Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Unrestricted and controlled identification of loss functions: Possibility and impossibility results

Robert P. Lieli [a,*], Maxwell B. Stinchcombe [b], Viola M. Grolmusz [a]

[a] *Department of Economics, Central European University, Budapest, Hungary*
[b] *Department of Economics, University of Texas at Austin, United States*

ARTICLE INFO

Keywords:
Loss functions
Bregman loss functions
GPL loss functions
Osband's principle
Identification
Point forecasts

ABSTRACT

The property that the conditional mean is the unrestricted optimal forecast characterizes the Bregman class of loss functions, while the property that the α-quantile is the unrestricted optimal forecast characterizes the generalized α-piecewise linear (α-GPL) class. However, in settings where the forecaster's choice of forecasts is limited to the support of the predictive distribution, different Bregman losses lead to different forecasts. This is not true for the α-GPL class: the failure of identification is more fundamental. Motivated by these examples, we state simple conditions that can be used to ascertain whether loss functions that are consistent for the same statistical functional become identifiable when off-support forecasts are disallowed. We also study the identifying power of unrestricted forecasts within the class of smooth, convex loss functions. For any such loss ℓ, the set of losses that are consistent for the same statistical functional as ℓ is a tiny subset of this class in a precise mathematical sense. Finally, we illustrate the identification problem that is posed by the non-uniqueness of consistent losses for the moment-based loss function estimation methods proposed in the literature.

1. Introduction

1.1. Overview

Under the assumption that point forecasts are constructed so as to minimize expected loss, the mean of the conditional distribution of the target variable is the unrestricted optimal forecast for any Bregman loss. (In statistical parlance, Bregman losses are *consistent for the mean*.) The converse of this statement is also true: if the conditional mean is the optimal forecast under a given loss function for a sufficiently rich set of distributions, then that loss function must belong to the Bregman class. These loss functions go back to Bregman (1967) and Savage (1971); more recently, Banerjee, Guo, and Wang (2005), Gneiting (2011a) and Patton (2011, 2016) have used Bregman losses to make various points about the construction and evaluation of point forecasts.

There are also other loss functions for which the optimal point forecast is given by a well-known statistical functional other than the mean. In the case of asymmetric absolute loss, the optimal forecast is a fixed quantile of the predictive distribution, and the quantile is determined by the marginal losses for positive vs. negative forecast errors. Each asymmetric absolute loss function for which the α-quantile is the optimal forecast belongs to a larger class of generalized α-piecewise linear (α-GPL) loss functions, a class that is characterized by the property that each member of the class induces the same α-quantile as the optimal forecast (see Saerens, 2000, for the characterization result; and Gneiting, 2011b; Komunjer, 2005; Lieli & Stinchcombe, 2013, for identification issues).

The Bregman class and the α-GPL classes pose a challenge for the literature that is concerned, either directly or indirectly, with recovering loss functions from forecasts,

[*] Corresponding author.
E-mail addresses: lielir@ceu.edu (R.P. Lieli), maxwell.stinchcombe@austin.utexas.edu (M.B. Stinchcombe), Grolmusz_Viola@phd.ceu.edu (V.M. Grolmusz).

https://doi.org/10.1016/j.ijforecast.2018.11.007
0169-2070/

realizations, and relevant covariates (e.g. Capistran, 2008; Elliott, Komunjer, & Timmermann, 2005, 2008; Patton & Timmermann, 2007). Problems arise because, even if the researcher were to know the predictive conditional distribution used by the forecaster, members of the Bregmann class and/or members of one of the α-GPL classes seem observationally equivalent. Furthermore, the statistical literature on the elicitability of functionals (e.g. Fissler, 2017; Fissler & Ziegel, 2016; Gneiting, 2011b; Steinwart, Pasin, Williamson, & Zhang, 2014) indicates that this sort of identification problem is rather widespread, and is not limited to the mean or the quantiles of a distribution.

This paper shows that it may be optimal for forecasters who are characterized by different Bregman loss functions to make different forecasts when they are required to make a choice from a limited set of options; e.g., they must provide a buy, hold, or sell recommendation for a stock, rather than a continuous price forecast. However, the same is not true for α-GPL forecasters. We know from Lieli and Stinchcombe (2013; henceforth LS) that the latter case is exceptional: the class of loss functions for which identification (up to scale) is not possible even under forecast restrictions is tiny in a precise mathematical sense. However, what about the identifying power of unrestricted forecasts? We show here that, within the class of smooth, convex loss functions defined over a compact domain, unrestricted forecasts identify each such loss up to a tiny (but non-trivial) equivalence class. Thus, while from a theoretical standpoint unrestricted forecasts have considerable power to distinguish between (convex) losses, in practice these equivalence classes still present an identification problem. In theory, observing forecasts made subject to restrictions can resolve some of this ambiguity.

1.2. Technical contributions

This paper employs and extends the identification theory in LS in order to distinguish among various degrees of observational equivalence, and to point out that complete observational equivalence is quite rare. Throughout, we work with continuous loss functions $\ell(\hat{y}, y)$ defined on a compact domain with the property that the loss is zero if the forecast $\hat{y}$ coincides with the realization y, and is strictly positive otherwise. Thus, we assume that the forecaster's unique best prediction in the face of certainty that a specific y will happen is to predict that particular y.

Our contributions are as follows:

(i) We show that different Bregman losses are distinguishable if off-support forecasts are excluded from consideration. This is because the mean of a distribution need not be in the support of a distribution, and different Bregman losses will prescribe different on-support forecasts as replacements. Thus, Bregman loss functions are at least potentially distinguishable, say, in controlled, experimental settings, where one can vary the predictive distribution, including its support, to an arbitrary extent.

(ii) In contrast, we prove that the α-GPL losses remain observationally equivalent even if off-support forecasts are excluded. This happens because the α-quantiles are always (at least partly) on-support, rendering this extra constraint ineffective.

(iii) We show that the "observational-equivalence-even-under-restrictions" property of the α-GPL class is highly exceptional. Based on LS, we reiterate the point that the set of loss functions can be partitioned into a "large" generic class of losses that are distinguishable when off-support forecasts are ruled out, and a "tiny" non-generic class of losses that are not.[1] We then make the theory of LS more accessible by stating simple conditions, which do not appear in *ibid.*, that can be used to ascertain whether a loss function belongs to the generic potentially identified class.

(iv) Finally, we also formally describe the identifying power of unrestricted forecasts within the class of convex, twice continuously differentiable loss functions. For any such loss function ℓ, the set of losses that are consistent for the same statistical functional as ℓ makes up at most a tiny subset of this class. This result is non-trivial, because Osband's principle (Gneiting, 2011a; Steinwart et al., 2014) implies that the tiny equivalence class of loss functions that are consistent for the same functional as ℓ contains materially different loss functions (rather than just scalar multiples of ℓ).

1.3. Related literature

Our paper has many points of contact with recent research on the elicitability of statistical functionals (Ehm, Gneiting, Jordan, & Krueger, 2016; Fissler, 2017; Fissler & Ziegel, 2016; Gneiting, 2011a; Steinwart et al., 2014). A central question in that literature is: given a statistical functional that represents some numerical property of a distribution, does there exist a loss function that is consistent for that functional, i.e., for which that functional is the optimal forecast? Then, the recoverability of loss functions from unrestricted forecasts is related to the further question of whether the consistent loss function is unique. As was established by Steinwart et al. (2014), for example, the answer in very general settings is *no*: given a loss function ℓ_0 that satisfies some (mild) restrictions, one can use Osband's principle (Fissler & Ziegel, 2016; Gneiting, 2011a; Steinwart et al., 2014) to generate all other loss functions that are consistent for the same functional as ℓ_0 (see Corollary 9 of Steinwart et al., 2014).

The contributions explained in points (i) to (iv) above fit into the context of this literature as follows. Items (i) and (ii) make the novel point that loss functions that are consistent for the same functional (such as Bregman losses) may still be distinguishable in controlled or restricted forecasting environments where off-support forecasts are disallowed. However, this restriction still does not identify different α-GPL losses. As was noted under item (iii), we provide general — and practically convenient — conditions for checking whether the on-support restriction can distinguish between losses that are consistent for the same

[1] The adjectives "large" and "tiny" have precise technical meanings. In particular, the latter combines a topological and a measure theoretic notion of what it means for a subset of an infinite-dimensional space to be small. In a finite-dimensional setting, "tiny" implies "Lebesgue measure zero". See Section 4.2 for details.

functional. Finally, item (iv) examines the scope of the non-invertibility of the mapping from loss functions to statistical functionals. While the inverse image of a statistical functional that represents an optimal forecasting rule is not a single point in the class of smooth, convex loss functions, it is still a tiny set in this class. If f is a continuous function from reals to reals but is not invertible, then this result is similar to saying that the non-invertibility cannot be due to f being constant over some interval.

