Probability forecasts of binary events are often gathered from multiple models and averaged to provide inputs regarding uncertainty in important decision-making problems. Averages of well-calibrated probabilities are underconfident, and methods have been proposed to make them more extreme. To aggregate probabilities, we introduce a large class of ensembles that are generalized additive models. These ensembles are based on Bayesian principles and can help us learn why and when extremizing is appropriate. Extremizing is typically viewed as shifting the average probability farther from one-half; we emphasize that it is more suitable to define extremizing as shifting it farther from the base rate. We also introduce the notion of anti-extremizing to learn if it might sometimes be beneficial to make average probabilities less extreme. Analytically, we find that our Bayesian ensembles often extremize the average forecast, but sometimes anti-extremize instead. On two publicly available datasets, we demonstrate that our Bayesian ensemble performs well and anti-extremizes in about 20% of the cases. It anti-extremizes much more often when there is bracketing with respect to the base rate among the probabilities being aggregated than with no bracketing, suggesting that bracketing is a promising indicator of when we should consider anti-extremizing.

Key words: Forecast aggregation; linear opinion pool; generalized linear model, extremizing and anti-extremizing, bracketing, probit ensemble.

Date: September 5, 2019

1. Introduction

Many organizations face forecasting challenges that involve binary events. These forecasts are critical to decisions such as the approval of credit (probability of default), the recommendation of a drug (probability of having a disease), or the choice of a national security response (probability of a geopolitical event occurring). Often a set of models generates probabilities for such events. To aggregate the individual probabilities, we offer a new method based on Bayesian principles.
This method generalizes several proposed aggregators in the literature, exhibits some structural advantages, and may be more accurate in practice.

Since its introduction by Stone (1961), many researchers have found the linear opinion pool to be an attractive way to aggregate forecasts (DeGroot 1974, Genest and Zidek 1986, DeGroot and Mortera 1991). The most popular way to aggregate forecasts is to take the average of the models’ forecasts, which is a linear opinion pool with equal weights (Clemen and Winkler 1986, Larrick and Soll 2006). In aggregating binary-events forecasts, Winkler and Poses (1993, p. 1533) state, “Simple averages of forecasts seem to work as well as or better than fancier combining methods.”

Nonetheless, Hora (2004) and Ranjan and Gneiting (2010) show that when the forecasts being aggregated are well-calibrated, the linear opinion pool is underconfident. For binary events, the linear opinion pool is, on average, not extreme enough. To address this problem, Ranjan and Gneiting (2010) propose a method that extremizes the linear opinion pool by pushing it closer to its nearest extreme (either one or zero). Others have employed schemes to extremize the average forecast (Karmarkar 1978, Turner et al. 2014, Satopää et al. 2014). Extremizing methods, such as the logit aggregator, are now used as benchmarks in practice (IARPA Geopolitical Forecasting Challenge 2018).

The existing methods, however, are heuristics that do not follow from Bayesian principles. Consequently, they risk being sub-optimal. The idea that a decision maker can use Bayesian reasoning to aggregate forecasts goes back at least to Winkler (1968) and Morris (1974). “[T]he probability assessment is a representation of his state of information; to the decision maker, the probability assessment is information.” (Morris 1974, p. 1241) Many other researchers have proposed models along these lines (Lindley et al. 1979, French 1980, Clemen 1987).

Dawid et al. (1995) make an important contribution to the literature on aggregating binary-event forecasts. They introduce the first set of aggregation methods based on Bayesian principles and a fundamental condition regarding calibration. The condition says that if the decision maker uses forecasts from only one calibrated model, he adopts that model’s forecast as his own. Recently, Satopää et al. (2016), building on the work of Dawid et al. (1995) and generalizing an example in Ranjan and Gneiting (2010), introduce a probit aggregator that is consistent with Bayesian principles and model calibration. We call such an aggregator a Bayesian ensemble in the spirit of Chipman et al. (2007).

In this paper, we derive a new class of Bayesian ensembles, which we call conjugate-pair Bayesian ensembles. This class includes some existing ensembles, such as the probit aggregator. Each ensemble in the class is derived from a joint distribution of the information models may have. Throughout,
we assume this joint distribution is described by a conjugate pair of distributions. The use of familiar conjugate pairs helps us build intuition for how a Bayesian ensemble emerges from the information models may contain. While we are not the first to approach forecast aggregation as an information aggregation problem—notable examples in this vein include Winkler (1981) and more recently Satopää et al. (2016)—we are the first to explore the material effect prior information can have on extremizing in a binary-event context. In this context, prior information takes the form of a prior-predictive probability. One can think of a prior-predictive probability as an event’s base rate frequency implied by a forecasting model and known by the decision maker, and we will use the terms “prior-predictive probability” and “base rate” interchangeably. For example, when forecasting rain in Phoenix, Arizona, both the decision maker and the aggregation procedure use the same base rate at which rain occurs daily, say 10%.

Our approach is fully Bayesian and involves model aggregation, but it should not be confused with Bayesian model selection, which implicitly assumes that one of the models being aggregated is the true model. We are almost always in a world where there is no true model, or if one insists on thinking in terms of a true model, it is not one of the models in our ensemble (Winkler et al. 2018). Therefore, our setting is information aggregation rather than model selection. Forecast aggregation is called stacking in the machine learning literature, and Yao et al. (2018) use stacking to average distributions in a Bayesian model selection setting. In our empirical study we use stacking to form our Bayesian ensemble by aggregating the information the models provide.

In practice, there is often not sufficient training data from which to identify an event’s base rate or the accuracy of each forecast being combined. We show that for these situations, because our model includes underlying uncertainties for the models’ probabilities, the reported probabilities themselves can be used to develop good estimates of the parameters in the ensemble. In particular, we derive unbiased estimators of the prior-predictive probability and models’ sample sizes. We call the ensemble based on these estimates a reference Bayesian ensemble.

We provide new insights into when it is appropriate for the aggregate forecast to extremize the average forecast. In our setting, it often makes sense to extremize the average forecast, but sometimes it does not. In some cases, it can make sense to anti-extremize, or make the aggregate forecast less extreme than the average forecast. The term anti-extremize is new; we introduce this term to highlight the importance of making the aggregate forecast less extreme than the average forecast when appropriate.

For the most part, the decision maker naturally wants to form an aggregate forecast that extremizes the average forecast because each individual forecast contains some weight on the prior informa-
tion. For example, with two models both reporting forecasts on the same side of the prior-predictive probability, the average forecast double counts the prior information and is pulled too far back in the direction of the prior-predictive probability. If the models’ forecasts, however, bracket the prior predictive probability, i.e. are on opposite sides of the prior-predictive probability, this pull toward the prior-predictive probability may not be enough or the average forecast may end up on the wrong side of the prior-predictive probability. Larrick and Soll (2016) introduced the notion of bracketing when averaging point forecasts, and Grushka-Cockayne et al. (2017) extended it to the averaging of quantiles.

In an empirical study of two large datasets, we apply our Bayesian ensemble and show that it performs well in comparison with several leading aggregation methods. It anti-extremizes in about 20% of the cases in each dataset, demonstrating that anti-extremizing can be a desirable option under some circumstances. Moreover, it anti-extremizes much more often when bracketing with respect to the base rate occurs among the probabilities being combined than in cases where such bracketing is not present. This suggests that such bracketing is a promising indicator of when we should give careful consideration to anti-extremizing.

This paper is concerned with aggregating probabilities from models. Schemes to extremize the average of subjective forecasts from experts have been proposed (Erev et al. 1994, Ariely et al. 2000, Shlomi and Wallsten 2010, Mellers et al. 2014, Baron et al. 2014). There is also an extensive literature indicating that subjective probabilities from experts have often been found to be overconfident (e.g., Wallsten et al. 1997). When experts tend to be overconfident, averages of their probabilities will become less overconfident and possibly underconfident. With fairly strong overconfidence, averages of just a few experts may still be overconfident (e.g., Gaba et al. 2017), in which case some anti-extremizing could be appropriate. Of course, models can be overconfident too, especially where there is overfitting, just as some experts are well-calibrated. Regardless of the source of the forecasts (models, experts, machine learning, perhaps some of each), sometimes extremizing could be a good idea and other times anti-extremizing would do better. Thus, our suggestion of bracketing (no bracketing) as an indicator of when anti-extremizing (extremizing) can be considered might apply more widely than just to averages of models’ forecasts.

