MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRCS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

Solving the Forecast Combination Puzzle

David T. Frazier, Ryan Covey, Gael M. Martin, and Donald Poskitt

Working Paper No. 18/23
October 2023

Solving the Forecast Combination Puzzle

David T. Frazier*[1], Ryan Covey[1], Gael M. Martin[1], and Donald Poskitt[1]

[1]Department of Econometrics and Business Statistics, Monash University, Clayton VIC 3800, Australia

Abstract

The forecast combination puzzle is the commonly encountered empirical result whereby predictions formed by combining multiple forecasts in complex ways do not out-perform more naive, e.g. equally-weighted, approaches. While various solutions for the cause of the puzzle exist in the literature, these solutions are limited in their scope and applicability. In contrast, we demonstrate a general solution to the puzzle by showing that this phenomenon is a direct consequence of the methodology used to produce forecast combinations. In particular, we show that tests which aim to discriminate between the predictive accuracy of competing forecast combination strategies have low power, and can lack size control, leading to an outcome that favours the naive approach. In addition, we demonstrate that the low power of such predictive accuracy tests in the forecast combination setting can be completely avoided if more efficient strategies are used in the production of the combinations. We illustrate these findings both in the context of forecasting a functional of interest and in terms of predictive densities. A short empirical example using daily financial returns exemplifies how researchers can avoid the puzzle in practical settings.

Keywords: Optimal Forecast Combinations, Tests for Forecast Accuracy, Probabilistic Forecasting, Scoring Rules, S&P500 Forecasting, One-step Versus Two-step Estimation

JEL Classification: C18, C12, C53

*Corresponding author: david.frazier@monash.edu

1 Introduction

Since their inception (Stone, 1961 and Bates and Granger, 1969), forecast combination methods have garnered a dedicated following due to their flexibility and accuracy (Timmermann, 2006; Aastveit et al., 2019). Such methods also align with the zeitgeist of modern econometric thought, in that they are designed to accommodate the fact that not all data sources are created equal, and that the models we are working with in economics are at best an approximation to reality.

Generally, forecast combinations are constructed by producing forecasts for individual, or constituent models, and then combining them via some combination function or weighting scheme. *Point* forecast combinations are typically constructed by taking a weighted average of point forecasts produced by the constituent models (Bates and Granger, 1969; Stock and Watson, 2004; Timmermann, 2006; Smith and Wallis, 2009; Claeskens et al., 2016). In the case of *distributional* forecast combinations, two commonly used approaches are the linear opinion pool (Stone, 1961; Hall and Mitchell, 2007; Geweke and Amisano, 2011; Opschoor et al., 2017; Martin et al., 2021), and the beta-transformed linear opinion pool (Ranjan and Gneiting, 2010; Gneiting and Ranjan, 2013; Satopää et al., 2014; Baran and Lerch, 2018). The weighted average, linear pool and beta-transformed linear pool are all *combination functions* that map a set of forecasts, produced using constituent models, to a single forecast combination.

Even though the last fifty years has seen these methods rise to prominence among empirical forecasters (Makridakis et al., 2018; Thorey et al., 2018; Wang et al., 2018; Makridakis et al., 2020; Taylor, 2020), key issues regarding the use and abuse of the methods remain. One of the most interesting issues is the so-called 'forecast combination puzzle', which is a stylized fact that states that predictions produced using complicated combinations of different forecasts, e.g., via optimizing the weights in the combination using some criterion function that encapsulates some aspect of forecast accuracy, do not generally outperform

simpler procedures (such as equal-weighted combinations). For example, see Stock and Watson (2004), Smith and Wallis (2009), Makridakis et al. (2018, 2020) for empirical evidence of this phenomenon.

Explanations for the puzzle range from the increased sampling variability of complex weighting schemes (Stock and Watson, 2004; Claeskens et al., 2016), to the similar performance of equally-weighted and optimally-weighted combinations (Elliott, 2011), and to bias in the averages loss functions (Chan and Pauwels, 2018). However, all of the above explanations are specific to linear combinations of point forecasts evaluated according to mean squared forecast error. To date, there is no single universally accepted answer that sufficiently explains this puzzle across both point and distributional forecasts, or across different performance measures. See Graefe et al. (2014) and Wang et al. (2022) for a historical summary of the puzzle.

Herein, we analyze the forecast combination puzzle in general terms, and through the lens of tests of superior forecast accuracy (White, 2000; Hansen, 2005). We demonstrate that the standard two-step approach to producing forecast combinations results in tests that have *no power* against a large class of alternatives, including fixed-, random-, and drifting-weighting schemes. This result is a consequence of the following unintuitive feature of forecast combinations: when produced in the standard manner, the usual test statistic employed to gauge differences in forecasts is such that estimated combination weights *imparts no sampling variability* into the statistic at first-order. This finding extends and renders rigorous previous results documented in a variety of contexts, including by Stock and Watson (2004), and Smith and Wallis (2009). In particular, Smith and Wallis (2009) argue that "the forecast combination puzzle rests on a gain ... that has no practical significance", and our results make rigorous this statement by showing that, in general settings, tests of superior forecast accuracy cannot distinguish forecast accuracy across a large class of different combination schemes.

More generally, we show that in order for the difference between two (sets of) forecast combinations to be meaningful, the combination weights must be surprisingly disparate. Critically, the distance between two combination weights needed to yield a test with non-trivial power depends entirely on the chosen loss and the variability in the constituent model forecasts. Therefore, when comparing two forecast combinations obtained using *the same* constituent models *but different* combination weights, the variability in the constituent forecasting models can easily swamp large differences in the forecast distribution due to differences in the combination weights. Consequently, a statistically significant difference between two forecast combinations is unlikely to eventuate in practice unless we have: 1) large sample sizes, much larger than typically used in empirical applications; and 2) constituent model forecasts that have low variability, which ultimately requires that the estimators of the underlying parameters of the constituent models also have low variability. Consequently, in empirical applications that require the use of high-variance constituent model forecasts, it is *unlikely* that we will be able to detect differences amongst different forecast combination methods.

We demonstrate that the poor behavior of forecast accuracy tests in this setting results from the test statistic having a non-standard asymptotic distribution under the null hypothesis of no inferior predictive accuracy. That is, while the critical values for such tests are typically based on the standard normal distribution, in the case of forecast combination methods, an appropriately scaled version of the test statistic converges in distribution to a generalized chi-squared distribution. Furthermore, this result persists across a large class of forecast combination methods, including those with time-varying weights.

Lastly, we demonstrate that the forecast combination puzzle can be circumvented in cases where it is feasible to produce forecast combinations in a single step. That is, except in trivial cases, forecast combinations produced in a single step do not exhibit the forecast combination puzzle, and will always yield superior forecast accuracy over standard, or

equally-weighted, forecast combination schemes. In this way, we build on an extensive investigation of one- and two-step forecast combinations in Zischke et al. (2022), which also provides support for the one-step approach. We reiterate that it is the two-step approach to producing forecast combinations that is the standard approach adopted in the literature on (frequentist) forecast combinations and, hence, the reason why the combination puzzle has been so empirically prevalent. Revisiting the S&P500 returns density prediction example considered in Geweke and Amisano (2011), we show that: 1) the forecast combination puzzle is evident in their original example for certain classes of volatility models; and 2) the puzzle is entirely resolved if we use forecast combinations that are built in a single step.

2 A Brief Motivating Example

We first motivate our analysis by reconsidering the results and discussion in Smith and Wallis (2009), subsequently elaborated on by Claeskens et al. (2016), and demonstrating that the forecast combination puzzle extends far beyond their initial analysis.

2.1 Revisiting Smith and Wallis (2009) and Claeskens et al. (2016)

Our goal is to produce a point forecast for the random variable Y_t at time $t = T + 1$ using observed data $\{y_t : 1 \leqslant t \leqslant T\}$ and models f_j, $j = 1, 2$. Denote the point predictions from these models by $\tilde{y}_{jt}$, and define the linear combination of point predictions by $\tilde{y}_t^{\eta} = \eta\tilde{y}_{1t} + (1 - \eta)\tilde{y}_{2t}$ where $0 \leqslant \eta \leqslant 1$, and assume that the parameters underlying the models f_1 and f_2 have been estimated in a first stage. Forecast accuracy is measured via mean squared forecast error (MSFE).

Under the assumptions in Smith and Wallis (2009) and Claeskens et al. (2016), e.g., $\tilde{y}_{jt}$ is unbiased with constant variance, σ_j^2, and covariance with $\tilde{y}_{it}$, $\sigma_{ji}, j, i = 1, 2$, the optimal forecast combination weight η is obtained by solving $\min_{\eta\in[0,1]} \mathbb{E}\left\{y_t - \tilde{y}_t^{\eta}\right\}^2$ and is given by

$\eta^\star = [\sigma_2^2 - \sigma_{12}]/(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$.

Following notational conventions (see, e.g., West, 1996): for a sample of size $T+1$, we split this sample into R in-sample observations used for model fitting, and P out-of-sample observations for forecast evaluation, so that $R + P = T + 1$. Under this evaluation regime, Smith and Wallis (2009) use the out-of-sample MSFE,

$$\sigma_\eta^2 = P^{-1} \sum_{t=R+1}^{T+1} (y_t - \tilde{y}_t^\eta)^2,$$

to compare two different forecast combinations: 1) the equally-weighted combination, with $\eta = 1/2$; 2) the sample estimate, $\tilde{\eta}$, of the optimal combination, $\eta^\star$. The key finding of Smith and Wallis (2009) is that, under their assumptions, when $\eta^\star = 1/2$, i.e. when the optimal approach ('in population') is to actually equally weight the two forecasts,

$$\sigma_{\eta=1/2}^2 - \sigma_{\tilde{\eta}}^2 \approx -P^{-1} \sum_{t=R+1}^{T+1} (\tilde{y}_t^{\tilde{\eta}} - \tilde{y}_t^{\eta=1/2})^2 \leqslant 0.$$

That is, when the optimal combination weight is $\eta^\star = 1/2$ the additional noise introduced via the estimation of η produces a more variable forecast, so that the fixed-weight scheme displays superior performance. This finding leads Smith and Wallis (2009) to conclude that "The parameter estimation effect [of the weights] is not large, nevertheless it explains the forecast combination puzzle."

We note that the analysis of Claeskens et al. (2016), which is also conducted in the setting of point forecasting under MSE loss with unbiased individual forecasts, produces findings that are more general than, but ultimately commensurate with, those in Smith and Wallis (2009). In particular, Claeskens et al. (2016) find that equally-weighted combinations can lead to smaller MSE than combinations based on optimally-estimated weights. Claeskens et al. (2016) argue that this increase in MSE over the equal-weighted combination is due to the additional bias and variance in the forecast that occurs from estimating the combination weight, i.e., the parameter estimation effect alluded to by Smith and Wallis (2009).

2.2 Extending Smith and Wallis (2009) and Claeskens et al. (2016)

While the analyses in Smith and Wallis (2009) and Claeskens et al. (2016) are insightful, their findings are not generalizable to related situations. Both analyses are based on linear combinations of point forecasts, with unbiased constituent forecasts that do not depend on unknown parameters.[1] In addition, in the analysis of Smith and Wallis (2009), the optimal MSFE weight is assumed to be $\eta^\star = 1/2$, which coincides with the default (equally-weighted) combination that typically underpins the forecast combination puzzle. Finally, this analysis does not immediately extend to other loss functions.

However, there is now mounting evidence to suggest that the forecast combination puzzle extends beyond this stylized setup. To this end, we explore scenarios in which we allow for a range of values for $\eta^\star$, and a range of values for the fixed weight η that *differ* from $\eta^\star$, and we do so for both point forecast combinations *and* distributional forecast combinations, and where these forecast combinations are estimated using different loss functions.

Following Section 3.1 of Smith and Wallis (2009), consider that the true DGP is from the AR(2) family,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \ \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2),$$

and that our forecasts are based on a linear pool $f^{(t)}$ of two normal constituent distributional forecasts $f_1^{(t)}$ and $f_2^{(t)}$:

$$f_1^{(t)}(y) = N\{y; \gamma_1 y_{t-1}, 1\}, \quad f_2^{(t)}(y) = N\{y; \gamma_2 y_{t-2}, 1\}, \quad f^{(t)}(y) = \eta f_1^{(t)}(y) + (1-\eta) f_2^{(t)}(y),$$

where $N\{x; \mu, \Sigma\}$ denotes the normal pdf evaluated at x with mean μ and variance Σ, γ_1 and γ_2 are the parameters of the constituent models, and η is the weight assigned to the first model.

[1] It is perhaps more accurate to say that the impact of having to estimate unknown parameters in the constituent models does not feature in their analysis.

We will estimate the parameters of two forecast combinations. First, we consider the distributional forecast combination given by the linear pool above, and estimate the parameters $(\eta,\ \gamma_1,\ \gamma_2)$ by minimizing the log loss (or equivalently by maximizing the log likelihood). Second, we consider the point forecast combination given by the *expectation* of the distributional forecast combination, and estimate the parameters by minimizing the MSFE. The second combination is identical to the combination considered by Smith and Wallis (2009) and discussed in the previous section. Parameters are estimated in the standard two-step fashion, so that γ_1 and γ_2 are chosen to minimize the selected loss of the first and second constituent models, respectively, and η is estimated by minimizing the loss of the combination given the aforementioned estimates for γ_1 and γ_2. Section 3.1 discusses such optimal forecast combinations in more detail.

For a given (point or distributional) combination, the parameters ϕ_1, ϕ_2 and σ_ϵ^2 can be chosen so that a desired value of $\eta^\star$ is achieved (we refer to Appendix D.1.1 for details). We then test the null of *no inferior forecast accuracy* across a variety of benchmark forecasts constructed using a range of fixed weights $\eta \in \{0.25, 0.5, 0.75\}$ (which includes the equally-weighted benchmark $\eta = 0.5$) against the alternative optimally-weighted combination.

Figure 1 presents the rejection frequency (y-axis) of the test across each fixed weight (rows) in the case of the MSFE (for the point forecast combination, left-hand column) and the log loss (for the distributional forecast combination, right-hand column). We present these results for DGP parameter values corresponding to a variety of pseudo-true weights $\eta^\star \in \{0, 0.25, 0.5, 0.75, 1\}$ (colors) and across a grid of values for the sample size $T + 1$ (x-axis), comprising the $R = (T + 1)/2$ in-sample observations followed by $P = (T + 1)/2$ out-of-sample observations. For the sake of brevity, we leave a more detailed discussion of the specific implementation details for this exercise to Appendix D.1.

Analyzing each figure we see that, across both loss functions, the probability of rejection under the null hypothesis of no inferior forecast accuracy of the benchmark (that is, where

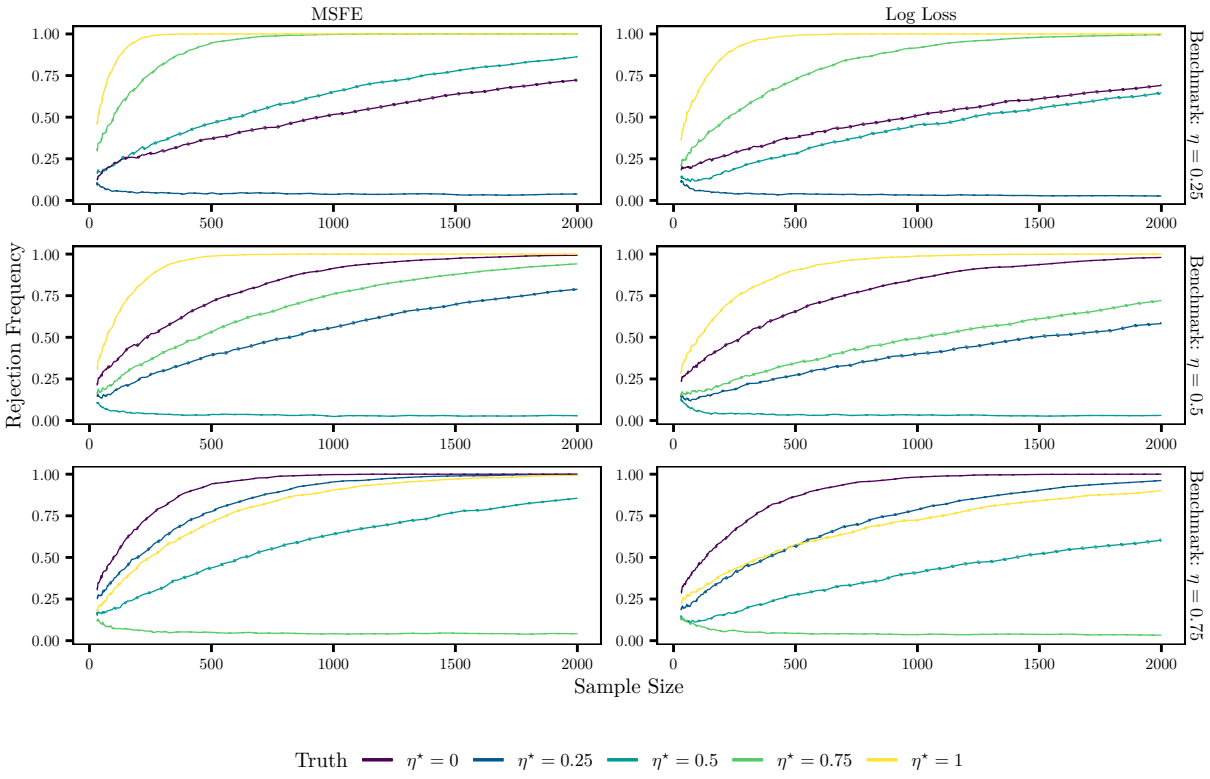

Figure 1: Estimates (solid) and their 95% confidence intervals (dotted) of the Rejection Frequency (y-axis) for the hypothesis test of no inferior predictive accuracy of a benchmark forecast combination with fixed weights (rows) against the alternative optimal forecast combination. The test is conducted with observations drawn from DGPs across a range of pseudo-true weights (colors), and across a grid of sample sizes (x-axis). Results for a point forecast combination with optimal weights minimizing the MSFE are given in the first column, and results for a distributional forecast combination with optimal weights minimizing the log loss are given in the second column.

the $\eta^\star$ value of the color equals the η value of the row) is (virtually) zero, and certainly less than 0.05, the nominal size of the test. Analyzing the rejection rates in all panels, which are given by the cases where the η values of the color are different from the $\eta^\star$ value of the row, suggests that the forecast combination puzzle persists even when the distance $|\eta^\star - \eta|$ between $\eta^\star$ and the incorrect fixed weight η is large. For example, a test of no inferior predictive accuracy of the equally-weighted combination (middle row) against the optimally-weighted combination where the truth is $\eta^\star = 0.25$ (dark blue) has a rejection

frequency smaller than 50% across all sample sizes less than 1000, whether we are using the MSFE (left-hand panel) or the log loss (right-hand panel). Hence, the power of such tests may be quite low in practice even if there are meaningful differences between forecasts.

3 The Accuracy of 'Optimal' Forecast Combinations

Now using more formal notation, let $Y_1, \cdots, Y_T, \ldots$ denote a sequence of random variables generated from the probability triple $(\mathsf{Y}, \mathscr{A}, G)$, where $\mathsf{Y} \subseteq \mathbb{R}^d$, $d \geqslant 1$. Since the true measure G is unknown in general, we postulate a class of models $\mathcal{Q}$ on Y, which we identify by their distribution functions $Q \in \mathcal{Q}$.

Given an observed sample, $y_1, \ldots, y_{T+1}$, our goal is to predict some feature of Y_{T+h} at a forecast horizon of $h \geqslant 1$. Since the true model is unknown, there is no single source of truth with which to predict features of interest. In such cases, entertaining multiple models is a valid approach that can produce reliable predictions, with the most common approach being to produce forecast combinations, as we have highlighted.

3.1 Defining 'Optimal' Forecast Combinations

Denote by $F_{\gamma_1} : \mathsf{Y} \times \Gamma_1 \to C[0,1]$ a probability measure on $(\mathsf{Y}, \mathscr{A})$ indexed by the parameter $\gamma_1 \in \Gamma_1 \subseteq \mathbb{R}^{d_{\gamma_1}}$ and itself lying in the family $\mathcal{F}(\Gamma_1) := \{F_{\gamma_1} : \gamma_1 \in \Gamma_1\}$. For any $1 \leqslant n \leqslant T$, let Ω_n denote the information available to the forecaster at time n, and denote the predictive measure based on time-n information as $F_{\gamma_1}^{(n)} := F_{\gamma_1}(\cdot : \Omega_n)$.