1.4. Practical relevance and further contributions

There are at least two reasons for caring about distinctions between loss functions that yield the same unrestricted optimal forecasts. The first concerns behaviors in settings that differ from the idealized forecasting problem. The second concerns the interpretation of the estimated loss function parameters reported in the literature.

Even when multiple loss functions yield the same unrestricted optimal forecasts, there are settings in which such loss functions lead to different optimal forecaster behaviors. This point is illustrated well by Patton (2016), who shows that different Bregman (α-GPL) loss functions only rank competing forecasts the same way if the forecasts come from correctly specified models for the conditional mean (the conditional α-quantile), the information sets are nested, and the estimation error is negligible. As he puts it, "the presence of misspecified models, parameter estimation error, or nonnested information sets, leads generally to sensitivity [of the ranking] to the choice of (consistent) loss function".

In a couple of influential papers, Elliott et al. (2005, 2008) propose a method for estimating loss functions based on moment conditions that are derived from the first-order condition of the forecaster's problem. They assume that the loss function belongs to a specific parametric family such as the absolute (lin-lin) loss with asymmetry parameter α. The final contribution of this paper, which is less technical but of considerable practical importance, is to demonstrate the problem posed by the non-uniqueness of consistent losses for interpreting estimated loss function parameters. In addition to theoretical arguments based on Osband's principle, we also revisit the original budget forecasting application of Elliott et al. (2005) and show that it is impossible to draw inferences from the estimated α about the shape of the loss function unless one puts *complete* trust in the lin-lin specification over other GPL losses. Nevertheless, our technical results show that α-GPL losses are observationally equivalent even in controlled environments, so one must rely on economic theory to motivate any additional identifying assumptions. One possible way of justifying the lin-lin specification would be to argue that the loss is a function of the forecast error only, but such arguments are typically missing from applications.

1.5. Organization

The paper proceeds as follows. Section 2 defines the Bregman and α-GPL loss functions and reviews the LS identification theory. Section 3 proves the results, making it easier to apply LS, and treats the two classes under consideration. Section 4 studies the identifying power of unrestricted forecasts for smooth, convex losses, and shows that they are identified up to a tiny equivalence class. Section 5 demonstrates that, while such equivalence classes may be small theoretically, they can still cause ambiguities for moment-based loss function estimation methods. Section 6 concludes and states some open questions. Short, simple proofs that provide insights into our results are given in the main text, while the proof of the identification result for smooth convex loss functions is in Appendix A.

2. Formal setup and the LS identification theory

The question studied by LS is: given the point forecasts published by an expected loss minimizing forecaster, and the distributions used in their construction, is it possible to identify the forecaster's loss function nonparametrically? This setup is a best-case scenario, in that the econometrician is assumed to know the predictive distributions.

More specifically, let $D = [a, b] \subset \mathbb{R}$ be a compact interval. The forecaster is endowed with a jointly continuous loss function $\ell(\hat{y}, y)$ defined over $D \times D$, where the first argument is the forecast and the second is the realization. At time t, the forecaster wants to forecast the value of a random variable Y_{t+1}, taking values from D. The conditional distribution of Y_{t+1} given the information available to a forecaster at time t is denoted p_t. The forecaster issues a point forecast $f_t \in F_t$ of Y_{t+1} by minimizing the expected loss,

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

The compact set $F_t \subseteq D$ is the set of allowable forecasts and Br is, mnemonically, the "best response". If $F_t = D$, then the forecast is unrestricted; more generally, F_t might be a smaller set, but it is assumed throughout that $p_t(F_t) = 1$. In general, allowing $F_t \neq D$ greatly enhances the amount of information that is revealed about ℓ.

In addition to joint continuity of the loss functions, we make the normalization $\ell(y, y) = 0$, and, more substantively, assume the following property.

Definition 1. A loss function ℓ exhibits "no bias in case of certainty" (abbreviated as *nbcc*) if $\ell(\hat{y}, y) > 0$ for $\hat{y} \neq y$. The set of (jointly) continuous loss functions with the *nbcc* property is denoted $\mathcal{C}_{nbcc}$.

The terminology is justified by the fact that if Y_{t+1} is known to take on a given value y, then the *unique* unrestricted optimal forecast for any *nbcc* loss function is y. This restriction is satisfied by most commonly-used losses, and can also be motivated by deriving loss functions as the best forecast input into an underlying decision problem, as per Granger and Machina (2006).

We make a particular study of two subclasses of $\mathcal{C}_{nbcc}$, the Bregman loss functions, denoted $\mathcal{L}_{Breg}$, and the α-GPL loss functions, denoted $\mathcal{L}^{\alpha}_{GPL}$. The loss functions in $\mathcal{L}_{Breg}$ are those of the form

$$\ell(\hat{y}, y) = [\phi(y) - \phi(\hat{y})] - \phi'(\hat{y})(y - \hat{y}), \tag{2}$$

where $\phi(\cdot)$ is strictly convex and twice continuously differentiable. For each $\alpha \in (0, 1)$, the loss functions in $\mathcal{L}^{\alpha}_{GPL}$ are those of the form

$$\ell(\hat{y}, y) = [1(y < \hat{y}) - \alpha][\psi(\hat{y}) - \psi(y)], \tag{3}$$

where $\psi(\cdot)$ is any continuous, strictly increasing function.[2] The derivative condition for $\hat{y}$ being a minimum for the Bregman class is

$$\frac{d}{d\hat{y}}\left[\int \Big((\phi(y) - \phi(\hat{y})) - \phi'(\hat{y})(y - \hat{y})\Big)\, dp(y)\right] \tag{4}$$

$$= -\phi'(\hat{y}) - \phi''(\hat{y})(E\,Y - \hat{y}) + \phi'(\hat{y}) = 0. \tag{5}$$

Since $\phi''(\hat{y}) > 0$, this is solved uniquely by $\hat{y} = E\,Y$, and this is independent of ϕ. From similar considerations, any element of the α-quantile for Y minimizes the α-GPL loss functions, and this is independent of ψ.

Definition 2. Let $\Delta(D)$ denote the set of distributions over D. Two loss functions ℓ and ℓ'

- are **unrestrictedly** forecast equivalent if for all $p \in \Delta(D)$, $Br\,(p \mid D, \ell) = Br\,\big(p \mid D, \ell'\big)$, and
- are **completely** forecast equivalent if for all $p \in \Delta(D)$ and all compact $F \subset D$ satisfying $p(F) = 1$, $Br\,(p \mid F, \ell) = Br\,\big(p \mid F, \ell'\big)$.

Forecasters with unrestrictedly equivalent loss functions always deliver the same forecasts when there are no controls on the set of allowable forecasts.[3] For example, any two Bregman losses are unrestrictedly forecast equivalent, but a Bregman loss and a 1/2-GPL loss are not, because they give different (unrestricted) forecasts for skewed distributions. In contrast, forecasters with completely equivalent loss functions deliver the same forecasts even when they are restricted to choose forecasts from the support of the distribution of Y_{t+1}. Trivially, loss functions in $\mathcal{C}_{nbcc}$ that are scalar multiples of each other are completely forecast equivalent. A central aim of this paper is to provide criteria that allow us to distinguish between loss functions that are and are not completely forecast equivalent. We will see that, although both the Bregman losses and the α-GPL losses are unrestrictedly equivalent, the Bregman losses are not completely equivalent, while the α-GPL losses are.

The definition of complete forecast equivalence entails a rather strong and non-standard condition. Ultimately, the theoretical justification for this concept is provided "ex-post" by the results proven by LS. As Bregman losses, GPL losses, and various other examples in LS show, constructing loss functions that are unrestrictedly forecast equivalent is not a particularly difficult task (we will return to this problem in Section 4.1). Then, the question that naturally arises is whether there are additional conditions on the forecaster's environment that could make forecasters with such losses distinguishable. As we will explain below, the restriction involved in the definition of complete forecast equivalence *almost always* does the job, in a precise technical sense.[4]

Furthermore, the concept of complete forecast equivalence is not entirely devoid of practical relevance, even in observational settings. For example, consider a variable $Y \in \{-1, 0, 1\}$ that indicates whether a given stock will underperform, match, or outperform the market over some period. A financial analyst could choose to report a continuous forecast from the $[-1, 1]$ interval or to construct probability forecasts of the events, but it is more customary to issue a direct buy/hold/sell recommendation, which could be interpreted as a point forecast restricted to the support of Y. Alternatively, the financial analyst could report to customers a forecast of the stock's excess return (perhaps with risk-adjustment), but again, it is more common to see "discrete" recommendations instead of such continuous forecasts.

