Econometric Theory, 2012, Page 1 of 28.
doi:10.1017/S0266466612000461

ON THE RECOVERABILITY OF FORECASTERS' PREFERENCES

ROBERT P. LIELI
Central European University
and
Magyar Nemzeti Bank

MAXWELL B. STINCHCOMBE
University of Texas–Austin

We study the problem of identifying a forecaster's loss function from observations on forecasts, realizations, and the forecaster's information set. Essentially different loss functions can lead to the same forecasts in all situations, though within the class of all continuous loss functions, this is strongly nongeneric. With the small set of exceptional cases ruled out, generic nonparametric preference recovery is theoretically possible, but identification depends critically on the amount of variation in the conditional distributions of the process being forecast. There exist processes with sufficient variability to guarantee identification, and much of this variation is also necessary for a process to have universal identifying power. We also briefly address the case in which the econometrician does not fully observe the conditional distributions used by the forecaster, and in this context we provide a practically useful *set* identification result for loss functions used in forecasting binary variables.

1. INTRODUCTION

This paper examines how much can be learned about an expected loss minimizing/expected utility maximizing forecaster's loss function from a sequence of forecasts, the corresponding sequence of realizations, and a sequence of covariates used in producing the forecasts. If the forecaster's loss function is known to belong in a given family, can such information uniquely identify the one used by the forecaster? How large can this family be for identification to be possible? Giving a complete answer to this question involves answering the question of how much of the best response correspondence must be observed to completely recover arbitrary continuous preferences.

We care about recovering loss functions for two related reasons. First, knowing the trade-off between underpredicting versus overpredicting is informative about

We thank David Hendry, Ivana Komunjer, Michael McCracken, Ulrich Müller, and two anonymous referees for useful comments. All errors are our responsibility. Address correspondence to Robert P. Lieli, Department of Economics, Central European University, Nádor u. 9, 1051 Budapest, Hungary; e-mail: lielir@ceu.hu.

behavior in its own right. Second, when the loss function is the reduced form of a deeper, structural model, one might identify the model's parameters by first identifying the loss function. If this is impossible, then the underlying structural model will also remain unidentified.

The existing econometric literature on the problem of recovering forecaster preferences is rather slim. Among the few papers that address this question explicitly are Elliott, Komunjer, and Timmermann (2003) and Elliott, Komunjer, and Timmermann (2005), and, partly, Patton and Timmermann (2005) and Patton and Timmermann (2007). More recently, Elliott, Komunjer, and Timmermann (2008) and Capistran and Timmermann (2009) have provided evidence of asymmetric loss in professional and institutional forecasts of real output growth and inflation. There is a much larger related literature concerned with testing the rationality of forecasts that dates back to at least Mincer and Zarnowitz (1969). Empirical work in this area typically relies on the assumption that the forecaster's objective is to minimize mean squared error loss. Indeed, the square loss function is technically convenient and has very sharp observable implications concerning the properties of optimal forecasts, including unbiasedness, uncorrelatedness of one-step-ahead forecast errors, increasing forecast error variance as the forecast horizon expands, etc.[1]

However, as argued by Granger (1969), economic forecasts are often produced in an environment where the square loss function or, more generally, any symmetric loss function, does not adequately capture the costs resulting from overprediction vs. underprediction. Under general loss functions, all of the optimality properties listed earlier can be lost, and, as pointed out by Elliott et al. (2003, 2005), purported tests of forecast rationality based on the implications of minimizing mean squared error are more appropriately viewed as joint tests of forecaster rationality and the mean square loss function.

Given that the notion of forecast rationality is inextricably linked to the objective that the forecaster is presumably trying to achieve, there are two ways to study individual forecasters' behavior. If interest continues to center on testing for optimizing (rational) behavior, then one needs to explore further the properties of optimal forecasts under general classes of loss functions that allow for asymmetries and functional forms other than square loss. This approach is outlined by Patton and Timmermann (2005, 2007). Alternatively, one can focus on the inverse problem: maintain the assumption of optimizing behavior and identify and estimate a loss function, or a class of loss functions, consistent with the properties of the observed forecasts. This is the viewpoint of Elliott et al. (2005) and of this paper.

Our approach to the preference recovery problem is substantially more general than that of Elliott et al. (2005). In particular, we study identification for the class of continuous loss functions that strictly prefer a forecast of $\hat{y}$ when it is known that $\hat{y}$ will be the next realization instead of any particular parametric specification. The cost of this generality is that our identification results are more abstract and do not directly translate into a strategy for estimation and inference.

The benefit is a comprehensive theory of the informational requirements for generic nonparametric identification.

More concretely, let $\mathscr{L}$ be the set of loss functions that depend continuously on both the forecast and the realization and strictly prefer on-target forecasts to any error. Expected loss minimization associates with each $\ell \in \mathscr{L}$ a forecasting rule, i.e., a map from forecast densities into point forecasts (e.g., for square loss, the optimal point forecast is the mean of the forecast density, whereas for absolute loss it is the median). We show that some forecasting rules are not unique to a specific element of $\mathscr{L}$—there are essentially different loss functions that produce the same point forecasts under all circumstances. As $\mathscr{L}$ contains indistinguishable elements, it is unidentified as a model of the forecaster preferences. However, the set of "problematic" loss functions in $\mathscr{L}$ is nongeneric, i.e., very small in a strong mathematical sense. Theorem 1 shows that outside of this nongeneric subset of $\mathscr{L}$, different loss functions have different forecasting rules unless they are generalized affine transformations of each other. We explicitly characterize the nongeneric set, and in Proposition 3.1, we give a sufficient condition for belonging to the generic set that is easy to verify.

Even with the exceptional cases ruled out, nonparametric identification still depends on sufficient variability in the sequence of forecast densities used by the forecaster. For example, the model {square loss, absolute loss} is unidentified if all forecast densities are symmetric despite the fact that the forecasting rules produce different results for skewed distributions. Theorem 1 in Section 3 also shows that for sequences of distributions with sufficient variability, one can guarantee identification of any element of $\mathscr{L}$ outside of the problematic ones, whereas Proposition 3.2 shows that much of the "sufficient variation" is also necessary to achieve such universal identification. Of particular note is the required variation in the support of the forecast densities seen in Examples 3.4 and 3.5.

The foregoing identification results rely on the assumption that the sequence of forecast densities used by the forecaster is also observed by the econometrician. This assumption is restrictive—if the forecaster conditions the forecast density on variables not observed by the econometrician, the forecast density is unidentified (to the econometrician). In this case nonparametric identification of the loss function is a more challenging problem. We provide a practically useful set identification result in the special case when the variable to be forecast is binary and sketch a more general identification strategy applicable to smooth (differentiable) utility functions.

The rest of the paper is organized as follows. Section 2 presents the forecasting environment along with some elementary results. Section 3 introduces the notion of nonparametric identification and states the main result of the paper. Section 4 addresses the problem of unobserved covariates and treats the binary case in particular. Section 5 concludes. Technical material and proofs are collected in the Appendixes.

2. THE FORECASTING ENVIRONMENT

We first describe the environment in which the econometrician observes the forecaster and then describe the set of loss functions under consideration.

2.1. The Data Generating Process and Forecaster Behavior

The variable to be forecast at time t is a random scalar Y_{t+1}, defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values in a compact set $D \subset \mathbb{R}$. The forecaster possesses a jointly continuous loss function on $D \times D$, $(\hat{y}, y) \mapsto \ell(\hat{y}, y)$. The first argument, $\hat{y}$, denotes the value of the forecast, and the second, y, the actual realization of Y_{t+1}. We will denote by $\Delta(D)$ the set of (Borel) probability measures defined on D, equipped with the Prokhorov metric.[2]

The information held by the forecaster at time t is given by a filtration $\mathcal{F}_t \subset \mathcal{F}$. Let p_t be an $\mathcal{F}_t$-measurable random element of $\Delta(D)$, representing either the conditional distribution of Y_{t+1} given $\mathcal{F}_t$, or possibly some estimate of it. We assume that f_t, the forecast of Y_{t+1} reported at time t, minimizes expected loss, i.e., belongs to the set

$$Br\,(p_t \mid F_t, \ell) := \arg\min_{\hat{y} \in F_t} \int \ell\left(\hat{y}, y\right) p_t\,(dy)\,, \tag{1}$$

where the notation Br stands for "best response" and $F_t \subset D$ is the set of allowable forecasts at time t. Assumption 1 restricts this set.

Assumption 1. $F_t = \text{supp}(p_t)$ for all t.[3]

One could allow for off-support forecasts, but we will not do so for a number of reasons. First, it would lead to a framing problem in that the support of p_t, if smaller than D, could be embedded in many larger sets F_t. Second, as we will show in Section 3.3, forecasters with two different loss functions can be indistinguishable without variation in F_t. If off-support forecasts were permissible, it would be hard to justify any variation in the set of allowable forecasts. (For example, one might simply set $F_t = D$ for all t.) Under Assumption 1, F_t varies "naturally" to the extent that supp(p_t) varies. Third, even if one allows F_t to be strictly larger than supp(p_t) and there is exogenous variation in F_t, Proposition 2.3 shows that one cannot tease out more information about ℓ from the observed forecasts than under Assumption 1.

We abbreviate $Br\,(p_t \mid \text{supp}(p_t), \ell)$ as $Br\,(p_t \mid \ell)$. Assumption 1 notwithstanding, some definitions, examples, and technical arguments will necessitate considering $F_t \supsetneq \text{supp}(p_t)$. For these cases, we retain the more elaborate notation defined in (1). Under Assumption 1, and if the process $\{Y_t\}$ is stationary, it is without loss of generality to set D as the support of the marginal distribution of Y_t. If $\{Y_t\}$ is not stationary, then D is the closure of $\cup_t \text{supp}(Y_t)$.

The following assumption specifies the econometrician's information set.

Assumption 2. The econometrician observes

(a) the sequence of distributions $\{p_t\}_{t=1}^{\infty}$ and
(b) a sequence of point forecasts $\{f_t\}_{t=1}^{\infty}$ satisfying equation (1).

Assumption 2(a) is motivated by the structure of an expected loss minimizer's problem. Such forecasters (act as if they) first form p_t, the conditional distribution of Y_{t+1} given $\mathcal{F}_t$, and second use p_t to pick a forecast f_t in a way that depends on the loss function ℓ. The mapping $\mathcal{F}_t \mapsto p_t$ contains no information about ℓ and is, in principle, also estimable by the econometrician if $\mathcal{F}_t$ is fully observed. It is only from the forecast mapping $p \mapsto Br\,(p \mid \ell)$ that one can hope to learn about ℓ. Our main focus is whether or not this forecast mapping contains enough information for ℓ to be identified. We therefore start from the assumption that the econometrician knows the p_t at which this mapping is being evaluated. Our first example adds some structure to the process being forecast to illustrate and further motivate Assumption 2.

Example 2.1
Suppose that the forecaster's information set can be represented as $\mathcal{F}_t = \sigma(X_t, X_{t-1}, \ldots)$, where $X_t \in \mathbb{R}^{\ell}$ is a finite-dimensional vector, also defined on $(\Omega, \mathcal{F}, P)$, such that

(a) $\{(Y_{t+1}, X_t)\}$ forms a stationary, ergodic process and
(b) the conditional distribution of Y_{t+1} given $\mathcal{F}_t$ depends only on X_t.

The vector X_t may of course contain current and lagged values of Y. Let $p_x(y)$ denote the distribution of Y_{t+1} conditional on $X_t = x$ induced by P. Under the assumptions stated previously, the forecaster has perfect knowledge of the function $x \mapsto p_x$, and, therefore, one can identify the probability measure p_t appearing in problem (1) with p_{X_t}. If the econometrician's date t information set contains $Y_t, X_t, Y_{t-1}, X_{t-1}, \ldots$, then p_{X_t} is also identified by the econometrician.

Section 4 relaxes Assumption 2 to allow the forecaster to have a larger information set than is available to the econometrician.

2.2. The Loss Functions

Let $C = C(D \times D)$ denote the set of continuous functions on $D \times D$, where D is a nonempty, compact subset of $\mathbb{R}$.