2. Conjugate-Pair Bayesian Ensembles

In this section, we introduce our Bayesian ensemble. Before introducing this ensemble, we introduce the information structure that defines the joint distribution of the base models use to issue forecasts. After that, we present our main result on the form of the conjugate-pair Bayesian ensemble.
2.1. Information Structure

Below we introduce an information structure that describes the information the decision maker and base forecasting models have in order to make their forecasts. In this setting, each of \( k \geq 2 \) models generates a probability forecast for the binary event \( y \) given model \( i \)'s sample information \( x_i \). The event \( y \) equals 1 if the event occurs, or 0 if the event does not occur. We denote model \( i \)'s forecast \( P(y = 1 | x_i) \) by \( p_i \). After learning the forecasts from the \( k \geq 2 \) models, the decision maker aggregates the models' forecasts into a single forecast.

A decision maker’s aggregate forecast is a Bayesian ensemble, denoted by \( \hat{p} \), if it is the posterior-predictive probability \( P(y = 1 | p_1, \ldots, p_k) \) derived from the joint distribution of \( (p_1, \ldots, p_k, y) \) using Bayes’ Theorem. This aggregate forecast is optimal because using any other forecast would yield a worse score when evaluated by a proper scoring rule (Gneiting and Raftery 2007).

Suppose there is an exchangeable sequence of data points:

\[
\begin{align*}
(x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2}, \ldots, x_{N_{k-1}+1}, \ldots, x_{N_k}, x_{N_k+1}).
\end{align*}
\]

(1)

Model \( i \) sees only the sample \( x_i = (x_{N_i-1+1}, \ldots, x_{N_i}) \) of size \( n_i \) for \( i = 1, \ldots, k \), where \( N_i = \sum_{l=1}^{i} n_l \).

The final data point \( x_{N_k+1} \), abbreviated as \( x \), is related to the binary event \( y \), which is what the decision maker ultimately cares about. If \( x \) is in the event occurrence set \( A \), then \( y \) equals 1; otherwise, \( y \) equals 0. For example, if \( x \) is a Bernoulli random variable, the set \( A \) might simply be \{1\}. Alternatively, if \( x \) is a normal random variable, the set \( A \) might be the interval \((0, \infty)\).

Data points in the sequence are independent and identically distributed according to a likelihood from a regular, one-parameter exponential family with probability mass or density function \( f(x_j | \theta) = a(x_j) b(\theta) \exp(c(\theta)h(x_j)) \). The parameter \( \theta \) is distributed according to a conjugate prior \( f(\theta) = [K(\tau_0, \tau_1)]^{-1} [b(\theta)]^{\tau_0} \exp(c(\theta)\tau_1) \), where \( \tau_0 \) and \( \tau_1 \) are the prior’s hyperparameters and \( K(\tau_0, \tau_1) \) is its normalizing constant. The joint distribution of \( (\theta, x_1, \ldots, x_k, x) \) and the event occurrence set \( A \) are common knowledge among the decision maker and the models. The prior/likelihood pair of distributions that describes this joint distribution is called a conjugate pair (Raiffa and Schlaifer 1961, Bernardo and Smith 2000). Hence we call the Bayesian ensemble that a conjugate pair generates a conjugate-pair Bayesian ensemble.

Based on these assumptions, the following function generates a set of predictive distributions—one for the prior-predictive, one for each model’s posterior-predictive, and one for the decision maker’s posterior-predictive. We call this function the predictive generating function:

\[
F_n(t) = \int_{x \in A} a(x) \frac{K(\tau_0 + n + 1, t + h(x))}{K(\tau_0 + n, t)} dx,
\]

(2)
where \( n \) is the sample size and \( t \) is the sufficient statistic (Bernardo and Smith 2000). Note that when the random variable \( x \) is discrete, the integral in (2), and other integrals like it throughout the paper, naturally become sums. With \( n = 0 \) and \( t = \tau_1 \), \( F_0(\tau_1) \) is equal to the common prior-predictive probability \( P(y = 1) \), denoted by \( p_0 \). With \( n = n_i \) and \( t = t_i \) where \( t_i = \sum_{j=N_i-1}^{N_i} h(x_j) \), \( F_{n_i+n_s}(\tau_1 + t_i) \) is equal to model \( i \)'s posterior-predictive probability \( P(y = 1|\mathbf{x}_i, \mathbf{x}_s) \).

With \( n = N_k \) and \( t = \sum_{i=1}^{k} t_i \), \( F_{N_k}(\tau_1 + \sum_{i=1}^{k} t_i) \) is equal to the decision maker’s posterior-predictive probability \( P(y = 1|\mathbf{x}_1, \ldots, \mathbf{x}_k) \), as if he had access to all the models’ sample information. In reality, the decision maker learns \( p_i \) from model \( i \), and not \( \mathbf{x}_i \). He can, however, deduce \( t_i \), the sufficient statistic for \( \mathbf{x}_i \), from \( p_i \) if \( F_n \) is invertible. If it is invertible, then \( t_i = F_{n_i}^{-1}(p_i) - \tau_1 \).

### 2.2. Conjugate-Pair Ensemble

The main result below gives the form of any conjugate-pair Bayesian ensemble. The form is a generalized additive model in the models’ probabilities and a generalized linear model in the models’ sufficient statistics. For this result, we need the following two definitions. A generalized additive model links the conditional expectation of a quantity of interest \( y \) to an additive function of some covariates \( (q_1, \ldots, q_k) \): \( E[y|q_1, \ldots, q_k] = g^{-1}(g_0 + g_1(q_1) + \cdots + g_k(q_k)) \) where \( g \) is the link function, \( g_0 \) is a constant, and each \( g_i \) for \( i = 1, \ldots, k \) is a smooth function (Hastie and Tibshirani 1986). A generalized linear model is a general additive model where each \( g_i \) is a linear function (Nelder and Wedderburn 1972). Proofs of this and other results appear in the Online Supplement.

**Proposition 1 (Ensemble Form).** Assume the predictive generating function \( F_n(t) \) in (2) is strictly monotonic in \( t \). Then the conjugate-pair Bayesian ensemble of the models’ probabilities is the generalized additive model

\[
\hat{p} = P(y = 1|p_1, \ldots, p_k) = F_{N_k} \left( (k-1)F_0^{-1}(p_0) + \sum_{i=1}^{k} F_{n_i}^{-1}(p_i) \right), \tag{3}
\]

where \( p_0 = F_0(\tau_1) \), \( p_i = F_{n_i}(\tau_1 + t_i) \), and \( t_i = \sum_{j=N_i-1}^{N_i} h(x_j) \). Also, the Bayesian ensemble of the models’ sufficient statistics is the generalized linear model

\[
P(y = 1|t_1, \ldots, t_k) = F_{N_k} \left( \tau_1 + \sum_{i=1}^{k} t_i \right). \tag{4}
\]

This result provides a large class of Bayesian ensembles. The class is as large as the class of regular, one-parameter exponential families. Any ensemble in this class is a generalized additive model of forecast probabilities that first transforms the models’ probabilities into their corresponding information states, then linearly combines these information states, and finally transforms the combined information states back into the probability space. Generalized additive model for aggregation have
existed in the literature, e.g., Dawid et al. (1995), among others, consider some special cases of the form (3).

Below we provide three examples of conjugate-pair Bayesian ensembles. The second one is a variant of the models studied in Ranjan and Gneiting (2010) and Satopää et al. (2016). The other two are new. The Appendix provides details on the derivation of each example’s ensemble, which rely on Proposition 1.