In the majority of forecasting settings, the practitioner entertains a collection of $K < T$ possible statistical models that can each describe, with varying accuracy, the movements of the stochastic process $\{Y_t : t \leqslant T\}$. We consider that each model is specified using a (semi-) parametric family $\mathcal{F}(\Gamma_j)$, which depends on $\Gamma_j \subseteq \mathbb{R}^{d_{\gamma_j}}$ unknown parameters for each $j = 1, \ldots, K$. For $\mathcal{M} = \times_{j=1}^{K} \mathcal{F}(\Gamma_j)$, $\Gamma_j \subseteq \mathbb{R}^{d_{\gamma_j}}$, for each $j = 1, \ldots, K$, denoting the

collection of all K constituent models, we can combine these models to produce a forecast combination. To this end, and following Gneiting and Ranjan (2013), we consider the combination function $C_\eta : \mathsf{Y} \times \mathcal{M} \times \mathcal{E} \to C[0,1]$, where $(F_{\gamma_1}, \cdots, F_{\gamma_K}) \mapsto C_\eta(F_{\gamma_1}, \cdots, F_{\gamma_K})$, and $\eta \in \mathcal{E} \subseteq \mathbb{R}^{d_\eta}$. Common choices for the combination family $\mathcal{C} := \{C_\eta : \eta \in \mathcal{E}\}$ include the linear pool (see, e.g., Geweke and Amisano, 2011), the Beta-transformed linear pool (Gneiting and Ranjan, 2013), and their nonparametric extension in Bassetti et al. (2018).

Given the family of combination functions, $\mathcal{C}$, and the member model family $\mathcal{M}$, the class of probability measures used for prediction is the composition of the two: define $\gamma := (\gamma_1', \ldots, \gamma_K')'$, $\theta := (\eta', \gamma')'$, $\Theta := \mathcal{E} \times \Gamma_1 \times \cdots \times \Gamma_K$ and consider $Q_\theta : \mathsf{Y} \times \mathcal{C} \times \mathcal{M} \to C[0,1]$ defined by $Q_\theta = C_\eta \circ M_\gamma$. We denote by $\mathcal{Q}$ the class $\{Q_\theta : \theta \in \Theta\}$.

Generally, the parameters of Q_θ are unknown and must be estimated. Throughout, we consider that the forecaster wishes to obtain 'optimal' forecasts in the spirit of Gneiting and Raftery (2007), Gneiting and Ranjan (2011), and Martin et al. (2021). Herein, we take optimal to mean that the distributions we choose to take out-of-sample are produced by targeting a loss function that measures precisely the features of Y_{T+h} that are of interest. Following Gneiting and Ranjan (2011), we consider a decision-theoretic framework for such forecasts. Let $\mathsf{Y} \subseteq \mathbb{R}^d$ denote the *observation domain*, and for some $k \geqslant 1$ let $\mathsf{A} \subseteq \mathbb{R}^k$ denote the *action space*. We consider that the forecaster is interested in measuring either the predictive accuracy for a functional of the distribution of Y_{T+h}, or the entire distribution.

3.1.1 Consistent Scoring Functions

Recall that $\mathcal{Q}$ is a class of distributions on Y, and consider a functional $U : \mathcal{Q} \mapsto \mathsf{A}, \quad Q \mapsto U[Q] \subseteq \mathsf{A}$ that maps a distribution $Q \in \mathcal{Q}$ to a subset $U[Q]$ of the action space. A *scoring function* is a measurable map $S : \mathsf{A} \times \mathsf{Y} \to [0, \infty)$. We orient the scoring functions so that a lower score implies a more accurate forecast. The scoring function $S(\cdot, \cdot)$ *is* $\mathcal{Q}$*-consistent*

for a functional $U[\cdot]$ if

$$\text{for all } x \in \mathsf{A},\ Q \in \mathcal{Q}: \quad \mathbb{E}_Q\left\{S\left(U[Q], Y\right)\right\} \leqslant \mathbb{E}_Q\left[S(x, Y)\right],$$

where Y is a random variable with distribution Q, and here and throughout we assume that any stated expectation exists and is finite. We say that $S(\cdot, \cdot)$ is $\mathcal{Q}$-*strictly consistent* for $T[\cdot]$ if $\mathbb{E}_Q\left\{S\left(U[Q], Y\right)\right\} = \mathbb{E}_Q\left[S(x, Y)\right] \implies x = U[Q]$. If $U[\cdot]$ admits a strictly consistent scoring function, then it is called *elicitable*. Here and throughout, we consider that the functional we are interested in is *elicitable*. For several functionals such as mean, quantiles and expectiles, this is the case, but there are many functionals, such as variance or expected shortfall, that are not elicitable, at least on their own. Some of these functionals are elicitable jointly with others, which is the case for the pairs (mean, variance) and (value at risk, expected shortfall), for example (Fissler and Ziegel, 2016).

Assume the forecaster is interested in the case where $U[\cdot]$ is a particular functional of the distribution of the random variable Y_{T+h}, and we have available information Ω_T. Then, we are interested in generating predictions at time $T + h$, $h \geqslant 1$, for the functional $U[F^{(T)}_{Y_{T+h}}]$, where we recall that $F^{(T)}_{Y_{T+h}}$ signifies the distribution of the random variable Y_{T+h} conditional on information Ω_T. To this end, we follow, among others, Patton (2020) and assume that any particular combination model $Q_\theta \in \mathcal{Q}$ admits a model for $U[F^{(T)}_{Y_{T+h}}]$ of the form $m(Z_T; \theta) = U[Q^{(T)}_\theta]$, where $Z_T \in \Omega_T$ denotes observable variables in the conditioning set and $m : \Omega_T \times \Theta \to \mathsf{A}$ is known up to the unknown θ. We can then produce point forecasts for $U[F^{(T)}_{Y_{T+h}}]$ using $m(Z_T; \theta)$, and by replacing the unknown θ with

$$\widehat{\theta}_T := \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{t=1}^{T} S\left[m(Z_t; \theta), y_{t+1}\right].$$

3.1.2 Proper Scoring Rules

The above approach allows one to produce distributions that generate reliable 'point forecasts' for a given functional, but there is no reason for these distributions to be accurate

in any other respect. In cases where we want distributional forecasts that are accurate as complete representations of the uncertainty surrounding an unobserved random variable, we can estimate our parameters according a proper scoring rule.

A negatively-orientated *proper scoring rule* is a function $S : \mathcal{Q} \times \mathsf{Y} \mapsto \mathbb{R}$ such that

$$\mathbb{E}_G \left[S(G, Y) \right] \leqslant \mathbb{E}_G \left[S(Q, Y) \right],$$

for all $Q, G \in \mathcal{Q}$. A *strictly* proper scoring rule is a proper scoring rule that is minimized by G alone.[2] While scoring rules can be used to measure the accuracy of the predictive distribution or densities, they can also be used to measure the accuracy of certain features of the distribution: e.g., quantiles, or predictive intervals (see, Gneiting and Raftery, 2007). Throughout the remainder, when we refer to scoring rules, it is meant that the action space is either the class of densities/distributions and not some functional of it such as quantiles or intervals; we keep this distinction as accuracy for such latter quantities can be readily measured using (consistent) scoring functions.

Similar to the case of scoring functions, producing density forecasts requires estimating θ in the combination model $Q_\theta \in \mathcal{Q}$. In this case, we can define an estimator of this parameter by minimizing the expected scoring rule over our observed sample

$$\widehat{\theta}_T := \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{t=1}^{T} S[Q_\theta^{(t)}, y_{t+1}].$$

3.1.3 Estimating Forecast Combinations

Regardless of whether one is producing 'point' forecasts of a functional or distributional forecasts, producing the 'optimal predictive' is categorically the same. Therefore, to simplify the presentation, we jointly treat both approaches. Given a collection of realized

[2]Scoring rules have a deep connection to decision theory, and we do not review this literature here (see, e.g., Pesaran and Skouras, 2002 and Granger and Machina, 2006 for a discussion in the context of economic and financial forecasting).

observations $\{y_t : 1 \leqslant t \leqslant n\}$, $n \leqslant T$, we search for the most accurate combination predictive distribution $Q_\theta \in \mathcal{Q}$ under the given loss function $L : \mathcal{Q} \times \mathsf{Y} \to \mathbb{R}$.

The optimal forecast combination can generally be produced in two possible ways. In cases where it is feasible to estimate the parameters jointly, the combination predictive can be produced by estimating the unknown model parameters,

$$\widehat{\theta}_n := \underset{\theta \in \theta}{\operatorname{argmin}}\ L_n(\theta) \equiv \underset{\theta \in \theta}{\operatorname{argmin}} \sum_{t=1}^{n} \ell_{t+1}(\theta), \text{ where } \ell_{t+1}(\theta) := L[Q_\theta^{(t)}, y_{t+1}], \tag{1}$$

and the forecast distribution at time $T + h$ can be taken as $Q_{\widehat{\theta}_n}^{(T)}$. However, due to the dimensionality of θ, it is often difficult to estimate θ jointly. Generally, forecast combinations are estimated in two steps: first, the constituent model forecasts are produced, and then the combination weights are estimated conditional on the constituent model forecasts; see, e.g., Hall and Mitchell (2007); Geweke and Amisano (2011); Gneiting and Ranjan (2013). Throughout we assume that such two-step forecast combinations are carried out in the spirit proposed in Gneiting and Raftery (2007). Namely, for each constituent model we estimate γ_j, $j = 1, 2, \cdots, K$, via the estimator

$$\tilde{\gamma}_{jn} = \underset{\gamma_j \in \Gamma_j}{\operatorname{argmin}} \sum_{t=1}^{n} L[Q_\theta^{(t)}, y_{t+1}]. \tag{2}$$

Collecting the $\tilde{\gamma}_{jn}$ into $\tilde{\gamma}_n := (\tilde{\gamma}_{1n}', \ldots, \tilde{\gamma}_{Kn}')'$, we can then estimate the combination parameters η by maximizing $Q_\theta \mapsto L_n(\theta)$, conditional on $\gamma = \tilde{\gamma}_n$, which yields

$$\tilde{\theta}_n = (\tilde{\eta}_n', \tilde{\gamma}_n')' = \underset{\theta \in \Theta}{\operatorname{argmin}}\ L_n(\eta, \gamma) \text{ s.t. } \gamma = \tilde{\gamma}_n. \tag{3}$$

Once $\tilde{\theta}_n$ have been calculated the predictive distribution $Q_{\tilde{\theta}_n}^{(T)}$ can be used to produce predictions at time $T + h$.

Throughout the remainder, we refer to the predictive distributions $Q_{\widehat{\theta}_n}^{(T)}$ and $Q_{\tilde{\theta}_n}^{(T)}$ as the one-step and two-step predictive combinations, respectively.

Under standard regularity conditions, the extremum estimators $\widehat{\theta}_n$ and $\tilde{\theta}_n$ will converge to well-defined probability limits, which we denote by $\theta^0 := (\eta^{0\prime}, \gamma^{0\prime})'$, and $\theta^\star := (\eta^{\star\prime}, \gamma^{\star\prime})'$,

respectively;[3] i.e., under our regularity conditions we will have

$$\widehat{\theta}_n \to_p \theta^0 := (\eta^{0'}, \gamma^{0'})' \text{ and } \tilde{\theta}_n \to_p \theta^\star := (\eta^{\star'}, \gamma^{\star'})', \tag{4}$$

where $\to_p$ denotes convergence in probability. It is well-known, see Newey and McFadden (1994) for details, that one-step and two-step estimators do not coincide in general, so that throughout we assume that $\theta^0 \neq \theta^\star$. Indeed, in most empirical settings significant differences between one- and two-step estimators exist, and so we restrict our attention to this case. For a general discussion on two-step estimation see Newey and McFadden (1994); for a more modern treatment see Frazier and Renault (2017); and for a particular approach to measuring the impact of two-step estimation in the context of distributional forecast combinations see Zischke et al. (2022).

3.2 Testing the Accuracy of Forecast Combinations

3.2.1 Evaluation Scheme

Following West (1996), White (2000), and many others, we measure out-of-sample predictive accuracy using loss differences of forecasts over a given out-of-sample period. For simplicity, and for consistency with the earlier exposition of the analysis, we let $h = 1$ denote the horizon over which we will make predictions, but note that our results can also accommodate $h \geqslant 1$ at the cost of additional notation. For ease of exposition, we re-introduce the notation introduced in Section 2.1, and assume the sample consists of $T+1$ total observations, which we partition into R in-sample periods, and P out-of-sample periods, across which we evaluate the predictions, where $R + P = T + 1$.

In the approach of West (1996), the information used to estimate θ is increased by one unit for each prediction; i.e., at time $t = R+1$, R observations are used to estimate θ, at time $t = R+2$, $R+1$ observations are used, and so forth. This formulation is useful as it allows

[3]For more precise definitions of θ^0 and $\theta^\star$, we refer the interested reader to Appendix B.

one to update the parameter estimates as new information becomes available. However, the resulting out-of-sample loss difference at time $T+1$ is then a complex combination of all previous estimators, and also has variability due to the P out-of-sample observations themselves. Consequently, disentangling variability due to parameter uncertainty, from the innate variability of the average loss difference becomes difficult, and obtaining clear intuition regarding the contribution to each of these pieces to the behavior of the out-of-sample average loss difference becomes difficult.

Furthermore, it is critical for us to understand the precise impact of parameter uncertainty on forecast accuracy since this effect, while not large, "explains the puzzle" according to Smith and Wallis (2009). Therefore, we consider a simple framework that cleanly dissects the two types of sampling variability, loss and parameter estimation. This is accomplished by estimating the unknown parameters once using R observations, with the resulting estimators then held fixed over the P out-of-sample periods;[4] the P out-of-sample periods are then used for evaluation only; further, we will also maintain that the in-sample period, R, and the out-of-sample period, P, are in rough proportion.[5]

Assumption 1 (Maintained)**.** *$R, P \to \infty$ as $T \to \infty$, and $0 < c := \lim_T R/P < \infty$.*

3.2.2 Tests of Forecasting Accuracy

Following White (2000) and Hansen (2005), we measure the accuracy of forecasts by testing the null hypothesis of *no inferior forecast performance* using the average loss difference over the P out-of-sample periods. Consider that we wish to test the accuracy of a benchmark forecast distribution $Q^{(T)}_{\vartheta^b}$, indexed by unknown parameters ϑ^b, against an alternative

[4]More specifically, instead of the first prediction being based on an estimator θ_R, obtained using observations $1, \ldots, R$, and the next based on θ_{R+1}, and so forth, we only consider estimators based on R observations (with $R \to \infty$ as $T \to \infty$).

[5]We refer to Section 8 of Clark and McCracken (2013) for a discussion on the benefits and disadvantages of various splitting schemes for forecast evaluation.

distribution $Q_{\theta^a}^{(T)}$, indexed by unknown parameters θ^a. The null hypothesis of *no inferior forecast accuracy* of the benchmark ($Q_{\vartheta^b}^{(T)}$) over the alternative ($Q_{\theta^a}^{(T)}$) is

$$\mathrm{H}_0 : \mathbb{E}(L[Q_{\vartheta^b}^{(T)}, Y_{T+1}]) \leqslant \mathbb{E}(L[Q_{\theta^a}^{(T)}, Y_{T+1}]). \tag{5}$$

To test the null in (5), we approximate the above expectation, and the unknown ϑ^b, θ^a, using sample counterparts. For any $t > R$, and any sequence of consistent estimators $\vartheta_R, \theta_R \in \Theta$ of ϑ^b, θ^a, respectively, define the average loss difference statistic

$$\Delta_P(\vartheta_R, \theta_R) := P^{-1} \sum_{t=R+1}^{T+1} d_t(\vartheta_R, \theta_R), \quad d_t(\vartheta_R, \theta_R) = \ell_t(\vartheta_R) - \ell_t(\theta_R),$$

where we recall that $\ell_t(\theta) = L[Q_\theta^{(t-1)}, y_t]$. The null hypothesis in (5) can then be tested using the standardised statistic

$$D_P(\vartheta_R, \theta_R) := \hat{\Omega}_R^{-1/2} \sqrt{P} \Delta_P(\vartheta_R, \theta_R), \tag{6}$$

where $\hat{\Omega}_R$ is a consistent estimator of the asymptotic variance of $\Delta_P(\vartheta^b, \theta^a)$.

4 Solving the Forecasting Combination Puzzle

Before stating and proving our general results, we note here the following notations used throughout the remainder of the paper. For a probability measure P, a random variable X and a random sequence X_n, we write $\mathbb{E}_P[X]$ to denote the expectation of X under P, $\operatorname{plim}_n X_n$ to denote the probability limit of X_n as $n \to \infty$ (if it exists), and $X_n \Rightarrow X$ if X_n converges in distribution to X. For some positive sequence R_n, the notations $X_n = o_p(R_n)$ and $X_n = O_p(R_n)$ have their usual definitions, see Van Der Vaart (1998) for a textbook treatment. We say that $X_n \asymp R_n$ if there exists constants c and C such that $cR_n \leqslant X_n \leqslant CR_n$ for all n large enough (with probability one). The gradient and hessian of a functional f of x is written $\nabla_x f(x)$ and $\nabla_{xx}^2 f(x)$, respectively, where for x on the boundary of the domain of f these symbols denote the left or right derivatives, whichever of those exist at x.

4.1 A General Phenomenon

The results in Section 2.2 demonstrate that even when the optimal combination weight (in population) is very different from a fixed, hypothetical combination weight, the standard approach to testing for differences in forecasting accuracy does not result in meaningful rejection rates; i.e., even though the hypothetical combination weight is inferior to the optimal combination weight, the resulting testing procedure does not reliably detect differences. This finding holds across two particular loss functions and a host of different optimal combination weights. In this section, we demonstrate that this phenomenon is present in any class of forecast combinations produced in the standard, i.e., two-step, manner. Thus we give, for the first time, a truly generic explanation for the puzzle that is agnostic to the chosen loss, and which is valid under standard regularity conditions.

To state this result, consider the setting where we are given *a known, i.e., hypothesised* combination weight η_T^δ, which we can always represent as

$$\eta_T^\delta := \eta^\star + \delta_T,$$

where $\eta^\star$ is as defined in (4), and $\eta^\star$ and δ_T are individually unknown - all that is known is η_T^δ. Consider that $\{\delta_T : T \geqslant 1\}$ evolves according to the following scheme: for $\delta \in \mathcal{E} \subset \mathbb{R}^{d_\eta}$ a bounded vector,

$$\delta_T = \delta/T^\xi, \text{ where } \xi \in [0, 1/4), \quad \xi = 1/4, \quad \xi \in (1/4, \infty). \tag{7}$$

The above class of sequences will allow us to evaluate the behavior of the standard testing framework across a wide range of hypothesised combination weights η_T^δ. In particular, the case $\xi = 0$ yields fixed alternatives, while $\xi = 1/2$ yields the class of canonical Pitman sequences. As we shall see, the behavior of the test depends crucially on the value of ξ in (7).

Our benchmark forecast is $Q_{\vartheta^b}^{(T)}$ with $\vartheta^b = (\eta_T^\delta, \gamma^\star)$, and we wish to test the null hypothesis that this benchmark forecast has *no inferior forecast performance* relative to the

alternative forecast $Q^{(T)}_{\theta^a}$ with $\theta^a = (\eta^\star, \gamma^\star)$. The null hypothesis in (5) then becomes

$$\mathrm{H}_0 : \ \mathbb{E}(L[Q^{(T)}_{(\eta^\delta_T, \gamma^\star)}, Y_{T+1}]) \leqslant \mathbb{E}(L[Q^{(T)}_{(\eta^\star, \gamma^\star)}, Y_{T+1}]).$$

The null hypothesis H_0 is then tested using the statistic in (6), where the infeasible θ^a is replaced by the feasible estimator, $\tilde{\theta}_R = (\tilde{\eta}'_T, \tilde{\gamma}'_R)'$, and the infeasible ϑ^b by its feasible counterpart $\vartheta^\delta_R = (\eta^\delta_T, \tilde{\gamma}'_R)'$, where $\eta^\delta_T = \eta^\star + \delta_T$, with δ_T varying as in (7). The rejection region for the test is

$$W_P(\alpha) := \{D_P : D_P(\vartheta^\delta_R, \tilde{\theta}_R) > \Phi^{-1}(1-\alpha)\}, \quad \alpha \in (0,1), \tag{8}$$

where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard normal distribution. The following theorem describes the behavior of the test under the class of sequences in (7).[6]

Theorem 1. *Let Assumptions 1-4, in Appendix B be satisfied.*
(i) If $\delta_T = \delta/T^\xi$, *with* $\xi \in [0, 1/4)$, *and* $|\delta' \nabla_\eta \mathcal{L}(\eta^\star + \delta, \gamma^\star)| > 0$, *then* $\lim_P \Pr\{W_P(\alpha)\} = 1$.
(ii) If $\delta_T = \delta/T^{1/4}$, *then* $\lim_P \Pr\{W_P(\alpha)\} > \alpha$ *or* $\lim_P \Pr\{W_P(\alpha)\} < \alpha$, *depending on* δ.
(iii) If $\delta_T = \delta/T^\xi$, *with* $\xi \in (1/4, \infty]$, *then* $\lim_T \Pr\{W_P(\alpha)\} = 0$ *for all* $\alpha \in (0,1)$.