DEFINITION 2.1. *We say that $\ell, \ell' \in C(D \times D)$ are*

(i) ***affine equivalent****, written $\ell \sim_{aff} \ell'$, if there exist a continuous $y \mapsto g(y)$ and $r > 0$ such that $\ell'(\hat{y}, y) = r \cdot \ell(\hat{y}, y) + g(y)$, and*
(ii) ***forecast equivalent*** *or* ***best response equivalent****, written $\ell \sim_{Br} \ell'$, if $Br\,(p \mid \ell) = Br\,(p \mid \ell')$ for all $p \in \Delta(D)$; i.e., ℓ and ℓ' always produce the same forecasts.*

PROPOSITION 2.1. *Affine equivalence implies forecast equivalence.*

Proof. If $\ell \sim_{aff} \ell'$, then $\int_F \ell'(\hat{y}, y) p(dy) = r \cdot \left[\int_F \ell(\hat{y}, y)\, p(dy)\right] + \int_F g(y)\, p(dy)$, where $F = \text{supp}(p)$. As $r > 0$ and the second term does not depend on $\hat{y}$, $\hat{y}^* \in F$ solves $\min_{\hat{y}} \int \ell(\hat{y}, y)\, p(dy)$ iff it also solves $\min_{\hat{y}} \int \ell'(\hat{y}, y)\, p(dy)$. ■

Example 3.3 shows that forecast equivalence does *not* imply affine equivalence when D contains three (or more) points. However, Example 3.1 shows that the two notions are equivalent if D has exactly two points, i.e., in the case where the variable to be forecast is binary.

Each affine equivalence class contains what we will call a *canonical form.*

DEFINITION 2.2. *The* ***canonical form*** *of $\ell \in C(D \times D)$ is defined as $\ell^c(\hat{y}, y) = \ell(\hat{y}, y) - \ell(y, y)$. The set of loss functions in canonical form is denoted $\mathscr{C}(D \times D)$ or simply $\mathscr{C}$.*

Clearly, $\ell \sim_{aff} \ell^c$ (set $r = 1$ and $g(y) = -\ell(y, y)$), and so $\ell \sim_{Br} \ell^c$. Hence, a loss function and its canonical form are indistinguishable given any data set on forecasts, realizations, and covariates. The canonical form is characterized by the property $\ell^c(y, y) = 0$ for all $y \in D$. Without loss of generality, we restrict attention to loss functions in canonical form, i.e., replace $C(D \times D)$ with $\mathscr{C}(D \times D)$. To simplify notation, we will drop the superscript c in referring to the elements of $\mathscr{C}$. Note that for ℓ and ℓ' in $\mathscr{C}$, $\ell \sim_{aff} \ell'$ if and only if $\ell' = r \cdot \ell$ for some $r > 0$.

One cannot possibly determine how the forecaster makes trade-offs between two possible forecasts if one (or both) of the forecasts is never made. The following condition rules out the existence of completely dominated forecasts. Here $\delta_y \in \Delta(D)$ denotes point mass on y; i.e., $\delta_y(E) = 1_E(y)$.

DEFINITION 2.3. *We say that $\ell(\hat{y}, y) \in \mathscr{C}(D \times D)$ has* ***no bias in case of certainty (nbcc)*** *if for all $y \in D$, $Br\left(\delta_y \mid D, \ell\right) = \{y\}$.*

By the "no-off-support-forecasts" postulated in Assumption 1, if the forecaster places unit mass on a given value of Y_{t+1}, then they are constrained to report that value as their forecast. For loss functions satisfying the *nbcc* property, this constraint is not binding—even if the set of allowable forecasts is extended to D, the unique optimal forecast for $p_t = \delta_y$ is y. For example, in the framework of Granger and Machina (2006), where the loss associated with a forecast-realization pair $(\hat{y}, y)$ derives from an underlying decision problem, it is automatically true that $y \in Br\left(\delta_y \mid D, \ell\right)$. In this case the *nbcc* condition simply adds the requirement that the optimal forecast be unique. The *nbcc* property could also be replaced with the Morris and Ui (2004, Prop. 2) condition that every forecast be strictly optimal for some δ_y.

The set of canonical loss functions with the *nbcc* property will be written as $\mathscr{C}_{nbcc} = \mathscr{C}_{nbcc}(D \times D)$. The following simple proposition ties this set to what is often the definition of a loss function in the traditional forecasting literature.

PROPOSITION 2.2. $\ell \in \mathscr{C}_{nbcc} \Leftrightarrow [\ell(\hat{y}, y) \geq 0$ *with equality iff* $\hat{y} = y]$.

Proof. Immediate from the definition of the canonical form, the *nbcc* property, and the fact that $\int \ell(\hat{y}, z)\delta_y(dz) = \ell(\hat{y}, y)$. ■

If $\#D = M < \infty$, then each $\ell(\hat{y}, y)$ in $\mathscr{C}_{nbcc}(D \times D)$ can be represented as an $M \times M$ matrix with zeros on its main diagonal and $M^2 - M$ strictly positive entries off the diagonal. In this case nonparametric preference recovery is a finite-dimensional problem. If D is an infinite set then $\mathscr{C}_{nbcc}(D \times D)$ is of course infinite-dimensional.

Forecast (or best response) equivalence requires that forecasts agree for all probabilities p when $\mathrm{supp}(p)$ is the set from which the forecaster must choose. The following gives a stronger-seeming, but equivalent formulation that will be useful subsequently.

PROPOSITION 2.3. *For* $\ell, \ell' \in \mathscr{C}_{nbcc}(D \times D)$, $\ell \sim_{Br} \ell'$ *if and only if* $Br\,(p \mid F, \ell) = Br\,(p \mid F, \ell')$ *for all compact* $F \subset D$ *and* $p \in \Delta(F)$.

This shows that the mapping $(p, F) \mapsto Br\,(p \mid F, \ell)$ contains no more information about the underlying loss function than the mapping $p \mapsto Br\,(p \mid \ell)$; i.e., as mentioned previously, off-support forecasts do not carry extra identifying power.

3. NONPARAMETRIC IDENTIFICATION

We begin with definitions and examples. The examples demonstrate that identifiability can fail either because the sequence $\{p_t\}$ does not vary enough to separate a given loss function from the other possible loss functions or because essentially different loss functions can be forecast equivalent. Theorem 1 in Section 3.2 shows that such loss functions are "knife-edge" or nongeneric phenomena—tiny changes in the loss functions make them identifiable. After ruling out this small set of loss functions, we show that sequences $\{p_t\}$ with a huge amount of variation are guaranteed to identify *any* member of the generic set. The examples and results in Section 3.3 show that much of this variation is also necessary for such universal identification to obtain.

3.1. Definition and Examples

Let the set $\mathscr{L} \subset \mathscr{C}_{nbcc}$ represent a model of the forecaster's loss function. Suppose that the model is correctly specified in that the forecaster's true loss function, ℓ°, is known to belong to $\mathscr{L}$.[4] The identification problem can be stated as follows. Given an infinite sequence of point forecasts and the underlying conditional distributions, is it possible to pick out uniquely from $\mathscr{L}$ the loss function used in generating the forecasts? (In the present setting uniqueness means up to affine equivalence.) More formally, we supply the following definition.

DEFINITION 3.1. *A loss function $\ell^\circ \in \mathscr{L}$ is **identified (in $\mathscr{L}$) under** $\{p_t\}_{t=1}^\infty \subset \Delta(D)$ if for any sequence of forecasts $f_t^\circ \in Br\,(p_t \mid \ell^\circ)$ the set*

$$\bigcap_{t\geq 1} \left\{ \ell \in \mathscr{L} : f_t^\circ \in Br\,(p_t \mid \ell) \right\} \tag{2}$$

*contains only positive scalar multiples of ℓ°, and it is **potentially identified** if it is identified under some sequence $\{p_t\}$.*

*A model $\mathscr{L}$ is **identified under** $\{p_t\}$ if every $\ell^\circ \in \mathscr{L}$ is identified under $\{p_t\}$, in which case we say that $\{p_t\}$ is a **universal identifying sequence for $\mathscr{L}$**. A model is **potentially identified** if it is identified under some $\{p_t\}$.*

Identification of a loss function in a model can fail for two reasons. First, it is possible that $\{p_t\}$ does not vary enough to pin down a given element of $\mathscr{L}$, i.e., the forecaster is not observed in a sufficiently diverse set of circumstances. Example 3.1 characterizes how much variability is needed when the variable to be forecast is binary. Second, the model $\mathscr{L}$ might not be potentially identified because it is too large. Example 3.3 shows that already when D contains three points, there may exist $\ell, \ell' \in \mathscr{L}$ such that the two functions are not affine equivalent and for *any* sequence $\{p_t\}$ the set (2) contains both ℓ and ℓ'. This is a more fundamental failure of identification because essentially different loss functions are indistinguishable from each other no matter how rich the available data set is.

Example 3.1
Let $Y \in D = \{0, 1\}$. An *nbcc* loss function $\ell(\hat{y}, y)$ in canonical form is specified by the two strictly positive numbers $\ell(1,0)$ and $\ell(0,1)$, where $\ell(0,0) \equiv \ell(1,1) \equiv 0$. At time t, the forecast is $f_t \in \operatorname{argmin}_{f\in\{0,1\}}[\ell(f,0)p_t(0) + \ell(f,1)p_t(1)]$ so that $f_t = 1(p_t(1) \geq c_\ell)$, where $c_\ell = 1/(1+\ell(1,0)/\ell(0,1)) \in (0,1)$ is the optimal cutoff for predicting 1 vs. 0.[5] Clearly, the cutoff c_ℓ is the most that can be recovered from a loss function because $\ell \sim_{Br} \ell'$ iff $c_\ell = c_{\ell'}$. Further, $c_\ell = c_{\ell'}$ iff $\ell \sim_{aff} \ell'$. The model $\mathscr{L} = \{\ell, \ell'\}$, $\ell \nsim_{aff} \ell'$, is identified under $\{p_t\}$ iff at least one of the p_t belongs to the interval between the associated cutoffs; a given ℓ° is identified in $\mathscr{L} = \mathscr{C}_{nbcc}$ iff there are subsequences $\{p_{t'}\}$ and $\{p_{t''}\}$ such that $p_{t'} \uparrow c_{\ell^\circ}$ and $p_{t''} \downarrow c_{\ell^\circ}$; and $\mathscr{L} = \mathscr{C}_{nbcc}$ is identified under $\{p_t\}$ iff $\{p_t(1)\}$ is dense in $[0,1]$, equivalently, $\{p_t\}$ is a universally identifying sequence for $\mathscr{C}_{nbcc}$ iff $\{p_t(1)\}$ is dense.

The next two examples show that for general D best response equivalence does not imply affine equivalence; hence, $\mathscr{C}_{nbcc}(D \times D)$ fails to be potentially identified. The first example pertains to asymmetric absolute loss, one of the canonical loss functions in the forecasting literature.[6]

Example 3.2
Let $D = [a,b] \subset \mathbb{R}$, $a < b$, and $\alpha \in (0,1)$. For $\phi : \mathbb{R} \to \mathbb{R}$ a continuous, strictly increasing function, the following loss functions belong to $\mathscr{C}_{nbcc}(D \times D)$:

$$\ell\left(\hat{y}, y\right) = \left[\alpha - 1\left(\hat{y} \geq y\right)\right]\left(y - \hat{y}\right) \quad \text{and} \quad \ell_\phi\left(\hat{y}, y\right) = \left[\alpha - 1\left(\hat{y} \geq y\right)\right] \times \left(\phi(y) - \phi\left(\hat{y}\right)\right). \tag{3}$$

Here $\ell \not\sim_{aff} \ell_\phi$ unless for some some $r > 0$, $(y - \hat{y}) = r \cdot (\phi(y) - \phi(\hat{y}))$, and for any distribution $p \in \Delta(D)$, $Br\,(p \mid \ell)$ consists of a set of α-quantiles, q_α, associated with p. The following considerations show that for any ϕ and p, $Br\left(p \mid \ell_\phi\right) = Br\,(p \mid \ell)$:

(a) the forecaster's objective under ℓ_ϕ is $\min_{\hat{y}} \int [\alpha - 1(\hat{y} \geq y)](\phi(y) - \phi(\hat{y}))p(dy)$;

(b) this can be rewritten as $\min_{\hat{z}} \int [\alpha - 1(\hat{z} \geq z)](z - \hat{z})p'(dz)$ where $p'(A) := p(\phi^{-1}(A))$ is the image law of p under the mapping ϕ;

(c) the solutions to the problems in (b) are the α-quantiles, q'_α, associated with p'; and

(d) q'_α is an α-quantile of p' iff $\phi^{-1}(q'_\alpha)$ is an α-quantile of p.

The second example uses a three-point domain and provides a deeper insight into the reasons why potential identification can fail in $\mathscr{C}_{nbcc}(D \times D)$. This example is the key to our theory of generic nonparametric identification.