The following definition relates complete forecast equivalence to identifiability.

Definition 3. A set of loss functions $\mathcal{L} \subset \mathcal{C}_{nbcc}$ is *potentially identified* if no two members of $\mathcal{L}$ are completely forecast equivalent unless they are scalar multiples.

If a forecaster's loss function is *known* to belong to a potentially identified class, then eventually it can be distinguished from every other element of the class (i.e., completely recovered) by observing forecasts that are produced in a sufficiently diverse set of environments. Of necessity, this diversity of environments includes sufficient variation in both the conditional distribution of the target variable and the set of allowable forecasts. We will show (in Section 3) that the class of Bregman loss functions is potentially identified, but α-GPL classes are not. Any two α-GPL losses are *completely* forecast equivalent.

The α-GPL classes demonstrate that $\mathcal{C}_{nbcc}$ itself is not potentially identified. However, this problem is not at all widespread — LS show that $\mathcal{C}_{nbcc}$ can be decomposed into a 'tiny' non-generic set of 'bad' loss functions, $\mathcal{B}$, and a 'large' generic set of 'good' loss functions, $\mathcal{G} = \mathcal{C}_{nbcc} \setminus \mathcal{B}$, with the property that the class $\mathcal{G}$ is potentially identified.[5] The definitions of $\mathcal{G}$ and $\mathcal{B}$ are based on the "three point boundary problem".

Definition 4. A loss function $\ell \in \mathcal{C}_{nbcc}$ has a *three-point boundary problem* at the three-point set $F = \{y_1, y_2, y_3\} \subset D$ if $Br\,(p \mid F, \ell) = F$ for some distribution p that satisfies $p(F) = 1$ and $p(\{y_i\}) = 0$ for some $y_i \in F$.

[2] Gneiting (2011a) only assumes the $\phi(\cdot)$ function for the Bregman class to be convex, and $\phi'(\hat{y})$ is any element of the subgradient of $\phi(\cdot)$ at $\hat{y}$. Gneiting (2011b) only assumes the $\psi(\cdot)$ function in the α-GPL classes to be non-decreasing.

[3] The statistical terminology for two loss functions being unrestrictedly forecast equivalent is that they are consistent for the same statistical functional.

[4] As was pointed out by a referee, one could go beyond the complete forecast equivalence concept used here and require forecasters to issue the same forecasts even when some *on-support* forecasts are ruled out. This restriction would give more identifying power, but it would be more difficult to motivate than the concept used here.

[5] The adjective "tiny" refers to a class that is not only small in the topological sense of the Baire category, but also "shy", which is an infinite-dimensional version of being a Lebesgue null set. For interpretations of shyness, see Stinchcombe (2001). We provide formal definitions and further discussion in Section 4.2.

In other words, given three distinct points y_1, y_2 and y_3 in D, if some predictive distribution p puts mass 1 on two of these points, say y_1 and y_3, but the forecaster is indifferent between reporting any of the points y_1, y_2 or y_3 as the forecast, then the underlying loss function has a three point boundary problem. LS then define the set $\mathcal{G}$ as follows.

Definition 5. Let $\mathcal{G}$ denote the collection of loss functions $\ell \in \mathcal{C}_{nbcc}$ for which there exists some dense $D' \subset D$ with the property that ℓ has no three point boundary problem at any three point subset of D' (the set D' may depend on ℓ).

Thus, loss functions in $\mathcal{G}$ can be "freed" from any three point boundary problems by restricting them to a suitable dense subset of D. Theorem 1 of LS shows that $\mathcal{G}$ is potentially identified and that $\mathcal{B} := \mathcal{C}_{nbcc} \setminus \mathcal{G}$ is "tiny". Example 3.3 and the subsequent discussion in the same paper explain how this condition is related to the failure of identification. Definitions 4 and 5 are both rather abstract; Section 3 provides a more practical formulation of the three point boundary problem and an easy-to-check sufficient condition for a given loss function to belong to $\mathcal{G}$.

3. Controlled identification results

Our first goal is to show that the class of Bregman loss functions is potentially identified, $\mathcal{L}_{Breg} \subset \mathcal{G}$, while no α-GPL class is, $\mathcal{L}^{\alpha}_{GPL} \subset \mathcal{B}$. To this end, we begin by showing that this is plausible by way of restricted forecast examples. We then provide two propositions that make it easier to decide whether or not a given loss function $\ell \in \mathcal{C}_{nbcc}$ belongs to the set $\mathcal{G}$. These results do not appear explicitly in LS. We end this section with our result that characterizes the identifying power of unrestricted forecasts for smooth, strictly convex loss functions.

3.1. Plausibility

Patton (2016) considers the parametric family of Bregman losses

$$\ell(\hat{y}, y; a) = \frac{2}{a^2}(e^{ay} - e^{a\hat{y}}) - \frac{2}{a}e^{a\hat{y}}(y - \hat{y}), \quad a \neq 0. \tag{6}$$

Suppose that these losses are used in forecasting a binary variable Y with $supp(Y) = \{0, 1\}$. If we set $D = F = [0, 1]$, then the unrestricted optimal forecast is $p(1) = E\,Y$ for all values of $a \neq 0$, where $p(1)$ is the (conditional) probability that $Y = 1$. As this holds for all $a \neq 0$, it is not possible to identify the forecaster's loss function using their unrestricted forecast.

In contrast, if the forecaster is restricted to forecast in $F = \{0, 1\}$, then the optimal forecast is $\hat{y} = 1$ if

$$p(1) > c_a = \frac{\ell(1, 0; a)}{\ell(1, 0; a) + \ell(0, 1; a)} = \frac{1}{1 - e^{-a}} - \frac{1}{a}; \tag{7}$$

it is either 0 or 1 if equality holds, and it is $\hat{y} = 0$ if the inequality is reversed. It can be shown that the cutoff c_a is strictly between 0 and 1 and is an increasing function of a. Thus, for any $a < a'$, the losses $\ell(\hat{y}, y; a)$ and $\ell(\hat{y}, y; a')$ induce different forecasts if $c_a < p(1) < c_{a'}$. This means that the different members of this class *can* be identified in this *controlled* environment, given sufficient variation in $p(1)$.

Suppose now that the loss function ℓ belongs to $\mathcal{L}^{\alpha}_{GPL}$; that is, $\ell(\hat{y}, y) = (1(y < \hat{y}) - \alpha)(\psi(\hat{y}) - \psi(y))$ for a strictly increasing $\psi(\cdot)$. If we set $D = F = [0, 1]$, the unrestricted optimal forecast is $\hat{y} = 0$ if $p(1) < 1 - \alpha$, any number in F if $p(1) = 1 - \alpha$, and $\hat{y} = 1$ if $p(1) > 1 - \alpha$, an answer which does not depend on ψ.[6] In stark contrast to the previous case, the optimal forecast remains independent of ψ if F is restricted to $\{0, 1\}$; the one minor difference is that only 0 or 1 can be reported when $p(1) = 1 - \alpha$. Thus, the losses in $\mathcal{L}^{\alpha}_{GPL}$ are observationally equivalent in this setting as well.

Indeed, for any distribution p and compact F with $p(F) = 1$, the set

$$\arg\min_{\hat{y} \in F} \int (1(y < \hat{y}) - \alpha)(\psi(\hat{y}) - \psi(y))p(dy)$$

consists of the on-support α-quantile(s) of p, and possibly the off-support α-quantiles as in the example above. Restricting F to the support of p (or a somewhat larger set) either has no effect or eliminates the same off-support quantiles for any ψ. Hence, no information about ψ is revealed.

3.2. General results

We begin with preliminaries, then provide and prove results that make it easier to apply the controlled identification theory in LS. Next, we apply the results to the two classes under consideration.

If $F = \{y_1, y_2, y_3\}$ is a three point subset of $D = [a, b]$, then testing for a boundary problem at F involves setting one of $p(y_1) = 0$, $p(y_2) = 0$, and $p(y_3) = 0$ (what can potentially matter is whether the largest, the smallest or the middle point gets zero weight). It is these three possibilities that give rise to the conditions in Eqs. (8), (9), and (10) in the following proposition.