The above result implies that if two forecasts have combination weights that are at least $O(T^{-1/4+\varepsilon})$ apart, for $\varepsilon > 0$, then the standard testing approach can distinguish between the forecasts. Surprisingly, however, if the combination weights are at a distance of $O(T^{-1/4})$, the test can be arbitrarily over- or under-sized depending on the magnitude of $\{\delta_T : T \geqslant 1\}$. More surprisingly, if two sets of combination weights are within a distance of $O(T^{-1/4-\varepsilon})$ from one another, e.g., a parametric neighbourhood of width $O(T^{-1/2})$, then the test has no power to detect differences between the combination forecasts. The third

[6]It is possible to extend our results to the original framework of West (1996). However, the results are not as intuitive as those presented herein, and are much more cumbersome to dissect and interpret. Therefore, we adopt a fixed-windows estimation scheme to err on the side of simplicity and interpretability instead of technicality.

result also encompasses the case considered in the illustrative example in Section 2.1, in which $\eta_T^\delta = \eta^\star$ – i.e. the benchmark fixed weight (denoted by η therein) coincided with $\eta^\star$– and the test displayed zero empirical size (up to Monte Carlo error).

As an example of the phenomenon in Theorem 1(iii), consider that we have two competing combination forecasts defined by different combination weight schemes such that $\tilde{\eta}_1$ and $\tilde{\eta}_2$ have distinct asymptotic distributions, but $R^{1/2}(\tilde{\eta}_1 - \tilde{\eta}_2) = O_p(1)$, then the usual test of the null hypothesis of *no inferior predictive accuracy* will only detect differences between the two forecasts on rare occasions, and will detect no statistically significant differences between the forecasts with probability converging to one.[7]

Practically speaking, Theorem 1 demonstrates that even if the benchmark forecast, e.g., the equally weighted forecast, is far away from the optimally weighted combination forecast, then the standard testing approach is unlikely to reject the inadequacy of this benchmark (with probability converging to one). In particular, Theorem 1(i) demonstrates that the standard test will asymptotically reject the null only when $\sqrt{T}(\eta_T^\delta - \tilde{\eta}_R)$ diverges faster than $T^{1/4}$, i.e., when $\xi < 1/4$. Therefore, for all intents and purposes, the standard approach to testing for differences in combination forecasts cannot be trusted to deliver reliable conclusions in the majority of empirical situations where it is applied.

Part (ii) of Theorem 1 is non-standard: in classical hypothesis testing, drifting sequences of alternatives do not generally yield a consistent test, but they display at least *some power* against such hypotheses. The lack of power in this case is entirely a consequence of the two-step nature with which the combinations are produced.

[7]We recall that under the maintained assumption on R, P, T, we have that $R \asymp P \asymp T$, so that $O(R^{1/2}/T^{1/2}) = O(1)$.

4.2 The Root of the Puzzle

Theorem 1 demonstrates that when comparing between forecast combinations, tests of forecast accuracy behave in non-standard ways. However, it is important to understand the mechanism causing this behavior. Recall that standard forecast combinations are produced in two steps: first, we estimate the unknown model parameters γ via $\tilde{\gamma}_R$, then the combination weights are estimated via $\tilde{\eta}_R$. In this section, we show that the two-step nature by which the forecast combination $Q_\theta^{(T)}$ is produced results in an average loss difference whose limiting distribution is non-standard when $\xi > 1/4$.

To state the asymptotic distribution of the test statistic $\Delta_P(\vartheta^b, \theta^a)$ define

$$\mathcal{L}(\theta) := \operatorname*{plim}_{P\to\infty} L_P(\theta)/P, \quad \mathcal{M}_{\eta\eta} := \nabla_{\eta\eta}\mathcal{L}(\eta^\star, \gamma^\star), \quad \mathcal{M}_{\gamma\eta} := \nabla_{\gamma\eta}\mathcal{L}(\eta^\star, \gamma^\star),$$

$$V_{P,R} := -\mathcal{M}_{\eta\eta}^{-1/2}\{\nabla_\eta L_P(\eta^\star, \gamma^\star)/P + \mathcal{M}_{\eta\gamma}(\tilde{\gamma}_R - \gamma^\star)\}.$$

We note that the above exist under Assumptions 1-4 in Appendix B, and that $V_{P,R} = O_p(1/\sqrt{P})$.

Lemma 1. *Let Assumptions 1-4 in Appendix B be satisfied. If $\delta_T = \delta/T^\xi$, with $\xi \in (0, \infty]$, then, for $\vartheta_R^\delta = (\eta_T^\delta, \tilde{\gamma}_R)$,*

$$\Delta_P(\vartheta_R^\delta, \tilde{\theta}_R) = \frac{1}{2}\|V_{P,R}\|^2 - \frac{1}{2}\|\mathcal{J}^{1/2}\left(\eta_T^\delta - \eta^\star\right) - \mathcal{J}^{1/2}\left(\tilde{\eta}_R - \eta^\star\right) - V_{P,R}\|^2 + o_p(\|(\eta_T^\delta - \eta^\star)\|^2 \vee \|\tilde{\eta}_R - \eta^\star\|).$$

The expansion in Lemma 1 clarifies the mechanism behind the behavior exhibited in Theorem 1. The behavior of the test is driven by the behavior of $\sqrt{P}\cdot\Delta_P(\vartheta_R^\delta, \tilde{\theta}_R)$, see equation (6), however, Lemma 1 makes clear that if the sequence $\{\delta_T : T \geqslant 1\}$ goes to zero fast enough, the limit distribution of $\sqrt{P}\cdot\Delta_P(\vartheta_R^\delta, \tilde{\theta}_R)$ is degenerate. That is, since $V_{P,R} = O(1/\sqrt{P})$ and $(\tilde{\eta}_R - \eta^\star) = O_p(1/\sqrt{P})$ under Assumptions 1-4, scaling $\Delta_P(\vartheta_R^\delta, \tilde{\theta}_R)$ by $\sqrt{P}$ results in a degenerate test statistic unless $\operatorname{plim}_{P\to\infty}\sqrt{P}\|(\eta_T^\delta - \eta^\star)\|^2 = \operatorname{plim}_{P\to\infty}\sqrt{P}\|\delta_T\|^2 > 0$.

When $\delta_T = \delta/T^{\xi}$, $\xi \in (1/4, \infty]$, we have that $\operatorname{plim}_P \sqrt{P}\|(\eta_T^{\delta} - \eta^{\star})\|^2 = 0$, which yields the result in part (iii) of Theorem 1. If instead we have $\delta_T = T^{-1/4}$, the behavior of $\sqrt{P}\Delta_P(\vartheta_R^{\delta}, \tilde{\theta}_R)$ is driven by the magnitude of

$$\frac{1}{2}P^{1/4}(\eta^{\star} - \eta_T^{\delta})'\mathcal{M}_{\eta\eta}P^{1/4}(\eta^{\star} - \eta_T^{\delta}) = \delta'\mathcal{M}_{\eta\eta}\delta + o_p(1),$$

which yields the second result in Theorem 1. The above term also drives the power of the statistic $D_P(\vartheta_R, \tilde{\theta}_R)$ in the case where $\delta_T \in [0, 1/4)$ (i.e., part (i) of Theorem 1).[8]

To obtain the limit distribution of the out-of-sample average loss difference under the null hypothesis in (5), based on the benchmark forecast combination $Q_{\vartheta^{(b)}}^{(T)}$ with $\vartheta^{(b)} = (\eta_T^{\delta}, \gamma^{\star})$, and with η_T^{δ} as in (7), we require a few additional definitions. Let $X_P := \nabla_{\eta}L_P(\eta^{\star}, \gamma^{\star})/P$, $Z_{R,\gamma} := (\tilde{\gamma}_P - \gamma^{\star})$, so that we can write $V_{P,R} = -\mathcal{M}_{\eta\eta}^{-1/2}(X_P + \mathcal{M}_{\eta\gamma}Z_{R,\gamma})$. Under Assumptions 1-4 in Appendix B, we have that $\sqrt{P}X_P \Rightarrow X \sim N(0, \Sigma_X)$, where $\Sigma_X := \lim_P \operatorname{Var}\{\nabla_{\eta}L_P(\eta^{\star}, \gamma^{\star})/\sqrt{P}\}$ and $\sqrt{R}Z_{R,\gamma} \Rightarrow Z_{\gamma} \sim N(0, \Sigma_{\gamma})$, and where

$$\Sigma_Z := \lim_{R\to\infty} \operatorname{Var}[\{[\nabla_{\gamma\gamma}\mathcal{L}(\gamma_1^{\star})]^{-1}\nabla_{\gamma_1}L_R(\gamma_1^{\star})/\sqrt{R}, \ldots, [\nabla_{\gamma\gamma}\mathcal{L}(\gamma_1^{\star})]^{-1}\nabla_{\gamma_K}L_R(\gamma_K^{\star})/\sqrt{R}\}'].$$

Recall that $c := \lim_T R/P$, with $0 < c < \infty$, and, by Assumption 5 in Appendix B, for some matrix Q_V, we have that $\mathcal{V} := (\mathcal{V}_1', \mathcal{V}_2')' = \mathcal{L}_{\infty}(c\sqrt{R}V_{R,R}, \sqrt{P}V_{P,R}) = N(0, Q_V)$ is the limit (joint) law of the terms $c\sqrt{R}V_{R,R}$ and $\sqrt{P}V_{P,R}$.

Corollary 1. *Under Assumptions 1-5, the following are satisfied.*
(i) If $\delta_T \asymp \delta/T^{\xi}$, with $\xi \in (0, 1/2)$, then $P \cdot \Delta_P(\vartheta_R^{\delta}, \tilde{\theta}_R) \to +\infty$, with probability converging to one.
(ii) If $\delta_T = \delta/T^{1/2}$, and $\theta^{\star} \in \operatorname{Int}(\Theta)$, then

$$P \cdot \Delta_P(\vartheta_R^{\delta}, \tilde{\theta}_R) \Rightarrow \frac{1}{2}\|X + c^{-1}\mathcal{M}_{\eta\gamma}Z_{\gamma}\|_{\mathcal{M}_{\eta\eta}^{-1}}^2 - \frac{1}{2}\|\{1/(1+c)\}^{1/2}\delta - (\mathcal{V}_1 - \mathcal{V}_2)\|^2.$$

(iii) If $\delta_T = \delta/T^{\xi}$, with $\xi \in (1/2, \infty]$, and if $\theta^{\star} \in \operatorname{Int}(\Theta)$, then

$$P \cdot \Delta_P(\vartheta_R^{\delta}, \tilde{\theta}_R) \Rightarrow \frac{1}{2}\|X + c^{-1}\mathcal{M}_{\eta\gamma}Z_{\gamma}\|_{\mathcal{M}_{\eta\eta}^{-1}}^2 - \frac{1}{2}\|\mathcal{V}_1 - \mathcal{V}_2\|^2.$$

[8] We note, however, that Lemma 1 is not valid, as stated, in the case where $\delta_T \in [0, 1/4)$ since the remainder term in Lemma 1 is no longer negligible, since $\sqrt{P}\|\eta_T^{\delta} - \eta^{\star}\|^2$ will diverge. Nonetheless, a similar argument to that used to prove Lemma 1 can be used to deal with this case.

When $\delta_T \asymp \delta/T^{\xi}$ and $\xi \in (0, 1/2)$, the result of Corollary 1(i) demonstrates that under the null hypothesis in (5) with benchmark forecast combination $Q^{(T)}_{\vartheta^b}$, $\vartheta^{(b)} = (\eta^{\delta}_T, \gamma^{\star})$, and alternative forecast combination $Q^{(T)}_{\theta^a}$, with $\theta^a = \theta^{\star}$, the test statistic $P \cdot \Delta_P(\vartheta^{\delta}_R, \tilde{\theta}_R)$, where $\vartheta^{\delta}_R = (\eta^{\delta}_T, \tilde{\gamma}_R)$, diverges. Hence, if $\eta^{\star}$ and η^{δ}_T are sufficiently different, we can accurately learn differences between competing forecast combination methods, albeit with a different scaling of the test statistic. In contrast, when $\delta_T \asymp \delta/T^{\xi}$, with $\xi \in [1/2, \infty]$, Corollary 1(ii)-(iii) demonstrates that the asymptotic distribution of $P \cdot \Delta_P(\vartheta^{\delta}_R, \tilde{\theta}_R)$ is non-standard.

In the regime where $\delta_T \asymp \delta/T^{\xi}$, with $\xi \in [1/2, \infty]$, the test statistic $P \cdot \Delta_P(\vartheta^{\delta}_R, \tilde{\theta}_R)$ converges in distribution to a random variable with two components. The first component is a generalized chi-squared random variable, which does not admit a closed-form formula for its density or distribution function.[9] The second component is itself possibly comprised of two components: so long as $\xi \geqslant 1/2$, the second term depends on the difference of two mean-zero but correlated normal random variables, which captures the behavior of (a scaled version of) the out-of-sample loss difference due to differences in the combination weights, i.e., the η-components, and a centering term that captures the difference between η^{δ}_T and $\eta^{\star}$.[10] In the regime where $\xi = 1/2$, an additional term is present that captures the fact that $\sqrt{P}(\eta^{\delta}_T - \eta_{\star}) = (\sqrt{P/T})\sqrt{T}(\eta^{\delta}_T - \eta^{\star}) = \sqrt{P/(P+R)}\delta$, which, since $c = \lim_T R/P$ converges to $(1/(1+c))^{1/2}\delta$.

However, it is important to note that since the first component of the asymptotic distribution is a generalized chi-squared random variable, even if the second component were not present it would still be infeasible to obtain closed-form quantiles for the null distribution of the test statistic. Furthermore, the distribution of the test statistic depends on the

[9]Since the quadratic form $\|X + \mathcal{M}_{\eta\gamma} Z_{\gamma}\|^2_{\mathcal{M}^{-1}_{\eta\eta}}$ cannot be re-written as a quadratic form with an idempotent weighting matrix, the distribution is not chi-squared.

[10]This term results from taking a second-order Taylor expansion of the loss difference, and grouping terms appropriately.

loss used in the analysis, the choice of constituent models, and the specific combination function chosen. As such, there is no hope for a single set of generally applicable critical values, and any simulated critical values will need to be application-specific.

While it is not feasible to obtain closed-form quantiles for the test statistic when $\xi \in (1/2, \infty]$, since the second component is always negative, it is possible to deduce a conservative test that uses just the critical values of the generalized chi-distribution. That is, since

$$\frac{1}{2}\|X + c^{-1}\mathcal{M}_{\eta\gamma}Z_\gamma\|^2_{\mathcal{M}_{\eta\eta}^{-1}} - \frac{1}{2}\|\mathcal{V}_1 - \mathcal{V}_2\|^2 \leqslant \frac{1}{2}\|X + c^{-1}\mathcal{M}_{\eta\gamma}Z_\gamma\|^2_{\mathcal{M}_{\eta\eta}^{-1}}$$

we can use the quantiles of the generalized chi-distribution to deduce a conservative, but feasible, test of the the null that a benchmark forecast $Q^{(T)}_{\vartheta^b}$, with $\vartheta^{(b)} = (\eta_T^\delta, \gamma^\star)$, is not inferior to an alternative forecast $Q^{(T)}_{\theta^a}$, with $\theta^a = \theta^\star$. Such a test will be asymptotically conservative in general, but will not be too conservative so long as the differences between $\sqrt{R}V_{R,R}$ and $\sqrt{P}V_{P,R}$ are small.

To implement such a test, for a given benchmark forecast $Q^{(T)}_{\vartheta^{(b)}}$ based on known combination weight η_T^δ, e.g., the equally-weighted combination, we construct a critical value from the generalized chi-distribution $\frac{1}{2}\|X + c^{-1}\mathcal{M}_{\eta\gamma}Z_\gamma\|^2_{\mathcal{M}_{\eta\eta}^{-1}}$ via simulation. In particular, we start out by first simulating $h = 1, \ldots, H$ realisations for the random variables $X^{(h)}, Z^{(h)}$ from normal distributions where Σ_X, Σ_Z are replaced with consistent estimators based on $\eta_R^\delta, \tilde{\gamma}_R$, and then, for each $h = 1, \ldots, H$, we form the statistic $\Delta^{(h)} := \frac{1}{2}\|X^{(h)} + (P/R)^{1/2}\widehat{\mathcal{M}}_{\eta\gamma}Z^{(h)'}\|^2_{\widehat{\mathcal{M}}_{\eta\eta}^{-1}}$, where $\widehat{\mathcal{M}}_{\eta\eta}$ and $\widehat{\mathcal{M}}_{\gamma\eta}$ are the usual sample estimate counterparts of the matrices $\mathcal{M}_{\eta\eta}$ and $\mathcal{M}_{\gamma\eta}$ and calculated at $\eta_R^\delta, \tilde{\gamma}_R$. Sorting $\{\Delta^{(h)} : h = 1, \ldots, H\}$, we can obtain an α-level critical value by choosing the $\lceil(1-\alpha)H\rceil$-th smallest value, and rejecting the null when the observed value of the statistic $P \cdot \Delta_P(\vartheta_R^\delta, \tilde{\theta}_R)$ exceeds this value.

Denoting the above simulated critical value by $\widehat{\text{cv}}_{\lceil(1-\alpha)H\rceil}$, we can define a test of the null hypothesis of no inferior predictive accuracy of the benchmark model $Q^{(T)}_{\eta_T^\delta, \gamma^\star}$ against the alternative model $Q^{(T)}_{\theta^\star}$ that is estimated in two-steps, via the corresponding rejection

region

$$W_P^{2s}(\alpha) := \{\Delta_P : P \cdot \Delta_P(\vartheta_R^{\delta}, \tilde{\theta}_R) > \widehat{\text{cv}}_{\lceil (1-\alpha)H \rceil}\}, \quad \alpha \in (0,1). \tag{9}$$

Unlike the usual test, based on $W_P(\alpha)$, the test based on the rejection region $W_P^{2s}(\alpha)$ explicitly accounts for the two-step nature by which $P \cdot \Delta_P(\vartheta_R^{\delta}, \tilde{\theta}_R)$ is constructed. As we have already seen, failure to account for the two-step nature of the forecast combinations does not deliver a test with appropriate size control. Conversely, since a test based on $W_P^{2s}(\alpha)$ accounts for the two-step nature of the estimation, the test should deliver appropriate size control.

We note that, while it is feasible to use the above simulation method to construct an appropriate critical value for the testing differences between competing forecast combinations, we do not necessarily advocate for this approach in general. In particular, while the above procedure would deliver an appropriately sized test, it is unclear if the above test is the most powerful approach, and additional research is necessary to determine this. In addition, as we elaborate on in the following Section 4.3, in certain cases there is a simpler alternative to adopting the above simulated critical value that allows us to entirely avoid the forecast combination puzzle. In Section A in the Supplementary material, we return to the running example and demonstrate that the usual testing procedure, aligned with the critical values discussed in this section, delivers much more reliable result than those based on the usual critical values.

4.3 Avoiding the Puzzle

As noted, the use of a two-step forecast combination, i.e., $Q_{\tilde{\theta}_R}^{(T)}$, is much more common in practice than a one-step forecast combination, i.e., $Q_{\hat{\theta}_R}^{(T)}$, where $\widehat{\theta}_R$ is defined in (1). However, the analysis in Section 4 demonstrates that the forecast combination puzzle is entirely due to this (two-step) estimation approach. This then begs the question of whether it may be possible to avoid the puzzle altogether by changing the way forecast combinations

are produced.

In this section we compare the accuracy of the one-step forecast combination, $Q^{(T)}_{\widehat{\theta}_R}$ against the standard two-step forecast combination, $Q^{(T)}_{\tilde{\theta}_R}$. Since the two-step approach is the standard approach in the literature, we test that the *two-step (benchmark) approach is not inferior to the one-step (alternative) approach.* The following result demonstrates that if the one-step forecast combination approach is computationally feasible, it will always yield superior forecast performance.

Theorem 2. *Under Assumptions 1-5 in Appendix B,* $\Pr[D_P(\tilde{\theta}_T, \widehat{\theta}_T) > 0] \to 1$ *as* $n \to \infty$.

Theorem 2 yields the following immediate corollary on the forecast accuracy of a benchmark equally-weighted combination against the optimally estimated one-step combination: let $\theta^{ew}_R = (K^{-1}\iota', \tilde{\gamma}'_R)'$, with ι a K-dimensional vector of ones.

Corollary 2. *If* $\eta^0 \neq K^{-1}\iota$, *and Assumptions 1-5 in Appendix B are satisfied, then* $\Pr[D_P(\theta^{ew}_R, \widehat{\theta}_T) > 0] \to 1$ *as* $n \to \infty$.