Example 3.3
For $D = \{1, 2, 3\}$, let $[\ell(\hat{y}, y)] = \begin{bmatrix} 0 & 1 & 3 \\ 1 & 0 & 2 \\ 3 & 2 & 0 \end{bmatrix}$ and $[\ell'(\hat{y}, y)] = \begin{bmatrix} 0 & 3 & 4 \\ 3 & 0 & 1 \\ 4 & 1 & 0 \end{bmatrix}$ where the (i, j) entry in each matrix corresponds to $\ell(\hat{y}, y) = \ell(i, j)$. Here ℓ and ℓ' represent essentially different trade-offs between joint distributions on $(\hat{y}, y)$ pairs because they do not represent the same preferences—there are $p, q \in \Delta(D)$ and $\hat{y} \in D$ such that $(\hat{y}, p) \succ_\ell (\hat{y}, q)$ but $(\hat{y}, q) \succ_{\ell'} (\hat{y}, p)$. We show that ℓ and ℓ' are forecast equivalent; i.e., for every $F \subset D$ and every $p \in \Delta(F)$, $Br\,(p \mid F, \ell) = Br\left(p \mid F, \ell'\right)$. This in turn implies that the loss functions in the two-dimensional cone $N(\ell, \ell') := \{\alpha\ell + \beta\ell' : \alpha, \beta \geq 0, \alpha + \beta > 0\}$ are all forecast equivalent, and so no model containing linearly independent elements from $N(\ell, \ell')$ can be potentially identified.

For $\#F = 3$, i.e., $F = D$, the best response correspondences $p \mapsto Br\,(p \mid D, \ell)$ and $p \mapsto Br\left(p \mid D, \ell'\right)$ over the unit simplex are depicted in Figure 1. First note that

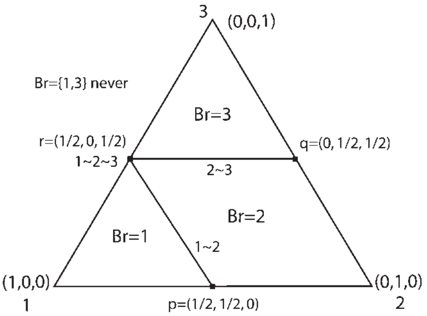

FIGURE 1. Best responses under ℓ and ℓ' when $F = D$.

- for $p = (p(1), p(2), p(3)) = (1/2, 1/2, 0)$, $Br\,(p \mid D, \ell) = Br\left(p \mid D, \ell'\right) = \{1, 2\}$,
- for $q = (0, 1/2, 1/2)$, $Br\,(q \mid D, \ell) = Br\left(q \mid D, \ell'\right) = \{2, 3\}$, and
- for $r = (1/2, 0, 1/2)$, $Br\,(r \mid D, \ell) = Br\left(r \mid D, \ell'\right) = \{1, 2, 3\}$.

Because the set of probabilities for which a given $\hat{y} \in D$ is an optimal forecast is closed and convex (see Lemma B.2 in Appendix B), $\{2, 3\} \subset Br\,(\pi \mid D, \ell)$ and $\{2, 3\} \subset Br\left(\pi \mid D, \ell'\right)$ for π along the line joining q and r; $\{1, 2\} \subset Br\,(\pi \mid D, \ell)$ and $\{1, 2\} \subset Br\left(\pi \mid D, \ell'\right)$ for π along the line joining p and r. Because *nbcc* is satisfied, moving π a small amount in the direction of any vertex from either of these indifference sets leads to a unique best forecast. Thus, $Br\,(\pi \mid D, \ell) = Br\left(\pi \mid D, \ell'\right)$ for all $\pi \in \Delta(D)$.

For $F \subsetneq D$, $\#F = 2$, note that if one deletes row i and column i from both ℓ and ℓ', then the resulting 2×2 matrices are positive scalar multiples, so trivially $Br\,(p \mid F, \ell) = Br\left(p \mid F, \ell'\right)$ for all $p \in \Delta(F)$.

There are four additional comments worth making about these two examples.

1. Example 3.3 does not depend on the choice of three-point set D.
2. There is an intimate connection between Examples 3.2 and 3.3. Let $F = \{y_1, y_2, y_3\} \subset [a, b]$, $y_1 < y_2 < y_3$ and consider the restriction of the loss functions in (3) to F. It is straightforward to check that the probability measure $p = (\alpha, 0, 1 - \alpha)$ makes the forecaster indifferent between forecasting y_1, y_2, or y_3 (if off-support forecasts are allowed). Thus, every three-point restriction of asymmetric absolute loss results in the same type of "bad" loss function exhibited in Example 3.3.
3. The "problem" with the loss functions in the cone $N(\ell, \ell')$ in Example 3.3 can be understood as follows. The point r at which the forecasts are completely indifferent imposes two linearly independent restrictions on the six numbers representing a given member of $N(\ell, \ell')$. Two additional points on the $1 \sim 2$ and $2 \sim 3$ indifference lines (say, p and q) give two more independent restrictions. However, as the point of "total indifference" falls on the boundary of the unit simplex, there are no other independent indifference conditions. Recovery of six numbers up to scale from four equations is not possible. A loss function in $\mathscr{C}_{nbcc}(D \times D)$ with $\#D = 3$ can fall into two other categories: the point of total indifference may be in the interior of the unit simplex (with three indifference lines meeting at this point); or it may not exist (there are only two indifference lines that do not intersect as in Figure 2 in Section 3.3). It turns out that in both of these cases there are five linearly independent indifference conditions allowing preference recovery up to scale.
4. The example can be easily rewritten so that instead of r being the point at which all three forecasts are optimal, it is either p or q (by *nbcc* only one of them can have this property). For each of these three choices, there is a two-dimensional cone of loss functions with the property that no two linearly

independent members are identifiable. More generally, let $p' = (a, 1-a, 0)$, $q' = (0, b, 1-b)$, and $r' = (c, 0, 1-c)$ where $0 < a, b, c < 1$. For each such (a, b, c), there are three, two-dimensional cones of loss functions with linearly independent members not being identifiable. Combining, the problematic loss functions are the union of three, five-dimensional manifolds. As rich as this class is, it is a very small subset of the six-dimensional space $\mathscr{C}_{nbcc}(D \times D)$. Further, the dimension of $\mathscr{C}_{nbcc}(D \times D)$ grows quadratically in $\#D$, whereas the dimension of the problematic manifold grows linearly, suggesting that this type of problem is strongly nongeneric when D is larger, though Example 3.2 shows that there can be an infinite-dimensional set of problematic loss functions when D is an interval.

3.2. Generic Identification

We now show that the failure of potential identification highlighted in Examples 3.2 and 3.3 is a "rare" occurrence for arbitrary compact $D \subset \mathbb{R}$, and that by ruling out such occurrences, identification is restored.

DEFINITION 3.2. *Let $F = \{x_1, x_2, x_3\}$ be a three-point subset of D. The loss function $\ell \in \mathscr{C}_{nbcc}(D \times D)$ has a* ***three-point boundary problem at*** *F if there exists $p \in \Delta(F)$ with $p(x_i) = 0$ for some $x_i \in F$ while $Br\,(p \mid F, \ell) = F$.*

DEFINITION 3.3. *Let $\mathscr{G} = \mathscr{G}(D \times D)$ denote the collection of ℓ in $\mathscr{C}_{nbcc}(D \times D)$ for which there exists a dense $D' \subset D$ such that ℓ does not have a three-point boundary problem at any $F \subset D'$.*

Loosely speaking, $\mathscr{G}(D \times D)$ collects all loss functions with a dense set of well-behaved three-point restrictions. For $\#D = 2$, $\mathscr{C}_{nbcc}$ vacuously satisfies Definition 3.3, so that, in this case $\mathscr{G} = \mathscr{C}_{nbcc}$, as was more directly observed in Example 3.1.

We will show that $\mathscr{C}_{nbcc}$ is a convex, topologically complete, separable metric space. A subset S of such a space is called **relatively shy** (Anderson and Zame, 2001) if it satisfies a generalization of being a Lebesgue null set. The subset S is **Baire small** (Baire, 1899) if it is the union of (at most) countably many closed sets with empty interior. A set is **totally small** if it is both relatively shy and Baire small.[7]

The following result is central to the paper.

THEOREM 1. *$\mathscr{G} = \mathscr{G}(D \times D)$ has the following properties.*

(i) $\mathscr{C}_{nbcc} \setminus \mathscr{G}$ is a totally small subset of $\mathscr{C}_{nbcc}$.

(ii) $\ell, \ell' \in \mathscr{G}$ are affine equivalent iff they are forecast equivalent. Indeed, for any $\ell, \ell' \in \mathscr{G}$, if $\ell \nsim_{aff} \ell'$, then there exists a nonempty open set $\mathcal{P} = \mathcal{P}_{\ell,\ell'} \subset \Delta(D)$ such that the set of $p \in \mathcal{P}$ for which $Br\,(p \mid \ell) \cap Br\,(p \mid \ell') = \varnothing$ is dense in $\mathcal{P}$.

(iii) If $\{p_t\}$ is a sequence of distributions with $\{p_t, \mathrm{supp}(p_t)\}$ dense in $\Delta(D) \times \mathcal{K}(D)$, where $\mathcal{K}(D)$ is the space of compact subsets of D, then $\mathscr{G}$ is identified under $\{p_t\}$.[8]

Part (i) is our statement that $\mathscr{G}$ is a large or generic subset of $\mathscr{C}_{nbcc}$. Part (ii) states that the way expected loss minimization "assigns" a best response correspondence $p \mapsto Br\,(p \mid \ell)$ to each $\ell \in \mathscr{C}_{nbcc}$ is essentially unique on $\mathscr{G}$ because the best response functions belonging to nonaffine equivalent loss functions differ from each other on a dense subset of an open set. Part (iii) shows that with a tremendous amount of variation in the conditional distribution of the outcome, $\mathscr{G}$ is sure to be identified, i.e., nonparametric identification of a generic set of preferences is theoretically possible. This is a consequence of the uniqueness of $Br\,(\cdot \mid \ell)$ over $\mathscr{G}$ combined with some continuity properties. In Section 3.3 we will briefly examine to what extent the variation stated in part (iii) is actually necessary for the identification of $\mathscr{G}$. An immediate corollary to part (iii) is that if the entire set $\mathscr{G}$ is identified under a sequence $\{p_t\}$, then each $\ell \in \mathscr{G}$ must be identified under some subsequence $\{p_{t'}^{(\ell)}\}$. Such a subsequence might have less variation than $\{p_t\}$; e.g., in the binary case, described in Example 3.1, $\{p_t(1)\}$ needs to be dense in $[0,1]$ for universal identification, whereas to identify a given loss function ℓ, one only needs a subsequence $\{p_{t'}^{(\ell)}(1)\}$ that is dense around c_ℓ (i.e., for every $\epsilon > 0$, $\{p_{t'}^{(\ell)}(1)\}$ is infinitely often in the interval $(c_\ell - \epsilon, c_\ell)$ and infinitely often in the interval $(c_\ell, c_\ell + \epsilon)$). Finally, we note that it is an open question whether or not there are sets strictly larger than $\mathscr{G}$ that are potentially identified; Theorem 1 gives sufficient conditions, but we do not know if they are necessary for general compact D.

In sum, the key part of Theorem 1, and the one that is perhaps the hardest to see, is that forecast equivalence implies affine equivalence in $\mathscr{G}(D \times D)$ for general compact D. The proof, detailed in Appendix D, proceeds as follows. First we show that the result holds for $\#D = 3$. We use induction to extend the claim to finite domains, where the definition of $\mathscr{G}$ means that no three-point restriction of a loss function belonging to it has a boundary problem. For arbitrary compact D, the general definition of $\mathscr{G}$ and the previous step imply that forecast equivalent loss functions are proportional to each other when restricted to any finite subset of D'; hence, they must be proportional on D' itself. As D' is dense in D and the loss functions are continuous, proportionality extends to D.

The set $\mathscr{G}(D \times D)$ has a rather abstract definition, and although being in $\mathscr{G}(D \times D)$ is a generic property, it would be good to know about specific classes of loss functions. The following definition requires that that no convex combination of any pair $\ell(\cdot, y_1)$ and $\ell(\cdot, y_2)$ be flat as a function of $\hat{y}$ on a set having positive Lebesgue measure.

DEFINITION 3.4. *For convex D, a loss function $\ell \in \mathscr{C}_{nbcc}(D \times D)$ is* ***almost nowhere an affine function of itself*** *if for all $y_1, y_2 \in D$, $y_1 \neq y_2$, for all $\beta \in (0,1)$, and for all $\kappa \in \mathbb{R}$, $\lambda(\{\hat{y} \in D : \beta\ell(\hat{y}, y_1) + (1-\beta)\ell(\hat{y}, y_2) = \kappa\}) = 0$, where λ is Lebesgue measure.*

The condition fails if $\ell(\cdot, y_1)$ is a negative affine transformation of $\ell(\cdot, y_2)$ on a nondegenerate interval for some y_1 and y_2, as is the case for all of the asymmetric absolute loss functions in Example 3.2 (take $\beta = \alpha$).