**Example 1 (Gamma/Poisson Pair).** Let $x_j$ given $\theta$ be drawn from a Poisson distribution with rate $\theta$. The conjugate prior for this likelihood is the gamma distribution with shape $\alpha$ and rate $\beta$, denoted by $\theta \sim \text{Ga}(\alpha, \beta)$. Suppose $A = \{0\}$ corresponds to the event that a piece of equipment, with exponentially distributed inter-arrival times of breakdowns, does not break down in the next year. With $A = \{0\}$, this conjugate pair leads to the Bayesian ensemble

$$
\hat{p} = \exp \left( - (k - 1) \frac{v_N}{v_0} \log(p_0) + \sum_{i=1}^{k} \frac{v_N}{v_{n_i}} \log(p_i) \right),
$$

where $v_n = \log((\beta + n)/(\beta + n + 1))$. Alternatively, we can write this ensemble as $\hat{p} = p_0^{-(k-1)v_N/v_0} \prod_i p_i^{v_N/v_{n_i}}$, which some authors call a logarithmic opinion pool. Dawid et al. (1995) use the term “logarithmic opinion pool” to describe what we call the logit aggregator, described later. We think a better name for this pool is a geometric opinion pool, since the form resembles a geometric mean of the probabilities.

**Example 2 (Normal/Normal Pair).** Let $x_j$ given $\theta$ be drawn from a normal distribution with mean $\theta$ and variance $\sigma^2$. The conjugate prior for this likelihood is another normal distribution with mean $\theta_0$ and variance $\sigma_0^2$. Suppose $A = (0, \infty)$ corresponds to the event that a new product makes a profit in its first year. With $A = (0, \infty)$, this conjugate pair leads to the Bayesian ensemble

$$
\hat{p} = \Phi \left( w_0 \Phi^{-1}(p_0) + \sum_{i=1}^{k} w_i \Phi^{-1}(p_i) \right),
$$

where $\Phi$ is the cumulative distribution function (cdf) of the standard normal distribution, $w_0 = -(k - 1) \sqrt{v_0/N_k}$, $w_i = \sqrt{v_{n_i}/v_N}$, and $v_n = (\sigma^2/\sigma_0^2 + n)(\sigma^2/\sigma_0^2 + n + 1)$. We call this ensemble a *probit ensemble* because the inverse link function in this generalized linear model is the standard normal cdf.

**Example 3 (Generalized-Gamma/Gumbel Pair).** Let $x_j$ given $\theta$ be drawn from a Gumbel distribution with location $\theta$ and scale $\sigma$. The conjugate prior for this likelihood is the reflection of the generalized gamma distribution in Ahuja and Nash (1967, Equation 2.7): $\exp(\theta/\sigma) \sim \text{Ga}(\alpha, \beta)$. 

Suppose $A = (-\infty, 0)$ corresponds to the event that a hedge-fund manager’s best investment makes a loss in some year. With $A = (-\infty, 0)$, this conjugate pair leads to the Bayesian ensemble
\[
\hat{p} = \left( \frac{- (k - 1) p_0^{v_0^2} / (1 - p_0^{v_0^2}) + \sum_{i=1}^{k} p_i^{v_{n_i}} / (1 - p_i^{v_{n_i}})}{1 - (k - 1) p_0^{v_0^2} / (1 - p_0^{v_0^2}) + \sum_{i=1}^{k} p_i^{v_{n_i}} / (1 - p_i^{v_{n_i}})} \right)^{1 / v_N k},
\]
where $v_n = 1 / (\alpha + n)$.

3. Extremizing/Anti-Extremizing

In this section, we introduce a definition of extremizing/anti-extremizing and analyze the example ensembles from the previous section. Our main interest is to understand when it is appropriate to extremize the average forecast and when it is not.

Our definition of extremizing is inspired by the definitions of sharpness in Winkler and Jose (2008) and Ranjan and Gneiting (2010). Sharper, or more extreme forecasts, are those farther away from their marginal event frequencies. Winkler and Jose (2008) state, “If climatology $c$ is used as the baseline probability for probability of precipitation forecasts, sharpness should be viewed in terms of shifts from $c$ toward zero or one instead of shifts from 0.5 toward zero or one.” In forecasting rain in Phoenix, Arizona where the historical daily frequency of rain is about 10%, a forecast of 40% would naturally be considered more extreme than a forecast of 30%.

For the following definition, we assume the average forecast is not equal to either the prior-predictive probability (i.e., the forecast based on the prior information only) or the aggregate forecast.

**Definition 1 (Extremizing/Anti-Extremizing).** The aggregate forecast extremizes the average forecast if it is farther away from the prior-predictive probability in the same direction as the average forecast. Otherwise, the aggregate forecast anti-extremizes the average forecast.

If the average forecast is equal to either the prior-predictive probability or the aggregate forecast, we say the aggregate forecast neither extremizes nor anti-extremizes the average forecast.

The definition of extremizing above differs from other definitions in the literature. For example, Baron et al. 2014, Satopää et al. (2014), and Satopää et al. (2016) say the aggregate forecast extremizes the average forecast if it is farther away from one-half in the same direction as the average forecast. Under this definition, a forecast of 40% for rain in Phoenix would be considered less extreme than a forecast of 30%.

3.1. Examples

For Examples 1-3, we compare the Bayesian ensemble $\hat{p}$, the average forecast $\bar{p} = (1/k) \sum_{i=1}^{k} p_i$, and the prior-predictive probability $p_0$. See Figure 1. In each panel, we hold $p_1$ fixed and plot $p_2$.
versus $\hat{p}$, $\bar{\hat{p}}$, and $p_0$. Thus, each plot of $p_2$ versus $\hat{p}$ represents a slice of the ensemble surface. We plot these slices because it is easier, at least initially, to identify the intervals of extremizing and anti-extremizing. Later we take a look down from above at the ensemble surface and visualize the collections of these intervals, which form regions.

To generate all three plots in Figure 1, we let $k = 2$, $n_1 = n_2 = 2$, $\alpha = 1$, $\beta = 2$, and $\sigma_0 = \sigma = 1$. For the gamma/Poisson pair in Figure 1(a), we let $t_1 = 0$, which implies $p_1 = 0.8$. For the normal/normal pair in Figure 1(b), we also let $\theta_0 \approx -1.4657$ and $t_1 \approx -0.8708$, which implies $p_0 = 0.15$ and $p_1 = 0.25$. For the generalized-gamma/Gumbel pair in Figure 1(c), we also let $t_1 = 1$, which implies $p_1 \approx 0.4219$.

In Figure 1(a), we see that the Bayesian ensemble always extremizes the average forecast. In Figures 1(b) and 1(c), the Bayesian ensemble often extremizes the average forecast, but there is an interval (in gray) in each plot where the Bayesian ensemble anti-extremizes the average forecast. The interval of anti-extremizing tends to be where the models’ forecasts are on opposite sides of the prior-predictive probability.

The intuition for why a Bayesian ensemble tends to extremize the average forecast is that the models’ forecasts each have some weight on the prior information, but when their forecasts are combined, the aggregate forecast only needs that weight once. This intuition explains the coefficient $-(k - 1)$ in front of the term $F_0^{-1}(p_0)$ in (3). In other words, extremizing tends to be the result of removing the redundant weight on the prior information. Because the models’ probabilities are first non-linearly transformed inside the probit ensemble, it does not always extremize the average forecast.
3.2. Probit Ensemble

Below we demonstrate the existence of some anti-extremizing with the probit ensemble. For this result, we assume models are exchangeable (each \( n_i = n_1 \)) and the normal/normal conjugate pair describes the joint distribution in our information structure. Consider Figure 2, which show plots of \( p_2 \) versus \( \hat{p} \), \( \bar{p} \), and \( p_0 \) for a given \( p_1 \). For Figure 2(a), we use the same settings as those used for Figure 1(b). For Figure 2(b), we use these same settings, except we decrease \( p_1 \) to 0.05.

In terms of extremizing/anti-extremizing, there are four cases shown in these plots (ignoring ties): (1) \( \hat{p} < \bar{p} < p_0 \) (extremizing), (2) \( \hat{p} < p_0 < \bar{p} \) (anti-extremizing), (3) \( p_0 < \hat{p} < \bar{p} \) (anti-extremizing), and (4) \( p_0 < \bar{p} < \hat{p} \) (extremizing). Because we held \( p_1 \) fixed for these plots, these four cases correspond to four intervals for \( p_2 \). As we decrease \( p_1 \) from 0.25 to 0.05, we notice that the anti-extremizing intervals widen. This is because there is more room for \( p_2 \) to be on the opposite side of the prior-predictive probability \( (p_0 = 0.15) \) when \( p_1 \) is below the prior-predictive probability. The next figure illustrates this point.