In certain cases, joint optimization over the parameters in the forecast combination (i.e., in one-step) will not produce a unique minima for the optimization program in (1); for instance, in cases where we wish to combine two mean point forecasts from separate regression models, then a one-step optimization may yield a set of optimal values. In such cases, the resulting point onto which the one-step estimator $\widehat{\theta}_n$ converges, i.e., θ^0, will not be unique. However, if one is only interested in forecasting accuracy, then this is immaterial for practice: since the forecasts are constructed from $L[Q^{(T)}_\theta, Y_{T+1}]$, and are evaluated using an out-of-sample estimator for $\mathbb{E}L[Q^{(T)}_\theta, Y_{T+1}]$, each of the optima will (asymptotically) produce the same level of accuracy for the forecast combination. Moreover, since the assumptions used to derive the results in Theorem 2 do not require point identification, the results on the accuracy of the one-step forecast combination given above will remain valid.

4.4 Revisiting Forecast Combinations

In Figure 2, we repeat the simulation of Section 2.2, but for different benchmark-alternative pairs (rows), with a view to illustrating the implications of Theorem 2 and Corollary 2, and thereby illustrating the benefits of the one-step combination.[11] As before, discussion of implementation details is left to Appendix D.2.

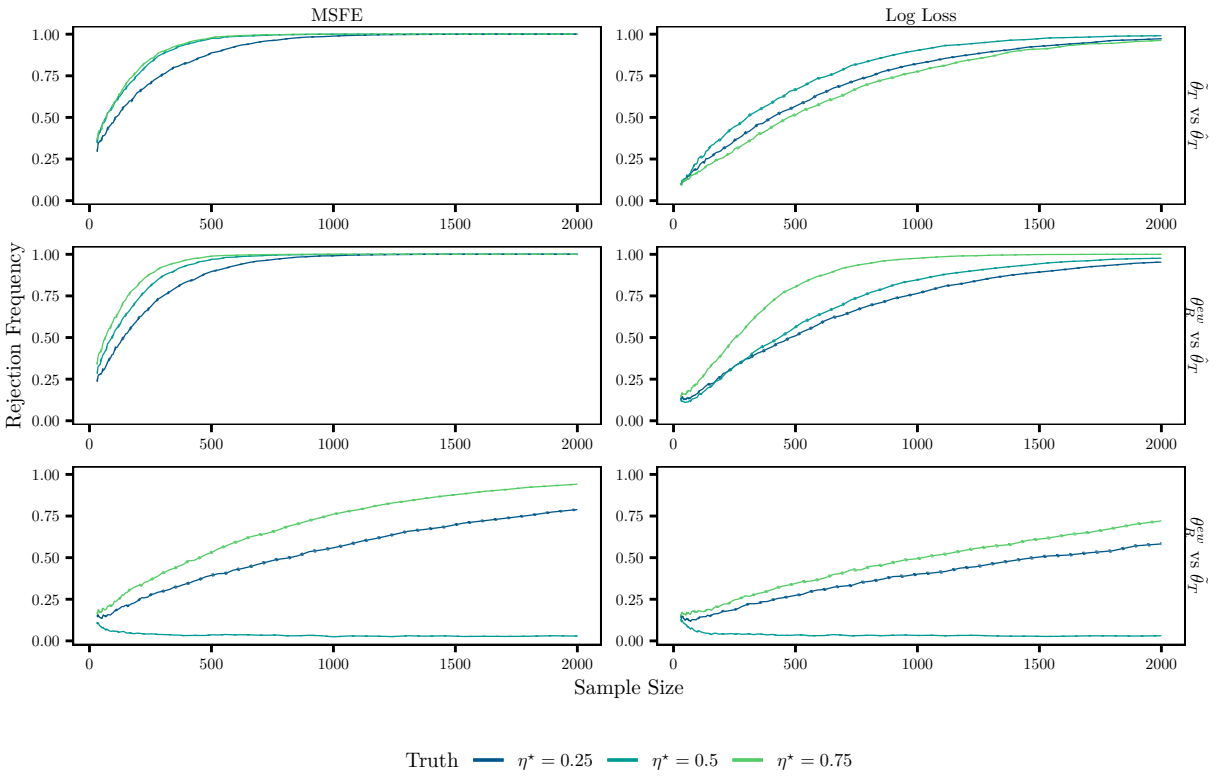

Figure 2: Estimates of the rejection frequency (y-axis) for the hypothesis test of no inferior predictive accuracy of a benchmark forecast combination against an alternative combination (rows, benchmark vs alternative). 95% confidence intervals for the rejection frequency only cover the thickness of the line, and are omitted. The test is conducted with observations drawn from DGPs across a range of pseudo-true weights (colours), and across a grid of sample sizes (x-axis). Results for a point forecast combination with optimal weights minimising the MSFE are given in the first column, and results for a distributional forecast combination with optimal weights minimising the log loss are given in the second column.

[11] We omit results for $\eta^\star = 0$ and $\eta^\star = 1$ to exclude cases where $\theta^0 = \theta^\star$, whereby the one- and two-step parameter estimators converge to the same values in the limit, violating Assumption 3.

In the first row, we test the benchmark two-step combination against the alternative one-step combination, and find that the rejection frequency (y-axis) quickly converges to one (in favour of the one-step combination) as the sample size (x-axis) increases, for combinations optimising both the MSFE (left-hand column) and the log loss (right-hand column) and for all values of $\eta^\star$ (colours). This reflects the result of Theorem 2, and supports preferring the one-step combination over its two-step counterpart when optimising the combination to maximise forecast performance.

In the second row, the hypothesis of no inferior predictive accuracy of the equally-weighted two-step benchmark is tested against the alternative one-step combination. Here, we again find that the rejection frequency (y-axis) rapidly converges to one as the sample size (x-axis) increases, for DGPs of *all* limiting two-step weights $\eta^\star$, including $\eta^\star = 0.5$, where θ_R^{ew} is the best-performing two-step combination. This is due to the fact that the one-step combination $\hat{\theta}_T$ converges to a higher-performing combination in the limit than is possible for *any* two-step combination, including θ_R^{ew} in the case where the optimal two-step weight $\eta^\star = 0.5$; this follows since even if $\eta^\star = 0.5$ under the two-step approach, the optimal one-step weight $\eta^0 \neq 0.5$.

Displayed in the bottom row are rejection frequencies for the test of no inferior predictive accuracy of the equally-weighted two-step benchmark against the optimally-weighted two-step alternative. Rejection frequencies for the same test were also displayed in the middle row of Figure 1, and here we see again that for the two-step combination this test is undersized and has low power, even when the equally-weighted vector θ_R^{ew} is far from optimal (dark blue and green), leading to the forecast combination puzzle. Comparing the bottom and middle rows of Figure 2, we find that the power of the test increases dramatically to resolve the puzzle when estimating parameters in one step (middle row) rather than two (bottom row). The increase in power from one-step estimation is seen across all sample sizes (x-axis), all pseudo-true weights (colours) and both losses (columns).

4.5 Revisiting forecast Combinations: Geweke and Amisano (2011)

In this section, we give an empirical example which demonstrates that one-step combinations resolve the forecast combination puzzle. Specifically, we follow Section 3 of Geweke and Amisano (2011) and consider a linear pool comprising the Gaussian exponential GARCH$(1,1)$ model ("EGARCH") and the GARCH$(1,1)$ model with i.i.d. Student t errors ("t-GARCH"); this pool is then used to produce one-step-ahead distributional forecasts of daily logarithmic S&P500 returns. All parameters (including combination weights) are estimated using returns for the 3783 trading days from years 1990 to 2004, inclusive (the "training set"). We then evaluate and compare the log-score-based forecasting performance of different combinations using returns for the out-of-sample period comprising the 3772 trading days from years 2005 to 2019, also inclusive (the "test set").

Three different ways of estimating the forecast combinations are compared: equally-weighted two-step estimation, optimally-weighted two-step estimation, and one-step estimation. In the two-step combinations, the EGARCH parameters are first chosen to maximize the average log score of the EGARCH one-step-ahead predictive distribution over the training set; likewise for the t-GARCH parameter estimates. To produce the equally-weighted two-step combination, we set the weights to 0.5. The combination weights for the optimally-weighted two-step combination are estimated by maximizing the log score of the combination across the training set, with the EGARCH and t-GARCH parameters fixed at their first-step values. The one-step combination parameters are jointly chosen to maximize the training-set average log score of the one-step-ahead predictive distribution of the combination density, in a single optimization program.

Table 1 contains the training-set average log scores of the three combinations (first column). Consistent with the forecast combination puzzle, even across the training-set the equally-weighted two-step combination outperforms the optimally-weighted two-step combination, since it has a higher average log score (second column). As suspected, both

two-step combinations are beaten by the one-step approach over the training-set.

In Table 2 we display the p-values (right column) for three tests of the null hypothesis that a benchmark combination (left column) is not inferior to an alternative combination (middle column). The test proceeds according to Section 3.2.2 using the loss differences pertaining to the log scores of the out-of-sample test set. The asymptotic variance of the loss difference is estimated using the method described in Section 4.1 of Okui (2010), with the quadratic spectrum kernel and $S = \sqrt{T}$. In the first row we fail to reject the null hypothesis that the equally-weighted two-step combination is not inferior to the optimally-weighted two step combination, reflecting the forecast combination puzzle. The tests displayed in the second and third rows unequivocally reject the null that the benchmark - either the equally- or the optimally-weighted two-step combination - is not inferior to the one-step combination, with both null hypotheses rejected at the 1% level. This perfectly reflects the theoretical results in Section 4.3: we can avoid the puzzle - and obtain a higher performing forecast - by estimating all parameters in one step, rather than the standard two steps.

Combination	Average Log Score
Equally-Weighted Two Step	3.3481
Optimally-Weighted Two Step	3.3459
One Step	3.3596

Table 1: The average log scores of three forecast combinations for S&P500 returns.

Benchmark	Alternative	p value
Equally-Weighted Two Step	Optimally-Weighted Two Step	0.8251
Equally-Weighted Two Step	One Step	5.675e-05
Optimally-Weighted Two Step	One Step	6.935e-12

Table 2: P-values (right column) for tests of the null hypothesis that a benchmark combination (left column) is not inferior to an alternative combination (right column). All combinations produce one-step-ahead distributional forecasts for S&P500 returns, and are evaluated on a log-score basis.

5 Conclusion

In this paper, we investigate the forecast combination puzzle through the lens of hypothesis testing approaches aimed at discriminating between the relative performances of equally-weighted and optimally-weighted forecast combinations. Forecast combination parameters are optimized according to a scoring function or scoring rule for point forecasting or distributional forecasting, respectively, and we thereby demonstrate that the forecast combination puzzle is a phenomenon that extends far beyond point forecasts optimized according to the MSFE.

Our theoretical analysis demonstrates that such hypothesis tests have at best low local power, and lack size control, when applied to hypothesis tests aiming to distinguish between the performance of equally- and optimally-weighted forecast combinations. We theoretically demonstrate that this perverse behavior is caused by the fact that the test *does not account for the two-step nature by which forecast combinations are produced*.

An extension to the Monte Carlo exercises of Smith and Wallis (2009) and Claeskens et al. (2016) illustrates how this lack of local power negatively impacts such tests. By producing the rejection frequencies of a variety of hypothesis tests of no inferior forecast accuracy of a fixed-weight benchmark against the optimally-weighted two-step alternative, we illustrate that such hypothesis tests require large sample sizes to reject in favour of the optimally-weighted combination, even when the (unknown) best-performing weights are very different from the vector of equal weights. This finding is seen repeatedly, for point and distributional forecast combinations optimized according to different scores, across a diverse range of DGP parameter values, and across several fixed-weighted (and not just equally-weighted) benchmark combinations. We also revisit a two-model distributional forecast combination of S&P500 returns in Geweke and Amisano (2011) to obtain empirical evidence that this phenomenon occurs in practice.

It is shown that, under mild assumptions, optimizing combination parameters in one

step will always eventually reject the null hypothesis that the equally-weighted benchmark is not inferior to the optimally-weighted combination. Repeating the Monte Carlo exercise for the one-step alternative under the optimally- and equally-weighted two-step benchmarks reveals that the low power that is characteristic of the forecast combination puzzle in the case of two-step forecast combinations is absent when the alternative optimally-weighted combination is estimated in one-step. In addition, we verify the superiority of one-step combinations in the S&P500 returns example considered in Geweke and Amisano (2011) by showing that the one-step density combination delivers superior predictive accuracy relative to two-step benchmark.

In this way, we argue that the root-cause behind the lack of evidence for the performance of optimally-weighted two-step combinations against their equally-weighted counterparts is an artefact of the way in which such combinations are generally produced, i.e., in two steps. Furthermore, we demonstrate that if it feasible to produce optimal combinations in a single step, the forecasting puzzle can be completely avoided. Hence, if the problem at hand is such that forecast combinations can be produced in a one-step fashion, the practitioner will (always) reap appreciable gains, in terms of forecast accuracy, by undertaking such a strategy.

Alternatively, if a one-step approach is infeasible, or if a two-step approach is preferable, we have demonstrates how the usual testing framework must be altered to accommodate the two-step nature by which the combinations were produced, to ensure reliable testing results. In particular, when using a two-step forecast combination approach, both the test statistic and the critical value employed must be altered in order to deliver a test that has correct size, and non-negligable power.

Before concluding, we note that there are many interesting cases where a forecast combination procedure may seem, at the outset, not to be produced in a two-step fashion, but which upon closer inspection reveals that such forecast combinations are actually produced

in (at least) a two-step fashion. As an illustrative example, consider the context of volatility forecasting using the HAR model (Corsi, 2009; Corsi et al., 2012). By viewing HAR model forecasts as the combination of lagged moving average models for realised variance, Clements and Vasnev (2021) document the existence of a forecasting combination puzzle in HAR models and show that such models do not generally perform better than a simple weighted average of the constituent forecasts.

Interestingly, the results of Clements and Vasnev (2021) seem to document the existence of a forecast combination puzzle without a 'first-stage' estimation step being required to produce the forecasts; at face value it then seems that such an example lies outside the scope of our general results. However, recall that the "observed value" of realised volatility is not a genuine realisation of "observed data", but a nonparametric estimator of integrated-variance, computed using inter-daily returns. That is, the very construction of the realised variance series constitutes a first-stage estimation step, and forecasts produced via HAR models can therefore be viewed as two-stage forecast combination methods: the first stage estimates the realised variance series, and the second the combination scheme. Hence, the lack of power HAR models exhibit to distinguish between equally and optimally weighted forecast combinations is also explained by our theoretical results. We leave a full study on such types of first-stage estimation steps for future research, but remark that, the heavy use of high-frequency returns, and realised variance in particular, in financial forecasting applications would seem to imply the existence of undiscovered combination puzzles.

SUPPLEMENTARY MATERIAL

The supplementary material contains our main assumptions, the proofs of all technical results, and additional numerical details demonstrating the performance of the critical value adjustment approach described in Section 4.2.

Acknowledgements

Frazier was supported by the Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE200101070), and Frazier and Martin were supported by the Australian Research Council's Discovery Project scheme (DP200101414).

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F., and van Dijk, H. K. (2019). The evolution of forecast density combinations in economics. In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press. 2

Baran, S. and Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, 34(3):477–496. 2

Bassetti, F., Casarin, R., and Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, 113(522):675–685. 11

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468. 2

Chan, F. and Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, 34(1):64–74. 3

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762. 2, 3, 5, 6, 7, 31

Clark, T. and McCracken, M. (2013). Advances in forecast evaluation. *Handbook of economic forecasting*, 2:1107–1201. 16

Clements, A. and Vasnev, A. L. (2021). Forecast combination puzzle in the har model. *Available at SSRN: https://ssrn.com/abstract=3875026 or http://dx.doi.org/10.2139/ssrn.3875026*. 33

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196. 33

Corsi, F., Audrino, F., and Renò, R. (2012). Har modeling for realized volatility forecasting. In *Handbook of Volatility Models and Their Applications*, pages 363–382. John Wiley & Sons, Inc, Hoboken, NJ, USA. 33

Elliott, G. (2011). Averaging and the optimal combination of forecasts. *Manuscript, Department of Economics, UCSD*. 3

Fissler, T. and Ziegel, J. F. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4):1680–1707. 12

Frazier, D. T. and Renault, E. (2017). Efficient two-step estimation via targeting. *Journal of Econometrics*, 201(2):212–227. 15

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141. 2, 5, 11, 14, 29, 31, 32

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*. 11, 13, 14

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422. 11

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782. 2, 11, 14

Graefe, A., Armstrong, J. S., Jones Jr, R. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54. 3

Granger, C. W. and Machina, M. J. (2006). Forecasting and decision theory. *Handbook of Economic Forecasting*, 1:81–98. 13

Hall, S. G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23(1):1–13. 2, 14

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380. 3, 16

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808. 2, 3

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74. 2, 3

Martin, G. M., Loaiza-Maya, R., Maneesoonthorn, W., Frazier, D. T., and Ramírez-Hassan, A. (2021). Optimal probabilistic forecasts: When do they work? *International Journal of Forecasting*. 2, 11

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245. 15

Okui, R. (2010). Asymptotically unbiased estimation of autocovariances and autocorrelations with long panel data. *Econometric Theory*, 26(5):1263–1304. 30

Opschoor, A., Van Dijk, D., and van der Wel, M. (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics*, 32(7):1298–1313. 2

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics*, 38(4):796–809. 12

Pesaran, M. H. and Skouras, S. (2002). Decision-based methods for forecast evaluation. *A Companion to Economic Forecasting*, pages 241–267. 13

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91. 2

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356. 2

Smith, J. and Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355. 2, 3, 5, 6, 7, 8, 16, 31

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430. 2, 3

Stone, M. (1961). The linear opinion pool. *Ann. Math. Statist*, 32:1339–1342. 2

Taylor, J. W. (2020). Forecast combinations for value at risk and expected shortfall. *International Journal of Forecasting*, 36(2):428–441. 2

Thorey, J., Chaussin, C., and Mallet, V. (2018). Ensemble forecast of photovoltaic power with online crps learning. *International Journal of Forecasting*, 34(4):762–773. 2

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196. 2

Van Der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press. 17

Wang, L., Wang, Z., Qu, H., and Liu, S. (2018). Optimal forecast combination based on neural networks for time series forecasting. *Applied Soft Computing*, 66:1–17. 2

Wang, X., Hyndman, R. J., Li, F., and Kang, Y. (2022). Forecast combinations: an over 50-year review. *International Journal of Forecasting*. 3

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, pages 1067–1084. 6, 15, 19

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126. 3, 15, 16

Zischke, R., Martin, G. M., Frazier, D. T., and Poskitt, D. S. (2022). The impact of sampling variability on estimated combinations of distributional forecasts. *arXiv preprint arXiv:2206.02376*. 5, 15

Supplementary Material: "Solving the Forecast Combination Puzzle"

Abstract

This supplementary material contains proofs of all technical results presented in the main paper, and additional numerical details demonstrating the performance of the critical value adjustment approach described in Section 4.2 of the main paper.

A Revisiting Two-Step Combinations

We return to the example setting discussed in Smith and Wallis (2009) and demonstrate numerically that the simulated critical value suggested in Section 4.2 of the main text delivers a testing procedure with (approximately) correct size under the null hypothesis of no inferior forecast accuracy, and also has higher power than the standard approach under the alternative.

To demonstrate the empirical size of the test, we return to the example in Section 2 and consider values of ϕ_1, ϕ_2 and σ^2 for the AR(2) family such that, under MSFE (respectively, log-loss), the optimal value of the combination weight, $\eta^\star$, obtained under MSFE (respectively, log-loss), is (approximately) equal to the benchmark combination weight of 1/2, i.e., the equally-weighted combination benchmark.[1] Under this DGP, we generate

[1]Under log score, a DGP that ensures that the optimal two-step forecast combination weight is (approximately) equal to 1/2 can be obtained by setting $(\phi_1, \phi_2, \sigma^2) = (.4000, -.4421, 1)^\top$ in the AR(2) model. For MSFE, taking $(\phi_1, \phi_2, \sigma^2) = (.4000, -.4070, 1)^\top$ delivers a DGP such that the optimal (two-step) forecast combination weight is (approximately) 1/2. The values of $(\phi_1, \phi_2, \sigma^2)$ in both cases were obtained

1000 replications across three different samples sizes $T = 1000, 2000$ and 5000. Across each of these datasets, we use the first $R = T/2$ observations for training, and the remaining $P = T - R$ observations for testing.

The first three rows of Table 1 compare the size of the standard test of no inferior forecast accuracy, based on $W_P(\alpha)$ in (8), against the testing approach that caters for the two-step nature of the combination forecast construction, via the use of $W_P^{2s}(\alpha)$ in (9). In the two-step approach, we use $B = 10000$ draws to simulate the critical value in all cases. As a comparator, we also present the empirical rejection frequency of the standard 't-test' of the null hypothesis that $\eta = 1/2$, based on the correct two-step standard error for $\tilde{\eta}_R$. This additional comparator serves as a benchmark of sorts, enabling us to diagnose, in some sense, if the rejection rates for the tests are due primarily to differences in the predictive ability of the combinations, or are due to differences in the combination weights themselves.