PROPOSITION 3.1. *For convex D, if* $\ell \in \mathscr{C}_{nbcc}(D \times D)$ *is almost nowhere an affine function of itself or each* $\ell(\cdot, y)$ *is strictly convex, then* $\ell \in \mathscr{G}(D \times D)$.

In many settings, it is easy to verify that nonconvex loss functions satisfy Definition 3.4. The strictly risk averse case covers generalized mean squared loss, i.e., $\ell(\hat{y}, y) = h(y)(\hat{y} - y)^2$, where $h(\cdot)$ is continuous and strictly positive. Generalized check functions, $\ell(\hat{y}, y) = |\hat{y} - y|^{\alpha} 1_{\{\hat{y} < y\}} + |\hat{y} - y|^{\gamma} 1_{\{\hat{y} \geq y\}}$ are also included—as long as $\alpha, \gamma > 1$.

As mentioned previously, asymmetric absolute loss violates Definition 3.4, and it is in fact excluded from $\mathscr{G}$ (by Example 3.2 this exclusion is necessary for identification to obtain). We do not however know whether or not all functions violating Definition 3.4 are excluded from $\mathscr{G}$.

3.3. Necessary Variation in $\{p_t\}$ for Universal Identification of $\mathscr{G}$

In Theorem 1 the denseness of the sequence $\{p_t, \operatorname{supp}(p_t)\}$ in $\Delta(D) \times \mathcal{K}(D)$ is sufficient to identify all loss functions in $\mathscr{G}(D \times D)$. It is natural to ask how much variation is *necessary* for $\{p_t\}$ to be a universal identifying sequence for $\mathscr{G}$.

In our nonparametric setting it is clearly not possible to recover fully the forecaster's loss function if parts of D never become optimal forecasts as p_t varies.[9] The following result indicates the amount of variation necessary for the set of observed forecasts to be dense in D.

PROPOSITION 3.2. *If* $D = [a,b] \subset \mathbb{R}$, $\mathcal{P} \subset \Delta(D)$ *is closed, and* $\cup_{p \in \mathcal{P}} Br(p \mid \ell) = D$ *for all* $\ell \in \mathscr{G}$, *then* $\delta_y \in \mathcal{P}$ *for all* $y \in D$.

To give some sense of the richness that this result entails, if each $p \in \mathcal{P}^{\circ}$ has a Lebesgue density and $\mathcal{P} := \operatorname{cl}(\mathcal{P}^{\circ})$ satisfies the conditions of Proposition 3.2, then $\mathcal{P}^{\circ}$ must be a complete class of distributions.

A sequence of conditional probabilities with full support D can have closure satisfying the variability requirement of Proposition 3.2 but still not be a universal identifying sequence for $\mathscr{G}$. To have this property, some variation in $\operatorname{supp}(p_t)$, i.e., the set of allowable forecasts, is also needed. The following examples are designed to highlight how this additional source of variation can reveal information about the forecaster's loss function not available otherwise.

- Example 3.4 gives, for $\#D = 3$, an open subset of loss functions in $\mathscr{G}$ that are unidentified under any sequence $\{p_t\}$ with $\operatorname{supp}(p_t) = D$ for all t. To identify loss functions in this open set, it is necessary to observe $Br(p \mid \ell)$ for distributions p supported on two-point subsets of D. For some of these p's, $Br(p \mid D, \ell) \cap Br(p \mid \ell) = \varnothing$, i.e. the forecaster would rather violate the support condition in Assumption 1.

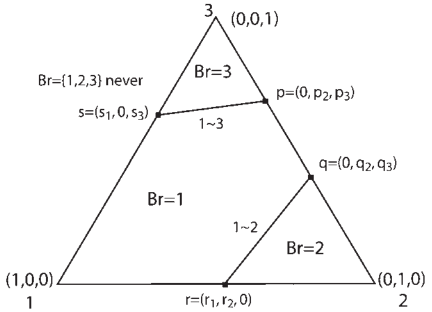

FIGURE 2. Best responses over D are fully determined by p, q, r, s.

- Example 3.5 gives, for D any compact convex subset of $(0,1]$, a class of utility functions parametrized by a single scalar, $\theta > 0$, with the following properties: each $\ell(\cdot,\cdot;\theta)$ belongs to $\mathscr{G}$; for $\theta \neq \theta'$, $\ell(\cdot,\cdot;\theta)$ and $\ell(\cdot,\cdot;\theta')$ are not forecast equivalent; yet, for distributions p supported on compact convex subsets of $(0,1]$, the optimal forecast $Br\,(p \mid \ell(\cdot,\cdot;\theta))$ does not depend on θ.

Example 3.4

Let $D = \{1,2,3\}$ and fix the probability measures $p = (0, p_2, p_3)$, $p_3 > 1/2$, $q = (0, q_2, q_3)$, $q_2 > 1/2$, $r = (r_1, r_2, 0)$, $r_1 > 0$, and $s = (s_1, 0, s_3)$, $s_1 > 0$ (see Figure 2). For $b > 0$ consider the *nbcc* loss function

$$\left[\ell_{p,q,r,s,b}(\hat{y}, y)\right] = \begin{bmatrix} 0 & 1 & b \\ \frac{r_2}{r_1} & 0 & \frac{q_2}{q_3} + b \\ \frac{s_3}{s_1} b & 1 + \frac{p_3}{p_2} b & 0 \end{bmatrix}, \tag{4}$$

where, as before, the (i,j) entry in the matrix corresponds to the loss of $(\hat{y}, y) = (i,j)$. This class of loss functions is constructed so that

$$Br\left(p \mid D, \ell_{p,q,r,s,b}\right) = Br\left(s \mid D, \ell_{p,q,r,s,b}\right) = \{1,3\},$$
$$Br\left(q \mid D, \ell_{p,q,r,s,b}\right) = Br\left(r \mid D, \ell_{p,q,r,s,b}\right) = \{1,2\}.$$

Observe that no $\ell_{p,q,r,s,b}$ has a three-point boundary problem so they all belong to $\mathscr{G}$; on the line segment between p and q, 1 is the unconstrained strict best response; i.e., for any $\alpha \in (0,1)$, $Br\left(\alpha p + (1-\alpha) q \mid D, \ell_{p,q,r,s,b}\right) = \{1\}$; for any distribution π supported on D, the best response $Br\left(\pi \mid D, \ell_{p,q,r,s,b}\right)$ is completely determined by the points p, q, r, and s and does not depend on the value of b. However, for binary distributions supported on $\{2,3\}$, the point between p and q at which the forecaster is indifferent between reporting 2 or 3 is uniquely determined by b so that, as in Example 3.1, b can be recovered by a suitable sequence $\{p_t\}$. Finally, for any given $\ell_{p,q,r,s,b}$, there exists some $\epsilon > 0$ such that for all $\ell' \in \mathscr{C}_{nbcc}(D \times D)$ with $\ell'(1,2) = 1$ and $\|\ell - \ell'\|_\infty < \epsilon$, ℓ' is also of the form (4). This shows that the $\ell_{p,q,r,s,b}$ constitute an open subset of $\mathbb{R}^5_{++}$ and also implies that positive scalar multiples of the $\ell_{p,q,r,s,b}$ constitute an open subset of $\mathbb{R}^6_{++} = \mathscr{C}_{nbcc}(D \times D)$.

Example 3.5
For $\theta > 0$, consider the class of loss functions

$$\ell\left(\hat{y}, y; \theta\right) = -\frac{1}{1+\theta}\hat{y}^{-(1+\theta)}y + \frac{1}{2+\theta}\hat{y}^{-(2+\theta)}y^2 + \frac{y^{-\theta}}{(1+\theta)(2+\theta)}$$

defined on $D \times D$, where D is a compact convex subset of $(0, 1]$. Nothing can be learned about θ by observing choices made by the forecaster over the set D—consider the forecaster's problem for p supported on D:

$$\min_{\hat{y}\in D}\int \ell\left(\hat{y}, y; \theta\right)\, p\,(dy) = \min_{\hat{y}\in D}\left[-\frac{1}{1+\theta}\hat{y}^{-(1+\theta)}\mathbb{E}Y + \frac{1}{2+\theta}\hat{y}^{-(2+\theta)}\mathbb{E}Y^2 + \text{const.}\right],$$

where Y is a random variable with distribution p. Setting the first-order conditions equal to zero yields $-\hat{y}^{-(2+\theta)}\mathbb{E}Y + \hat{y}^{-(3+\theta)}\mathbb{E}Y^2 = 0$, which one solves for $\hat{y}^* = \mathbb{E}Y^2/\mathbb{E}Y$, and the second-order conditions for a minimum are satisfied for all $\theta > 0$. To see that this class has *nbcc*, for $p = \delta_r$, note that $\mathbb{E}Y^2/\mathbb{E}Y = r$. To see that this class belongs to $\mathscr{G}(D \times D)$, and so is potentially identified, note that their canonical forms are never piecewise affine functions of themselves and apply Proposition 3.1.

It is worth observing how delicate the problems with identification in Example 3.5 are—small changes in the specification of the loss function would obviate them.

4. IDENTIFICATION WHEN SOME COVARIATES ARE UNOBSERVED

When the forecaster uses covariates unavailable to the econometrician, there are two ways to proceed: in the case of smooth loss functions, one can find subclasses that are identified by the observation that the expected derivative with respect to the forecast is equal to 0; alternatively, one can look for partial (or set) identification results, and this turns out to be quite easy in the binary case.

4.1. A General Strategy for Smooth Loss Functions

It is implicit in Assumption 2 that the econometrician can observe the forecaster's entire information set based on which the forecast is produced. We now relax this assumption and allow the forecaster to possess private information.

Consider the setup described in Example 2.1. We now suppose that the covariates observed by the forecaster can be partitioned as $X_t = (Z_t, Z_t')$, where the econometrician observes Z_t but not Z_t'. Hence, the econometrician identifies the distribution of Y_{t+1} conditional on Z_t but not conditional on (Z_t, Z_t').

Suppose that the density $p_{Z_t,Z_t'}$ is supported on some interval $[a_t, b_t] \subset [a, b] = D$, $a_t < b_t$, and let $\mathscr{C}_{nbcc}^{(1)}$ denote the collection of *nbcc* loss functions that are continuously differentiable in $\hat{y}$ over D. If the forecaster's true utility function ℓ° belongs to $\mathscr{C}^{(1)}_{nbcc}$, then forecasts f_t° falling strictly between a_t and b_t must satisfy the first-order condition

$$\int \ell^\circ_{\hat{y}}\left(f_t^\circ, y\right) p_{Z_t, Z_t'}(y) dy = 0 \quad \text{a.s.} \tag{5}$$

By the law of iterated expectations, equality (5) remains valid if $p_{Z_t,Z_t'}$ is replaced by p_{Z_t}. Then, adopting Definition 3.1, ℓ° is identified if the set

$$\bigcap_t \left\{ \ell \in \mathscr{C}^{(1)}_{nbcc} : \int \ell_{\hat{y}}\left(f_t^\circ, y\right) p_{Z_t}(y) dy = 0 \right\}$$

contains scalar multiples of ℓ° only.

It is easy to make this identification strategy technically more precise, e.g., by accommodating forecasts that are not interior solutions, etc. We could then ask similar questions to those underlying Theorem 1: How large a subset of $\mathscr{C}^{(1)}_{nbcc}$ is potentially identified? How much variation in $\{p_{Z_t}\}$ is sufficient for universal identification within such a set? How much is necessary? We hope to return to these questions in future research.

4.2. A set Identification Result for the Binary Case

In addition to Example 2.1, consider the setup in Example 3.1. Again, X_t is partitioned as $X_t = (Z_t, Z_t')$, where the econometrician observes Z_t but not Z_t'. Hence, for any given value z of Z_t, only $p_z := p_z(1) = P(Y_{t+1} = 1 | Z_t = z)$ is identified by the econometrician. However, the observed forecasts are based on $p_{z,z'} := p_{z,z'}(1) = P(Y_{t+1} = 1 \mid Z_t = z, Z_t' = z')$; specifically, $f_t = 1(p_{Z_t,Z_t'} \geq c_\ell)$.

Let $Q_z = P(p_{Z_t,Z_t'} \geq c_\ell \mid Z_t = z)$ be the proportion of the time a forecast of 1 is observed when $Z_t = z$. Although the distribution of the random variable $p_{Z_t,Z_t'}$ over the interval $[0,1]$ is unknown, by the law of iterated expectations it must satisfy $\mathbb{E}[p_{Z_t,Z_t'} \mid Z_t] = p_{Z_t}$ with probability 1. Let $\mathcal{R}_z$ denote the set of all distributions R on $[0,1]$ satisfying $\int_{[0,1]} r\, R(dr) = p_z$. For z and c_ℓ fixed, the following constraints on Q_z must be satisfied:

$$\inf_{R \in \mathcal{R}_z} \int_{[c_\ell, 1]} R(dr) \leq Q_z \leq \sup_{R \in \mathcal{R}_z} \int_{[c_\ell, 1]} R(dr).$$

For $p_z < c_\ell$, these bounds imply $Q_z \in [0, p_z/c_\ell]$; for $p_z \geq c_u$, $Q_z \in [(p_z - c_\ell)/(1 - c_\ell), 1]$.[10] Combining the two constraints and solving for c_ℓ yields $c_u \in [(p_z - Q_z)/(1 - Q_z), p_z/Q_z]$.