For Figure 3, we take a view looking down from above on the surface of the probit ensemble. We keep the same settings as those used for Figure 2. In Figure 3(a), we see the regions that correspond to the four cases of extremizing (in white) or anti-extremizing (in gray) in Figure 2 as we vary \( p_1 \) along the entire unit interval. The slices that are Figure 2(a) and 2(b) are shown as the dashed line in Figure 3(a).

In Figure 3(b), we show two rectangles (in gray) as well as the outlines of the four extremizing/anti-extremizing regions. These two rectangles, which we call opposing-sides rectangles, correspond to pairs of \( p_1 \) and \( p_2 \) that are on opposite sides of the prior-predictive probability. The left rectangle has \( p_2 < p_0 < p_1 \), and the right rectangle has \( p_1 < p_0 < p_2 \). When the forecasts are on opposite sides of the prior-predictive probability, we see that the probit ensemble often anti-extremizes the average forecast. Being on opposite sides of the prior-predictive probability is not quite necessary for anti-extremizing, but it is close. There is tiny area of anti-extremizing in Figure 3(b) where both models’ forecasts are very near the prior-predictive probability, but are not on opposite sides of it.

Before moving on, we note that there are two other possible cases of anti-extremizing that do not occur in the two previous figures: (5) \( \bar{p} < \hat{p} < p_0 \) (anti-extremizing) and (6) \( \bar{p} < p_0 < \hat{p} \) (anti-extremizing). Because of the symmetry involved in the probit ensemble, we can generate these other two cases if we simply reverse the signs of \( \theta_0 \) and \( t_1 \) in the settings for Figure 2. Note that the fifth and sixth cases do occur in Figure 1(c) for the generalized-gamma/Gumbel pair.
Below we present results on the existence of extremizing/anti-extremizing for the probit ensemble in (5) with exchangeable models (each $n_i = n_1$).
Proposition 2 (Anti-Extremizing With the Probit Ensemble). Assume \( p_1 \leq \ldots \leq p_k \), \((q_1, \ldots, q_{j-1}) = (p_1, \ldots, p_{j-1})\), \( q_j = kp_0 - \sum_{i=1}^{k-1} p_i \), and \((q_{j+1}, \ldots, q_k) = (p_j, \ldots, p_{k-1})\) so that \( 0 < q_1 \leq \ldots \leq q_k < 1 \). Also, assume \((q_i + q_{k-i+1})/2 < 1/2\) for \( i = 1, \ldots, k \). Then \( p_0 < 1/2, \sum_{i=1}^{k-1} p_i/k < p_0 < (1 - \sum_{i=1}^{k-1} p_i)/k, \sum_{i=1}^{k-1} p_i < 1/2, \) and the probit ensemble anti-extremizes the average forecast according to case 2 \((\hat{p} < p_0 < \bar{p})\) for all \( p_k \) in the interval \((q_j, \Phi((1 - w_0)\Phi^{-1}(p_0))/w_k - \sum_{i=1}^{k-1} \Phi^{-1}(p_i)))\), which is an interval of positive length.

According to this result, anti-extremizing tends to happen when \( p_1, \ldots, p_{k-1} \) are all small \((\sum_{i=1}^{k-1} p_i < 1/2)\) and \( p_k \) is large \((p_k > kp_0 - \sum_{i=1}^{k-1} p_i)\). If \( \sum_{i=1}^{k-1} p_i < 1/3 \) and \( 2 \sum_{i=1}^{k-1} p_i/k < p_0 \), then \( \sum_{i=1}^{k-1} p_i < p_0 < p_k \), which is a case when bracketing occurs because the smallest \( k - 1 \) forecasts are on the opposite side of the prior-predictive probability compared to the largest forecast. This indicates that anti-extremizing might be beneficial when the base models bracket with respect to the base rate.

For this next result, we consider the special condition where \( p_0 = 1/2 \) and show that the probit ensemble always extremizes the average forecast under this condition. This special condition of \( p_0 = 1/2 \) is the only one considered in Satopää et al. (2016), which allowed them to show that their probit ensemble always extremizes.

Proposition 3 (No Anti-Extremizing With the Probit Ensemble). If \( p_0 = 1/2 \) and \( \bar{p} \neq 1/2 \), for all model probabilities, either \( \hat{p} < p_0 < \bar{p} \) (extremizing toward zero) or \( p_0 < \bar{p} < \hat{p} \) (extremizing toward one) holds.

Based on the above two results, we see the importance of considering (when appropriate) a prior-predictive probability not equal to 1/2. For times when the prior-predictive probability is not equal to 1/2, an ensemble, such as the probit ensemble, will not always extremize the average forecast. If we do always extremize the average forecast in these situations, we may introduce unnecessary overconfidence via our aggregation procedure. One question that arises in practice is how to set the prior-predictive probability when the decision maker is quite removed from the details of the forecasting context or when the aggregator is a machine. In such automated settings, we will want some way to estimate the prior-predictive probability. This estimation, in part, is the subject of the next subsection.

3.3. Estimation of an Ensemble

Below we give a result for estimating \( p_0 \) by \( \bar{p} \). We show that this estimate is unbiased. This estimator is attractive not only because it is unbiased, but also because it can be formed without a training set. Only the forecasts for the present quantity of interest are needed to form this estimate of \( p_0 \).
Proposition 4 (Estimation of the Prior-Predictive Probability). Assume a conjugate pair describes the joint distribution in our information structure. Then, $\bar{p}$ is an unbiased estimate of $p_0$, i.e., $E[\bar{p}] = p_0$.

Based on this estimate, we propose the following probit ensemble, which we call the reference probit ensemble, for use in practice:

$$\hat{p} = \Phi \left( - (k - 1) \sqrt{\frac{v_0}{v_{kn_1}}} \Phi^{-1}(\bar{p}) + \sum_{i=1}^{k} \sqrt{\frac{v_{n_1}}{v_{kn_1}}} \Phi^{-1}(p_i) \right),$$

where $v_n = (n_0 + n)(n_0 + n + 1)$. This ensemble assumes the base models are exchangeable. Without more information, we might let $n_0 = n_1 = 1$.

With the reference probit ensemble, it is always the case that we neither extremize nor anti-extremize the average forecast. This is because we set $p_0 = \bar{p}$ by construction. This construction, however, does not mean we will not see a version of extremizing (or lack thereof) based on the traditional definitions of Baron et al. 2014, Satopää et al. (2014), and Satopää et al. (2016). With this ensemble, we will see situations beyond their version of extremizing ($\hat{p} < \bar{p} < 1/2$ and $1/2 < \bar{p} < \hat{p}$). We will likely see cases such as $\bar{p} < \hat{p} < 1/2$ or $\bar{p} < \hat{p} < 1/2$. These are cases one would not see with the probit ensemble where $p_0 = 1/2$.

Assuming $n_0 = 0$ in our information structure, we know $p_0 = 1/2$. This is the case of a non-informative prior distribution for $\theta$. Therefore, the reference probit ensemble with a non-informative prior becomes

$$\hat{p} = \Phi \left( \sum_{i=1}^{k} \sqrt{\frac{v_{n_1}}{v_{kn_1}}} \Phi^{-1}(p_i) \right),$$

where $v_n = n(n + 1)$. Without more information for this ensemble, we might let $n_1 = 1$.

In Figure 4, we can see the difference between these two reference probit ensembles, where $p_1 = 0.25$ and $p_2 = 0.35$. Figure 4(a) depicts a reference probit ensemble with an informative prior ($n_0 = n_1 = 1$). And Figure 4(b) depicts a reference probit ensemble with a non-informative prior ($n_0 = 0$ and $n_1 = 1$). For each reference probit ensemble, we see an interval (in gray) where the ensemble does not extremize the average forecast (in the traditional sense). We also notice that the reference probit ensemble with an informative prior is less extreme than the one with a non-informative prior in the sense that the reference probit ensemble with an informative prior is shrunk more toward the average forecast.