Proposition 1. *A loss function $\ell \in \mathcal{C}_{nbcc}$ has a three point boundary problem at $F = \{y_1, y_2, y_3\} \subset D$ if and only if one of the following conditions is satisfied:*

$$\begin{aligned} g_1(y_1, y_2, y_3) &:= \ell(y_2, y_3)\ell(y_3, y_2) - \ell(y_2, y_3)\ell(y_1, y_2) \\ &\quad - \ell(y_1, y_3)\ell(y_3, y_2) = 0 \qquad (8) \\ g_2(y_1, y_2, y_3) &:= \ell(y_1, y_3)\ell(y_3, y_1) - \ell(y_1, y_3)\ell(y_2, y_1) \\ &\quad - \ell(y_2, y_3)\ell(y_3, y_1) = 0 \qquad (9) \\ g_3(y_1, y_2, y_3) &:= \ell(y_1, y_2)\ell(y_2, y_1) - \ell(y_1, y_2)\ell(y_3, y_1) \\ &\quad - \ell(y_3, y_2)\ell(y_2, y_1) = 0. \qquad (10) \end{aligned}$$

Proof. Let Y be a random variable with distribution p and suppose that $p(Y \in F) = 1$ for some $F = \{y_1, y_2, y_3\}$. By definition, ℓ has a three point boundary problem at F if and only if one of the following three conditions hold:

- $p(y_1) := p(Y = y_1) = 0$ and $y_1, y_2, y_3 \in Br\,(p \mid F, \ell)$, or
- $p(y_2) := p(Y = y_2) = 0$ and $y_1, y_2, y_3 \in Br\,(p \mid F, \ell)$, or
- $p(y_3) := p(Y = y_3) = 0$ and $y_1, y_2, y_3 \in Br\,(p \mid F, \ell)$.

[6] A number x is an α-quantile of the distribution p if $p((-\infty, x)) \leq \alpha$ and $p((\infty, x]) \geq \alpha$.

We show that the second case, $p(y_2) = 0$, is equivalent to Eq. (9); i.e., $g_2(y_1, y_2, y_3) = 0$. The arguments for the remaining cases are parallel.

Suppose that $p(y_1) + p(y_3) = 1$. The forecaster is indifferent between forecasting $Y = y_1$ and $Y = y_3$ iff the two forecasts yield the same expected loss, that is, iff

$$\ell(y_1, y_3)p(y_3) = \ell(y_3, y_1)p(y_1), \tag{11}$$

where we also use the fact that $\ell(y, y) = 0$. Similarly, the forecaster is indifferent between forecasting $Y = y_1$ and $Y = y_2$ iff

$$\ell(y_1, y_3)p(y_3) = \ell(y_2, y_1)p(y_1) + \ell(y_2, y_3)p(y_3). \tag{12}$$

Using $p(y_1) + p(y_3) = 1$, one can solve for $p(y_1)$ and $p(y_3)$ using Eq. (11) and substitute the resulting expressions into Eq. (12). This yields Eq. (9). □

The next result uses Proposition 1 to construct an easier-to-check sufficient condition for a loss function to belong to the identified set $\mathcal{G}$. Let D_0^3 denote the set of triples with pairwise distinct coordinates in the interior of D^3. Each point in D_0^3 defines a three-point set $F = \{y_1, y_2, y_3\}$ that is to be tested for the boundary problem as follows.

Proposition 2. *If the sets $g_j^{-1}(0) = \{(y_1, y_2, y_3) \in D_0^3 : g_j(y_1, y_2, y_3) = 0\}$, $j = 1, 2, 3$, have Lebesgue measure zero in $\mathbb{R}^3$, then the loss function ℓ belongs to $\mathcal{G}$.*

Proof. Let U_i, $i = 1, 2, \ldots$ be i.i.d. uniform random variables with support D. Then, with probability one, $\{U_i\}_{i=1}^{\infty}$ is dense in D; U_n, U_m and U_k are distinct for any distinct n, m, k, and $g_1(U_n, U_m, U_k) \neq 0$, $g_2(U_n, U_m, U_k) \neq 0$, $g_3(U_n, U_m, U_k) \neq 0$. Hence, for almost all realizations of the sequence $\{U_i\}_{i=1}^{\infty}$, the dense set $D' := \{U_i\}_{i=1}^{\infty} \subset D$ will satisfy the requirement that ℓ has no three-point boundary problem at any $\{y_1, y_2, y_3\} \subset D'$. □

The following lemma, which is a special case of Theorem 1 of Ponomarev (1987), states the conditions under which the inverse image of a measure zero set is a measure zero set. A simple corollary of this lemma is particularly useful for verifying the conditions of Proposition 2.

Lemma 1. *Let O be an open subset of $\mathbb{R}^n$ and $f : O \to \mathbb{R}$ be a continuously differentiable function on O. If $\nabla f(x) \neq 0$ almost everywhere in O, then $f^{-1}(A)$ has Lebsegue measure zero in $\mathbb{R}^n$ whenever A has Lebesgue measure zero in $\mathbb{R}$.*

Corollary 1. *If the set $\{x \in O : \frac{\partial}{\partial x_i} f(x) = 0\}$ has Lebesgue measure zero in $\mathbb{R}^n$ for any $i \in \{1, \ldots, n\}$, then $f^{-1}(A)$ has measure zero in $\mathbb{R}^n$ whenever A has measure zero in $\mathbb{R}$.*

3.3. Controlled identification is possible for Bregman losses

We can now show formally that Bregman losses are potentially identified.

Proposition 3. *For any Bregman loss ℓ, the sets $g_j^{-1}(0)$, $j = 1, 2, 3$ have Lebesgue measure zero in $\mathbb{R}^3$, and hence $\ell \in \mathcal{G}$.*

Proof. We will apply Corollary 1 with $O = D_0^3$, $f = g_2$, $i = 2$, and $A = \{0\}$. That is, we need to show that the set of triples in D_0^3 for which

$$\begin{aligned}\frac{\partial}{\partial y_2} g_2(y_1, y_2, y_3) = {} & -\ell(y_1, y_3)\ell_{\hat{y}}(y_2, y_1) \\ & -\ell(y_3, y_1)\ell_{\hat{y}}(y_2, y_3) = 0\end{aligned}$$

is a measure zero subset of $\mathbb{R}^3$, where $\ell_{\hat{y}}$ denotes the partial derivative of ℓ with respect to its first argument. By the definition of Bregman loss, $\ell_{\hat{y}}(\hat{y}, y) = -\phi''(\hat{y})(y - \hat{y})$, so that

$$\begin{aligned}\frac{\partial}{\partial y_2} g_2(y_1, y_2, y_3) = {} & \ell(y_1, y_3)\phi''(y_2)(y_1 - y_2) \\ & +\ell(y_3, y_1)\phi''(y_2)(y_3 - y_2).\end{aligned}$$

As $\phi''(y_2) > 0$,

$$\begin{aligned}\frac{\partial}{\partial y_2} g_2(y_1, y_2, y_3) = 0 \Leftrightarrow {} & \ell(y_1, y_3)y_1 \\ & -[\ell(y_1, y_3) + \ell(y_3, y_1)]y_2 \\ & +\ell(y_3, y_1)y_3 = 0.\end{aligned} \tag{13}$$

Let $h(y_1, y_2, y_3) = \ell(y_1, y_3)y_1 - [\ell(y_1, y_3) + \ell(y_3, y_1)]y_2 + \ell(y_3, y_1)y_3$. Based on the equivalency stated in Eq. (13), we can complete the proof by showing that the zeros of h in D_0^3 are a measure zero set. Applying Corollary 1 again, it is sufficient to argue that $\frac{\partial}{\partial y_2} h(y_1, y_2, y_3)$ is nonzero almost everywhere in D_0^3. Indeed,

$$\frac{\partial}{\partial y_2} h(y_1, y_2, y_3) = -[\ell(y_1, y_3) + \ell(y_3, y_1)] < 0,$$

given that y_1 and y_3 are distinct. Hence, $\frac{\partial}{\partial y_2} h$ has no zeros in D_0^3. The test functions g_1 and g_3 can be dealt with using analogous arguments. □

The implication of Proposition 3 is that any two Bregman losses can be told apart by restricting the set of allowable forecasts to the support of the predictive distribution.

3.4. Controlled identification is not possible for α-GPL losses

Using Proposition 1 only, we will now show that GPL losses have a boundary problem with *any* three-point set chosen from (the interior of) D. This fact not only violates the sufficient condition for $\ell \in \mathcal{G}$ stated in Proposition 2, but, as we will explain below, actually implies that GPL losses are not in $\mathcal{G}$.