As the function $z \mapsto Q_z$ is identified by the econometrician, c_ℓ is set-identified as a point in the interval

$$\bigcap_t \left[\frac{p_{Z_t} - Q_{Z_t}}{1 - Q_{Z_t}}, \frac{p_{Z_t}}{Q_{Z_t}} \right] = \left[\sup_z \frac{p_z - Q_z}{1 - Q_z}, \inf_z \frac{p_z}{Q_z} \right]. \tag{6}$$

In the extreme case when no covariates are observed by the econometrician, i.e., $p_z = p = P(Y_{t+1} = 1)$ and $Q_z = Q = P(f_t = 1) = P(p_{Z'_t} \geq c_\ell)$, the interval given in equation (6) may still have nontrivial identifying power as long as (p, Q) is off of the diagonal of the unit square. With some of the covariates observed by the econometrician, and sufficient variability in p_{Z_t} and Q_{Z_t}, it is even possible that the identified set is a singleton.

At the other extreme, when Z'_t is empty, then Q_z is either 0 or 1, depending on the value of z. Let $\mathcal{Z}_0 = \{z : p_z < c_\ell\}$ and $\mathcal{Z}_1 = \{z : p_z \geq c_\ell\}$. Then $Q_z = 0$ for $z \in \mathcal{Z}_0$ and $Q_z = 1$ for $z \in \mathcal{Z}_1$. With the convention that division by zero produces plus or minus infinity, the interval in (6) reduces to $\left[\sup_{z\in\mathcal{Z}_0} p_z, \inf_{z\in\mathcal{Z}_1} p_z\right]$. Thus, the loss function is (point) identified if $\sup_{z\in\mathcal{Z}_0} p_z = \inf_{z\in\mathcal{Z}_1} p_z$. This identification result is of course the same as in Example 3.1.

5. CONCLUSION

In this paper we studied the problem of recovering forecaster preferences from a sequence of forecasts and the underlying sequence of forecast densities—the latter identified by the realizations of the variable of interest and the full set of covariates in the forecaster's information set. We showed that within a large nonparametric class of loss functions, defined by the "no bias in case of certainty" condition, optimal forecasting rules are not necessarily unique. Nevertheless, the set of loss functions responsible for this multiplicity turns out to be very small. We explicitly characterize a *generic* set of preferences for which nonparametric recovery is possible provided that the forecaster is observed making choices in response to a sufficiently wide variety of conditional distributions. Because of continuity properties of optimal forecasting rules, a dense sequence of distributions with a dense sequence of supports is guaranteed to identify all generic loss functions (though any particular loss function might be identifiable with substantially less variation).

In practice, it may be that not all variables in the forecaster's information set are observed by the econometrician. This means that the forecast density used by the forecaster will be unidentified. In the special case when the variable to be forecast is binary, we give a practically useful partial identification result for identifying the forecaster's preferences, but developing formal results for the more general cases is left for future research.

It is instructive to contrast these identification results with identification in a parametric setting (cf. Elliott et al., 2003, 2005). If the forecaster's utility function depends on a d-dimensional vector of parameters, d finite, then observing forecasts for d or more linearly independent forecast densities will, under general conditions, identify the unknown parameter, at least locally. Thus, unless the parametric model is exactly correct, the unknown parameter will often be overidentified in practice, necessitating the use of a minimum distance criterion for estimation. Making a parametric assumption greatly reduces the amount of information needed to achieve identification at the cost of providing a possibly imperfect approximation to the true loss function.

More generally, the strength of the various identifying assumptions one could impose on loss functions in addition to the *nbcc* property can be measured by how much the necessary variation needed for universal identification is reduced relative to the general case discussed in Section 3.3. This is also an interesting direction for future research.

NOTES

1. The following Web site, hosted at the Philadelphia Federal Reserve, lists a number of papers concerned with testing the rationality of economic forecasts: http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.cfm (accessed on 2 April 2012; updated from a link cited by Patton and Timmermann, 2005).

2. The Prokhorov distance metrizes the usual weak convergence of probabilities; see Appendix A.

3. The support of a probability p is the smallest *closed* set $C \subset D$ with $p(C) = 1$.

4. If $\mathscr{L}$ is a small subset of $\mathscr{C}_{nbcc}$ (e.g., a parametric model), then one might be worried about the possibility that $\mathscr{L}$ is misspecified. The models we consider will be either $\mathscr{C}_{nbcc}$ itself or large, nonparametric subsets of $\mathscr{C}_{nbcc}$. We will therefore maintain the assumption of correct specification.

5. For simplicity only, we assume that ties lead to $f_t = 1$.

6. This example is based on the criterion function proposed by Komunjer and Vuong (2010) to define a family of conditional quantile estimators. We thank an anonymous referee for this reference.

7. More detailed definitions and a discussion are in Appendix.

8. Denseness in $\mathcal{K}(D)$ is in terms of the Hausdorff metric; see Appendix A. A sequence $\{p_t\}$ can be dense in $\Delta(D)$ without any variation in the support of its elements; therefore, requiring the denseness of the support sets makes the condition stronger.

9. If there is a set $G \subset D$, open relative to D, such that points in G are never optimal forecasts under $\{p_t\}$, then one can alter the loss function on G in a way that further increases loss for $\hat{y} \in G$ (uniformly in y) and preserves continuity. Such a transformation will not change the best responses observed under $\{p_t\}$.

10. The solution arises by considering two-point measures in $\mathcal{R}_z$.

REFERENCES

Anderson, R.M. & W.R. Zame (2001) Genericity with infinitely many parameters. *Advances in Theoretical Economics* 1, 1–62.

Baire, R. (1899) Sur les Fonctions de Variables Réelles (thèse). *Annali di Matematica Pura ed Applicata* (ser. 3) 3, 1–123.

Capistran, C. & A. Timmermann (2009) Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking* 41, 365–396.

Corbae, D., M.B. Stinchcombe, & J. Zeman (2009) *An Introduction to Mathematical Analysis for Economic Theory and Econometrics*. Princeton University Press.

Elliott, G., I. Komunjer, & A. Timmermann (2003) Estimating Loss Function Parameters. Working Paper, Department of Economics, University of California, San Diego.

Elliott, G., I. Komunjer, & A. Timmermann (2005) Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies* 72, 1107–1125.

Elliott, G., I. Komunjer, & A. Timmermann (2008) Biases in macroeconomic forecasts: Irrationality or asymmetric loss. *Journal of the European Economic Association* 6, 122–157.

Granger, C.W.J. (1969) Prediction with a generalized cost of error function. *Operational Research Quarterly* 20, 199–207.

Granger, C.W.J. & M. Machina (2006) Forecasting and decision theory. In G. Elliott, C.W.J. Granger, & A. Timmermann (eds.), *Handbook of Economic Forecasting*. Elsevier.

Komunjer, I. & Q. Vuong (2010) Semiparametric efficiency bound in time series models for conditional quantiles. *Econometric Theory* 26, 383–405.

Mincer, J. & V. Zarnowitz (1969) The evaluation of economic forecasts. In J. Mincer (ed.), *Economic Forecasts and Expectations*. Columbia University Press.

Morris, S. & T. Ui (2004) Best response equivalence. *Games and Economic Behavior* 49, 260–287.

Patton, A.J. & A. Timmermann (2005) Testable Implications of Forecast Optimality. Working paper, Department of Economics, London School of Economics and Political Science.

Patton, A.J. & A. Timmermann (2007) Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* 102, 1172–1184.

APPENDIX A: Notation and Definitions

We introduce additional notation used in the proofs but not given in the main text. In addition, we state the definitions of some technical concepts used in the main text or the proofs.

1. For $y \in D$ and $\epsilon > 0$, $B_\epsilon(y) = \{y' \in D : |y' - y| < \epsilon\}$ is the ϵ-ball around y.
2. For $A \subset D$, $A^\epsilon = \cup_{y \in A} B_\epsilon(y)$, denotes the ϵ-ball around the set A.
3. The **Prokhorov distance** between two distributions $p, q \in \Delta(D)$ is given by $\rho(p,q) = \inf\{\epsilon \geq 0 : \forall A \subset D,\ p(A) \leq q(A^\epsilon) + \epsilon$ and $q(A) \leq p(A^\epsilon) + \epsilon\}$. The distance between point masses, $\rho(\delta_y, \delta_{y'})$, is equal to $\min\{|y - y'|, 1\}$, and $\rho(p_n, p) \to 0$ iff $\int h\, dp_n \to \int h\, dp$ for all bounded continuous h. See, e.g., Corbae, Stinchcombe, and Zeman (2009, Ch. 9.3) for further properties.
4. For $A \subset \Delta(D)$, $A^\epsilon = \cup_{p \in A} B^\rho_\epsilon(p)$ denotes the Prokhorov ϵ-ball around the set A.
5. $\mathcal{K}(D)$ denotes the class of nonempty, compact subsets of D.
6. $d_H(A, B) = \inf\{\epsilon \geq 0 : A \subset B^\epsilon,\ B \subset A^\epsilon\}$ denotes the **Hausdorff distance** between compact sets.
7. A compact-valued correspondence $\Gamma : \mathbb{R} \to \mathcal{K}(\mathbb{R})$ is **upper hemicontinuous** if for every $r \in \mathbb{R}$ and every $\epsilon > 0$ there exists $\delta > 0$ such that $[d(r, r') < \delta] \Rightarrow [\Gamma(r') \subset [\Gamma(r)]^\epsilon]$.
8. For $\ell \in \mathscr{C}$ and $F \subset D$ the restriction of ℓ onto $F \times F$ is denoted as $\ell_{|F \times F}$.

APPENDIX B: Auxiliary Results

We now establish some basic properties of the mapping $(p, F) \mapsto Br(p \mid F, \ell)$.

LEMMA B.1. *For any $\ell \in \mathscr{C}_{nbcc}$, the mapping $(p, F) \mapsto Br(p \mid F, \ell) \in \mathcal{K}(F)$ is upper hemicontinuous at each $(p, F) \in \Delta(D) \times \mathcal{K}(D)$.*

Proof. Let $F \in \mathcal{K}(D)$ and $p \in \Delta(D)$ (p may or may not belong to $\Delta(F)$). Define $L(\hat{y}, p, F) := \int_D [-\ell(\hat{y}, y)] p(dy)$. This function is jointly continuous on $D \times \Delta(D) \times \mathcal{K}(D)$; in fact, it is constant in the argument F. The optimization problem of interest is $\max_{\hat{y} \in F} L(\hat{y}, p, F)$, where we consider (p, F) as parameters of the problem. As $F \subset D$ is compact and is trivially a continuous function of (F, p), it is immediate from the theorem of the maximum (e.g., Corbae et al., 2009, Thm. 4.10.2) that

$Br(p \mid F, \ell) = \arg\max_{\hat{y}\in F} L(\hat{y}, p, F)$ is compact and upper hemicontinuous at each $(p, F) \in \Delta(D) \times \mathcal{K}(D)$. ■

LEMMA B.2. *Let $F \subset D$, F compact, and $\hat{y} \in F$. The set of probability measures for which $\hat{y}$ is optimal as a forecast, i.e., the set $\{p \in \Delta(F) : \hat{y} \in Br(p \mid F, \ell)\}$, is closed and convex.*

Proof. Closure comes from Lemma B.1. For convexity, suppose that for all $\hat{y}' \in F$,

$$\int \ell\left(\hat{y}, y\right) dp(y) \leq \int \ell\left(\hat{y}', y\right) dp(y) \quad \text{and} \quad \int \ell\left(\hat{y}, y\right) dq(y) \leq \int \ell\left(\hat{y}', y\right) dq(y).$$

Multiplying the first inequality by $\alpha \in [0, 1]$, the second inequality by $1 - \alpha$, adding the two together, and using the linearity of the integral yields $\int \ell(\hat{y}, y)\, d(\alpha p + (1-\alpha)q)(y) \leq \int \ell(\hat{y}', y)\, d(\alpha p + (1-\alpha)q)(y)$. ■

The following lemma relies on the *nbcc* property.