Satopää et al. (2014) and Turner et al. (2014) propose the logit aggregator, a variant of the logarithmic opinion pool (French 1980, Dawid et al. 1995). This aggregator is a generalized additive model without an intercept: $p_{logit} = F_{Lo(0,1)}(\sum_{i=1}^{k} (a/k) F_{Lo(0,1)}^{-1}(p_i))$ where $F_{Lo(t,\psi)}(z) = 1/(1+
$e^{-\psi^2 z^2}$ is the cdf of the logistic distribution with location $l$ and scale $\psi$ and $F^{-1}_{Log(l,\psi)}(p) = l + \psi \log(p/(1-p))$ is the inverse of its cdf. Ranjan and Gneiting (2010) and Satopää et al. (2016) study versions of the probit ensemble in Example 2 with $p_0 = 1/2$. When $p_0 = 1/2$, the probit ensemble reduces to a generalized additive model with no intercept. These two ensembles are depicted in Figure 5(b). The probit ensemble $\hat{p}$ in Figure 5(b) has the same settings as the one in Figure 1(c).
(except for $\theta_0 = 0$ and $t_1 = 0$). This probit ensemble and the logit aggregator with $a = 1.25$ are virtually indistinguishable.

4. Generalized Probit Ensemble

Below we introduce the generalized probit ensemble. For this ensemble, we first assume the information states being combined (e.g., sufficient statistics) are jointly normally distributed, as in the probit ensemble of Example 2, but with a general correlation structure. Second, we assume a generalized linear model of these information states. This second assumption is informed by the Bayesian ensemble in Proposition 1 and, in particular, by its generalized linear model form in (4).

At the outset, the decision maker places beliefs directly on the models’ information states—states that may result from models using some information in common. He assumes the information states $x = (x_1, \ldots, x_k)'$ are jointly normally distributed with mean $\mu$ and covariance matrix $\Sigma$, denoted $x \sim N(\mu, \Sigma)$. The covariance matrix has elements $\sigma_{ij}$, and the correlation between $x_i$ and $x_j$ is defined as $\rho_{ij} = \sigma_{ij}/(\sqrt{\sigma_{ii}\sigma_{jj}})$.

The decision maker also assumes the conditional distribution of $y$ given the information states is given by the generalized linear model of the information states

$$P(y = 1|x_1, \ldots, x_k) = F_{z_0}(\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_k x_k),$$

where the inverse link function $F_{z_0}$ is the cdf of a continuous random variable $z_0$ with the entire real line as its support. In addition, he assumes the inverse link function $F_{z_0}$ is the standard cdf from a location-scale family, i.e., the random variable $z = l + \psi z_0$ has the cdf $F_{z_0}((z - l)/\psi)$ where $l$ is the location parameter and $\psi$ is the scale parameter. Finally, he assumes every model $i$ provides the probability $p_i = P(y = 1|x_i)$.

In our generalized linear model, we do not restrict ourselves here to the standard normal link function in a generalized linear model of the base models’ information states. We propose a member of any location-scale family for the inverse link function and later focus on the exponential-power distribution. The exponential-power inverse link function has the flexibility to be either the standard normal near one extreme or the uniform on the other extreme. Therefore, the generalized probit ensemble can take the form of either an inverse s-shape or a linear shape, as needed.

**Proposition 5 (Generalized Probit Ensemble).** Given the assumptions in this section, the optimal way to aggregate models’ forecasts is with the generalized probit ensemble

$$\hat{p} = P(y = 1|p_1, \ldots, p_k) = F_{z_0}\left(\beta_0 F_{z_0}^{-1}(\sqrt{\sigma_{x_0}}(p_0)) + \sum_{i=1}^{k} \beta_i F_{z_0}^{-1}(\sqrt{\sigma_{x_0}}(p_i))\right),$$

(9)
where \( p_0 = F_{z_0 + \sqrt{v_0}x_0}(m_0) \), \( p_i = F_{z_0 + \sqrt{v_i}x_0}(m_0 + \beta_i^{-1}\alpha_i(x_i - \mu_i)) \), \( m_0 = \alpha_0 + \sum_{i=1}^{k} \alpha_i \mu_i \),

\[
\beta_0 = 1 - \sum_{i=1}^{k} \beta_i, \quad \text{and} \quad \beta_i = \frac{\alpha_i \sqrt{\sigma_{ii}}}{\sum_{j=1}^{k} \alpha_j \sqrt{\sigma_{jj} \rho_{ij}}} \quad \text{for} \quad i = 1, \ldots, k.
\]

Also, the cdf \( F_{z_0 + \sqrt{v_i}x_0} \) is the cdf of the sum of the independent random variables \( z_0 \) and \( \sqrt{v_i}x_0 \) where \( z_0 \) is the standard random variable from a location-scale family, \( x_0 \) is a standard normal random variable,

\[
v_0 = \sum_{i=1}^{k} \sum_{j=1}^{k} \alpha_i \alpha_j \sigma_{ij}, \quad \text{and} \quad v_i = \sum_{j \neq i} \sum_{j' \neq i} \alpha_j \alpha_{j'} \sigma_{jj'} - \frac{(\sum_{j \neq i} \alpha_j \sigma_{ij})^2}{\sigma_{ii}} \quad \text{for} \quad i = 1, \ldots, k.
\]

Immediately we see that the generalized probit ensemble in (9) is a generalized additive model of the base models’ probabilities. The main benefit of this ensemble is that we do not need to work with a conjugate pair’s predictive distribution, which can be difficult to do in all but few cases.

Another immediate consequence of the result is that each model is calibrated. A model is calibrated if \( P(y = 1 | p_i) = p_i \) (French 1986, Murphy and Winkler 1987). As was the case in the previous section, this calibration means that the decision maker, if he received a forecast from only one model, would use the same probability the model provided. Note that if the model is not calibrated, then the decision maker could re-calibrate the model’s reported probability using the re-calibration function \( R_i(p_i) \) so that \( P(y = 1 | x_i) = R(p_i) \). All results related to the generalized probit ensemble hold with \( R(p_i) \) in place of \( p_i \), if a model is not calibrated.

To interpret the coefficients in the ensemble, it is helpful to consider the case of exchangeable models. If the models are exchangeable, then the weights are given by

\[
\beta_0 = 1 - \frac{k}{(k-1)\rho + 1} < 0 \quad \text{and} \quad \beta_i = \frac{1}{(k-1)\rho + 1} > 0 \quad \text{for} \quad i = 1, \ldots, k,
\]

Note that for the exchangeable distribution of information states to be a proper distribution, we must have \( \rho > -1/(k-1) \), which ensures that the matrix \( \Sigma \) is positive definite.

### 4.1. Exponential-Power Inverse Link Function

Below we propose a family of inverse link functions to use when fitting the generalized probit ensemble to data. The family is based on the exponential-power distribution, also known as a generalized error or generalized Gaussian distribution (Subbotin 1923, Box and Tiao 1973, Mineo and Ruggieri 2005, Zhang et al. 2012). This family contains the normal and Laplace distributions as special cases. Before we give the family’s general form, we look at the normal distribution.
**Example 4 (Normal Distribution).** Let the inverse link function $F_{z_0}$ in Proposition 4 be standard normal cdf. In this case, $F_{z_0 + \sqrt{v_0}x_0}(u) = \Phi(u/\sqrt{1+v_i})$ so that $F_{z_0 + \sqrt{v_0}x_0}^{-1}(p_i) = \sqrt{1+v_i}\Phi^{-1}(p_i)$. The probit ensemble, according to the assumptions in this section, is given by

$$\hat{p} = \Phi\left(\beta_0\sqrt{1+v_0}\Phi^{-1}(p_0) + \sum_{i=1}^{k} \beta_i\sqrt{1+v_i}\Phi^{-1}(p_i)\right). \quad \square$$

To estimate this probit ensemble’s coefficients $\beta_i = \beta_i\sqrt{1+v_i}$ from data, one can fit a generalized linear model of $y$ on $(\Phi^{-1}(p_1), \ldots, \Phi^{-1}(p_k))$ using a standard normal inverse link function.