The results in the table demonstrate that, as already highlighted in Section 2, the standard approach has zero size in (nearly) all cases when the benchmark combination weight, 1/2 in this case, is close to the optimal combination weight. In contrast, even with an asymptotically conservative critical value, accounting for the two-step nature by which the forecasts were produced, via $W_P^{2s}(\alpha)$ in (9), results in a test that has sizes that are much closer to the nominal level. Depending on the chosen loss, for the finite samples used, the test based on $W_P^{2s}(\alpha)$ can be slightly over- or under-sized, but delivers results that are much closer to the nominal level than does the standard approach.

The second set of three rows in Table 1, compares the empirical power of the alternative testing approaches, under both loss functions, under a DGP that is a relatively small dis-

numerically by maximizing the corresponding loss function using a sample size of 10 million observations generated from the DGP.

tance away from the DGP that delivers an optimal combination weight that is equal to the benchmark combination weight of 1/2.[2]Across both loss functions, our testing approach has vastly higher power than the standard approach, with the standard testing approach based on the log-loss having zero power under this minor deviation from the null hypothesis, and with its power under the MSFE being only around one percent. Moreover, the magnitude of the power of the correctly sized test is broadly similar (for any given sample size) under the two losses. Interestingly enough, the comparator 't-test' has quite high power under log loss, but very different, and much lower power under MSFE.

Consequently, the results in Table 1 demonstrate empirically that the if one wishes to conduct a test of forecast accuracy based on approaches that use forecast combinations, accounting for the two-step nature by which these forecasts were produced will be critical in producing tests with good power and reliable size.

B Assumptions and Proofs

B.1 Assumptions and Discussion

We wish to treat cases where the model combination weights, η, are allowed to lie on the boundary of the parameter space, and we wish to be agnostic about the asymptotic distribution of the estimated parameters in the constituent models. Before stating our maintained assumptions, we recall that $L_n(\theta) = \sum_{t=1}^{n-1} \ell_t(\theta)$, for some loss function $\ell_{t+1}(\theta) =$

[2]Under log-loss, the DGP generating the observed data is fixed at $(\phi_1, \phi_2, \sigma^2) = (.40, -.50, 1)^\top$, while under MSFE we take the DGP generating the data to be $(\phi_1, \phi_2, \sigma^2) = (.40, -.45, 1)^\top$. That is, in both cases, the DGP deviates from the version that delivered equivalence between the equal-weighted combination and the optimally-weighted combination by moving the second root of the AR(2) process, ϕ_2, by -.05.

Size	MSFE			Log-Loss		
	Two-Step	T-Test	Stand	Two-Step	T-Test	Stand
$T = 1000$	0.0202	0.0348	0.0000	0.0530,	0.1690	0.0000
$T = 2000$	0.0240	0.0330	0.0004	0.0440,	0.3250	0.0000
$T = 5000$	0.0322	0.0392	0.0000	0.0510,	0.5480	0.0000
Power	MSFE			Log-Loss		
	Two-Step	T-Test	Stand	Two-Step	T-Test	Stand
$T = 1000$	0.1544	0.1426	0.0122	0.1960	0.3960	0.0000
$T = 2000$	0.3038	0.2536	0.0108	0.3670	0.6980	0.0000
$T = 5000$	0.6546	0.5450	0.0102	0.6870	0.9570	0.0000

Table 1: Monte Carlo rejection rates for testing the null hypothesis in (5) under the size and power DGPs for the loss functions MSFE and Log-Loss. The table compares the rejection rates of the standard test (Stand) and the two-step version based on the rejection region in (9) (Two-Step). The column T-Stat gives the empirical rejection rates of the T-test that $\eta = 1/2$.

$L[P_\theta^{(t)}, y_{t+1}]$.

Assumption 1. *The parameter space Θ is compact, and can be written as a Cartesian product of intervals with the form $[0, c]$ or $[-c, c]$, for some $0 < c < \infty$ that can change dimension-by-dimension.*

Remark 1. The assumption that Θ is a Cartesian product is not onerous since the parameter space is a product of closed intervals, and since the requirement that the boundary is 'on the left', and at zero, is without loss of generality: if the j-th element of the parameter vector originally satisfies $\theta_j \in [c_{1j}, c_{2j}]$, then we can always consider the translated parameter $\vartheta_j = \theta_j - c_1$, which lies in $[0, c_{2j} - c_{1j}]$. For instance, this assumption is immediately satisfied for combination weights in the case of linear pools. Further, this assumption satisfies the conditions on the parameter space necessary to apply the results of Andrews (1999).

Since our goal is not inference on $\gamma^\star$, we maintain the following high-level regularity condition for the estimated parameters in the constituent models.

Assumption 2. *The population criterion $\sum_{j=1}^K \mathcal{L}(\gamma_k)$ exists and is uniquely minimized at $\gamma^\star = (\gamma_1^\star, \ldots, \gamma_K^\star)'$; (ii) $\sqrt{n}(\tilde{\gamma}_n - \gamma^\star) = O_p(1)$.*

Together with Assumption 1, the following conditions give consistency of the one- and two-step estimators $\widehat{\theta}_n, \tilde{\theta}_n$ to their corresponding limit optimizers.

Assumption 3. *There exists a function $\theta \mapsto \mathcal{L}(\theta)$ such that: (i) $\sup_{\theta\in\Theta} |L_n(\theta)/n - \mathcal{L}(\theta)| = o_p(1)$; (ii) There exist a non-empty set $\Theta_{\mathrm{I}} \subset \Theta$ with a finite number of elements such that, for each $\theta^0 \in \Theta_{\mathrm{I}}$, the map $\theta \mapsto \mathcal{L}(\theta)$ is minimized at $\theta^0 \in \Theta_{\mathrm{I}}$, i.e., $\Theta_{\mathrm{I}} := \operatorname{arginf}_\Theta \mathcal{L}(\theta)$; $\eta \mapsto \mathcal{L}(\eta, \gamma^\star)$ is uniquely minimized at $\eta^\star \in \mathcal{E}$, and $\theta^\star \notin \Theta_{\mathrm{I}}$.*

Assumption 3 differs from the usual point-identification assumption imposed in classical extremum estimation problems. Instead, we only require that the one-step estimator is set-identified, and that the set contains only a finite collection of values. This is helpful for treating situations where the forecast combination delivers unidentified parameter estimates due to the existence of multiple roots in the criterion function. However, from the point of forecasting accuracy whether Θ_{I} is a singleton is irrelevant so long as the forecast accuracy associated with the collection of points in Θ_{I} is constant, which is precisely what Assumption 3 stipulates.

The following assumption allows us to deduce quadratic expansions that we use to establish the behavior of tests of forecasting accuracy. In what follows, we recall that $\nabla_\theta \mathcal{L}(\theta)$ denotes left/right (hereafter, l/r) derivatives or standard derivatives, depending on the context.

Assumption 4. *(i) For some $\delta > 0$, $\vartheta \in \Theta_{\mathrm{I}} \cup \{\theta^\star\}$, and all $\|\theta - \vartheta\| < \delta$, $L_n(\theta)$ and $\mathcal{L}(\theta)$ admit second-order l/r partial derivatives in θ that are continuous with respect to components that can be perturbed to the l/r; (ii) for each $\vartheta \in \Theta_{\mathrm{I}}$, $\nabla_\theta L_n(\vartheta) = O_p(1)$ and $[\nabla_\eta L_n(\eta, \gamma^\star); \nabla_\gamma L_n(\gamma^\star)]/\sqrt{n} = O_p(1)$; (iii) the matrices $\nabla_{\eta\eta}\mathcal{L}(\eta, \gamma^\star)|_{\eta=\eta^\star}$ and $\nabla_{\theta\theta}\mathcal{L}(\theta)|_{\theta=\theta^\star}$ are positive-definite; (iv) For any $\delta_n = o(1)$, $\sup_{\vartheta\in\Theta_{\mathrm{I}}} \sup_{\|\theta-\vartheta\|\leq\delta_n} \|\nabla_{\theta\theta} L_n(\theta)/n - \nabla_{\theta\theta}\mathcal{L}(\theta)\| = o_p(1)$.*

Remark 2. The assumptions regarding the continuous l/r derivatives can be relaxed in cases by replacing the differentiability conditions with a 'stochastic differentiability' condition that depend on the specific type of loss function, and which are well-known in the literature on empirical processes (see, e.g., van der Vaart et al., 1996). Since this complication is not our main interest, we maintain the stronger conditions on the l/r derivatives.

Assumptions 1-4 are sufficient to deduce consistency, as well as the rate of convergence,

for the one-step estimator of the parameters in the forecast combination. The following result follows similar arguments to Theorem 2 Andrews (1999), but requires slight alternations since we do note assume the existence of a unique minimum for $\mathcal{L}(\theta)$. In what follows, define $\mathcal{M}_{\theta\theta}(\theta) := \nabla_{\theta\theta}\mathcal{L}(\theta)$, with derivatives for η and γ defined accordingly.

Lemma 1. *Suppose Assumptions 1, 3 and 4 are satisfied, then for some $\vartheta \in \Theta_{\mathrm{I}}$ the following are satisfied.*

(i) $\sqrt{n}(\widehat{\theta}_n - \vartheta) = O_p(1)$,

(ii) For $\lambda_n = \sqrt{n}(\widehat{\theta}_n - \vartheta)$, *and* $Z_n := -\mathcal{M}_{\theta\theta}(\vartheta)^{-1}\nabla_\theta L_n(\vartheta)/\sqrt{n}$,

$$L_n(\widehat{\theta}_n) - L_n(\vartheta) = -\frac{1}{2}Z_n'\mathcal{M}_{\theta\theta}(\vartheta)Z_n + \frac{1}{2}(\lambda_n - Z_n)'\mathcal{M}_{\theta\theta}(\vartheta)(\lambda_n - Z_n) + o_p(1).$$

A result like Lemma 1 can also be deduced for the two-step estimator of the combination weights.

Lemma 2. *Under Assumptions 1-4,*

(i) $\sqrt{n}(\tilde{\eta}_n - \eta^\star) = O_p(1)$.

(ii) For $\mathcal{J} = \mathcal{M}_{\eta\eta}(\theta^\star)$, $\kappa_n = \mathcal{J}^{1/2}\sqrt{n}(\tilde{\eta}_n - \eta^\star)$, *and* $V_n := -\mathcal{J}^{-1/2}\{\nabla_\eta L_n(\theta^\star)/\sqrt{n} + \mathcal{M}_{\eta\gamma}(\theta^\star)\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)\}$,

$$L_n(\tilde{\theta}_n) - L_n(\eta^\star, \tilde{\gamma}_n) = -\frac{1}{2}\|V_n\|^2 + \frac{1}{2}\|\kappa_n - V_n\|^2 + o_p(1).$$

Lemma 2 demonstrates that the two-step criterion has a quadratic expansion that depends on the centering variable V_n, which is a linear function of $\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)$. This centering sequence is a consequence of the two-step nature with which the criterion has been optimized, which treats $\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)$ as a fixed quantity. In contrast, a joint expansion of $L_n(\tilde{\theta}_n)$ about $\theta^\star$ produces a linear expansion at first-order.

Lemma 3. *If Assumptions 1-4 are satisfied, then*

$$n^{-1/2}\{L_n(\tilde{\eta}_n, \tilde{\gamma}_n) - L_n(\eta^\star, \gamma^\star)\} = \nabla_\gamma \mathcal{L}(\theta^\star)' \sqrt{n}(\tilde{\gamma}_n - \gamma^\star) + o_p(1).$$

Remark 3. Lemma 3 demonstrates that the dominant term in the expansion of $\{L_n(\tilde{\eta}_n, \tilde{\gamma}_n) - L_n(\eta^\star, \gamma^\star)\}$ is the linear term $\nabla_\gamma \mathcal{L}(\theta^\star)' \sqrt{n}(\tilde{\gamma}_n - \gamma^\star)$. That is, the behavior of $\sqrt{n}(\tilde{\eta}_n - \eta^\star)$ is irrelevant in determining the behavior of $L_n(\tilde{\theta}_n)/\sqrt{n}$. Lemma 3 is a consequence of the two-step nature of the estimator $\tilde{\theta}_n$.

Remark 4. The above result demonstrates that the estimation, and limiting behavior, of the combination weights has no impact on D_P defined in (6): the asymptotic behavior of D_P only depends on the pseudo-true value $\eta^\star$, and the variation in $\tilde{\gamma}_R$. So long as $\sqrt{R}(\tilde{\eta}_R - \eta^\star) = O_p(1)$, and even if $\eta^\star$ is on the boundary of the parameter space, the variability in the estimates of the parameters governing the constituent forecasts ultimately determines the behavior of D_P.

The following remaining assumptions are used to control the behavior of the average loss difference.

Assumption 5. *The variable* $Z_n = \left[L_n(\theta^\star)/\sqrt{n} - L_n(\theta^0)/\sqrt{n} - \{\mathcal{L}(\theta^\star) - \mathcal{L}(\theta^0)\}, \sqrt{n}(\tilde{\gamma}_n - \gamma^\star)\right]'$, *is such that* $Z_n \Rightarrow N(0, Q)$. *Further, for* $V_{P,R} := \nabla_\eta L_P(\theta^\star)/P + \mathcal{M}_{\eta\gamma}(\tilde{\gamma}_R - \gamma^\star)$, *we have that* $(\sqrt{R}V_{R,R}, \sqrt{P}V_{P,R}) \Rightarrow N(0, Q_V)$ *for some positive semi-definite covariance matrix* Q_V.

B.2 Proofs of Main Results

Proof of Theorem 1. By part (i) of Lemma 2, $\sqrt{R}(\tilde{\eta}_R - \eta^\star) = O_p(1)$, under Assumption 1. Now, using Assumption 4(i), and Theorem 6 in Andrews (1999), expand the loss difference around $\theta_R^\star = (\eta_R^\star, \tilde{\gamma}_R')'$, $\eta_R^\star = \delta_R + \eta^\star$, where applicable, all derivatives are again taken to be left/right partial derivatives: for a sequence of intermediate values $\bar{\eta}_R$ satisfying

$\|\bar{\eta}_R - \eta_R^\star\| \leq \|\tilde{\eta}_R - \eta_R^\star\|$, we obtain

$$\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) = \{\nabla_\eta L_P(\eta_R^\star, \tilde{\gamma}_R)/P\}'\sqrt{P}(\tilde{\eta}_R - \eta_R^\star) + \frac{1}{2}\sqrt{P}(\tilde{\eta}_R - \eta_R^\star)'\{\nabla^2_{\eta\eta} L_P\left(\bar{\eta}_R, \tilde{\gamma}_R\right)/P\}(\tilde{\eta}_R - \eta_R^\star).$$

From compactness of Θ, Assumption 1, and differentiability of $\nabla_\eta L_R(\eta, \gamma^\star)$ in η, Assumption 4(i), the following are satisfied via the usual arguments:

$$\|\nabla_\eta L_R(\eta_R^\star, \tilde{\gamma}_R)/P - \nabla_\eta \mathcal{L}(\eta_R^\star, \tilde{\gamma}_R)\| = o_p(1), \quad \|\nabla^2_{\eta\eta} L_P\left(\bar{\eta}_R, \tilde{\gamma}_R\right)/P - \mathcal{M}_{\eta\eta}(\bar{\eta}_R, \tilde{\gamma}_R)\| = o_p(1).$$

From the above convergence, and the continuity of the left/right derivatives in Assumption 4, we can conclude that

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) = \nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star)'\sqrt{P}(\tilde{\eta}_R - \eta_R^\star) + (\tilde{\eta}_R - \eta_R^\star)'\mathcal{M}_{\eta\gamma}(\eta_R^\star, \gamma^\star)'\sqrt{P}(\tilde{\gamma}_R - \gamma^\star) + o_p(1) \\ + \frac{1}{2}\sqrt{P}(\tilde{\eta}_R - \eta_R^\star)'\mathcal{M}_{\eta\eta}\left(\bar{\eta}_R, \gamma^\star\right)(\tilde{\eta}_R - \eta_R^\star) + o_p(1). \qquad (1)\end{aligned}$$

Case (1). Consider that $\delta_R = \delta/T^\xi$, for $\xi \in [0, 1/4)$, then

$$\sqrt{P}(\tilde{\eta}_R - \eta_R^\star) = \sqrt{P}(\tilde{\eta}_R - \eta^\star) + \delta_R\sqrt{P} = O_p(1) + \delta_R\sqrt{P},$$

where the $O_p(1)$ follows since, by Lemma 2, $\|\tilde{\eta}_R - \eta^\star\| = O_p(1/\sqrt{P})$.

Consider two cases: $\xi = 0$, and $\xi \in (0, 1/4)$. In the case where $\xi = 0$, from the definition of $\eta_R^\star$, for some $\varepsilon > 0$, there exists an $T(\delta)$ large enough such that for all $T > T(\delta)$, $\|\eta^\star - \eta_R^\star\| \geq \gamma$ with $\gamma = \delta - \varepsilon > 0$. Hence, by Assumption 3(ii) and the differentiability in 4(i) $\operatorname{plim}_{T\to\infty} \|\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star)\| > 0$ and $\operatorname{plim}_{T\to\infty} |\delta_R'\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star)| > 0$. Alternatively, in the case where $\xi \in (0, 1/4)$, there exists a $T(\delta)$ large enough such that for all $T > T(\delta)$, $\|\eta^\star - \eta_R^\star\| \leq \gamma$, and we have $\operatorname{plim}_{T\to\infty} |\delta_R'\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star)| = 0$.

Applying the above into equation (1) we have that

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) &= \sqrt{P}\delta_R'\{\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star) + o_p(1) + M^\star_{\eta\gamma}\left(\eta_R^\star, \gamma^\star\right)\sqrt{P}(\tilde{\gamma}_R - \gamma^\star) + \frac{1}{2}M^\star_{\eta\eta}\left(\bar{\eta}_R, \gamma^\star\right)\delta_R\} \\ &= \sqrt{P}\delta_R'\{\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star) + O_p(1/\sqrt{P})\} + \frac{\sqrt{P}}{2}\delta_R' M^\star_{\eta\eta}\left(\bar{\eta}_R, \gamma^\star\right)\delta_R \\ &\geq \sqrt{P}\delta_R'\{\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star) + O_p(1/\sqrt{P})\} + \frac{\sqrt{P}}{2}\inf_{\eta\in\mathcal{E}}\left\{\delta_R' M^\star_{\eta\eta}\left(\eta, \gamma^\star\right)\delta_R\right\},\end{aligned}$$

where the $O_p(1)$ term in the second equation follows by Assumption 2. Taking absolute values, and applying the reverse triangle inequality then yields

$$|\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star)| \geq \sqrt{P}||\delta_R' \nabla_\eta \mathcal{L}(\eta^\star + \delta_R, \gamma^\star)| - |\inf_{\eta \in \mathcal{E}} \left\{\delta_R' M_{\eta\eta}^\star(\eta, \gamma^\star)\, \delta_R\right\}| + o_p(1) \tag{2}$$

In the case where $\xi = 0$, it can been seen directly that the term inside the absolute value is non-zero, and so the RHS diverges as $P \to \infty$. In the case where $\xi \in (0, 1/4)$, rewrite

$$\sqrt{P}\delta_R = \sqrt{P}\delta/T^\xi = \sqrt{\frac{P}{P+R}}\frac{\sqrt{P+R}}{T^\xi}\delta = \sqrt{(1+c)^{-1}}T^{1/2-\xi} + o(1),$$

where $c = \lim_T R/P$ is as in Assumption 1. Consider the second term on the RHS of (2). Since $\xi < 1/4$, the second term is proportional to $\sqrt{P}\|\delta_R\|^2 = CT^{1/2-\xi} \to +\infty$ as $T \to +\infty$. Consequently, the RHS of (2) also diverges so long as $\xi \in (0, 1/4)$.

