sqrt{2})/(1 - 1/\sqrt{2})$.

Proof. Under the assumption of information projection, we have the following Pythagorean decomposition:

$$D(f\|g_{\mu,\sigma}) = D(f\|g_{\mu^*,\sigma^*}) + D(g_{\mu^*,\sigma^*}\|g_{\mu,\sigma})$$

for some $-\infty < \mu^* < \infty$, and $\sigma^* > 0$. It follows that $D(f\|g_{\mu^*,\sigma^*}) \le D(f\|g)$ and $D(g_{\mu^*,\sigma^*}\|g) \le D(f\|g)$. Under the assumptions on g, the latter inequality implies that when $D(f\|g) \le c/2$, we have

$$(\mu^*)^2 + (\sigma^* - 1)^2 \le \frac{D(f\|g)}{c} \le \tfrac{1}{2}.$$

Then $|\mu^*| \le 1/\sqrt{2}$ and $1 - 1/\sqrt{2} \le \sigma^* \le 1 + 1/\sqrt{2}$. Under Condition 4 on g, with a simple linear transformation, we have

$$D(g_{\mu^*,\sigma^*}\|g_{\mu,\sigma}) = D(g\|g_{(\mu-\mu^*)/\sigma^*,(\sigma/\sigma^*)}) \le B_{s_0,T}\left(\frac{(\sigma^*-\sigma)^2}{(\sigma^*)^2} + \frac{(\mu^*-\mu)^2}{(\sigma^*)^2}\right),$$

where $s_0 = (1 - 1/\sqrt{2})/b$ and $T = (a + 1/\sqrt{2})/(1 - 1/\sqrt{2})$ (because $1/b(1 + 1/\sqrt{2}) \le \sigma/\sigma^* \le b/(1 - 1/\sqrt{2})$ and $|\mu^* - \mu|/\sigma^* \le (a + 1/\sqrt{2})/(1 - 1/\sqrt{2})$). By triangle inequality,

$$\frac{(\sigma^*-\sigma)^2}{(\sigma^*)^2} \le \frac{2(\sigma^*-1)^2 + 2(\sigma-1)^2}{(\sigma^*)^2} \le \frac{2(\sigma^*-1)^2 + 2(\sigma-1)^2}{(1-1/\sqrt{2})^2},$$

$$\frac{(\mu^*-\mu)^2}{(\sigma^*)^2} \le \frac{2(\mu^*)^2 + 2\mu^2}{(\sigma^*)^2} \le \frac{2(\mu^*)^2 + 2\mu^2}{(1-1/\sqrt{2})^2}.$$

Thus when $D(f\|g) \le c/2$, we have

$$\frac{(\sigma^*-\sigma)^2}{(\sigma^*)^2} + \frac{(\mu^*-\mu)^2}{(\sigma^*)^2} \le \frac{2\left(\frac{1}{c}D(f\|g) + (\sigma-1)^2 + \mu^2\right)}{(1-1/\sqrt{2})^2}.$$

Then

$$D(g_{\mu^*,\sigma^*}\|g_{\mu,\sigma}) \le c'(D(f\|g) + (\sigma-1)^2 + \mu^2),$$

where $c' = 2B_{s_0,T}/(c(1 - 1/\sqrt{2})^2)$. Together with earlier inequalities, we have that when $D(f\|g) \le c/2$,

$$\begin{aligned} D(f\|g_{\mu,\sigma}) &\le D(f\|g) + c'(D(f\|g) + (\sigma-1)^2 + \mu^2) \\ &\le (1 + c')(D(f\|g) + (\sigma-1)^2 + \mu^2). \end{aligned}$$

This completes the proof of Lemma 2. ■

For the next result, consider the setup in the introduction to this paper. Let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be a filtration. Assume that the errors e_i form a martingale difference relative to the filtration; i.e., for $i \geq 1$, e_i is $\mathcal{F}_i$ measurable and $E(e_i|\mathcal{F}_{i-1}) = 0$ almost surely. Clearly e_i has mean zero and conditional variance $v_i = E(e_i^2|\mathcal{F}_{i-1})$ for $i \geq 1$. Let δ be a forecasting procedure producing forecasts $\hat{y}_1, \hat{y}_2, \ldots$ at time 1,2, and so on. We "estimate" v_i by $\hat{v}_i = [1/(i-1)]\sum_{l=1}^{i-1}(Y_l - \hat{y}_l)^2$ prior to observing Y_i. The following proposition gives a property of this estimator.

PROPOSITION 3. *Assume that e_i have uniformly bounded fourth moments and $\sup_{i\geq 1} E|\hat{y}_i - m_i|^2 < \infty$. In addition, we assume that v_i converges to a positive random variable v with probability one. If δ is consistent (as defined in Section 2.5), then $\hat{v}_i - v_i$ converges in probability to zero.*

Remark. Clearly the conditions on the errors are satisfied if they are i.i.d. with a finite fourth moment.