LEMMA B.3. *Let $\ell \in \mathscr{C}_{nbcc}$, $F \subset D$, F compact, $p \in \Delta(F)$, and $\hat{y} \in Br(p \mid F, \ell)$. If $\hat{y}$ is an optimal forecast at p, then there are q's arbitrarily close to p for which $\hat{y}$ is the unique optimal forecast.*

Proof. We must show that for all $\epsilon > 0$, there exists $q \in \Delta(F)$, $q \neq p$, $\rho(p,q) < \epsilon$ such that $Br(q \mid F, \ell) = \{\hat{y}\}$. Because $\hat{y} \in Br(p \mid F, \ell)$, $\int \ell(\hat{y}, y)\, dp(y) \leq \int \ell(\hat{y}', y)\, dp(y)$ for all $\hat{y}' \neq \hat{y}$. Because ℓ has *nbcc*, $\int \ell(\hat{y}, y)\, d\delta_{\hat{y}}(y) < \int \ell(\hat{y}', y)\, d\delta_{\hat{y}'}(y)$ for all $\hat{y}' \neq \hat{y}$. Therefore, for any $\alpha \in (0, 1)$, $\int \ell(\hat{y}, y)\, d(\alpha\delta_{\hat{y}} + (1-\alpha)p)(y) < \int \ell(\hat{y}', y)\, d(\alpha\delta_{\hat{y}} + (1-\alpha)p)(y)$ for all $\hat{y}' \neq \hat{y}$. For α sufficiently close to 0, $\rho(\alpha\delta_{\hat{y}} + (1-\alpha)p, p) < \epsilon$. ■

LEMMA B.4. *For $\ell \in \mathscr{C}_{nbcc}(D \times D)$ and $F \subset D$ compact, $Br(p \mid F, \ell) \neq Br(p \mid F, m)$ for some $p \in \Delta(F)$ iff there exists a nonempty open $\mathcal{P} \subset \Delta(F)$ with $Br(p \mid F, \ell) \cap Br(p \mid F, m) = \varnothing$ for all $p \in \mathcal{P}$.*

Proof. ($\Leftarrow$) As $Br(p \mid F, \ell)$ and $Br(p \mid F, m)$ are nonempty for any p and F, $Br(p \mid F, \ell) \cap Br(p \mid F, m) = \varnothing$ implies that $Br(p \mid F, \ell) \neq Br(p \mid F, m)$.

($\Rightarrow$) Now suppose that for a given $F \subset D$, F compact, and $p \in \Delta(F)$, $Br(p \mid F, \ell) \neq Br(p \mid F, m)$. Interchanging ℓ and m if necessary, there exist $\epsilon > 0$ and $\hat{y} \in Br(p \mid F, \ell)$ with $d(\hat{y}, Br(p \mid F, m)) > 2\epsilon$. By upper hemicontinuity, there exists some $\delta > 0$ such that $q \in B^{\rho}_{\delta}(p)$ implies $Br(q \mid F, m) \subset [Br(p \mid F, m)]^{\epsilon}$. Because $\ell \in \mathscr{C}_{nbcc}$, by Lemma B.3 there is $q' \in B^{\rho}_{\delta}(p)$ such that $Br\left(q' \mid F, \ell\right) = \{\hat{y}\}$. By upper hemicontinuity again, for some $\delta' > 0$, $B^{\rho}_{\delta'}(q') \subset B^{\rho}_{\delta}(p)$, and for every $q'' \in B^{\rho}_{\delta'}(q')$, $Br\left(q'' \mid F, \ell\right) \subset [Br\left(q' \mid F, \ell\right)]^{\epsilon} = (\hat{y} - \epsilon, \hat{y} + \epsilon)$ and, by construction, $Br\left(q'' \mid F, m\right) \subset [Br(p \mid F, m)]^{\epsilon}$. As $d(\hat{y}, Br(p \mid F, m)) > 2\epsilon$, for every q'' in the ρ-open set $B^{\rho}_{\delta'}(q')$, $Br\left(q'' \mid F, \ell\right)$ and $Br\left(q'' \mid F, m\right)$ are disjoint. ■

APPENDIX C: Notions of Smallness and Genericity

We now describe in some detail the notion of smallness used in Theorem 1. Relative shyness has a rather involved definition. If C is a separable, topologically complete, convex subset of a toplogical vector space, we say that a set $S \subset C$ is shy relative to C if for all $c \in C$, all neighborhoods U_c of c, and all $\epsilon > 0$, there exists a compactly supported $\eta \in \Delta(C)$ such that $\eta(U_c \cap [\epsilon C + (1-\epsilon)c]) = 1$ and $(\forall x \in V)[\eta(S' + x) = 0]$. A useful sufficient condition for shyness is finite shyness, which takes η to be the continuous affine image of the uniform distribution on the unit ball in $\mathbb{R}^k$ for some k. See Anderson and Zame (2001) for further details.

DEFINITION C.1. *If C is a separable, topologically complete, convex subset of a topological vector space, we say that a set $S \subset C$ is* ***totally small*** *if it is both Baire small and relatively shy (with respect to C). A set is* ***totally large*** *(or generic in C) if it is the complement of a totally small set.*

Remarks.

1. A metric space (X, d) is topologically complete if there is some metric e that induces the same topology as d, and X is complete under e.
2. In our context the ambient vector space is $\mathscr{C}(D \times D)$ equipped with the topology induced by the uniform (sup) metric, i.e., $d_\infty(\ell, m) = \sup_{\hat{y}, y \in D} |\ell(\hat{y}, y) - m(\hat{y}, y)|$. This space is shown to be topologically complete subsequently. The convex subset of interest relative to which shyness is considered is $C = \mathscr{C}_{nbcc}$.
3. As stated in the text, a set is Baire small if it is the countable union of closed sets with no interior.
4. A subset of a finite-dimensional topologically complete C is relatively shy if and only if it is a Lebesgue null subset of the affine hull of C. More generally, relatively shy sets have empty relative interior, and the class of relatively shy sets is closed under countable union (again, see Anderson and Zame, 2001, for details). Such properties are then inherited by totally small sets; e.g., the class of totally small sets is closed under countable unions; a totally small set has no interior (relative to C), and if C is finite-dimensional, a totally small set must have Lebesgue measure zero, though this is not sufficient.
5. The combination of Baire smallness and Lebesgue measure zero is a strictly stronger criterion for smallness than either of the two taken separately, and the same is true for Baire smallness and relative shyness. Consider the following example. Let $\{q_n : n \in \mathbb{N}\}$ enumerate the points in $\mathbb{R}^k$ with rational coordinates. Let $E(\epsilon) := \cup_n B_{\epsilon/2^n}(q_n)$, an open dense set having Lebesgue measure $K\epsilon^k$ for some $K > 0$. Therefore, $F(\epsilon) := [E(\epsilon)]^c$ is a closed set with no interior. It follows that $\cup_n F(1/n)$ is a Baire small set having full Lebesgue measure and $\cap_n E(1/n)$ is a Baire large set having Lebesgue measure zero.

To guarantee that relative shyness and Baire smallness are useful, we must show the following result.

LEMMA C.1. *$\mathscr{C}_{nbcc}(D \times D)$ is separable and topologically complete.*

Proof. If D is finite, separability and topological completeness are trivial, though the proof for general compact D can also be applied. The outline of the proof is as follows.

(a) We define a suitable metric d on $\mathscr{C}_{nbcc}(D \times D)$ for compact D.
(b) Then we show that it is topologically equivalent to the uniform (sup norm) metric d_∞.
(c) Then we show that $\mathscr{C}_{nbcc}(D \times D)$ is complete in the newly defined metric.

Step (a): Defining a metric for $\mathscr{C}_{nbcc}(D \times D)$ for compact D. For $k \in \mathbb{N}$ and $\ell \in \mathscr{C}_{nbcc}(D \times D)$, let $r_k(\ell) = \min_{|\hat{y}-y| \geq 1/k} |\ell(\hat{y}, y)|$, and for $\ell, m \in \mathscr{C}_{nbcc}(D \times D)$, let $f_k(\ell, m) = \min\{1, |1/(r_k(\ell)) - 1/(r_k(m))|\}$. Define

$$d(\ell, m) = d_\infty(\ell, m) + \textstyle\sum_k \frac{1}{2^k} f_k(\ell, m).$$

Here $\ell(\hat{y}, y) = 0$ iff $\hat{y} = y$ implies $r_k(\ell) > 0$, so d is well defined, and it is immediate that d is a metric.

Step (b): Topological equivalence. We first show that for each k and all ℓ, m, $|r_k(\ell) - r_k(m)| \leq d_\infty(\ell, m)$. To see this, note that the set $\{(\hat{y}, y) \in D \times D : |\hat{y} - y| \geq 1/k\} \subset \mathbb{R}^2$ is compact; hence, we can choose from it $(\hat{y}_0, y_0)$ such that $|m(\hat{y}_0, y_0)| = r_k(m)$. Further, we can write

$$\begin{aligned} r_k(\ell) \leq \left|\ell\left(\hat{y}_0, y_0\right)\right| &= \left|\ell\left(\hat{y}_0, y_0\right) - m\left(\hat{y}_0, y_0\right) + m\left(\hat{y}_0, y_0\right)\right| \\ &\leq \left|\ell\left(\hat{y}_0, y_0\right) - m\left(\hat{y}_0, y_0\right)\right| + \left|m\left(\hat{y}_0, y_0\right)\right| \\ &\leq d_\infty\left(\ell, m\right) + r_k(m). \end{aligned}$$

Reversing the roles of ℓ and m yields

$$\left|r_k(\ell) - r_k(m)\right| \leq d_\infty\left(\ell, m\right). \tag{C.1}$$

Because $d(\ell, m) \geq d_\infty(\ell, m)$, if $d(\ell_n, \ell) \to 0$, $d_\infty(\ell_n, \ell) \to 0$. Suppose that $d_\infty(\ell_n, \ell) \to 0$ and pick $\epsilon > 0$. We must show that there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $d(\ell_n, \ell) < \epsilon$. Pick N_1 such that for all $n \geq N_1$, $d_\infty(\ell_n, \ell) < \epsilon/3$. Pick K such that $\sum_{k>K} 1/2^k < \epsilon/3$. Finally, using (C.1), pick N_2 such that for all $n \geq N_2$ and for all $k \leq K$, $f_k(\ell_n, \ell) < \epsilon/(3K)$. For all $n \geq \max\{N_1, N_2\}$,

$$d\left(\ell_n, \ell\right) = d_\infty\left(\ell_n, \ell\right) + \textstyle\sum_{k \leq K} \frac{1}{2^k} f_k\left(\ell_n, \ell\right) + \sum_{k>K} \frac{1}{2^k} f_k\left(\ell_n, \ell\right) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}.$$

Step (c): d-completeness. Let ℓ_n be a d-Cauchy sequence in $\mathscr{C}_{nbcc}(D \times D)$, hence a d_∞-Cauchy sequence in $C(D \times D)$. Because C is d_∞-complete, there exists a $\ell \in C$ such that $d_\infty(\ell_n, \ell) \to 0$. All that is left to show is that $\ell \in \mathscr{C}_{nbcc}$. Because $\ell_n(\hat{y}, y) \geq 0$ for all $(\hat{y}, y)$, and $d_\infty(\ell_n, \ell) \to 0$ implies $\ell_n(\hat{y}, y) \to \ell(\hat{y}, y)$ pointwise, it follows that $\ell(\hat{y}, y) \geq 0$ for all $(\hat{y}, y)$. Suppose that $\ell \notin \mathscr{C}_{nbcc}$, i.e., $\ell(\hat{y}_0, y_0) = 0$ for some $\hat{y}_0 \neq y_0$. Let k_0 be the smallest value of k with $|\hat{y}_0 - y_0| \geq 1/k$. As $\ell_n(\hat{y}_0, y_0) \to 0$, it follows that $r_k(\ell_n) \to_n 0$ for all $k \geq k_0$; in fact, $\sup_{k \geq k_0} r_k(\ell_n) = r_{k_0}(\ell_n) \to_n 0$. Therefore, for any fixed integer $n \in \mathbb{N}$ there exists $J \in \mathbb{N}$ so that $f_k(\ell_n, \ell_{n+j}) = 1$ for all $j \geq J$ and $k \geq k_0$. Hence, for all j large enough, $d(\ell_n, \ell_{n+j}) \geq \sum_{k \geq k_0} 1/2^k f_k(\ell_n, \ell_{n+j}) \geq 1/2^{k_0}$, contradicting ℓ_n being a d-Cauchy sequence. ∎

APPENDIX D: Remaining Proofs

Proof of Proposition 2.3. (⇐) Suppose that $Br\,(p \mid F, \ell) = Br\left(p \mid F, \ell'\right)$ for all compact $F \subset D$ and $p \in \Delta(F)$. For any $p \in \Delta(D)$, $F = \text{supp}(p)$ is a compact set, so that $\ell \sim_{Br} \ell'$.