**Example 5 (Exponential-Power Distribution).** Let the inverse link function be the cdf $F_{z_0}$ with density

$$f_{z_0}(z) = \frac{1}{2\eta^{1/\eta}\Gamma(1+1/\eta)}\exp\left(-\frac{1}{\eta} \left| \frac{z - l}{\psi} \right|^{\eta}\right).$$

This is the density of an exponential-power distribution. We denote this distribution by $z_0 \sim EP(l, \psi, \eta)$ where $l$ is the mean, $\eta^{2/\eta}\psi^2\Gamma(3/\eta)/\Gamma(1/\eta)$ is the variance, and $\eta > 0$ is the power parameter. For a fixed power parameter, this distribution is a member of a location-scale family. \quad \square

For the ensemble based on the exponential-power distribution, the inside function is $F_{z_0 + \sqrt{v_0}x_0}$, which is not tractable (Soury and Alouini 2015). Nonetheless, we can approximate this inside function closely by matching the first two moments. We choose $v_i'$ in $\sqrt{1+v_i'}z_{EP(0,1,\eta)}$ so that the variance of this random variable equals the variance of $z_{EP(0,1,\eta)} + \sqrt{v_i'x_0}$. Their means are both zero by construction. Consequently, the approximate generalized probit ensemble with an exponential-power inverse link function, denoted by $\hat{p}$, is given by

$$\hat{p} = F_{EP(0,1,\eta)}\left(\beta_0\sqrt{1+v_0'}F_{EP(0,1,\eta)}^{-1}(p_0) + \sum_{i=1}^{k} \beta_i\sqrt{1+v_i'}F_{EP(0,1,\eta)}^{-1}(p_i)\right). \quad (10)$$

This ensemble is useful in applications because one can estimate the coefficients $\beta_i' = \beta_i\sqrt{1+v_i'}$ by fitting a generalized linear model of $y$ on $(F_{EP(0,1,\eta)}^{-1}(p_1), \ldots, F_{EP(0,1,\eta)}^{-1}(p_k))$ using the exponential-power inverse link function $F_{EP(0,1,\eta)}$. For $\eta = 2$, the exponential-power distribution becomes the normal distribution, and the resulting ensemble is a probit ensemble. For $\eta = 1$, the exponential-power distribution becomes the Laplace distribution. For $\eta$ strictly less (greater) than 2, the distribution has fat (thin) tails. As $\eta \to \infty$, the distribution of $z_{EP(l,\psi,\eta)}$ goes to a uniform distribution on $(l - \psi, l + \psi)$ and the resulting ensemble approaches a linear ensemble, like the one in the beta/Bernoulli ensemble of Dawid at al. (1995).
5. Empirical Study

In this section, we present two empirical studies in which we fit the generalized probit ensemble and compare its out-of-sample forecasting performance to that of several alternative models. The challenge in the first study is to predict defaults on loans acquired by Fannie Mae in 2007, just before the 2008 Great Recession. These data are available at https://loanperformancedata.fanniemae.com/lppub/index.html. In the Fannie Mae data, there are 1,056,724 records on acquired loans along with 20 independent variables, such as the borrower’s credit score, the home’s loan-to-value, and borrower’s debt-to-income ratio. This year of acquisitions had the highest rate of defaults at 8.5% in the period 2000-2015. The second study’s challenge is to predict bad used-car buys by Carvana (a large used-car retailer) during the period 2009-2010. These data are available at https://www.kaggle.com/c/DontGetKicked/data. In the Carvana data, there are 72,983 records on used-car purchases and 34 independent variables, such as the vehicle’s age, the vehicle’s odometer, the vehicle’s model, the buyer’s id number, and the auction’s location. This prediction challenge was part of a Kaggle competition called Don’t Get Kicked!, which finished in January of 2012. The base rate of defective cars in the training set from Kaggle is 12.3%.

5.1. Training and Stacking the Models

In the machine learning literature, forecast aggregation is known as stacking (Wolpert 1992, Breiman 1996, Smyth and Wolpert 1999, Dzeroski and Ženko 2004). The idea is that the predictions from several base models become features in a second-stage stacker model that combines the predictions to form an ensemble forecast. Breiman (1996) and Smyth and Wolpert (1999) both consider stacker models that are convex combinations (or linear opinion pools) of base models’ probabilities. They choose optimal weights in a linear opinion pool that maximize the likelihood on a training set. Linear stacking works especially well on continuous quantities of interest, but not as well on binary events of interest. Our experience with leading machine learning algorithms in the two prediction challenges considered here is that the best algorithm (xgboost) is often difficult to beat.

For each study, we trained three base models using the covariates available in the datasets. Each base model is a leading statistical or machine learning algorithm or part of a competition-winning ensemble. The first model is a regularized logistic regression (RLR), the lasso proposed by Tibshirani (1996). The second model is the random forest (RF) introduced by Breiman (2001). The third model is the extreme gradient boosted trees model called xgboost (XGB) (Chen and Guestrin 2016). This model is an extension of the gradient boosted trees model introduced by Friedman (2001). We could add more base models to our ensemble and compare our ensemble to
other ensembles beyond the linear opinion pool, but our goal here is to demonstrate the plausibility of our approach, in a parsimonious manner. The xgboost model, by itself, represents a difficult benchmark to beat. It was a part of 17 winning solutions published on Kaggle in 2015, and it was used by every team in the top 10 at KDD Cup 2015 (Chen and Guestrin 2016).

To ensure that an ensemble was trained on out-of-sample probabilities from the base models, we employed a two-step process for building a stacker model. First, we randomly split the data into 10 equal folds and used these folds for both steps. In Step 1, the base models were fit to the first nine folds of data using the available covariates (e.g., credit score and loan-to-value) and the outcomes of the binary event of interest (e.g., loan defaults). Then the base models were used to predict the binary event of interest in the tenth fold. Next the base models were trained on Folds 1-8 and 10 and used to predict the binary events of interest in the Fold 9. This process continued until each fold was held out once and out-of-sample predictions were made for it. In Step 2, the ensemble, or stacker model, was trained on the out-of-sample predictions made by the base models in Step 1. In particular, the stacker model was trained on Folds 1-9 and tested (or evaluated) on Fold 10. This training and testing was done 10 times, with each fold serving as the hold-out sample once. For more details on how each of these models was trained to avoid overfitting, see the Online Supplement. All data and code are available upon request from the authors.

5.2. Scoring the Models’ Forecasts

To evaluate out-of-sample forecasts, we use the asymmetric log score (ALS), a positively-oriented (higher score is better) score. The asymmetric log score of a probability forecast \( p \) for a binary event \( y \) is given by \( ALS(y, p) = (LL(c, y) - LL(p, y))/LL(c, I_{p>c}) \) where \( LL(p, y) = -(y \log(p) + (1 - y) \log(1 - p)) \) for \( 0 < p < 1 \) and \( I_{p>c} \) equals 1 if \( p > c \) and equals 0 otherwise (Winkler 1994). This score is “adjusted for the difficulty of the forecast task ... with the value of \( c \in (0, 1) \) adapted to reflect a baseline probability” (Gneiting and Raftery 2007, p. 365). In the results we report below, \( c \) is taken to be the base rate of occurrence of \( y \) (denoted \( \bar{y} \)) in the training set.

5.3. Results

Table 1 reports the average scores of out-of-sample predictions from the three base models (\( p_{RLR} \), \( p_{RF} \), and \( p_{XGB} \)), three existing aggregation models (the linear opinion pool with equal weights \( \bar{p} \), the beta-transformation of the linear opinion pool with optimal weights \( p_{BLOP} \) from Ranjan and Gneiting (2010), and the logit aggregator \( p_{logit} \)), and the best approximate generalized probit ensemble with an exponential-power inverse link function \( \tilde{p} \). The best \( \tilde{p} \) has the power parameter \( \eta^* \) that minimizes the average \( ALS \) out-of-sample. The estimates of \( \eta^* \) are 40 and 10 in the two
In both studies, the best approximate generalized probit ensemble outperforms all other models on the average of the 10 cross-validation folds. Similar results were obtained when using alternative scoring rules such as the log loss or area under the curve.