Proof. Expanding squares, we have

$$\frac{1}{i-1}\sum_{l=1}^{i-1}(Y_l - \hat{y}_l)^2 = \frac{1}{i-1}\sum_{l=1}^{i-1}(m_l - \hat{y}_l)^2 + \frac{1}{i-1}\sum_{l=1}^{i-1}e_l^2 + \frac{2}{i-1}\sum_{l=1}^{i-1}(m_l - \hat{y}_l)e_l.$$

It follows then that

$$\begin{aligned} &P\left(\left|\frac{1}{i-1}\sum_{l=1}^{i-1}((Y_l - \hat{y}_l)^2 - v_l) - \frac{1}{i-1}\sum_{l=1}^{i-1}(m_l - \hat{y}_l)^2\right| \geq \epsilon\right) \\ &\quad \leq P\left(\left|\frac{1}{i-1}\sum_{l=1}^{i-1}(e_l^2 - v_l)\right| \geq \epsilon/2\right) + P\left(\left|\frac{2}{i-1}\sum_{l=1}^{i-1}(m_l - \hat{y}_l)e_l\right| \geq \epsilon/2\right) \\ &\quad \leq \frac{E\left(\sum_{l=1}^{i-1}(e_l^2 - v_l)\right)^2}{(i-1)^2\epsilon^2/4} + \frac{E\left(\sum_{l=1}^{i-1}(m_l - \hat{y}_l)e_l\right)^2}{(i-1)^2\epsilon^2/16}. \end{aligned}$$

Note that

$$E(e_l^2 - v_l) = E(E(e_l^2 - v_l)|\mathcal{F}_{l-1}) = 0$$

and for $l_1 > l_2$,

$$\begin{aligned} E(e_{l_1}^2 - v_{l_1})(e_{l_2}^2 - v_{l_2}) &= E(E(e_{l_1}^2 - v_{l_1})(e_{l_2}^2 - v_{l_2})|\mathcal{F}_{l_1-1}) \\ &= E((e_{l_2}^2 - v_{l_2})E(e_{l_1}^2 - v_{l_1})|\mathcal{F}_{l_1-1}) \\ &= 0. \end{aligned}$$

It follows that

$$E\left(\sum_{l=1}^{i-1}(e_l^2 - v_l)\right)^2 \leq \sum_{l=1}^{i-1}\mathrm{Var}(e_l^2) \leq (i-1)\sup_{1\leq l<\infty} Ee_l^4$$

and similarly

$$E\left(\sum_{l=1}^{i-1}(m_l-\hat{y}_l)e_l\right)^2 \le \sum_{l=1}^{i-1}(Ee_l^2)E(m_l-\hat{y}_l)^2 \le (i-1)\sqrt{\sup_{1\le l<\infty} Ee_l^4}\sup_{1\le l<\infty} E(m_l-\hat{y}_l)^2.$$

Therefore

$$P\left(\left|\frac{1}{i-1}\sum_{l=1}^{i-1}((Y_l-\hat{y}_l)^2-v_l)-\frac{1}{i-1}\sum_{l=1}^{i-1}(m_l-\hat{y}_l)^2\right|\ge\epsilon\right)$$
$$\le \frac{4\sup_{1\le l<\infty} Ee_l^4}{(i-1)\epsilon^2}+\frac{16\sqrt{\sup_{1\le l<\infty} Ee_l^4}\sup_{1\le l<\infty} E(m_l-\hat{y}_l)^2}{(i-1)\epsilon^2}. \qquad \textbf{(A.16)}$$

It follows that $[1/(i-1)]\sum_{l=1}^{i-1}(Y_l-\hat{y}_l)^2 - [1/(i-1)]\sum_{l=1}^{i-1}v_l - [1/(i-1)]\times\sum_{l=1}^{i-1}(m_l-\hat{y}_l)^2$ converges in probability to zero. Because δ is a consistent forecasting procedure, by Markov inequality, we have for any $\epsilon>0$,

$$P\left(\frac{1}{i-1}\sum_{l=1}^{i-1}\frac{(\hat{y}_l-m_l)^2}{v_l}\ge\epsilon\right)\le\frac{\frac{1}{i-1}\sum_{l=1}^{i-1}E\frac{(\hat{y}_l-m_l)^2}{v_l}}{\epsilon}\to 0 \quad \text{as } i\to\infty.$$

Together with the assumption that v_l converges almost surely to $v>0$, it can be easily shown that $[1/(i-1)]\sum_{l=1}^{i-1}(m_l-\hat{y}_l)^2\to 0$ in probability. Again because v_l converges to v with probability one, we have $[1/(i-1)]\sum_{l=1}^{i-1}v_l\to v$ and $[1/(i-1)]\sum_{l=1}^{i-1}v_l - v_i\to 0$ almost surely. Together with (A.16), we have

$$\frac{1}{i-1}\sum_{l=1}^{i-1}(Y_l-\hat{y}_l)^2-v_i\to 0 \quad \text{in probability.}$$

This completes the proof of Proposition 3.