(⇒) Now suppose that there exist some compact $F \subset D$ and $p \in \Delta(F)$ such that $Br\,(p \mid F, \ell) \neq Br\left(p \mid F, \ell'\right)$. Then, by Lemma B.4, there is an open set $\mathcal{P} \subset \Delta(F)$ such that $Br\,(p \mid F, \ell) \cap Br\,\left(p \mid F, \ell'\right) = \varnothing$ for all $p \in \mathcal{P}$. Pick any $p \in \mathcal{P}$ and any $q \in \Delta(F)$ with support F. For $\alpha > 0$ sufficiently small, $p_\alpha := \alpha q + (1-\alpha)p$ is also contained in $\mathcal{P}$, and its support is also F. Hence, $Br\,(p_\alpha \mid \mathrm{supp}(p_\alpha), \ell) \neq Br\,\left(p_\alpha \mid \mathrm{supp}(p_\alpha), \ell'\right)$. ■

Proof of Theorem 1(i). We first show that for finite D, $\mathscr{C}_{nbcc} \setminus \mathscr{G}$ is Lebesgue-negligible in $\mathscr{C}_{nbcc}$. Given $\#D = M$, we show in particular that the closure of $\mathscr{C}_{nbcc} \setminus \mathscr{G}$ has Lebesgue measure 0 as a subset of (the negative orthant of) $\mathbb{R}^{M^2-M}$. If $M = 2$ then $\mathscr{C}_{nbcc} \setminus \mathscr{G}$ is empty. Let $M \geq 3$, and pick an arbitrary three-point subset $F = \{y_1, y_2, y_3\}$ from D. Restricted to $F \times F$, any $u \in \mathscr{C}_{nbcc}$ can be represented by six positive numbers, a through f, ordered clockwise as

$\hat{y}\downarrow$			
y_1	0	a	b
y_2	f	0	c
y_3	e	d	0
$y \rightarrow$	y_1	y_2	y_3

.

According to Definition 3.3, if ℓ fails to be in $\mathscr{G}$, then there must be a $y_i \in F$ and $p \in \Delta(F)$ with $p(y_i) = 0$ such that $Br\,(p \mid F, \ell) = F$. Suppose, for the sake of concreteness, that $y_i = y_2$, so that $p = (\alpha, 0, (1-\alpha))$ for some $\alpha \in (0,1)$. Note that $Br\,(p \mid F, \ell) = F$ iff $b(1-\alpha) = f\alpha + c(1-\alpha) = e\alpha$, equivalently, iff

$$\begin{bmatrix} 0 & (1-\alpha) & (\alpha-1) & 0 & 0 & -\alpha \\ 0 & 0 & (1-\alpha) & 0 & -\alpha & \alpha \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{D.1}$$

For each $\alpha \in [0,1]$, let $S_\alpha(y_2)$ be the set of $(a,b,c,d,e,f) \in \mathbb{R}^6$ satisfying (D.1). Because the 2×6 matrix is of full row rank for all α, each $S_\alpha(y_2)$ is a four-dimensional linear subspace of $\mathbb{R}^6$. Because α smoothly parametrizes the $S_\alpha(y_2)$, $S(y_2) := \cup_\alpha S_\alpha(y_2)$ is a closed manifold of dimension at most 5. Because $\mathscr{C}_{nbcc}$ is convex and has nonempty interior in $\mathbb{R}^6$, $S(y_2) \cap \mathscr{C}_{nbcc}$ is negligible in $\mathscr{C}_{nbcc}$, i.e., has Lebesgue null closure. This argument, repeated two more times, covers the cases where $y_i = y_1$ and $y_i = y_3$. Finally, the result follows as there are only finitely many subsets of D that are of size 3, and a finite union of negligible sets is negligible.

We now show that for infinite compact D, $\mathscr{C}_{nbcc} \setminus \mathscr{G}$ is totally small. Let D' be an arbitrary, countable dense subset of D (which exists because D is compact). There is a countable class of three-point sets, $F \subset D'$. For each F, consider the three five-dimensional closed manifolds $S(y_1)$, $S(y_2)$, $S(y_3)$ constructed in the last step. The set of ℓ in $\mathscr{C}_{nbcc}$ such that $\ell_{|F \times F}$ falls in $S(y_1) \cup S(y_2) \cup S(y_3)$, and hence fails the conditions in Definition 3.3, is finitely shy and Baire small ($\ell_{|F \times F}$ is six-dimensional if ℓ varies freely). By Lemma C.1, $\mathscr{C}_{nbcc}$ is topologically complete, implying that the countable union of totally small sets is totally small. ■

Proof of Theorem 1(ii). Suppose that the first claim in part (i) has already been shown, i.e., for any $\ell, m \in \mathscr{G}$, $\ell \not\sim_{aff} m$, it is true that $\ell \not\sim_{Br} m$. Then, by Proposition 2.3 and Lemma B.4, there exist $F \subset D$, F compact, and $\mathcal{P} \subset \Delta(F)$, $\mathcal{P}$ open, such that $Br\,(p \mid F, \ell) \cap Br\,(p \mid F, m) = \varnothing$ for all $p \in \mathcal{P}$. Fix any $p \in \mathcal{P}$ and $\epsilon > 0$. We must show that for some $p' \in B^{\rho}_{\epsilon}(p)$, $Br\left(p' \mid \text{supp}(p'), \ell\right) \cap Br\left(p' \mid \text{supp}(p'), m\right) = \varnothing$. To construct such p', pick any $q \in \Delta(F)$ with full support F and some $\alpha \in (0,1)$ small enough such that $p' := \alpha q + (1-\alpha)p$ is contained in $\mathcal{P}$ and $\rho(p, p') < \epsilon$. Therefore, $Br\left(p' \mid \ell\right)$ and $Br\left(p' \mid m\right)$ are disjoint, and the proof is complete.

We will now show that if $\ell, m \in \mathscr{G}(D \times D)$, $[\ell \sim_{Br} m]$ implies $\ell \sim_{aff} m$. The outline of the proof is as follows:

(I) Show the result for $\#D = 3$;

(II) use induction to show the result for $\#D = M < \infty$;

(III) use continuity and denseness of D' in D to show the result when D is a general compact set.

Part (I): #D=3. To keep the notation simple, let $D = \{1,2,3\}$. Each $\ell \in \mathscr{C}_{nbcc}(D \times D)$ can be represented as six positive numbers, a through f, ordered clockwise as

$$\begin{array}{c|ccc} \hat{y}\downarrow & & & \\ \hline 1 & 0 & a & b \\ 2 & f & 0 & c \\ 3 & e & d & 0 \\ \hline y\to & 1 & 2 & 3 \end{array}.$$

As discussed in comment 3 after Example 3.3, each $\ell \in \mathscr{C}_{nbcc}(D \times D)$ can be classified according to whether $Br\,(p \mid D, \ell) = D$ can happen for some $p \in \Delta(D)$. There are three mutually exclusive cases:

Case 0. There is an $i \in D$ such that if $p(i) = 0$ then $Br\,(p \mid D, \ell) = \{1,2,3\}$. As this case is ruled out in the definition of $\mathscr{G}$, we need only consider the next two cases.

Case 1. There exists $p = (p(1), p(2), p(3))$, $p(i) > 0$, $i = 1,2,3$, such that $Br\,(p \mid D, \ell) = \{1,2,3\}$.

Case 2. There is no $p \in \Delta(D)$ such that $Br\,(p \mid D, \ell) = D$. Equivalently, there is an $i \in D$ such that for all α in some nonempty open interval $(r,s) \subset (0,1)$, if $p(i) = 0$, $p(j) = \alpha$, and $p(k) = 1-\alpha$, then $Br\,(p \mid D, \ell) = \{i\}$ and, further, $Br\,(p \mid D, \ell) = \{i, j\}$ when $\alpha = s$ and $Br\,(p \mid D, \ell) = \{i, k\}$ when $\alpha = r$. This case is depicted in Figure 2.

Case 1. In this case there exist distributions satisfying

$$\begin{aligned} p &= (p_1, p_2, 0) \quad \text{with } Br\,(p \mid D, \ell) = \{1,2\}, \\ q &= (q_1, 0, q_2) \quad \text{with } Br\,(q \mid D, \ell) = \{1,3\}, \\ r &= (r_1, r_2, r_3) \quad \text{with } Br\,(r \mid D, \ell) = \{1,2,3\}, \\ s &= (0, s_1, s_2) \quad \text{with } Br\,(s \mid D, \ell) = \{2,3\}, \end{aligned}$$

where p_1, p_2, q_1, q_2, etc., are all strictly positive. We will show that the indifference conditions implicit in these best response sets determine ℓ up to a multiplicative constant. If $m \in \mathscr{C}_{nbcc}$ is another loss function with $\ell \sim_{Br} m$ then by Proposition 2.3, $Br\,(\pi \mid D, \ell) = Br\,(\pi \mid D, m)$ for $\pi = p, q, r, s$. Therefore, m must be a scalar multiple of ℓ, i.e., $\ell \sim_{aff} m$.

Let us normalize a to 1. Combining this normalization with the five equalities that come from the indifference conditions for p, q, r, and s (there are two conditions associated with r), we obtain the following six linear equations in the six unknowns, a, b, c, d, e, and f:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & -p_1 \\ 0 & q_2 & 0 & 0 & -q_1 & 0 \\ r_2 & r_3 & 0 & -r_2 & -r_1 & 0 \\ 0 & 0 & r_3 & -r_2 & -r_1 & r_1 \\ 0 & 0 & s_2 & -s_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{normalization} \\ Br\,(p \mid D,\ell) = \{1,2\} \\ Br\,(q \mid D,\ell) = \{1,3\} \\ Br\,(r \mid D,\ell) = \{1,2,3\} \\ Br\,(r \mid D,\ell) = \{1,2,3\} \\ Br\,(s \mid D,\ell) = \{2,3\} \end{array}. \tag{D.2}$$

We will show that the determinant of the 6 × 6 coefficient matrix is nonzero, meaning that there is exactly one *normalized* $\ell \in \mathscr{C}_{nbcc}$ with the best response sets determined by p, q, r, and s. To do this, we first expand into cofactors along the top row, which has only one nonzero entry, 1. In the remaining 5 × 5 matrix, we again expand into cofactors along the top row, which has only one nonzero entry, $-p_1$. Thus, we arrive at needing to show that

$$\det \begin{bmatrix} q_2 & 0 & 0 & -q_1 \\ r_3 & 0 & -r_2 & -r_1 \\ 0 & r_3 & -r_2 & -r_1 \\ 0 & s_2 & -s_1 & 0 \end{bmatrix} = q_2 \det \begin{bmatrix} 0 & -r_2 & -r_1 \\ r_3 & -r_2 & -r_1 \\ s_2 & -s_1 & 0 \end{bmatrix} - r_3 \det \begin{bmatrix} 0 & 0 & -q_1 \\ r_3 & -r_2 & -r_1 \\ s_2 & -s_1 & 0 \end{bmatrix} \neq 0.$$

After expanding the 3 × 3 matrices, this is $q_2 r_1 s_1 r_3 + r_3 q_1 r_2 s_2 - q_1 s_1 r_3^2$. Because $r_3 > 0$, we take it out as a common factor so that we need to show $q_2 r_1 s_1 + q_1 r_2 s_2 - q_1 s_1 r_3 \neq 0$. In this last expression, replace q_2 with $(1 - q_1)$ and s_2 with $(1 - s_1)$ and rearrange, arriving at needing to show $r_1 s_1 (1 - q_1) + r_2 q_1 (1 - s_1) + r_3 q_1 s_1 \neq 0$. Because each term in this sum is strictly positive, this is indeed the case.

Case 2. For concreteness, suppose that Assumption 1 binds on edge E_2 of the unit simplex. There are probability distributions such that

$$\begin{aligned} &p = (p_1, p_2, 0) && \text{with } Br\,(p \mid D,\ell) = \{1,2\}, \\ &q = (q_1, 0, q_2) && \text{with } Br\,(q \mid D,\ell) = \{1,2\}, \\ &t = (t_1, 0, t_2) && \text{with } Br\,(t \mid D,\ell) = \{2\} \quad \text{and} \quad Br\,(t \mid \operatorname{supp}(t),\ell) = Br\,(t \mid \ell) = \{1,3\}, \\ &r = (r_1, 0, r_2) && \text{with } Br\,(r \mid D,\ell) = \{2,3\}, \quad \text{and} \\ &s = (0, s_1, s_2) && \text{with } Br\,(s \mid D,\ell) = \{2,3\}, \end{aligned}$$

where p_1, p_2, q_1, q_2, etc., are all strictly positive. The strategy of proof is the same as in Case 1; i.e., we will show that the indifference conditions implicit in these best response sets determine ℓ up to scale.