Proposition 4. *For any α-GPL loss ℓ, the set $\cup_{j=1}^3 g_j^{-1}(0)$ is equal to D_0^3.*

Proof. Pick a point in D_0^3 with, say, $y_1 < y_2 < y_3$. The definition of GPL loss implies: $\ell(y_1, y_3) = -\alpha[\psi(y_1) - \psi(y_3)]$, $\ell(y_3, y_1) = (1 - \alpha)[\psi(y_3) - \psi(y_1)]$, $\ell(y_2, y_1) = (1 - \alpha)[\psi(y_2) - \psi(y_1)]$ and $\ell(y_2, y_3) = -\alpha[\psi(y_2) - \psi(y_3)]$. Substituting into the equation $g_2(y_1, y_2, y_3) = 0$ and dividing through by $\alpha(1 - \alpha)$ yields

$$\begin{aligned}&[\psi(y_3) - \psi(y_1)]^2 + [\psi(y_1) - \psi(y_3)][\psi(y_2) - \psi(y_1)] \\ &\quad +[\psi(y_2) - \psi(y_3)][\psi(y_3) - \psi(y_1)] = 0.\end{aligned}$$

As ψ is strictly increasing, $\psi(y_3) > \psi(y_1)$, and therefore the equation simplifies further to

$$[\psi(y_3) - \psi(y_1)] - [\psi(y_2) - \psi(y_1)] + [\psi(y_2) - \psi(y_3)] = 0,$$

which holds for all y_1, y_2, y_3. Any other ordering of the coordinates will set some g_j identically to zero. □

The result $\cup_{j=1}^{3} g_j^{-1}(0) = D_0^3$ means that, no matter how one picks three distinct points $F = \{y_1, y_2, y_3\}$ from the interior of D, the underlying loss function has a three-point boundary problem at F. Clearly, there is no dense subset D' of D with the property that three-point subsets of D' are free from the boundary problem. Thus, by the definition of $\mathcal{G}$, GPL losses do not belong in $\mathcal{G}$.

4. Unrestricted identification results

The Bregman losses show that restricted forecasts (controlled environments) can reveal strictly more information about loss functions than unrestricted forecasts. We will now focus on the class of smooth, convex loss functions and characterize the identifying power of unrestricted forecasts within this class in a general and mathematically precise way.

The setup is as follows. Let $\mathcal{D}^2_{\text{conv}}$ be defined as the set of loss functions in $\mathcal{C}_{nbcc}$ such that (i) for each $y \in D = [a, b]$, the second derivative of ℓ w.r.t. $\hat{y}$ exists on some open interval containing D; (ii) $\ell_{\hat{y}\hat{y}}(\hat{y}, y)$ is jointly continuous over $D \times D$; and (iii) $\ell_{\hat{y}\hat{y}}(\hat{y}, y) > 0$ at all $(\hat{y}, y)$ pairs in $D \times D$. For each loss function $\ell \in \mathcal{D}^2_{\text{conv}}$, we define

$$\mathcal{B}(\ell) = \Big\{\ell^\dagger \in \mathcal{C}_{nbcc} : Br\left(\ell^\dagger \mid D, p\right) = Br\left(\ell \mid D, p\right) \text{ for all } p \in \Delta(D)\Big\}.$$

This is the set of loss functions in $\mathcal{C}_{nbcc}$ that are consistent for the same statistical functional as ℓ.

We are concerned with the size or richness of the sets $\mathcal{B}(\ell)$ and $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$. We first observe that $\mathcal{B}(\ell)$ does not just contain multiples of ℓ; in general, $\mathcal{B}(\ell)$ is a nontrivial, infinite-dimensional equivalence class that is akin to a Bregman or GPL class. Nevertheless, $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ is a tiny subset of $\mathcal{D}^2_{\text{conv}}$ in a precise mathematical sense that will be described below. Loosely speaking, this means that smooth, convex losses that are consistent for the same functional are rare; given any one of them, unrestricted forecasts can distinguish it from "almost every" other loss in $\mathcal{D}^2_{\text{conv}}$.

4.1. Generating unrestrictedly forecast equivalent losses

The fact that $\mathcal{B}(\ell)$ does not just contain scalar multiples follows from *Osband's principle* (after Osband, 1985).[7] Given an initial loss function $\ell(\hat{y}, y)$, the idea is to generate unrestrictedly forecast equivalent losses via the integral

$$\ell^\dagger(\hat{y}, y) := \int_y^{\hat{y}} \ell_{\hat{y}}(t, y) w(t) dt, \tag{14}$$

where $w(t) > 0$ is a continuously differentiable weight function. When $\ell(\cdot, y)$ is convex, so is $\int \ell(\cdot, y)\, dp(y)$, but $\ell^\dagger(\cdot, y)$ need not be, unless $w(\cdot)$ satisfies further conditions (c.f. Fissler, 2017, Ch. 4). The next set of arguments shows that for any $w(\cdot)$, the first-order condition $\frac{d}{d\hat{y}} \int \ell^\dagger(\cdot, y)\, dp(y) = 0$ has the same unique solution as the corresponding condition for ℓ, and the second-order condition for a minimum is also satisfied at the unique solution.

- Integrating $\ell^\dagger$ with respect to a distribution $p \in \Delta(D)$, taking the derivative with respect to $\hat{y}$, and then interchanging the two operations (see Lemma 2(ii) in Appendix A.1) yields
$$\frac{d}{d\hat{y}} \int \ell^\dagger(\hat{y}, y) dp(y) = \int \ell^\dagger_{\hat{y}}(\hat{y}, y) dp(y), \tag{15}$$
and therefore,
$$\frac{d}{d\hat{y}} \int \ell^\dagger(\hat{y}, y)\, dp(y) = w(\hat{y}) \int \ell_{\hat{y}}(\hat{y}, y)\, dp(y). \tag{16}$$
- Pulling the derivative out of the integral on the r.h.s. (see again Lemma 2(ii) in Appendix A.1) gives
$$\frac{d}{d\hat{y}} \int \ell^\dagger(\hat{y}, y)\, dp(y) = w(\hat{y}) \frac{d}{d\hat{y}} \int \ell(\hat{y}, y)\, dp(y), \tag{17}$$
for all $\hat{y} \in (a, b)$. Eq. (17) shows that the first-order condition for $Br\left(p \mid D, \ell^\dagger\right)$ is uniquely satisfied at the unique solution to the first-order condition for $Br\left(p \mid D, \ell\right)$.
- We check that the second-order condition holds by taking the derivative in Eq. (17):
$$\frac{d}{d\hat{y}} w(\hat{y}) \int \ell_{\hat{y}}(\hat{y}, y)\, dp(y) = w'(\hat{y}) \int \ell_{\hat{y}}(\hat{y}, y)\, dp(y) + w(\hat{y}) \int \ell_{\hat{y},\hat{y}}(\hat{y}, y)\, dp(y).$$
When the first order condition holds, the first term is zero, and the second term is strictly positive.

By varying the weight function $w(\cdot)$, one can generate an entire class of forecast equivalent loss functions to ℓ, a class in which the first-order conditions determine the unrestricted optimal forecast uniquely.[8]

4.2. Unrestrictedly forecast equivalent losses are rare in $\mathcal{D}^2_{\text{conv}}$

Our contribution to the theory of losses that are consistent for the same statistical functional is to show that, despite the freedom in choosing $w(\cdot)$, the set $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ is still a tiny subset of $\mathcal{D}^2_{\text{conv}}$ according to a number of technical but easy-to-interpret criteria. Before stating our result, we provide a brief discussion of what it means for a subset of an infinite-dimensional function space to be "tiny". In particular, our definition combines two ideas. One is purely topological and has a long history in real analysis; the other is a measure theoretic notion that has been developed more recently and that generalizes the properties of Lebesgue null sets to infinite-dimensional settings, where the Lebsegue measure is not defined.

[7] We thank a referee for pointing this out.

[8] For example, starting from the square loss $(\hat{y} - y)^2$, Bregman losses can be generated by integrating $2w(t)(t - y)$. Working in a more general setting, Steinwart et al. (2014) demonstrate that all order-sensitive unrestrictedly forecast equivalent loss functions can be generated this way.

Definition 6 (*a*). If X is a (topologically) complete metric space, $B \subset X$ is *Baire small* if it is closed and has an empty interior or if it is the countable union of such sets. (b) If X is a (topologically) complete convex metric space, $B \subset X$ is 1-*shy* if there exists a 1-dimensional line segment L in X such that $\mu_L(B+x) = 0$ for all $x \in X$, where μ_L is the uniform distribution on L.

Remarks.

(i) The standard terminology for Baire small sets in real analysis is the rather nondescript "first category" set. A non-trivial textbook example is as follows. Let $C = C[0, 1]$ denote the set of continuous functions on $[0, 1]$ that are endowed with the sup metric. Let $D(x) = \{f \in C : f \text{ is differentiable at } x\}$. Then, $D = \cup_{x\in[0,1]} D(x)$ is Baire small in C. For the proof, see for example Ross (2013, Ch. 7). Hence, nowhere differentiable functions are the "typical" elements of C.

(ii) A shy set extends the concept of a Lebesgue null set; if some property holds for all $x \in X$ outside of a shy set, then that property can be said to hold "almost everywhere". The general definition of shyness is more involved; see Hunt, Sauer, and Yorke (1992) or Anderson and Zame (2001) for detailed treatments. The definition of 1-shyness given here is based on sufficient conditions for shyness.