Case (2). Applying the same steps as in case (1), we arrive at the inequalities

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) \leq& \sqrt{P}\delta_R'\{\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star) + O_p(1/\sqrt{P})\} + \sup_{\eta \in \mathcal{E}} \frac{\sqrt{P}}{2}\delta_R' M_{\eta\eta}^\star(\bar{\eta}_R, \gamma^\star)\, \delta_R \\ \geq& \sqrt{P}\delta_R'\{\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star) + O_p(1/\sqrt{P})\} + \inf_{\eta \in \mathcal{E}} \frac{\sqrt{P}}{2}\delta_R' M_{\eta\eta}^\star(\bar{\eta}_R, \gamma^\star)\, \delta_R\end{aligned}$$

As discussed in Case (1), when $\delta_R = \delta/T^\xi$, we have that $\sqrt{P}\delta_R = C\delta T^{1/2-\xi}$. However, in the case where $\xi \geq 1/4$, we have that

$$\nabla_\eta \mathcal{L}(\eta_R^\star, \gamma^\star) = \mathcal{M}_{\eta\eta}(\eta^\star, \gamma^\star)\delta_R + o(\|\delta_R\|)$$

Applying this into the above inequalities yields

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) \leq& \sqrt{P}\delta_R'\mathcal{M}_{\eta\eta}(\eta^\star, \gamma^\star)\delta_R + O_p(\|\delta_R\|) + o(\sqrt{P}\|\delta_R\|^2) + \sup_{\eta \in \mathcal{E}} \frac{\sqrt{P}}{2}\delta_R' M_{\eta\eta}^\star(\bar{\eta}_R, \gamma^\star)\, \delta_R \\ \geq& \sqrt{P}\delta_R'\mathcal{M}_{\eta\eta}(\eta^\star, \gamma^\star)\delta_R + O_p(\|\delta_R\|) + o(\sqrt{P}\|\delta_R\|^2) + \inf_{\eta \in \mathcal{E}} \frac{\sqrt{P}}{2}\delta_R' M_{\eta\eta}^\star(\bar{\eta}_R, \gamma^\star)\, \delta_R\end{aligned} \tag{3}$$

Further, since $\sqrt{P}\|\delta_R\|^2 = C > 0$,

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) - C^2\delta'\mathcal{M}_{\eta\eta}(\eta^\star, \gamma^\star)\delta \leq & O_p(\|\delta_R\|) + o(\sqrt{P}\|\delta_R\|^2) + \sup_{\eta\in\mathcal{E}} \frac{\sqrt{C^2}}{2}\delta'\mathcal{M}_{\eta\eta}\left(\eta, \gamma^\star\right)\delta \\ \geq & O_p(\|\delta_R\|) + o(\sqrt{P}\|\delta_R\|^2) + \inf_{\eta\in\mathcal{E}} \frac{\sqrt{C^2}}{2}\delta'\mathcal{M}_{\eta\eta}\left(\eta, \gamma^\star\right)\delta.\end{aligned}$$

Since $\mathcal{M}_{\eta\eta}(\eta, \gamma^\star)$ is continuous in η, the last terms in the inequalities are bounded. Letting $S_1 = \sup_{\eta\in\mathcal{E}} \frac{\sqrt{C^2}}{2}\delta'\mathcal{M}_{\eta\eta}\left(\eta, \gamma^\star\right)\delta$ and $S_2 = \inf_{\eta\in\mathcal{E}} \frac{\sqrt{C^2}}{2}\delta'\mathcal{M}_{\eta\eta}\left(\eta, \gamma^\star\right)\delta$ the above analysis has demonstrated that

$$S_2 + o_p(1) \leq \sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) - C^2\delta'\mathcal{M}_{\eta\eta}(\eta^\star, \gamma^\star) \leq S_1 + o_p(1).$$

Hence, depending on S_1 and S_2, we have show that $\lim_T \Pr\left(|\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star)| > 0\right) \lesseqgtr \alpha$.

Case (3). With $\delta_R = \delta/T^\xi$, and $\xi > 1/4$, we see that $\sqrt{P}\|\delta_R\|^2 = C\|T^{1/2-2\xi}\|$. Hence, when $\xi > 1/4$, $\sqrt{P}\|\delta_R\|^2 = o(1)$. Applying this into the upper bound in equation (3) implies that

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \theta_R^\star) \leq & \sqrt{P}\delta_R'\mathcal{M}_{\eta\eta}(\eta^\star, \gamma^\star)\delta_R + O_p(\|\delta_R\|) + o(\sqrt{P}\|\delta_R\|^2) + \sup_{\eta\in\mathcal{E}} \frac{\sqrt{P}}{2}\delta_R' M_{\eta\eta}^\star\left(\bar{\eta}_R, \gamma^\star\right)\delta_R \\ & \leq O_p(\sqrt{P}\|\delta_R\|^2)\{1 + o_p(1)\} + O_p(\|\delta_R\|) \\ & = o_p(1)\end{aligned}$$

□

Proof of Lemma 1. Recalling that

$$\Delta_P(\vartheta_T^\delta, \tilde{\theta}_R) = P^{-1}\{L_P(\eta_T^\delta, \tilde{\gamma}_R) - L_P(\tilde{\eta}_R, \tilde{\gamma}_R)\} = -P^{-1}\left\{L_P(\tilde{\eta}_R, \tilde{\gamma}_R) - L_P(\eta_T^\delta, \tilde{\gamma}_R)\right\},$$

the result follows by applying Lemma 2, in particular equation (10) obtained in the proof. For $\delta_T \asymp \delta/T^\xi$ with $\xi \in (0, \infty]$, equation (10) implies that

$$\begin{aligned}\Delta_P(\vartheta_T^\delta, \tilde{\theta}_R) = & \left(\sqrt{P}(\tilde{\eta}_R - \eta_T^\delta)'\mathcal{J}^{1/2}\left[-\mathcal{J}^{-1/2}\left\{\nabla_\eta L_P(\theta^\star)/\sqrt{P} + \mathcal{M}_{\eta\gamma}\sqrt{P}(\tilde{\gamma}_R - \gamma^\star)\right\}\right]\right)/P \\ & - \left\{\frac{1}{2}\sqrt{P}(\eta_T^\delta - \tilde{\eta}_R)'\mathcal{J}\sqrt{P}(\eta_T^\delta - \tilde{\eta}_R)\right\}/P + R_P(\theta)/P,\end{aligned}$$

which, recalling the definition of $V_{P,R}$, can be re-arranged as

$$\Delta_P(\vartheta_T^\delta, \tilde{\theta}_R) = (\tilde{\eta}_R - \eta_T^\delta)'\mathcal{J}^{1/2}V_{P,R} - \frac{1}{2}(\eta_T^\delta - \tilde{\eta}_R)'\mathcal{J}(\eta_T^\delta - \tilde{\eta}_R) + R_P(\theta)/P$$

Adding and subtracting $\frac{1}{2}\|V_{P,R}\|^2$, we can re-arrange the above equation as

$$\Delta_P(\vartheta_T^\delta, \tilde{\theta}_R) = \frac{1}{2}\|V_{P,R}\|^2 - \frac{1}{2}\|\mathcal{J}^{1/2}\left(\eta_T^\delta - \eta^\star\right) - \mathcal{J}^{1/2}\left(\tilde{\eta}_R - \eta^\star\right) - V_{P,R}\|^2 + R_P(\theta)/P \quad (4)$$

Under the maintained assumption that $R \asymp P$, from equation (11) in the proof of Lemma 2,

$$\begin{aligned} R_P(\vartheta_P)/P &\leq P^{-1}o_p\left\{1 + \|\sqrt{P}(\eta_T^\delta - \eta^\star)\| + \|\sqrt{P}\left(\eta_T^\delta - \eta^\star\right)\|^2 + \|\sqrt{P}(\tilde{\theta}_R - \theta^\star)\|^2\right\} \\ &\equiv o_p\left\{1 + P^{-1/2}\|\eta_T^\delta - \eta^\star\| + \|\eta_T^\delta - \eta^\star\|^2 + \|(\tilde{\theta}_R - \theta^\star)\|^2\right\}, \end{aligned}$$

where the last term follows for any $\eta_T^\delta - \eta^\star = o_p(1)$.

□

Proof of Corrolary 1. The result follows by appropriately manipulating the expansion in (4). Multiplying the RHS of equation (4) by P, and re-arranging terms yields

$$\Delta_P(\vartheta_T^\delta, \tilde{\theta}_R) = -\frac{1}{2}\|\mathcal{J}^{1/2}\sqrt{P}\left(\eta_T^\delta - \eta^\star\right) - \mathcal{J}^{1/2}\sqrt{P}\left(\tilde{\eta}_R - \eta^\star\right) - \sqrt{P}V_{P,R}\|^2 + \frac{1}{2}\|\sqrt{P}V_{P,R}\|^2 + R_P(\theta)$$

Recalling the definition of $V_{P,R}$: (i) since $R \asymp P$, we have that $\sqrt{P}V_{P,R} = O_p(1)$, and a similar argument shows that $\|\sqrt{P}(\tilde{\eta}_R - \eta^\star)\| = \|\frac{\sqrt{P}}{\sqrt{R}}\sqrt{R}(\tilde{\eta}_R - \eta^\star)\| = O_p(1)$; (ii) since $\eta_T^\delta - \eta^\star = o_p(1)$ under the maintained assumptions, the remainder term vanishes as $T \to +\infty$.

Using the definitions of X_P and Z_P stated before the corollary, points (i) and (ii) allow us to rewrite the display equation as

$$\begin{aligned} P \cdot \Delta_P(\vartheta_T^\delta, \tilde{\theta}_R) = &-\frac{1}{2}\left\|\mathcal{J}^{1/2}\sqrt{P}(\eta_T^\delta - \eta^\star) - \left\{\mathcal{J}^{1/2}\sqrt{P}(\tilde{\eta}_R - \eta^\star) - \mathcal{J}^{-1/2}\left(\sqrt{P}X_P + c^{-1}\mathcal{M}_{\eta\gamma}\sqrt{R}Z_{R,\gamma}\right)\right\}\right\|^2 \\ &+ \frac{1}{2}\|\mathcal{J}^{-1/2}\left(\sqrt{P}X_P + c^{-1}\mathcal{M}_{\eta\gamma}\sqrt{R}Z_{R,\gamma}\right)\|^2 + o_p(1). \end{aligned}$$

Case (i). When $\delta_T \asymp \delta/T^{\xi}$, with $\xi \in (0, 1/2)$, we have that $\sqrt{P}(\eta_T^{\delta} - \eta^{\star}) \asymp \sqrt{P}T^{-\xi} \asymp T^{1/2-\xi} \to +\infty$ as $T \to +\infty$ under the maintained assumption on P, R, T. Consequently, since

$$\left\{\mathcal{J}^{1/2}\sqrt{P}(\tilde{\eta}_R - \eta^{\star}) - \mathcal{J}^{-1/2}\left(\sqrt{P}X_P + \mathcal{M}_{\eta\gamma}\sqrt{P}Z_{R,\gamma}\right)\right\} = O_p(1),$$

we have that $P \cdot \Delta_P(\vartheta_T^{\delta}, \tilde{\theta}_R) \to +\infty$ as $T \to +\infty$.

Case (ii). Firstly, we note that when $\theta^{\star} \in \mathrm{Int}(\Theta)$, standard results on the asymptotic behavior of two-step estimators can be used to show that

$$\mathcal{J}^{1/2}(\tilde{\eta}_R - \eta^{\star}) = -\mathcal{J}^{-1/2}\{\nabla_{\eta} L_R(\theta^{\star})/R + \mathcal{M}_{\eta\gamma}(\tilde{\gamma}_R - \gamma^{\star})\} + o_p(1/\sqrt{R})$$

Applying the definitions of X_P, and $Z_{R,\gamma}$, and the above equation implies that, recalling that $c = \lim_T P/R$,

$$\begin{aligned}\sqrt{P}\mathcal{J}^{1/2}(\tilde{\eta}_R - \eta^{\star}) - \sqrt{P}V_{P,R} =& -\frac{1}{c}\mathcal{J}^{-1/2}\{X_R/\sqrt{R} + \mathcal{M}_{\eta\gamma}\sqrt{R}Z_{R,\gamma}\} \\ &+ \mathcal{J}^{-1/2}\{X_P/\sqrt{P} + \frac{1}{c}\mathcal{M}_{\eta\gamma}\sqrt{R}Z_{R,\gamma}\} + o_p(1) \\ =& c^{-1}\sqrt{R}V_{R,R} - \sqrt{P}V_{P,R}\end{aligned}$$

Hence, if $(c^{-1}\sqrt{R}V_{R,R}, \sqrt{P}V_{P,R}) \Rightarrow \mathcal{V} = (\mathcal{V}_1', \mathcal{V}_2')'$, which is guaranteed under Assumption 5, then, by the continuous mapping theorem,

$$\mathcal{J}^{1/2}\sqrt{P}(\tilde{\eta}_R - \eta^{\star}) - \sqrt{P}V_{P,R} \Rightarrow \mathcal{V}_1 - \mathcal{V}_2.$$

When $\delta_T = \delta/T^{1/2}$, we have that $\sqrt{P}(\eta_T^{\delta} - \eta^{\star}) = \delta\sqrt{P}T^{-\xi} = \delta\left(\frac{P}{R+P}\right)^{1/2}$. From the maintained assumption on R, P, T, we have that

$$\left(\frac{P}{R+P}\right)^{1/2} \to \{1/(1+c)\}^{1/2} \quad \text{as } T \to +\infty.$$

Hence, applying the above two displayed equations yields

$$-\frac{1}{2}\|\mathcal{J}^{1/2}\sqrt{P}\left(\eta_T^{\delta} - \eta^{\star}\right) - \mathcal{J}^{1/2}\sqrt{P}\left(\tilde{\eta}_R - \eta^{\star}\right) - \sqrt{P}V_{P,R}\|^2 \Rightarrow \|\{1/(1+c)\}^{1/2}\delta - \mathcal{V}_1 + \mathcal{V}_2\|^2.$$

Further, recalling the definition of $V_{P,R}$, and since $\sqrt{P}V_{P,R} \to \mathcal{V}_2$, we have that

$$+\frac{1}{2}\|\mathcal{J}^{-1/2}\left(\sqrt{P}X_P + c^{-1}\mathcal{M}_{\eta\gamma}\sqrt{R}Z_{R,\gamma}\right)\|^2 \equiv \frac{1}{2}\|\sqrt{P}V_{P,R}\|^2 \Rightarrow \|\mathcal{V}_2\|^2 \equiv \frac{1}{2}\|X + c^{-1}\mathcal{M}_{\eta\gamma}Z_\gamma\|^2$$

Since $(c^{-1}\sqrt{R}V_{R,R}, \sqrt{P}V_{P,R}) \Rightarrow \mathcal{V} = (\mathcal{V}_1', \mathcal{V}_2')'$, any continuous transformation of these components also converges in distribution. Hence, the stated result follows.

Case (iii). The result follows precisely as in case (ii) by taking $\delta = 0$. □

Proof of Theorem 2. By Assumption 1, $P \asymp R$ and we can apply the conclusion of Lemma 3 to obtain the following expansion for the loss differences of the two-step combinations:

$$\begin{aligned}\sqrt{P}\{L_P(\tilde{\theta}_R)/P - L_P(\theta^\star)/P\} &= P^{-1/2}\{L_P(\tilde{\theta}_R) - L_P(\theta^\star)\} \\ &= \sqrt{P}(\tilde{\eta}_R - \eta^\star)'\nabla_\eta L_P(\eta^\star, \gamma^\star)/P \\ &+ \sqrt{P}(\tilde{\gamma}_R - \gamma^\star)'\nabla_\gamma L_P(\eta^\star, \gamma^\star)/P + o_p(1 + \|\sqrt{P}(\tilde{\theta}_R - \theta^\star\|^2)\end{aligned}$$

However, By part (i) of Lemma 2, we have $\sqrt{P}(\tilde{\theta}_R - \theta^\star) \asymp \sqrt{R}(\tilde{\theta}_R - \theta^\star) = O_p(1)$ (under Assumption 1). Applying the above, and the fact that $\{\nabla_\eta L_P(\theta^\star)\}/P = o_p(1)$, yields

$$\begin{aligned}\sqrt{P}\{L_P(\tilde{\theta}_R)/P - L_P(\theta^\star)/P\} &= o_p\{\|\sqrt{P}(\tilde{\eta}_R - \eta^\star)\|\} + \nabla_\gamma\mathcal{L}(\theta^\star)\sqrt{P}(\tilde{\gamma}_R - \gamma^\star) + o_p(\|\sqrt{P}(\tilde{\gamma}_R - \gamma^\star)\|) \\ &= \nabla_\gamma\mathcal{L}(\theta^\star)\sqrt{P}(\tilde{\gamma}_R - \gamma^\star) + o_p(1). \qquad (5)\end{aligned}$$

From part (ii) of Lemma 1, with $R \asymp P$,

$$\sqrt{P}\{L_P(\widehat{\theta}_R)/P - L_P(\theta^0)/P\} = O_p(1/\sqrt{P}) \qquad (6)$$

Subtracting the two expansion in (5) and (6), we have that

$$\begin{aligned}\sqrt{P}\Delta_P(\tilde{\theta}_R, \widehat{\theta}_R) &= \sqrt{P}\left\{L_P(\tilde{\theta}_R)/P - L_P(\widehat{\theta}_R)/P\right\} \\ &= \sqrt{P}\left\{[L_P(\tilde{\theta}_R) - L_P(\theta^\star)]/P - [L_P(\widehat{\theta}_R) - L_P(\theta^0)]/P\right\} + \sqrt{P}\{L_P(\theta^\star)/P - L_P(\theta^0)/P\} \\ &= \sqrt{P}\left\{L_P(\theta^\star)/P - L_P(\theta^0)/P\right\} + O_p(P^{-1/2}) + o_p(P^{-1/2}) + \nabla_\gamma\mathcal{L}(\theta^\star)'\sqrt{P}(\tilde{\gamma}_R - \gamma^\star) \\ &= \sqrt{P}\left\{\mathcal{L}(\theta^\star) - \mathcal{L}(\theta^0)\right\} + O_p(P^{-1/2}) + [1 : \nabla_\gamma\mathcal{L}(\theta^*)']Z_P\end{aligned}$$

with Z_P as defined in Assumption 5. By the hypothesis in Assumption 5, $[1 : \nabla_\gamma \mathcal{L}(\theta^*)']Z_P$ is asymptotically normal with zero mean and variance $\Omega = [1 : \nabla_\gamma \mathcal{L}(\theta^*)']Q[1 : \nabla_\gamma \mathcal{L}(\theta^*)']'$.

Now, define $\tilde{Z}_P := [1 : \nabla_\gamma \mathcal{L}(\theta^*)']Z_P/\sqrt{\Omega}$, and consider the probability

$$\begin{aligned}\Pr\left[\Delta_P(\tilde{\theta}_R, \widehat{\theta}_R)/\sqrt{\Omega} \leq 0\right] &= \Pr\left[\tilde{Z}_P + O_p(P^{-1/2}) + \sqrt{P}\left\{\mathcal{L}(\theta^\star) - \mathcal{L}(\theta^0)\right\}/\sqrt{\Omega} \leq 0\right] \\ &= \Pr\left[\tilde{Z}_P + O_p(P^{-1/2}) \leq \frac{\sqrt{P}}{\sqrt{\Omega}}\left\{\mathcal{L}(\theta^0) - \mathcal{L}(\theta^\star)\right\}\right].\end{aligned}$$

Define $z_P := \sqrt{P}\left\{[\mathcal{L}(\theta^0) - \mathcal{L}(\theta^\star)]\right\}/\sqrt{\Omega}$ and note that, for any $P \geq 1$, $z_P < 0$, since, by Assumption 3, $\mathcal{L}(\theta^0) < \mathcal{L}(\theta^\star)$, and $z_P \to -\infty$ as $P \to \infty$.

Since $\tilde{Z}_P$ is asymptotically standard normal,

$$\begin{aligned}\Pr\left[\Delta_P(\tilde{\theta}_R, \widehat{\theta}_R)/\sqrt{\Omega} \leq 0\right] &= \Phi(z_k) + \left\{\Pr(\tilde{Z}_P \leq z_P) - \Phi(z_k)\right\} + \left\{\Pr[\tilde{Z}_P \leq z_P + o_p(1)] - \Pr(\tilde{Z}_P \leq z_P)\right\} \\ &\leq \Phi(z_k) + \sup_z |\Pr(\tilde{Z}_P \leq z_P) - \Phi(z_k)| + o(1).\end{aligned}$$

Fix $\varepsilon > 0$. Since $z_k \to -\infty$ as $P \to \infty$, for some P large enough we can conclude that $\Phi(z_k) \leq \varepsilon/2$. From the convergence $\tilde{Z}_P \Rightarrow N(0,1)$, and the continuity of $\Phi(z)$, Polya's Theorem implies that for some P large enough,

$$\sup_z \left|\Pr(\tilde{Z}_P \leq z) - \Phi(z)\right| \leq \varepsilon/2,$$

and for some P large enough $\Pr[\Delta_P(\tilde{\theta}_R, \widehat{\theta}_R)/\sqrt{\Omega} \leq 0] \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, the results follows. □

C Proofs of Key Lemmas

Proof of Lemma 1. To prove part (i), we first prove consistency of $\widehat{\theta}_n$ for some $\vartheta \in \Theta_{\mathrm{I}}$, which can be proven by verifying the sufficient conditions in Theorem 2 of Yuan and Jennrich (1998). By 4(i), the function $\nabla_\theta L_n(\vartheta) = O_p(1/\sqrt{n})$ for each $\vartheta \in \Theta_{\mathrm{I}}$, which satisfies Assumption 1 of Yuan and Jennrich (1998). From Assumption 4(i), for any $\vartheta \in \Theta_{\mathrm{I}}$,

there exists a neighbourhood $\mathcal{N}(\vartheta)$, such that for each $\theta \in \mathcal{N}(\vartheta)$, $L_n(\theta)$ has continuous second-order left/right partial derivatives, $\nabla_{\theta\theta} L_n(\theta)$, and by Assumption 4(iii), $\nabla_{\theta\theta} L_n(\theta)/n$ converges uniformly to $\nabla_{\theta\theta}\mathcal{L}(\theta)$, which is non-singular for each $\vartheta \in \Theta_{\mathrm{I}}$. Thus, Assumption 2 of Yuan and Jennrich (1998) is satisfied, and we can conclude that $\widehat{\theta}_n := \arg_{\theta\in\Theta}\{\nabla_\theta L_n(\theta) = 0\}$ converges to ϑ, for some $\vartheta \in \Theta_{\mathrm{I}}$.