Again, we normalize a to 1. Combining this normalization with the five equalities that come from the indifference conditions for p, q, r, s, and t, we have the following six linear equations in the six unknowns, a, b, c, d, e, and f:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & -p_1 \\ 0 & q_2 & -q_2 & 0 & 0 & -q_1 \\ 0 & t_2 & 0 & 0 & -t_1 & 0 \\ 0 & 0 & r_2 & 0 & -r_1 & r_1 \\ 0 & 0 & s_2 & -s_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{normalization} \\ Br\,(p \mid D,\ell) = \{1,2\} \\ Br\,(q \mid D,\ell) = \{1,2\} \\ Br\,(t \mid \operatorname{supp}(t),\ell) = \{1,3\} \\ Br\,(r \mid D,\ell) = \{2,3\} \\ Br\,(s \mid D,\ell) = \{2,3\} \end{array} \tag{D.3}$$

Again, we will show that the determinant of the 6 × 6 coefficient matrix is nonzero. We first expand into cofactors along the top row, which has only one nonzero entry, 1. In the remaining 5×5 matrix, we expand into cofactors along the third column, which has only one nonzero entry, $-s_1$. In the remaining 4×4 matrix, we expand along the top row, which has only one nonzero entry, $-p_1$. The remaining 3×3 matrix is

$$\begin{bmatrix} q_2 & -q_2 & 0 \\ t_2 & 0 & -t_1 \\ 0 & r_2 & -r_1 \end{bmatrix} \quad \text{which has determinant } q_2 \left[\begin{vmatrix} 0 & -t_1 \\ r_2 & -r_1 \end{vmatrix} + \begin{vmatrix} t_2 & -t_1 \\ 0 & -r_1 \end{vmatrix} \right] = q_2 \left[t_1 r_2 - t_2 r_1 \right].$$

Combining all of this, the determinant of the 6×6 matrix is $\kappa[t_1 r_2 - t_2 r_1]$, where $\kappa = -s_1 p_2 q_2 < 0$. The term $[t_1 r_2 - t_2 r_1]$ can be rewritten as $[t_1(1-r_1) - (1-t_1)r_1] = [t_1 - r_1]$. Because $t \neq r$, we know that $t_1 \neq r_1$.

Part (II): Induction on #*D*. The inductive hypothesis is that Theorem 1(i) holds for $\#D \leq M$, where M is some fixed integer greater than or equal to three. The inductive step is to show that the theorem also holds for $\#D = M+1$.

Let $D \subset \mathbb{R}$ with $\#D = M+1$ and suppose that for $\ell, m \in \mathscr{G}(D \times D)$, $\ell \sim_{Br} m$. To keep the notation simple, put $D = \{1, 2, \dots, M, M+1\}$. It follows immediately from Proposition 2.3 that $\ell \sim_{Br} m$ implies $\ell_{|F \times F} \sim_{Br} m_{|F \times F}$ for $F \subset D$.

Let $F_1 = \{1, \dots, M\}$ and $F_2 = \{2, \dots, M+1\}$. As $\ell_{|F_1 \times F_1} \sim_{Br} m_{|F_1 \times F_1}$, by the inductive hypothesis there exists a unique $r_1 > 0$ such that $\ell_{|F_1 \times F_1} = r_1 \cdot m_{|F_1 \times F_1}$. Also, as $\ell_{|F_2 \times F_2} \sim_{Br} m_{|F_2 \times F_2}$, by the inductive hypothesis there exists unique $r_2 > 0$ such that $\ell_{|F_2 \times F_2} = r_2 \cdot m_{|F_2 \times F_2}$. By considering the common part of $F_1 \times F_1$ and $F_2 \times F_2$, which includes points at which ℓ and m are both nonzero, we must have $r_1 = r_2 := r$. Therefore, $\ell(\hat{y}, y) = r \cdot m(\hat{y}, y)$ must hold for all $(\hat{y}, y) \in D \times D$ except possibly at the points $(1, M+1)$, $(M+1, 1)$; see Table 1. Replacing F_2 with $F_3 = \{1, 3, 4, \dots, M, M+1\} \subset D$ and repeating the arguments above shows that $\ell(\hat{y}, y) = r \cdot m(\hat{y}, y)$ holds even at these points.

TABLE 1. $D \times D$

	$(1,1)$	$(1,2)$	...	$(1,M)$	$(1,M+1)$	
$F_1 \times F_1$	$(2,1)$	$(2,2)$	...	$(2,M)$	$(2,M+1)$	
	$\vdots$	$\vdots$		$\vdots$	$\vdots$	
	$(M,1)$	$(M,2)$	...	(M,M)	$(M,M+1)$	$F_2 \times F_2$
	$(M+1,1)$	$(M+1,2)$	...	$(M+1,M)$	$(M+1,M+1)$	

Part (III): Compact *D*. Suppose that for $\ell, m \in \mathscr{G}(D \times D)$, $\ell \sim_{Br} m$. The previous two steps have shown that there exists a unique $r > 0$ such that for all finite $F \subset D'$, $\ell_{|F \times F} = r \cdot m_{|F \times F}$. This implies that $\ell_{|D' \times D'} = r \cdot m_{|D' \times D'}$. Because $D' \times D'$ is dense in $D \times D$ and ℓ and m are continuous, $\ell = r \cdot m$. ∎

Proof of Theorem 1(iii). Letting $F_t = \text{supp}(p_t)$, consider a dense sequence $\{(p_t, F_t)\}$ in $\Delta(D) \times \mathcal{K}(D)$. Fix $\ell^\circ \in \mathscr{G}$. By Proposition 2.3 and Lemma B.4, for any $\ell \in \mathscr{G}$, $\ell \nsim_{aff} \ell^\circ$, there exist $F' \subset D$, F' compact, and $p' \in \Delta(F')$ such that $Br\left(p' \mid F', \ell^\circ\right) \cap Br\left(p' \mid F', \ell\right) = \varnothing$. As both best response sets are compact, this implies $\epsilon := d_H(Br\left(p' \mid F', \ell^\circ\right), Br\left(p' \mid F', \ell\right)) > 0$. Because the mapping $(p, F) \mapsto Br(p \mid F, \ell)$ is upper hemicontinuous, there exists $\delta > 0$ such that $\rho(p, p') < \delta$, $d_H(F, F') < \delta$ implies $Br(p \mid F, \ell^\circ) \subset [Br\left(p' \mid F', \ell^\circ\right)]^{\epsilon/2}$ and $Br(p \mid F, \ell) \subset [Br\left(p' \mid F', \ell\right)]^{\epsilon/2}$.

As $\{(p_t, F_t)\}$ is dense, there exists some t such that $\rho(p_t, p') < \delta$ and $d_H(F_t, F') < \delta$. Hence, $Br\,(p_t \mid F_t, \ell^\circ) \cap Br\,(p_t \mid F_t, \ell) = \varnothing$. As ℓ was arbitrary (apart from $\ell \not\sim_{aff} \ell^\circ$), ℓ° is identified up to scale as $t \to \infty$. ■

Proof of Proposition 3.1. Let $\{R_n\}_{n\in\mathbb{N}}$ be an independent and identically distributed sequence of random variables, defined on a probability space $(\Omega, \mathcal{F}, P)$, with the distribution of each R_n having a continuous, strictly positive Lebesgue density on D. We will show that for any $\ell \in \mathscr{C}_{nbcc}(D \times D)$ that is nowhere a piecewise affine function of itself or satisfies each $\ell(\cdot, y)$ being strictly convex, there exists a probability 1 set of ω such that the dense set $\{R_n(\omega)\}$ serves as the D' in Definition 3.3. That is, for each three-point set $F \subset D'$ and $p \in \partial\Delta(F)$, it follows that $Br\,(p \mid F, \ell) \neq F$, where $\partial\Delta(F)$ denotes the boundary of the unit simplex.

Step 1. Because the R_n have a strictly positive density, there is an $\Omega' \in \mathcal{F}$ with $P(\Omega') = 1$ such that for all $\omega \in \Omega'$ and $n \neq n'$, $R_n(\omega) \neq R_{n'}(\omega)$, and $\{R_n(\omega)\}_{n\in\mathbb{N}}$ is dense in D.

Step 2. Let $\{n_1, n_2, n_3\}$ be one of the countably many subsets of $\mathbb{N}$ containing three distinct points, and condition on $R_{n_1} = y_1$ and $R_{n_2} = y_2$, $y_1 \neq y_2$. By *nbcc*, the unique probability on $\{y_1, y_2\}$ making y_1 and y_2 indifferent as forecasts is $\alpha\delta_{y_1} + (1-\alpha)\delta_{y_2}$ where $\alpha = \ell(y_1, y_2)/(\ell(y_1, y_2) + \ell(y_2, y_1)) \in (0, 1)$. Letting $\kappa = \alpha\ell(y_2, y_1) = (1-\alpha)\ell(y_1, y_2)$, both ℓ being nowhere a piecewise affine function of itself and each $\ell(\cdot, y)$ being strictly convex implies that

$$P\left(\left\{\omega : \alpha\ell\left(R_{n_3}, y_1\right) + (1-\alpha)\,\ell\left(R_{n_3}, y_2\right) = \kappa\right\}\right) = 0 \tag{D.4}$$

because R_{n_3} has a density with respect to Lebesgue measure. Because we conditioned on arbitrary $y_1 \neq y_2$, there is a probability 1 set of ω, call it $\Omega(n_1, n_2, n_3)$, for which there exists no $p \in \partial\Delta(F)$ with $Br\,(p \mid \ell, F) = F$ where $F = \{R_{n_1}, R_{n_2}, R_{n_3}\}$. Define $\Omega'' = \bigcap \Omega(n_1, n_2, n_3)$ where the intersection is taken over three-point subsets of $\mathbb{N}$ so that $P(\Omega' \cap \Omega'') = 1$.

Step 3. For all ω in $\Omega' \cap \Omega''$, and for all three-point subsets, $F = \{y_1, y_2, y_3\}$ of the dense set $D' = \{R_n(\omega)\}_{n\in\mathbb{N}}$, there is no $p \in \partial\Delta(F)$ with $Br\,(p \mid F, \ell) = F$. ■

Proof of Proposition 3.2. Without loss of generality, we set $D = [0, 1]$. Considering the loss function $\ell(\hat{y}, y) = (y - \hat{y})^2$, we know that $\delta_1 \in \mathcal{P}$ because the only p with 1 solving $\min_{\hat{y}\in\operatorname{supp}(p)} \int \ell(\hat{y}, y)\, p(dy)$ is $p = \delta_1$. Similarly, $\delta_0 \in \mathcal{P}$. Suppose that there exists $y_\circ \in (0, 1)$ such that $\delta_{y_\circ} \notin \mathcal{P}$. We will then construct a loss function $\ell \in \mathscr{G}$ such that $y_\circ$ does not belong to $Br\,(p \mid \ell)$ for any p in $\mathcal{P}$. This, however, contradicts the definition of $\mathcal{P}$.

As $\mathcal{P}$ is closed, there exists some $\epsilon > 0$ such that $\rho(\delta_{y_\circ}, \mathcal{P}) > \epsilon$. The definition of the Prokhorov metric then implies $p(B_\epsilon(y_\circ)) < 1 - \epsilon$ for all $p \in \mathcal{P}$. Further, because $\delta_0, \delta_1 \in \mathcal{P}$, decreasing ϵ if necessary, $B_\epsilon(y_\circ) = (y_\circ - \epsilon, y_\circ + \epsilon) \subset (0, 1)$. By continuous interpolation, there exists a loss function ℓ with the properties

(a) $\ell(1, y) = e^{r|y-1|} - 1$ for some r satisfying $0 < r < \log(1 + \epsilon^2)$.

(b) $\ell(y_\circ, y) = e^{s|y-y_\circ|} - 1$ for some $s > \frac{1}{\epsilon}\log(1 + \epsilon)$.

(c) ℓ is almost nowhere a piecewise affine function of itself and hence is in $\mathscr{G}$.

The choice of r ensures that $\ell(1,y) < \epsilon^2$ for $y \in D$. The choice of s ensures that $\ell(y_\circ, y) > \epsilon$ for $|y - y_\circ| > \epsilon$. Pick an arbitrary $p \in \mathcal{P}$. As established previously $p(B_\epsilon(y_\circ)) < 1 - \epsilon$ so that $p(D \setminus B_\epsilon(y_\circ)) \geq \epsilon$. In particular,

$$\int \ell\,(y_\circ, y)\; p(dy) = \int_{B_\epsilon(y_\circ)} \ell\,(y_\circ, y)\; p(dy) + \int_{D \setminus B_\epsilon(y_\circ)} \ell\,(y_\circ, y)\; p(dy) > \epsilon^2.$$

On the other hand, $\int \ell(1,y) p(dy) < \epsilon^2$ so that $y_\circ$ cannot be an optimal forecast. ■