(iii) Throughout the paper, we say that a set is "tiny" if it is both Baire small and shy, and that it is "large" or "generic" if its complement is tiny.

(iv) To illustrate the perhaps lesser-known concept of shyness in "action", we will show that, for a given x, the set $D(x)$ defined in item (i) above is 1-shy. A one-dimensional line segment in C is parameterized as $L(\beta) = \beta f_1 + (1-\beta)f_2$, where f_1 and f_2 are fixed elements of C and $\beta \in [0, 1]$. We need to show that there is a choice of f_1, f_2 such that the sets $A_r = \{\beta : L(\beta) \in D(x) + r\}$ have Lebesgue measure zero for all $r \in C$. Pick f_1 and f_2 such that $f_1 - f_2 \notin D(x)$. Suppose that there exist two distinct scalars $\beta \neq \beta'$ such that $\beta f_1 + (1-\beta)f_2 = d + r$ and $\beta' f_1 + (1-\beta')f_2 = d' + r$ for some $d, d' \in D(x)$. Taking the difference of the two equations and rearranging yields

$$f_1 - f_2 = \frac{1}{\beta - \beta'}(d - d'). \tag{18}$$

The function on the r.h.s. of Eq. (18) is differentiable at x, while the function on the l.h.s. is not, by virtue of the choices of f_1 and f_2. This is a contradiction, implying $\#A_r = 0$ or $\#A_r = 1$. In either case A_r has Lebesgue measure zero.

(v) The set D defined in item (i) above is shy in the general sense, but it is not 1-shy; see Hunt et al. (1992, Proposition 4). This shows that 1-shyness is a particularly stringent condition for smallness.

Relating Definition 6 to our setup, $\mathcal{D}^2_{\text{conv}}$ plays the role of the ambient space X, and $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ plays the role of the subset B that is of interest. We state the following result.

Proposition 5. *For each $\ell \in \mathcal{D}^2_{\text{conv}}$, the set $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ is 1-shy and Baire small (in $\mathcal{D}^2_{\text{conv}}$). That is, for each $\ell \in \mathcal{D}^2_{\text{conv}}$, the set $\mathcal{G}^2(\ell) := \mathcal{D}^2_{\text{conv}} \setminus \mathcal{B}(\ell)$ is a generic subset of $\mathcal{D}^2_{\text{conv}}$, and has the property that for every $\ell^\dagger \in \mathcal{G}^2(\ell)$, there exists $p \in \Delta(D)$ such that $Br\left(p \mid D, \ell^\dagger\right) \neq Br\left(p \mid D, \ell\right)$.*

Remarks.

(i) The proof of Proposition 5 is somewhat detailed, and is given in the Appendix A.

(ii) By Lemma B.4 of LS, if $Br\left(p \mid D, \ell^\dagger\right) \neq Br\left(p \mid D, \ell\right)$ for some p, there exists a non-empty open set of distributions containing p for which the same is true.

(iii) $\mathcal{B}(\ell)$ contains many non-convex losses. To see why, take $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ and define $\ell^\dagger$ as in Eq. (14). Then, by direct calculations,

$$\ell^\dagger_{\hat{y}\hat{y}}(\hat{y}, y) = w(\hat{y}) + (\hat{y} - y)w'(\hat{y}),$$

where the second term can be negative. In fact, if one allows $D = \mathbb{R}$, then for any choice of $w(\cdot)$ and $\hat{y}$ such that $w'(\hat{y}) \neq 0$, the inequality $\ell^\dagger_{\hat{y}\hat{y}}(\hat{y}, y) < 0$ holds for all y above or below some threshold. Hence, scalar multiples of the square loss are the only globally convex Bregman losses over $\mathbb{R}$, a point that follows from the more general results of Fissler (2017). If $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ is a one-dimensional set, it is of course shy and Baire small in $\mathcal{D}^2_{\text{conv}}$. However, D is a compact interval in our setup, which means that $\ell^\dagger_{\hat{y}\hat{y}} > 0$ over $D \times D$ for a multitude of weight functions $w(\cdot)$, even when ℓ is a square loss. Hence, $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ is not simply a one-dimensional set, and Proposition 5 is a thoroughly non-trivial result.

(iv) Together with Osband's principle, Proposition 5 offers novel insights into the structure of the set $\mathcal{D}^2_{\text{conv}}$. In particular, unrestricted forecast equivalence, viewed as an equivalence relation among loss functions, partitions $\mathcal{D}^2_{\text{conv}}$ into an uncountably large number of tiny equivalence classes. Given any loss in $\mathcal{D}^2_{\text{conv}}$, "almost all" other loss functions give rise to a different statistical functional from the unrestricted optimal forecast. Nevertheless, from a practical standpoint, $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ may still be a diverse set of losses.

(v) Proposition 5 is stronger than the identification results of LS in that we do not need to have control over which sets the forecasters choose, but is also weaker in two senses. First, the generic set $\mathcal{G}^2(\ell)$ depends on ℓ, and we do not know whether there exists a single generic $\mathcal{E} \subset \mathcal{D}^2_{\text{conv}}$ with the property that for each pair $\ell, \ell' \in \mathcal{E}$, $\ell \neq r \cdot \ell'$, $r > 0$, the two losses are consistent for different functionals. Second, we are restricting attention to loss functions that are both smooth and convex — a class that includes some, but not all Bregman loss functions on D, and that contains none of the α-GPL loss functions.

5. Implications of Osband's principle for moment-based loss function estimation

We illustrate the practical relevance of the abstract identification problems addressed in this paper by drawing out the consequences of Osband's principle for the moment-based loss function estimation method proposed by Elliott et al. (2005; henceforth EKT). EKT's framework achieves identification by assuming very specific classes of loss functions: asymmetric absolute loss (lin-lin) or asymmetric quadratic loss (quad-quad), with the asymmetry parameter α requiring estimation. Nevertheless, Osband's principle implies that, given α, there are many other losses that explain the data equally well (in the case of lin-lin, for example, other α-GPL losses). Thus, any conclusions that may be drawn about the shape of the underlying loss from the estimate of α are extremely sensitive to the specific parametrization used. We now make this argument formal and provide a partly empirical example.

5.1. The ambiguity of the EKT moment conditions

Let $\hat{Y}_{t+1}$ denote the forecast of the target variable Y_{t+1} that is made by the forecaster at time t. The information set that is available to the forecaster and on which the forecast is based is denoted by Ω_t; the predictive distribution p_t is the conditional distribution of Y_{t+1} given Ω_t. The loss function possessed by the forecaster is modeled by the econometrician as $\ell(\hat{y}, y; \theta)$, where θ is a finite-dimensional vector of parameters. If the model for the loss function is specified correctly, there exists some value θ_0 of the parameters such that the observed forecast $\hat{Y}_{t+1}$ minimizes the expectation of $\ell(\hat{y}, y; \theta_0)$ with respect to p_t, and hence satisfies the corresponding FOC:

$$\frac{d}{d\hat{y}} \int \ell(\hat{Y}_{t+1}, y; \theta_0) dp_t(y) = \int \ell_{\hat{y}}(\hat{Y}_{t+1}, y; \theta_0) dp_t(y) = 0. \tag{19}$$

While p_t itself is not observed by the econometrician, Eq. (19) and the law of iterated expectations imply

$$E[W_t \ell_{\hat{y}}(\hat{Y}_{t+1}, Y_{t+1}; \theta_0)] = 0, \tag{20}$$

where W_t is any random vector measurable with respect to Ω_t and for which the expectation is well-defined. Thus, as was put forward by EKT, one can estimate the loss function parameters θ_0 using moment conditions in the form of Eq. (20) without a full knowledge of Ω_t. All that is required is some instrument vector W_t that may plausibly be available to the forecaster at time t.