After performing 10-fold cross validation for each study, we estimate the coefficients in the generalized linear model for the best approximate generalized probit ensemble one last time on the entire dataset. See Table 2 for the results of these estimates. Not surprisingly, the ensemble puts the most weight on the best base model ($p_{XGB}$). We also report in Table 2 the base rate ($\bar{y}$) at which the binary events occur in each dataset. In addition, we provide the percentage of times in out-of-sample forecasting that the best generalized probit ensemble extremizes the linear opinion pool. Each time we compare the linear opinion pool to the base rate to identify which extreme is the linear opinion pool’s more relevant extreme.

Figure 6 depicts the best approximate generalized probit ensemble as a function of the best base model (xgboost), with the other two models set to twice the base rate. In each plot, we highlight the anti-extremizing region in gray. In these plots, the only time we see bracketing of the base model relative to the prior, happens in this grey region. Within this region, when $p_{XGB} < p_0$, the base models bracket the prior (since in these figures we set $p_{RLR} = p_{RF} = 2p_0$). Note that bracketing is not, however, a sufficient condition for anti-extremizing. In addition, because the estimated power parameters are high in these studies, we can see that the ensembles are nearly linear over much of the domain of $p_{XGB}$.

### Understanding Extremizing and Anti-Extremizing

In Table 2, extremizing (anti-extremizing) occurs in approximately 80% (20%) of the cases in both datasets. To add to the theoretical developments in Section 3, we next examine more closely those instances in which extremizing occurs and those instances in which anti-extremizing occurs.
### Table 2 Final Estimation of Best Approximate Generalized Probit Ensembles.

|                     | Fannie Mae | Carvana |
|---------------------|------------|---------|
| Constant            | 0.0496***  | 0.0653*** |
|                     | (0.0037)   | (0.0098) |
| Coefficient $\beta_{RLR}'$ | 0.3710*** | -0.0281 |
|                     | (0.0087)   | (0.0234) |
| Coefficient $\beta_{RF}'$ | 0.0810*** | 0.3864*** |
|                     | (0.0062)   | (0.0232) |
| Coefficient $\beta_{XGB}'$ | 0.6070*** | 0.7176*** |
|                     | (0.0093)   | (0.0241) |

| Power parameter $\eta^*$ | 40 | 9 |
| $\hat{p}$ extremizes $\bar{p}$ | 79.6% | 81.3% |
| Base rate $\bar{y}$ | 0.0849 | 0.1230 |
| No. of observations | 1,056,724 | 72,983 |

Standard errors in the coefficients are reported in parentheses. The superscript *** indicate significance at the 99.9% level.

---

**Figure 6 Best Approximate Generalized Probit Ensemble as a Function of the Best Base Model’s Forecast.**

This provides additional insight into when extremizing and anti-extremizing are appropriate in aggregating probabilities.

In Table 3(a), we report the percentage of cases in the Fannie Mae and Carvana datasets for each row-column combination. The rows represent four possible cases involving the three base models’ probabilities: (i) all three higher than the base rate, (ii) two above the base rate and one below, (iii) one above the base rate and two below, and (iv) all below the base rate. The columns represent our approximate generalized probit ensemble extremizing toward zero, anti-extremizing,
Table 3(a) Fannie Mae (1,056,724 cases) Carvana (72,983 cases)

| Joint %          | Extr to 0 | Anti-Extr | Extr to 1 | Row % | Extr to 0 | Anti-Extr | Extr to 1 | Row % |
|------------------|-----------|-----------|-----------|-------|-----------|-----------|-----------|-------|
| All Above        | 0%        | 9%        | 17%       | 26%   | 0%        | 5%        | 14%       | 19%   |
| 2 Above, 1 Below | 1%        | 5%        | 2%        | 8%    | 1%        | 4%        | 5%        | 10%   |
| 1 Above, 2 Below | 8%        | 4%        | 0%        | 12%   | 8%        | 6%        | 2%        | 16%   |
| All Below        | 51%       | 3%        | 0%        | 53%   | 51%       | 4%        | 0%        | 56%   |
| Column %         | 59%       | 20%       | 20%       | 100%  | 60%       | 19%       | 21%       | 100%  |

Table 3(b) Fannie Mae Carvana

| Joint %          | Extr to 0 | Anti-Extr | Extr to 1 | Row % | Extr to 0 | Anti-Extr | Extr to 1 | Row % |
|------------------|-----------|-----------|-----------|-------|-----------|-----------|-----------|-------|
| Bracketing       | 9%        | 9%        | 3%        | 20%   | 9%        | 10%       | 7%        | 25%   |
| No Bracketing    | 51%       | 12%       | 17%       | 80%   | 51%       | 9%        | 14%       | 75%   |
| Column %         | 59%       | 20%       | 20%       | 100%  | 60%       | 19%       | 21%       | 100%  |

Table 3 Contingency Tables (in Percentages) for the Type of Extremizing/Anti-Extremizing by the Approximate Generalized Probit Ensemble (Columns) and Agreement Among the Base Models’ Probabilities (Rows) in the Fannie Mae and Carvana Data.

and extremizing toward one. Table 3(b) is like Table 3(a) with only two rows, bracketing [Cases (ii) and (iii)] and no bracketing [Cases (i) and (iv)].

As noted above, anti-extremizing of the probit occurs in about 20% of the cases in our data. First, consider Fannie Mae. When there is bracketing, the probit anti-extremizes in 43% of the cases, while it anti-extremizes at a much lower rate of 15% when there is no bracketing. For Carvana, these rates are slightly lower at 36% and 12%, respectively, but their ratio is similar. This suggests that bracketing is a good indicator of when anti-extremizing might be worthwhile to consider.

Figure 7 gives three-dimensional scatter plots of the individual data points for Fannie Mae (1,056,724 data points) and Carvana (72,983 data points), showing cases of extremizing toward zero, anti-extremizing, and extremizing toward one separately. With this many data points, we cannot identify individual points but can see the general picture. For both Fannie Mae and Carvana, extremizing toward zero occurs about 60% of the time; when it does occur, the three models are all below the base rate 85% of the time and cannot be above the base rate. We see these tendencies visually in Figure 7, where the points cluster near (0,0,0) in 7(a) and 7(d). This clustering is reduced in 7(b) and more so in 7(e), and cannot occur in 7 (c) and 7(f), being cut off at the base rates. Clustering near (1,1,1), in contrast, is strongest in 7(c) and more so in 7(f) but still not as strong as the clustering near (0,0,0) in 7(a) and 7(d). It is much weaker in 7(b) and especially in 7(e) for Carvana, and nonexistent in 7(a) and 7(d).

Complementing Figure 7, stacked point plots, with each horizontal set of points representing one instance in the dataset (a specific car), with three points for the three base models’ probabilities \( p_{RLR}, p_{RF}, \) and \( p_{XGB} \), are given in Figure 8 for Carvana. Stacked point plots for Fannie Mae exhibit similar patterns and are omitted here for brevity. As in Figure 7, we cannot pick out individual sets of points, but can get an indication of how spread out the three probabilities tend
to be. Figure 8 illustrates how the variability among the three probabilities tends to be much lower with extremizing to zero than with extremizing to one. The low base rates (0.085 and 0.123) and the resulting larger distance from the base rates to one than to zero may partially explain the discrepancies in Figure 7 and 8. In turn, these discrepancies reinforce the importance of defining extremizing and anti-extremizing relative to the base rate, not to one-half.
6. Discussion and Conclusions

In this paper, we introduce a large class of Bayesian ensembles. Because the ensembles in this class are based on Bayesian reasoning, they can help us understand why and when extremizing is appropriate. For example, each of the base models includes the prior-predictive probability, or base rate. Thus, the base rate is counted multiple times, and this redundancy shifts the average probability closer to the base rate than it should be, yielding forecasts that are often called underconfident. Extremizing represents an attempt to correct for this overweighting of the base rate by shifting the ensemble farther away from the base rate. To the extent that there is considerable overlap in the prior information available for the base models, as we would expect to be the case in many situations, extremizing can be beneficial.