Given that $\widehat{\theta}_n = \vartheta + o_p(1)$, for some $\vartheta \in \Theta_{\mathrm{I}}$, the remainder of the result follows similar arguments to Theorem 1 of Andrews (1999). In particular, from Theorem 6 in Andrews (1999) for l/r differentiable functions, the following Taylor series expansion is valid: for some $\vartheta \in \Theta_{\mathrm{I}}$,

$$\begin{aligned} L_n(\theta) - L_n(\vartheta) &= \sqrt{n}(\theta-\vartheta)'\nabla_\theta L_n(\vartheta)/\sqrt{n} + \frac{1}{2}\sqrt{n}(\theta-\vartheta)'\left[\nabla_{\theta\theta} L_n(\bar{\theta})/n\right]\sqrt{n}(\theta-\vartheta)' \\ &= \sqrt{n}(\theta-\vartheta)'\nabla_\theta L_n(\vartheta)/\sqrt{n} + \frac{1}{2}\sqrt{n}(\theta-\vartheta)'\left[\nabla_{\theta\theta}\mathcal{L}(\vartheta)\right]\sqrt{n}(\theta-\vartheta) \\ &+ \frac{1}{2}\sqrt{n}(\theta-\vartheta)'\left[\nabla_{\theta\theta}\mathcal{L}_n(\bar{\theta})/n - \nabla_{\theta\theta}\mathcal{L}(\vartheta)\right]\sqrt{n}(\theta-\vartheta) \end{aligned} \tag{7}$$

for $\bar{\theta}$ an intermediate value satisfying $\|\vartheta - \bar{\theta}\| \leq \|\theta - \vartheta\|$. Clearly,

$$R_n(\theta) := \frac{1}{2}\sqrt{n}(\theta-\vartheta)'\left[\nabla_{\theta\theta}\mathcal{L}_n(\bar{\theta})/n - \nabla_{\theta\theta}\mathcal{L}(\vartheta)\right]\sqrt{n}(\theta-\vartheta) \leq \frac{1}{2}\|\sqrt{n}(\theta-\vartheta)\|^2\|\nabla_{\theta\theta}\mathcal{L}_n(\bar{\theta})/n - \nabla_{\theta\theta}\mathcal{L}(\vartheta)\|,$$

so that for any $\|\theta-\vartheta\| = o(1)$, we have that the remainder term in (7) is $O(\|\sqrt{n}(\theta-\vartheta)\|^2)$ by Assumption 4(iii).

Recalling the notations $\mathcal{M}_{\theta\theta}(\theta) = \nabla_{\theta\theta}\mathcal{L}(\theta)$, $\mathcal{J} = \mathcal{M}_{\theta\theta}(\vartheta)$, and $Z_n = -\mathcal{J}^{-1/2}\nabla_\theta L_n(\vartheta)/\sqrt{n}$, we obtain

$$L_n(\theta) - L_n(\vartheta) = -\sqrt{n}(\theta-\vartheta)'\mathcal{J}^{1/2}Z_n + \frac{1}{2}\sqrt{n}(\theta-\vartheta)'\mathcal{J}\sqrt{n}(\theta-\vartheta) + R_n(\theta).$$

From the definition of $\widehat{\theta}_n$, we have $L_n(\vartheta) \geq L_n(\widehat{\theta}_n)$, so that applying the above equation yields

$$0 \geq L_n(\widehat{\theta}_n) - L_n(\vartheta) = -\sqrt{n}(\widehat{\theta}_n-\vartheta)'\mathcal{J}^{1/2}Z_n + \frac{1}{2}\sqrt{n}(\widehat{\theta}_n-\vartheta)'\mathcal{J}\sqrt{n}(\widehat{\theta}_n-\vartheta) + R_n(\widehat{\theta}_n).$$

The above equation is precisely the (negative of) equation (7.3) in the proof of Theorem 1 in Andrews (1999) and the remainder of the proof follows the same argument.

Having shown that $\sqrt{n}(\widehat{\theta}_n - \vartheta) = O_p(1)$, for some $\vartheta \in \Theta_{\mathrm{I}}$, the proof of part (ii) of the stated result follows the same argument as in Theorem 2 part (b) of Andrews (1999).

□

Proof of Lemma 2. To prove the first stated result, we use Theorem 6 in Andrews (1999) for l/r differentiable functions to obtain the following Taylor series expansion:

$$\begin{aligned} L_n(\eta, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) &= \sqrt{n}(\eta - \eta^\star)' \nabla_\eta L_n(\eta^\star, \tilde{\gamma}_n)/\sqrt{n} + \frac{1}{2}\sqrt{n}(\eta - \eta^\star)' \left[\nabla_{\eta\eta} L_n(\bar{\eta}, \tilde{\gamma}_n)/n\right] \sqrt{n}(\eta - \eta^\star)' \\ &= \sqrt{n}(\eta - \eta^\star)' \left\{\nabla_\eta L_n(\theta^\star)/\sqrt{n} + n^{-1}\nabla_{\eta\gamma} L_n(\eta^\star, \bar{\gamma})[\sqrt{n}(\tilde{\gamma}_n - \eta^\star)]\right\} \\ &\quad + \frac{1}{2}\sqrt{n}(\eta - \eta^\star)' \left[n^{-1}\nabla_{\eta\eta} L_n(\bar{\eta}, \tilde{\gamma}_n)\right] \sqrt{n}(\eta - \eta^\star)' \end{aligned}$$

for $\bar{\eta}$, $\bar{\gamma}$, intermediate values satisfying $\|\eta^\star - \bar{\eta}\| \leq \|\eta - \eta^\star\|$ and $\|\gamma^\star - \bar{\gamma}\| \leq \|\tilde{\gamma}_n - \gamma^\star\|$, respectively. Letting

$$\begin{aligned} R_n(\eta, \tilde{\gamma}_n) :=& \sqrt{n}\,(\eta - \eta^\star)' \left\{[\nabla_{\eta\eta} L_n(\bar{\eta}, \tilde{\gamma}_n)/n] - [\nabla_{\eta\eta}\mathcal{L}(\bar{\eta}, \tilde{\gamma}_n)]\right\} \sqrt{n}\,(\eta - \eta^\star) \\ &- \sqrt{n}\,(\eta - \eta^\star)' \left\{[\nabla_{\eta\eta}\mathcal{L}(\theta^\star)] - [\nabla_{\eta\eta}\mathcal{L}(\bar{\eta}, \tilde{\gamma}_n)]\right\} \sqrt{n}\,(\eta - \eta^\star) \\ &+ \sqrt{n}(\eta - \eta^\star)\{n^{-1}\nabla_{\eta\gamma} L_n(\eta^\star, \bar{\gamma}) - \nabla_{\eta\gamma}\mathcal{L}(\theta^\star)\}\sqrt{n}(\tilde{\gamma}_n - \gamma^\star), \end{aligned} \tag{8}$$

we have

$$\begin{aligned} L_n(\eta, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) &= \sqrt{n}(\eta - \eta^\star)' \left\{\nabla_\eta L_n(\eta^\star, \tilde{\gamma}_n)/\sqrt{n} + \nabla_{\eta\gamma}\mathcal{L}(\theta^\star)\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)\right\} \\ &\quad + \frac{1}{2}\sqrt{n}(\eta - \eta^\star)' \left[\nabla_{\eta\eta}\mathcal{L}(\theta^\star)\right] \sqrt{n}(\eta - \eta^\star)' + R_n(\eta, \tilde{\gamma}_n). \end{aligned} \tag{9}$$

Recalling the notations $\mathcal{M}_{\eta\eta}(\theta) = \nabla_{\eta\eta}\mathcal{L}(\theta)$, $\mathcal{J} = \mathcal{M}_{\eta\eta}(\theta^\star)$, and $V_n = -\mathcal{J}^{-1/2}\{\nabla_\eta L_n(\theta^\star)/\sqrt{n} + \mathcal{M}_{\eta\gamma}(\theta^\star)\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)\}$, we obtain

$$L_n(\eta, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) = -\sqrt{n}(\eta - \eta^\star)'\mathcal{J}^{1/2}(\theta^\star)V_n + \frac{1}{2}\sqrt{n}(\eta - \eta^\star)'\mathcal{J}\sqrt{n}(\eta - \eta^\star) + R_n(\theta). \tag{10}$$

From the definition of $\tilde{\eta}_n$, we have $L_n(\eta^\star, \tilde{\gamma}_n) \geq L_n(\tilde{\eta}_n, \tilde{\gamma}_n)$, so that applying the above equation yields

$$0 \geq L_n(\tilde{\eta}_n, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) = -\sqrt{n}(\tilde{\eta}_n - \eta^\star)'\mathcal{J}^{1/2}V_n + \frac{1}{2}\sqrt{n}(\tilde{\eta}_n - \eta^\star)'\mathcal{J}\sqrt{n}(\tilde{\eta}_n - \eta^\star) + R_n(\tilde{\theta}_n).$$

Now, consider $R_n(\tilde{\theta}_n) = R_{1n}(\tilde{\theta}_n) + R_{2n}(\tilde{\theta}_n) + R_{3n}(\tilde{\theta}_n)$, corresponding to each of the three terms in (8). From the consistency of $\tilde{\theta}_n$, there exist some $\delta_n = o(1)$ such that $\|\tilde{\theta}_n - \theta^\star\| \leq \delta_n$ with probability converging to one. For the first term, we have

$$|R_{1n}(\tilde{\theta}_n)| \leq \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|^2 \sup_{\|\eta - \eta^\star\| \leq \delta_n} \|\nabla_{\eta\eta} L_n(\bar{\eta}, \tilde{\gamma}_n)/n - \nabla_{\eta\eta}\mathcal{L}(\bar{\eta}, \tilde{\gamma}_n)\| = o_p\left(\|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|^2\right)$$

where the equality follows from Assumption 4(iv). Similarly, for the third term we have

$$\begin{aligned} |R_{3n}(\tilde{\theta}_n)| &\leq \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|^2 \sup_{\|\eta - \eta^\star\| \leq \delta_n} \|\nabla_{\eta\gamma} L_n(\eta^\star, \bar{\gamma})/n - \nabla_{\eta\gamma}\mathcal{L}(\eta^\star, \gamma^\star)\| \\ &= o_p\left(\|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|^2 \vee \|\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)\|\right), \end{aligned}$$

where the $o_p(1)$ term follows by consistency of $\tilde{\gamma}_n$, the definition of the intermediate value, $\bar{\gamma}$, i.e., $\|\bar{\gamma} - \gamma^\star\| \leq \|\gamma^\star - \tilde{\gamma}_n\|$, and continuity of the second derivatives (Assumption 4(i)). For the second term, $|R_{2n}(\tilde{\theta}_n)|$, a similar argument to the above shows

$$|R_{2n}(\tilde{\theta}_n)| \leq o_p(1)(1 + \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|\|\sqrt{n}(\tilde{\gamma}_n - \gamma^\star)\|) = o_p(1 + \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|)O_p(1) = o_p(\|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|)$$

where the $O_p(1)$ term in the second equality follows by Assumption 2. Putting all the terms together yields

$$|R_n(\tilde{\theta}_n)| \leq o_p\{1 + \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\| + \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|^2\}. \tag{11}$$

Now, note that, for $\vartheta_n = (\eta^\star, \tilde{\gamma}_n)^\top$, we have that

$$n \cdot \Delta_n(\tilde{\theta}_n, \vartheta_n) \equiv L_n(\tilde{\eta}_n, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n).$$

Recalling $\mathcal{J} := \mathcal{M}_{\eta\eta}(\theta^\star)$, let $\kappa_n := \mathcal{J}^{1/2}\sqrt{n}(\tilde{\eta}_n - \eta^\star)$, where $\mathcal{J}^{1/2}$ exists by Assumption 4(iii). Using the definitions, V_n, κ_n and equation (11), we can re-arrange

$$0 \geq L_n(\tilde{\eta}_n, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) \equiv n \cdot \Delta_n(\tilde{\theta}_n, \vartheta_n)$$

as

$$\begin{aligned}
0 \geq n\Delta_n(\tilde{\theta}_n, \vartheta_n) &= -\kappa_n' V_n + \frac{1}{2}\kappa_n'\kappa_n + o_p\{1 + \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\| + \|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|^2\} \\
&= -O_p(\|\kappa_n\|) + \|\kappa_n\|^2/2 + o_p(1)\{1 + \|\mathcal{J}^{-1/2}\kappa_n\| + \|\mathcal{J}^{-1/2}\kappa_n\|^2\} \\
&= -\{O_p(1) - o_p(1)\}\|\kappa_n\| + \|\kappa_n\|^2/2 + o_p(\|\kappa_n\|^2) + o_p(1), \\
&= -\zeta_n\|\kappa_n\| + \|\kappa_n\|^2/2 + o_p(\|\kappa_n\|^2) + o_p(\|\kappa_n\|),
\end{aligned}$$

where ζ_n denotes the $\{O_p(1) - o_p(1)\}$ term. Rearranging the LHS of the above equation we have

$$\zeta_n^2 \geq [\|\kappa_n\|\{1 + o_p(1)\} - \zeta_n]^2 + o_p(1)$$

and we obtain $\|\kappa_n\| \leq O_p(1)$. Since $\mathcal{J}$ is non-singular, by Assumption 4(iv), there exists some $c > 0$ such that

$$O_p(1) \geq \|\kappa_n\| = \|\mathcal{J}^{1/2}\sqrt{n}(\tilde{\eta}_n - \eta^\star)\| \geq c\|\sqrt{n}(\tilde{\eta}_n - \eta^\star)\|$$

and part (i) of the result follows.

To establish part (ii), we use part (i) and the expansion

$$L_n(\eta, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) = -\sqrt{n}(\eta - \eta^\star)'\mathcal{J}^{1/2}V_n + \frac{n}{2}(\eta - \eta^\star)'\mathcal{J}(\eta - \eta^\star) + R_n(\tilde{\theta}_n) + o_p(\|\eta - \eta^\star\|),$$

which, from part (i), we can re-arrange as

$$\begin{aligned}
n \cdot \Delta_n(\tilde{\theta}_n, \vartheta_n) = L_n(\tilde{\eta}_n, \tilde{\gamma}_n) - L_n(\eta^\star, \tilde{\gamma}_n) &= -\kappa_n' V_n + \frac{1}{2}\kappa_n'\kappa_n + o_p(1)\{1 + \|\kappa_n\| + \|\kappa_n\|^2\} \\
&= \frac{1}{2}\|\kappa_n - V_n\|^2 - \frac{1}{2}V_n'V_n + o_p(1). \qquad (12)
\end{aligned}$$

The $o_p(1)$ term follows since, by part (i) of the result and Assumption 2, we have $\|\tilde{\theta}_n - \theta^\star\| = O_p(1)$, so that we have $R_n(\tilde{\theta}_n) = o_p(1)$.

□

Proof of Lemma 3. Again, Theorem 6 of Andrews (1999) allows us to expand $L_n(\eta, \gamma)$ around $\theta = \theta^\star$:

$$L_n(\eta,\gamma) - L_n(\eta^\star,\gamma^\star) = \sqrt{n}(\eta-\eta^\star)'\nabla_\eta L_n(\theta^\star)/\sqrt{n} + \sqrt{n}(\gamma-\gamma^\star)'\nabla_\gamma L_n(\theta^\star)/\sqrt{n} + \frac{n}{2}(\theta-\theta^\star)'\nabla_{\theta\theta}\mathcal{M}(\theta^\star)(\theta-\theta^\star) + R_n(\theta),$$

where the remainder term $R_n(\theta)$ is given by

$$R_n(\theta) = \frac{n}{2}(\theta-\theta^\star)'\left[\nabla_{\theta\theta}L_n(\bar{\theta})/n - \mathcal{L}(\bar{\theta})\right](\theta-\theta^\star) + \frac{n}{2}(\theta-\theta^\star)'\left[\nabla_{\theta\theta}\mathcal{L}(\bar{\theta}) - \mathcal{L}(\theta^\star)\right](\theta-\theta^\star).$$

and $\bar{\theta}$ an intermediate value that satisfies $\|\theta^\star - \bar{\theta}\| \leq \|\theta - \theta^\star\|$.

However, since $\|\sqrt{n}(\tilde{\theta}_n - \theta^\star)\| = O_p(1)$, by Lemma 2(a), Assumption 4(iv) and the definition of the intermediate value, we have that $|R_n(\tilde{\theta}_n)| \leq o_p\{1 + \|\sqrt{n}(\tilde{\theta}_n - \theta^\star)\|^2\}$. Consequently,

$$n^{-1/2}\{L_n(\tilde{\theta}_n) - L_n(\theta^\star)\} = \sqrt{n}(\tilde{\theta}_n - \theta^\star)'\frac{1}{n}\begin{pmatrix}\nabla_\eta L_n(\theta^\star)\\ \nabla_\gamma L_n(\theta^\star)\end{pmatrix} + \frac{n^{-1/2}}{2}\|\mathcal{M}_{\theta\theta}(\theta^\star)^{1/2}\sqrt{n}(\tilde{\theta}_n - \theta^\star)\|^2 + o_p(1/\sqrt{n})$$

$$= (\tilde{\eta}_n - \eta^\star)'\nabla_\eta L_n(\theta^\star)/\sqrt{n} + \sqrt{n}(\tilde{\gamma}_n - \gamma^\star)'\nabla_\gamma L_n(\theta^\star)/n + O_p(1/\sqrt{n}) + o_p(1).$$

From Assumption 4(ii), $\nabla_\eta L_n(\theta^\star)/\sqrt{n} = O_p(1)$, while from Lemma 2, $\|\tilde{\eta}_n - \eta^\star\| = o_p(1)$, and the first term above is $o_p(1)$.

By Assumption 4(iv), $\nabla_\gamma L_n(\theta^\star)/n$ converges in probability to $\nabla_\gamma \mathcal{L}(\theta^\star)$. Moreover, by 3(iii), $\theta^\star$ is only a solution to $\nabla_\eta \mathcal{L}(\theta) = 0$, so that $\nabla_\gamma \mathcal{L}(\theta^\star) \neq 0$. Therefore, we can conclude that

$$n^{-1/2}\{L_n(\tilde{\theta}_n) - L_n(\theta^\star)\} = o_p(1) + \sqrt{n}(\tilde{\gamma}_n - \gamma^\star)'\nabla_\gamma \mathcal{L}(\theta^\star).$$

□

D Numerical Implementation Details

D.1 Extending the Findings of Smith and Wallis (2009)

In this section, we detail the computational steps used to produce the results in Section 2.2, which extend the simulation exercise given in Section 3.1 of Smith and Wallis (2009). For convenience, the results are reproduced in Figure 1.

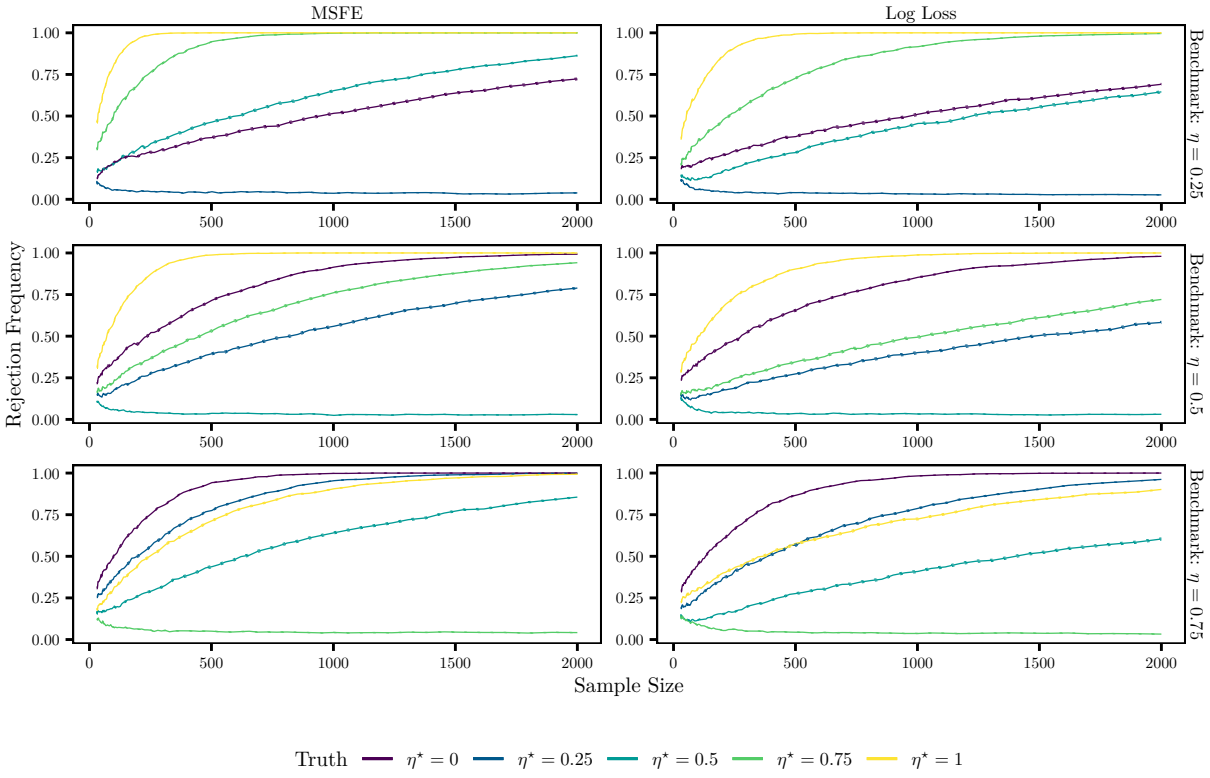

Figure 1: Estimates (solid) and their 95% confidence intervals (dotted) of the Rejection Frequency (y-axis) for the hypothesis test of no inferior predictive accuracy of a benchmark forecast combination with fixed weights (rows) against the alternative optimal forecast combination. The test is conducted with observations drawn from DGPs across a range of pseudo-true weights (colors), and across a grid of sample sizes (x-axis). Results for a point forecast combination with optimal weights minimizing the MSFE are given in the first column, and results for a distributional forecast combination with optimal weights minimizing the log loss are given in the second column.