While the estimation strategy described above cleverly deals with the problem that p_t is not observed by the econometrician, Osband's principle implies that strong additional assumptions are needed to identify the forecaster's loss function. In particular, Eq. (17) shows that, given a suitable weight function $w(\cdot)$, the loss function $\ell^{\dagger}(\hat{y}, y; \theta_0) = \int_y^{\hat{y}} w(t)\ell_{\hat{y}}(t, y; \theta_0)dt$ satisfies

$$\frac{d}{d\hat{y}} \int \ell^{\dagger}(\hat{Y}_{t+1}, y; \theta_0) dp_t(y) = \int w(\hat{Y}_{t+1}) \times \ell_{\hat{y}}(\hat{Y}_{t+1}, y; \theta_0) dp_t(y) = 0.$$

Again, for any Ω_t-measurable random vector W_t, this implies

$$E[W_t \ell^{\dagger}_{\hat{y}}(\hat{Y}_{t+1}, Y_{t+1}; \theta_0)] = E[W_t w(\hat{Y}_{t+1}) \ell_{\hat{y}}(\hat{Y}_{t+1}, Y_{t+1}; \theta_0)] = 0, \tag{21}$$

provided that the expectations exist. The identification problem arises because $W_t w(\hat{Y}_{t+1})$ is also Ω_t-measurable, since $\hat{Y}_{t+1}$ is, of necessity, a function of the information available at time t. Hence, the moment condition in Eq. (21) has two equally valid interpretations: it can be regarded as a consequence of the forecaster's FOC under either the loss $\ell(\hat{y}, y; \theta_0)$ and instrument choice $W_t w(\hat{Y}_{t+1})$ or, alternatively, the loss $\ell^{\dagger}(\hat{y}, y; \theta_0)$ and instrument choice W_t. Even if these moment conditions point-identify θ_0, the loss functions ℓ and $\ell^{\dagger}$ may look very different. A knowledge of θ_0 on its own says very little, if anything, about the shape of the underlying loss.

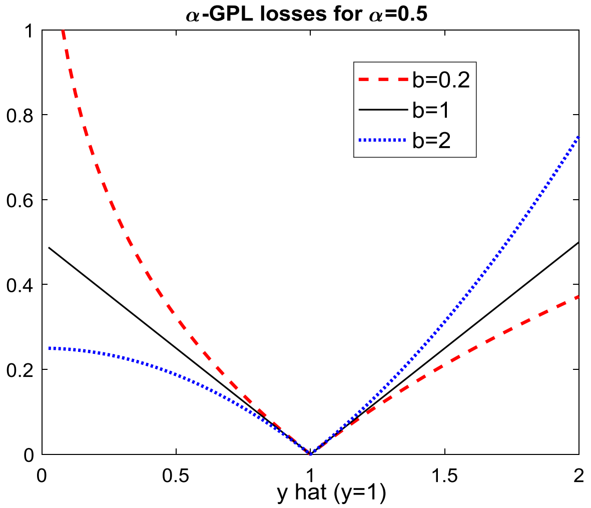

Fig. 1. The loss functions in Eq. (22) for different values of b and $\alpha = 1/2$. The realization y is fixed at 1 and the forecast $\hat{y}$ varies.

5.2. Empirical example

An example will help to reinforce the argument. Let us embed lin-lin losses into a larger class of GPL loss functions, introduced by Patton (2016). In particular, the function $\psi(t) = sgn(t)|t|^b$ is strictly increasing for $b > 0$, so

$$\ell(\hat{y}, y; \alpha) = [1(y - \hat{y} < 0) - \alpha] \cdot [sgn(\hat{y})|\hat{y}|^b - sgn(y)|y|^b], \quad \alpha \in (0, 1), \tag{22}$$

is indeed a collection of GPL losses for any $b > 0$. Setting $b = 1$ corresponds to lin-lin loss, but other values of b give rise to very differently shaped loss functions that may be asymmetric in either direction; see Fig. 1 for an illustration with α fixed at 0.5.

While each loss function exhibited in Fig. 1 has the property that the median of p_t is the optimal point forecast for any distribution p_t, these losses are not equivalent economically. For example, suppose that the forecaster is presented with the following question: "If $Y_{t+1} = 1$, how much would you be willing to pay to avoid a forecast error of size $+1$ versus one of size -1?" Clearly, a forecaster whose loss is given by the dashed line would respond differently from a forecaster whose loss is given by the dotted line.

Table 1
Estimated α parameters for various values of b.

	$b = 0.25$	$b = 0.5$	$b = 1$	$b = 2$	$b = 3$
$\hat{\alpha}$	0.46	0.46	0.45	0.40	0.33
s.e.	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
p-value ($\alpha = 0.5$)	[0.36]	[0.35]	[0.32]	[0.18]	[0.07]

Notes: Based on the IMF's current-year budget deficit forecasts for France. Sample period: 1980–2000; W_t = constant, lagged budget deficit (EKT's instrument 3). The GMM weighting matrix $\hat{S}$ is specified as per EKT. The case $b = 1$ corresponds to the original EKT estimator (their estimate of α is 0.54).

Let us now turn to the moment conditions for estimating α that result from Eq. (22). In this case, Eq. (20) takes the form

$$E\left\{ W_t|\hat{Y}_{t+1}|^{b-1} \times \Big[-\alpha 1(Y_{t+1} > \hat{Y}_{t+1}) + (1-\alpha)1(Y_{t+1} < \hat{Y}_{t+1}) \Big]\right\} = 0. \tag{23}$$

For any given $b > 0$, the generalized method of moments estimator of α that is derived from the sample analog of Eq. (23) is isomorphic to EKT's estimator with the instrument choice $W_t|\hat{Y}_{t+1}|^{b-1}$ (in EKT's setting, $b = 1$). We highlight the ambiguity in interpreting estimates of α by revisiting one of EKT's original applications involving annual budget deficit forecasts published by the IMF for various countries. Using Eq. (23), we re-estimate the α parameter for different values of b, while setting W_t equal to one of the original instruments considered by EKT.[9]

Table 1 shows a small set of result. While $\hat{\alpha}$ varies somewhat as a function of b, the null hypothesis that $\alpha = 0.5$ cannot be rejected formally at the 5% level in any of the cases (albeit the conclusion is borderline for $b = 3$).[10] If, following EKT, one takes lin-lin as the underlying loss, these estimates suggest no significant deviation from symmetry, with the different values of b simply corresponding to different instruments. Nevertheless, an alternative, and *a priori* equally plausible interpretation of Eq. (23) is that the underlying losses belong to the set in Eq. (22), with some value of b being different from 1. Plotting the loss functions that correspond to different $(b, \hat{\alpha})$ pairs in Table 1 would yield a picture that is similar to Fig. 1: the observed budget forecasts and realizations can also be rationalized by loss functions that are asymmetric in either direction!

The conservative interpretation of the α estimates reported in EKT's Table 2 (and our Table 1) is that they simply approximate the quantile of the forecaster's predictive distribution that is used as the point forecast. There is then a diverse class of α-GPL losses that can potentially rationalize this behavior. Hence, any conclusions drawn by EKT about the (a)symmetry of the underlying loss[11] are conditional on *complete* trust in the lin-lin specification. However, our results in Section 3 show that different α-GPL losses are observationally equivalent even in controlled environments, so it requires strong *theoretical* arguments to single out any particular subclass as the appropriate model of forecaster behavior.

A property that makes lin-lin losses special among GPL losses is that the forecaster's loss is a function of the forecast error $y - \hat{y}$ only, independently of the level of $\hat{y}$ or y. All of EKT's statements about symmetry depend critically on this implicit identifying assumption. In general, any loss function estimation exercise that uses the lin-lin (or quad-quad) specification should argue the point that a loss function that depends solely on the forecast error is appropriate for the situation at hand. Another approach that was proposed recently by Schmidt and Katzfuss (2018) is to give up on recovering the loss function altogether, and to focus simply on estimating the quantile or expectile of the predictive distributions that best corresponds to the observed point forecasts.

6. Conclusion

The main result of LS is that a generic subset of $\mathcal{C}_{nbcc}$, namely the set $\mathcal{G}$, is potentially identified. We can frame this statement as a "possibility theorem", saying that, although it may require a tremendous amount of variation in the conditional distributions faced by the forecaster, as well as in the set of allowable forecasts, eventually any loss function in $\mathcal{G}$ can be identified nonparametrically up to scale. While observational data on forecasts are unlikely to incorporate sufficient variation for this result to be of practical application, preference recovery is at least theoretically possible.

By showing that Bregman losses are part of the potentially identified set $\mathcal{G}$, this paper highlights the role of varying the set of allowable forecasts in identifying loss functions. If the forecasts are unrestricted, Bregman losses are a striking example of a very diverse set of loss functions being observationally equivalent. Nevertheless, this equivalence is broken if the predictions of the forecaster must belong to the support of their distributions. On the other hand, the α-GPL classes of loss functions show that even this type of variation may not be sufficient for distinguishing among all possible loss functions, albeit these counterexamples are part of a tiny set within $\mathcal{C}_{nbcc}$. Proposition 2 provides a novel, and applicable, method of checking whether a given loss function belongs to the "good" class of loss functions.

Motivated by these results, we have studied identification by unrestricted forecasts in a general setting. If ℓ is a member of $\mathcal{D}^2_{\text{conv}}$, the set of twice continuously differentiable, convex losses, then unrestricted forecasts distinguish it from all other members of $\mathcal{D}^2_{\text{conv}}$, with the exception of a subset of $\mathcal{D}^2_{\text{conv}}$ that is both shy and Baire

[9] As we do not have access to EKT's original dataset, we collected our own data for the same sample period (1980–2000).