In the literature on extremizing when combining probabilities, extremizing is generally viewed as shifting the average probability farther from one-half. An important contribution of our paper is to emphasize that it is more suitable to define extremizing as shifting the average probability farther from the base rate. If the base rate does happen to be one-half, Proposition 3 shows that it is indeed optimal to extremize for many commonly used aggregators. However, we posit that a base rate of one-half is not appropriate in most real-world binary situations, as illustrated in the two datasets analyzed in the empirical study in Section 5.

An even more important contribution of our paper is to introduce the notion of anti-extremizing, which involves shifting the average probability toward the base rate instead of farther away from the base rate. To highlight that a decision maker may sometimes want to anti-extremize, Proposition 2 demonstrates that it can be optimal to anti-extremize in our Bayesian ensembles. Moreover, in our empirical study, we show that the proposed Bayesian ensemble, approximated by a generalized probit model, performs well as an ensemble when tested out of sample. This ensemble anti-extremizes in about 20% of the cases in each of the datasets, as reported in Section 5.

A close examination of the cases in the empirical analysis reveals that our probit ensemble anti-extremizes much more often in cases where bracketing with respect to the base rate occurs among the probabilities being combined than in cases where such bracketing does not occur. This suggests that bracketing of the models’ probabilities is a promising indicator of when we should give careful consideration to anti-extremizing.

To summarize, we developed our ensemble within a Bayesian framework and derived theoretical results to show that anti-extremizing can be optimal for this ensemble. We argued that extremizing should be represented by shifting the average probability away from the base rate instead of the commonly used notion of extremizing by shifting it farther from one-half. We introduced the notion
of anti-extremizing, which involves moving toward the base rate, and derived results demonstrating that anti-extremizing can be optimal.

Finally, using two large datasets, we showed that the proposed Bayesian ensemble performed well when tested out of sample. Our ensemble combined probabilities from three powerful models: the regularized logistic regression, the random forest, and gradient boosted trees. Generalizing these results more broadly remains an open question for future study. We can state, however, that when some of the individual probabilities being combined are lower than the base rate and some higher than the base rate, an ensemble that anti-extremizes might be useful.

References

Ahuja, J C, S W Nash. 1967. The generalized Gompertz-Verhulst family of distributions. *Sankhyā: The Indian Journal of Statistics, Series A* 29(2) 141–156.

Ariely, D, W T Au, R H Bender, D V Budescu, C B Dietz, H Gu, T S Wallsten, G Zauberman. 2000. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology Applied* 6(2) 130–146.

Baron, J, B A Mellers, P E Tetlock, E Stone, L H Ungar. 2014. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis* 11(2) 133–145.

Bernardo, J M, A F M Smith. 2000. *Bayesian Theory*. John Wiley & Sons, Ltd, Chichester, England.

Box, G E P, G C Tiao. 1973. *Bayesian Inference in Statistical Inference*. Adison-Wesley, Reading, MA.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2) 123–140.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1) 5–32.

Chen, T, C Guestrin. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

Chipman, Hugh A., Edward I. George, Robert E. McCulloch. 2007. Bayesian ensemble learning. P B Schölkopf, J C Platt, T Hoffman, eds., *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 265–272.

Clemen, R T. 1987. Combining overlapping information. *Management Science* 33(3) 373–380.

Clemen, R T, R L Winkler. 1986. Combining economic forecasts. *Journal of Business & Economic Statistics* 4(1) 39–46.

Dawid, A P, M H DeGroot, J Morter. 1995. Coherent combination of experts’ opinions. *Test* 4(2) 263–313.

DeGroot, M H. 1974. Reaching a consensus. *Journal of the American Statistical Association* 69(345) 118–121.

DeGroot, M H, J Morter. 1991. Optimal linear opinion pools. *Management Science* 37(5) 546–558.
Džeroski, S, B Ženko. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine Learning* 54(3) 255–273.

Erev, I, T S Wallsten, D V Budescu. 1994. Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review* 101(3) 519–527.

French, S. 1980. Updating of belief in the light of someone else’s opinion. *Journal of the Royal Statistical Society, Series A* 143(1) 43–48.

French, S. 1986. Calibration and the expert problem. *Management Science* 32(3) 315–321.

Friedman, J H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5) 1189–1232.

Gaba, A, I Tsetlin, R L Winkler. 2017. Combining interval forecasts. *Decision Analysis* 14(1) 1–20.

Genest, C, J V Zidek. 1986. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science* 1(1) 114–135.

Gneiting, T, A E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477) 359–378.

Grushka-Cockayne, Y, K C Lichtendahl Jr, V R R Jose, R L Winkler. 2017. Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research* 65(3) 712–728.

Hastie, T, R Tibshirani. 1986. Generalized additive models. *Statistical Science* 1(3) 297–318.

Hora, S C. 2004. Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science* 50(5) 597–604.

Karmarkar, U S. 1978. Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance* 21(1) 61–72.

Larrick, R P, J B Soll. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science* 52(1) 111–127.

Lindley, D V, A Tversky, R V Brown. 1979. On the reconciliation of probability assessments. *Journal of the Royal Statistical Society, Series A* 142(2) 146–180.

Mellers, B, L Ungar, J Baron, J Ramos, B Gurcay, K Fincher, S E Scott, D Moore, P Atanasov, S A Swift, T Murray, E Stone, P E Tetlock. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science* 25(5) 1106–1115.

Mineo, A M, M Ruggieri. 2005. A software tool for the exponential power distribution: The normalp package. *Journal of Statistical Software* 12(4) 1–24.

Morris, P A. 1974. Decision analysis expert use. *Management Science* 20(9) 1233–1241.

Murphy, A H, R L Winkler. 1987. A general framework for forecast verification. *Monthly Weather Review* 115(7) 1330–1338.
Nelder, J A, R W M Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**(3) 370–385.

Raiffa, H, R Schlaifer. 1961. *Applied Statistical Decision Theory*. Harvard Business School, Boston, MA.

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

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

Satopää, V A, R Pemantle, L H Ungar. 2016. Modeling probability forecasts via information diversity. *Journal of the American Statistical Association* **111**(516) 1623–1633.

Shlomi, Y, T S Wallsten. 2010. Subjective recalibration of advisors’ probability estimates. *Psychonomic Bulletin & Review* **17**(4) 492–498.

Smyth, P, D Wolpert. 1999. Linearly combining density estimators via stacking. *Machine Learning* **36**(1-2) 59–83.

Soury, H, M S Alouini. 2015. New results on the sum of two generalized Gaussian random variables. *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1017–1021.

Stone, M. 1961. The opinion pool. *The Annals of Mathematical Statistics* **32**(4) 1339–1342.

Subbotin, M T. 1923. On the law of frequency of error. *Matematicheskii Sbornik* **31**(2) 296–301.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**(1) 267–288.

Turner, B M, M Steyvers, E C Merkle, D V Budescu, T S Wallsten. 2014. Forecast aggregation via recalibration. *Machine Learning* **95**(3) 261–289.

Wallsten, T S, D V Budescu, I Erev, A Diederich. 1997. Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making* **10**(3) 243–268.

Winkler, R L. 1968. The consensus of subjective probability distributions. *Management Science* **15**(2) B61–B75.

Winkler, R L. 1994. Evaluating probabilities: Asymmetric scoring rules. *Management Science* **40**(11) 1395–1405.

Winkler, R L, V R R Jose. 2008. Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17** 251–255.

Winkler, R L, V R R Jose, K C Lichtendahl Jr, Y Grushka-Cockayne. 2018. Contributed discussion to Yao et al. 2018. *Bayesian Analysis* **13**(3) 968–970.

Winkler, R L, R M Poses. 1993. Evaluating and combining physicians’ probabilities of survival in an intensive care unit. *Management Science* **39**(12) 1526–1543.

Wolpert, D H. 1992. Stacked generalization. *Neural Networks* **5**(2) 241–259.
Yao, Y, A Vehtari, D Simpson, A Gelman. 2018. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis* **13**(3) 917–1003.

Zhang, Z, S Wang, D Liu, M I Jordan. 2012. EP-GIG priors and applications in Bayesian sparse learning. *Journal of Machine Learning Research* **13** 2031–2061.