Recall that this simulation exercise is conducted based on data drawn from the following zero-mean AR(2) process:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t,\ \epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2),$$

where the data is used to estimate the parameters in the standard two-step fashion, by minimizing the log loss or the MSFE of a distributional or point forecast combination, respectively, based the following model:

$$f_1^{(t)}(y) = N\{y; \gamma_1 y_{t-1}, 1\},$$
$$f_2^{(t)}(y) = N\{y; \gamma_2 y_{t-2}, 1\},$$
$$f^{(t)}(y) = \eta f_1^{(t)}(y) + (1-\eta) f_2^{(t)}(y),$$

where $N\{x; \mu, \Sigma\}$ denotes the normal pdf evaluated at x with mean μ and variance Σ, γ_1 and γ_2 are the parameters of the constituent models, and η is the weight assigned to the first model.

To produce the results, we must first produce the corresponding DGP parameters given a selected value for the true optimal combination weight $\eta^\star$. Then for each such $\eta^\star$ value of interest, we draw samples from the associated induced DGP to conduct the suite of tests represented in Figure 1, and iterate on these steps to produce accurate estimates of the rejection frequency.

D.1.1 Obtaining the DGP Parameters

Given a desired value for $\eta^\star$, we conduct the following steps to obtain DGP parameters $(\phi_1, \phi_2, \sigma_\epsilon^2)$ for which the actual value for $\eta^\star$ is as desired.

1. Draw $z_t \overset{i.i.d.}{\sim} N(0,1)$ for $t = 1, 2, \ldots, 10^7$.

2. Let $y_0 = y_{-1} = 0$ and $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \sigma_\epsilon z_t$ for $t = 1, 2, \ldots, 10^7$.

3. Solve the following equality-constrained optimization program:

$$\min_{\phi_1,\phi_2,\sigma_\epsilon^2,\gamma,\eta} \quad -1.1\log(1+\phi_2) - \log(1-\phi_1-\phi_2) - 2\log(1+\phi_1-\phi_2) - 0.1\log(\phi_1^2+\phi_2^2) \tag{13}$$

$$\text{s.t.} \quad \eta = \eta^\star, \tag{14}$$

$$\text{Var}(y_t) = 1, \tag{15}$$

$$\frac{1}{10^7}\sum_{t=1}^{10^7} \frac{\partial}{\partial\eta} S(\eta,\gamma,y_t) = 0, \tag{16}$$

$$\frac{1}{10^7}\sum_{t=1}^{10^7} \frac{\partial}{\partial\gamma_j} S_j(\gamma_j,y_t) = 0,\ j=1,2, \tag{17}$$

where $S(\eta,\gamma,y_t)$ and $S_j(\gamma_j,y_t)$ is the score[3] of the forecast combination and constituent forecast j, respectively, of y_t.

We produce results for S being either the MSFE, in which case

$$\begin{aligned} S(\eta,\gamma,y_t) &= (\mathbb{E}_{f^{(t)}}[y_t] - y_t)^2 \\ &= (\eta\mathbb{E}_{f_1^{(t)}}[y_t] + (1-\eta)\mathbb{E}_{f_2^{(t)}}[y_t] - y_t)^2 \\ &= (\eta\gamma_1 y_{t-1} + (1-\eta)\gamma_2 y_{t-2} - y_t)^2, \\ S_j(\gamma_j,y_t) &= (\mathbb{E}_{f_j^{(t)}}[y_t] - y_t)^2 \\ &= (\gamma_j y_{t-j} - y_t)^2, \end{aligned}$$

or S being the log loss, so that

$$\begin{aligned} S(\eta,\gamma,y_t) &= -\log f^{(t)}(y_t) \\ &= -\log(\eta f_1^{(t)}(y_t) + (1-\eta)f_2^{(t)}(y_t)) \\ &= -\log\Big(\eta(2\pi)^{-1}\exp\Big(-\frac{1}{2}(y_t-\gamma_1 y_{t-1})^2\Big) + (1-\eta)(2\pi)^{-1}\exp\Big(-\frac{1}{2}(y_t-\gamma_2 y_{t-2})^2\Big)\Big), \\ S_j(\gamma_j,y_t) &= -\log f_j^{(t)}(y_t) \\ &= -\log(2\pi) - \frac{1}{2}(y_t-\gamma_j y_{t-j})^2. \end{aligned}$$

The criterion function given in Equation (13) above is designed to ensure that the DGP is strictly stationary and that $\eta^\star$ is identified. The DGP is stationary if and only

[3] A score, referring to either a scoring rule or scoring function, is a measure of forecast accuracy. See Section 3 for details.

if the roots of its characteristic function lie inside the unit circle, which corresponds to parameter values that satisfy $\phi_2 > -1$, $\phi_2 < 1 - \phi_1$ and $\phi_2 < 1 + \phi_1$. This space of parameter values is sometimes referred to as the 'stationarity triangle', because it forms a triangle in Euclidean space. Inspecting the first three logarithmic terms in Equation (13) reveals that the criterion function is undefined for parameter values that do not correspond to a stationary process, and that the criterion function approaches infinity as the vector (ϕ_1, ϕ_2) approaches the boundary of the stationarity triangle. The final logarithmic term in Equation (13) ensures that the criterion function approaches infinity as the vector (ϕ_1, ϕ_2) approaches zero, where $\eta^\star$ is not identified since $\gamma_1^\star = \gamma_2^\star = 0$. Figure 2 displays the stationarity triangle and contours of the criterion function, which is minimized at $(\phi_1, \phi_2) = (0.38, 0.14)$ when unconstrained. Note that the constants $(-4, -1, -2, -0.1)$ with which the logarithmic terms are multiplied in the criterion function are arbitrary, and any other collection of negative constants would also suffice.

The constraints in Equations (14) - (17) ensure that the desired value for the weight $\eta^\star$ is actually obtained by the resulting DGP, and that the unconditional variance of the DGP is identical across different values of $\eta^\star$. The value of the weight itself is chosen in Equation (14), and the unconditional variance of the DGP is fixed at unity by the constraint in Equation (15). After solving the Yule-Walker equations for the AR(2) process, we obtain the following expression for the unconditional variance of y_t as a function of the parameters $(\phi_1, \phi_2, \sigma_\epsilon^2)$:

$$\text{Var}(y_t) = \frac{\sigma_\epsilon^2(1 - \phi_2)}{(1 - \phi_1^2 - \phi_2^2)(1 - \phi_2) - 2\phi_1^2\phi_2}.$$

The constraint in Equation (16) is the first-order condition for optimizing the weights according to the chosen score, across a gigantic sample of 10^7 observations. By the law of large numbers for stationary processes, this average can approximate $\mathbb{E}_{DGP}[\partial S(\eta, \gamma, y_t)/\partial \eta]$

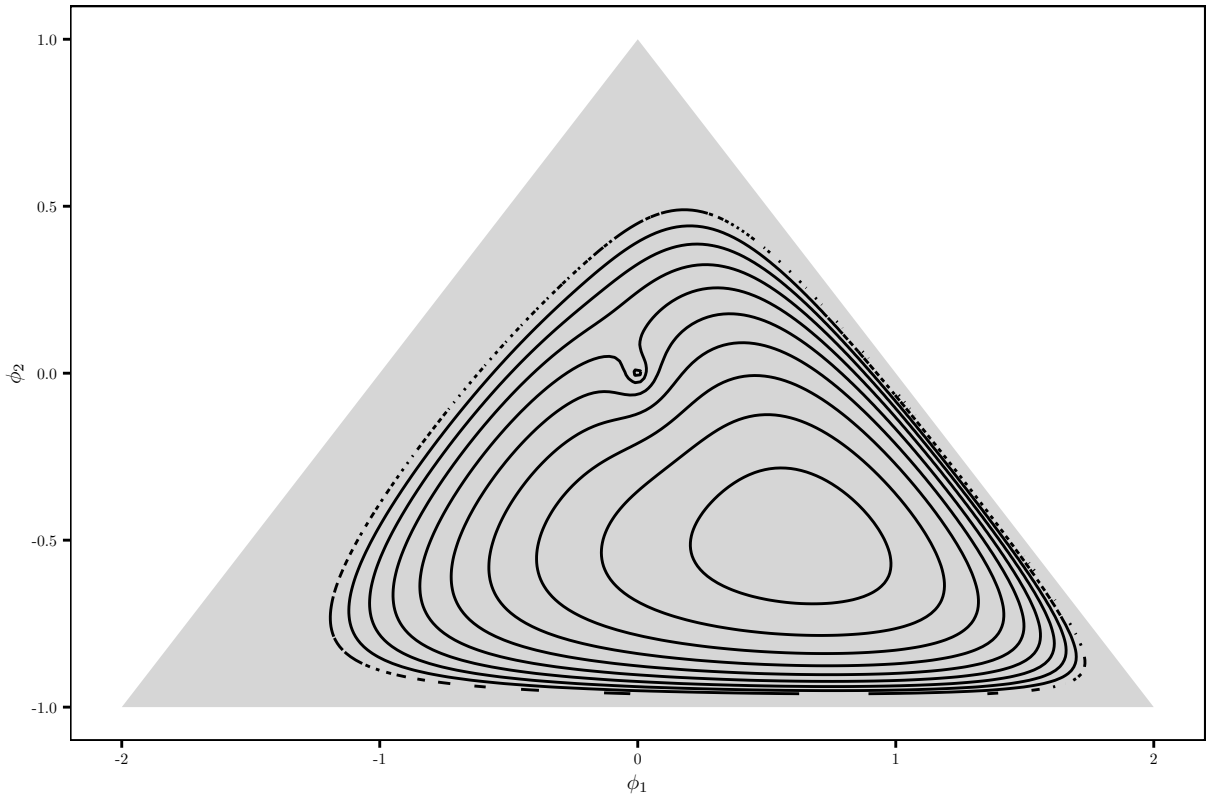

Figure 2: The set of values for the AR(2) parameters (ϕ_1, ϕ_2) for which the resulting stochastic process is strictly stationary (gray), and contours of the criterion function in Equation (13) in a neighborhood of its minimum at $(\phi_1, \phi_2) = (0.38, 0.14)$ (black).

to any desired degree of accuracy for a large enough sample size. Similarly, the constraint in Equation (17) ensures that the first-order condition for optimizing the constituent model parameters are satisfied according to the chosen score, across the same gigantic sample, and with a large enough sample size the average in the constraint for the j^{th} constituent model can approximate $\mathbb{E}_{DGP}[\partial S_j(\gamma_j, y_t)/\partial \gamma_j]$ to any desired degree of accuracy. Note that by Assumptions 1 - 4, $(\eta, \gamma) = (\eta^\star, \gamma^\star)$ is the solution to the problem of setting the above expectations to zero, just as the constraints (16) and (17) set the corresponding averages to zero.[4] Since the sample size of 10^7 is orders of magnitude larger than the largest hypothesis

[4]See Section 3.1.3 for more details on estimating optimal forecast combinations, where Equations (2) and (3) define the two-step estimated combinations for which (16) and (17) are the first-order conditions.

test sample size of 2000 considered in the results displayed in Figure 1, we assume that there is a negligible difference between the desired weight imposed in Equation (15) and the actual optimal weight $\eta^\star$ corresponding to the DGP which solves the optimization program.

We code the optimization program in R, and solve it using the *nloptr* package, a library of nonlinear optimization routines (Johnson, 2022). In particular, we solve the optimization problem using sequential quadratic programming, with an implementation based on Kraft (1988, 1994). See Gill et al. (2021) for a textbook treatment of sequential quadratic programming.

D.1.2 Estimating Rejection Frequency

Once we have a means of obtaining an AR(2) DGP with a selected pseudo-true weight $\eta^\star$, producing the required estimates of the rejection frequencies plotted in Figure 1 is straightforward. Briefly, for each pseudo-true weight $\eta^\star$ (colors), each benchmark weight η (rows), each score (columns), and each sample size (x-axis), we conduct 5000 hypothesis tests of no inferior forecast accuracy of the benchmark combination against the alternative optimal combination using 5000 samples of the required sample size drawn from a DGP with the required $\eta^\star$. The rejection frequency (y-axis) is then estimated as the proportion of those tests that reject the null hypothesis. In detail, the steps are as follows.

1. Obtain the DGP parameters $(\phi_1, \phi_2, \sigma_\epsilon^2)$ corresponding to each desired pseudo-true weight $\eta^\star \in \{0, 0.25, 0.5, 0.75, 1\}$ for each score $S \in \{\text{MSFE}, \log \text{loss}\}$ using the method detailed in Appendix D.1.1 above.

2. For each score $S = \{\text{MSFE}, \log \text{ loss}\}$, each pseudo-true weight $\eta^\star \in \{0, 0.25, 0.5, 0.75, 1\}$ and each benchmark weight $\eta \in \{0.25, 0.5, 0.75\}$:

   (a) Draw the sample $y_{1:2000}^{(1)}$ from the DGP corresponding to S and $\eta^\star$ that was found

in Step 1.

(b) For each such sample, consider the truncated samples $y^{(1)}_{1:T+1}$ for $T+1 = 30, 32, \ldots, 2000$, and for each of these conduct a hypothesis test of no inferior forecast accuracy of: a) the benchmark forecast combination with the weight fixed at η, against b) the alternative forecast combination with optimal weights, where c) forecast accuracy is measured according to S, and d) the in-sample and out-of-sample sizes are equal. Record $R^{(1)}_{T+1} = 1$ where the test is rejected at the 5% level, and $R^{(1)}_{T+1} = 0$ otherwise. This hypothesis test is described in Section 3.2.2, where $R = P = (T+1)/2$.

3. Repeat Step 2 5000 times, drawing 5000 samples $y^{(1)}_{1:T+1}, y^{(2)}_{1:T+1}, \ldots, y^{(5000)}_{1:T+1}$ and obtaining hypothesis test results $R^{(i)}_{T+1}$ for each $i = 1, 2, \ldots, 5000$, $T+1 = 30, 32, \ldots, 2000$, S, $\eta^\star$ and η.

4. Estimate the rejection frequencies with the statistic $\hat{r}_{T+1} = \frac{1}{5000}\sum_{i=1}^{5000} R^{(i)}_{T+1}$, and calculate the corresponding 95% confidence intervals using standard asymptotics for i.i.d. Bernoulli random variables.

D.2 Cause of the puzzle

In this section, we address the computational steps used to produce the results in Section 4.4, which are re-displayed in Figure 3 for convenience. These steps are almost identical to those described in Section D.1 above, with one difference. Whereas in Section D.1.2 we conduct the hypothesis test using a benchmark two-step combination with fixed weights $\eta \in \{0.25, 0.5, 0.75\}$ and an optimally-weighted two-step alternative, here we use the (benchmark, alternative) pairs $\{(\tilde{\theta}_T, \hat{\theta}_T), (\theta^{ew}_R, \hat{\theta}_T), (\theta^{ew}_R, \tilde{\theta}_T)\}$, which you can see in the labels for the rows of Figure 3.

References

Andrews, D. W. (1999). Estimation when a parameter is on a boundary. *Econometrica 67*(6), 1341–1383. 5, 7, 8, 16, 17, 20

Gill, P. E., W. Murray, and M. H. Wright (2021). *Numerical linear algebra and optimization*. SIAM. 26

Johnson, S. G. (2022). The nlopt nonlinear-optimization package. 26

Kraft, D. (1988). A software package for sequential quadratic programming. *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*. 26

Kraft, D. (1994). Algorithm 733: Tomp–fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software (TOMS) 20*(3), 262–281. 26

Smith, J. and K. F. Wallis (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics 71*(3), 331–355. 1, 21

van der Vaart, A., A. W. van der Vaart, A. van der Vaart, and J. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media. 6

Yuan, K.-H. and R. I. Jennrich (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis 65*(2), 245–260. 15, 16

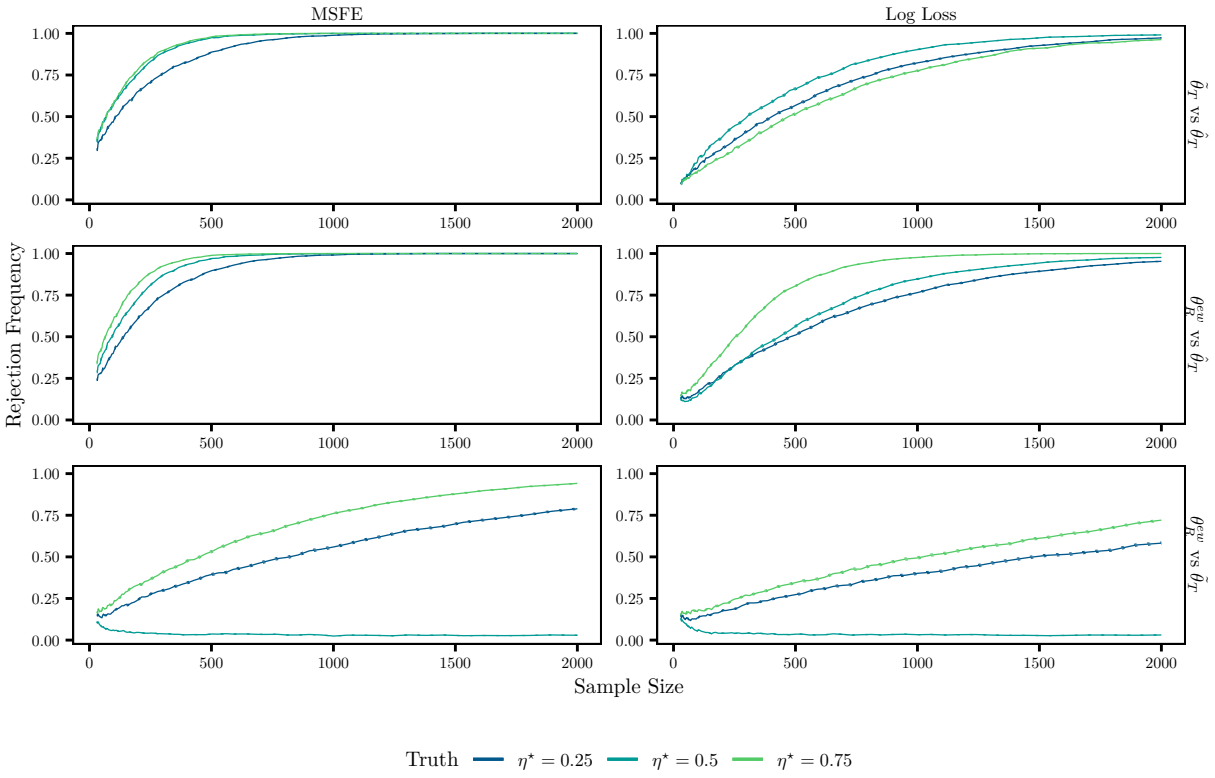

Figure 3: Estimates (solid) and their 95% confidence intervals (dotted) of the Rejection Frequency (y-axis) for the hypothesis test of no inferior predictive accuracy of a benchmark forecast combination against an alternative combination (rows, benchmark vs alternative). The test is conducted with observations drawn from DGPs across a range of pseudo-true weights (colors), and across a grid of sample sizes (x-axis). Results for a point forecast combination with optimal weights minimizing the MSFE are given in the first column, and results for a distributional forecast combination with optimal weights minimizing the log loss are given in the second column.