[10] If there are differences between the point estimates that go beyond small sample variation, a plausible explanation is that the family of losses in Eq. (23) is misspecified, in the broader sense that the point forecasts reported by the forecaster do not correspond to a fixed quantile of the underlying predictive distributions. Other possible interpretations include instrument invalidity (W_t is not in the forecaster's information set) or forecaster irrationality.

[11] One example: "[T]he point estimates of α suggest strong asymmetries in the loss function... For some countries they indicate that underpredictions of budget deficits are viewed as up to three times costlier than overpredictions". (EKT, p. 1117).

small. Nevertheless, this exceptional set may still contain losses other than scalar multiples of ℓ, so even though convex losses that are consistent for the same functional are rare in a theoretical sense, practical identification problems can (and do) remain. There are open theoretical questions as well: (i) Is there a *single* generic subset $\mathcal{E} \subset \mathcal{D}^2_{\text{conv}}$ such that no two members of $\mathcal{E}$ are unrestrictedly forecast equivalent (unless scalar multiples)? (ii) Can a single class of distributions smaller than $\Delta(D)$ distinguish each ℓ from every loss in $\mathcal{G}^2(\ell)$, and how large does this collection need to be? (iii) Can the results be extended to classes more general than $\mathcal{D}^2_{\text{conv}}$?

While these identification results have an abstract flavor, we have also argued that Osband's principle, i.e., the fundamental non-uniqueness of consistent losses, has important consequences for applied research that is aimed at estimating a forecaster's loss function from observed data. In particular, the conclusions drawn about the shape of the underlying loss can be very sensitive to the exact parametrization used, and therefore the properties of the adopted parametric model (such as convexity and forecast error dependence) should be carefully motivated based on subject matter theory. The loss function estimation exercises that we are aware of do not address this important point.

Acknowledgments

We thank Andrew Patton for useful comments and for bringing Bregman losses to our attention. Reviews from two anonymous referees led to substantial improvements. All errors are our responsibility.

Appendix A. Proof of Proposition 5

Appendix A.1 provides preliminary results, while Appendix A.2 gives the actual proof. Throughout, we normalize the interval $D = [a, b]$ to $D = [0, 1]$. Let $\Delta(D)$ denote the set of distributions over (subsets of) D; we equip $\Delta(D)$ with the Prokhorov metric. As Proposition 5 is concerned only with unrestricted forecasts, and because the optimal forecast is unique for $\ell \in \mathcal{D}^2_{\text{conv}}$ (see below), it will be convenient to replace the notation $Br\,(p \mid D, \ell)$ with $\hat{y}^*_\ell(p)$. We use various standard notations to denote partial derivatives.

A.1. Preliminary results

We make a number of simple observations about $\mathcal{D}^2_{\text{conv}}$.

Lemma 2. *Loss functions in $\mathcal{D}^2_{\text{conv}}$ have the following properties:*

(i) For all $\ell_1, \ell_2 \in \mathcal{D}^2_{\text{conv}}$ and all $r, s > 0$, $r\ell_1 + s\ell_2 \in \mathcal{D}^2_{\text{conv}}$, i.e., $\mathcal{D}^2_{\text{conv}}$ is a convex cone.

(ii) For every $\ell \in \mathcal{D}^2_{\text{conv}}$ and every $p \in \Delta(D)$,

$$\frac{d^j}{d\hat{y}^j}\int \ell(\hat{y}, y)p(dy) = \int \frac{\partial^j}{\partial \hat{y}^j}\ell(\hat{y}, y)p(dy), \quad j = 1, 2. \tag{24}$$

(iii) For every $\ell \in \mathcal{D}^2_{\text{conv}}$ and every $p \in \Delta(D)$, the solution to $\min_{\hat{y}} \int_D \ell(\hat{y}, y)\, dp(y)$ is a singleton $\hat{y}^*_\ell(p)$.

(iv) For every $\ell \in \mathcal{D}^2_{\text{conv}}$, the functional $p \mapsto \hat{y}^*_\ell(p)$ is continuous over $\Delta([0, 1])$.

(v) For every $\ell \in \mathcal{D}^2_{\text{conv}}$ and every distribution p that puts positive mass on $(0, 1)$, the interior of D, the solution $\hat{y}^*_\ell(p)$ satisfies $0 < \hat{y}^*_\ell(p) < 1$.

Proof. In what follows, let $F(\hat{y}) := \int_D \ell(\hat{y}, y)\, dp(y)$.

(i) Immediate.

(ii) Take $j = 1$ and $\hat{y}_n \to \hat{y}$. The l.h.s. of Eq. (24) is $\lim_n\{[F(\hat{y}_n) - F(\hat{y})]/(\hat{y}_n - \hat{y})\} = \lim_n \int\{[\ell(\hat{y}_n, y) - \ell(\hat{y}, y)]/(\hat{y}_n - \hat{y})\}dp(y)$. By the mean value theorem, the absolute value of the integrand is dominated by $g(y) := \max_{t\in D} |\ell_{\hat{y}}(t, y)|$, where $g(y)$ is continuous, and hence bounded over D. Applying the dominated convergence theorem completes the proof.

(iii) The function $F(\hat{y})$ is strictly convex on $[0, 1]$.

(iv) The function $(\hat{y}, p) \mapsto \int_D \ell(\hat{y}, y)\, dp(y)$ is jointly continuous over $[0, 1] \times \Delta([0, 1])$. The claim then follows from the work of Corbae, Stinchcombe, and Zeman (2009, Theorem 4.2.17) and point (iii).

(v) $\ell_{\hat{y}}(0, y) < 0$ for all $y \in (0, 1]$ because $\ell_{\hat{y}}(y, y) = 0$ and $\ell_{\hat{y}}(\hat{y}, y)$ strictly increases in $\hat{y}$. Similarly, $\ell_{\hat{y}}(1, y) > 0$ for all $y \in [0, 1)$. The claim then follows from the fact that $\frac{d}{d\hat{y}}F(\hat{y}) = \frac{\partial}{\partial \hat{y}}\ell(\hat{y}, 0)p(0) + \int_{(0,1)} \frac{\partial}{\partial \hat{y}}\ell(\hat{y}, y)\, dp(y) + \frac{\partial}{\partial \hat{y}}\ell(\hat{y}, 1)p(1)$. □

We equip $\mathcal{D}^2_{\text{conv}}$ with the following metric. Given a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $(x_1, x_2) \mapsto f(x_1, x_2)$, let $\frac{\partial^\alpha}{\partial x_1^\alpha} f(x_1^\circ, x_2^\circ)$, $\alpha = 1, 2, \ldots$, denote the first, second, etc., partial derivative of f with respect to its first argument at a point (x_1°, x_2°). For $\alpha = 0$, the same notation means f itself. We measure the distance between two functions in $\mathcal{D}^2_{\text{conv}}$ using the (Sobolev) metric

$$d(\ell, \ell') = \sum_{\alpha\in\{0,1,2\}} \left(\max_{(\hat{y},y)\in D\times D} \left|\frac{\partial^\alpha}{\partial \hat{y}^\alpha}\ell(\hat{y}, y) - \frac{\partial^\alpha}{\partial \hat{y}^\alpha}\ell'(\hat{y}, y)\right| \right).$$

Intuitively, if ℓ and ℓ' are close to each other in the Sobolev metric, then in addition to $|\ell(\hat{y}, y) - \ell'(\hat{y}, y)|$ being uniformly small, the distance between the partial derivatives w.r.t. $\hat{y}$ (up to order two) is also uniformly small. A metric space is complete topologically if it has an equivalent metric in which it is complete. With the given metric, $\mathcal{D}^2_{\text{conv}}$ is topologically complete.

For any given loss function ℓ in $\mathcal{D}^2_{\text{conv}}$, the unrestricted optimal forecast $\hat{y}^*_\ell(p)$ is a statistical functional that maps distributions on $[0, 1]$ into $[0, 1]$, and is continuous over $\Delta([0, 1])$. We let $C(\Delta([0, 1]); [0, 1])$ denote the set of such functionals, and equip it with the sup norm. We say that a loss function ℓ is consistent for a given (continuous) functional $\hat{y}(p)$ if the optimal forecast under ℓ is given by $\hat{y}(p)$ for any distribution p.

Hence, the forecaster's problem also defines a mapping from loss functions to functionals, $\gamma : \mathcal{D}^2_{\text{conv}} \to C(\Delta([0, 1]); [0, 1])$, where $\gamma(\ell) = \hat{y}^*_\ell(\cdot)$ gives the statistical functional for which the loss function ℓ is consistent. The mapping $\gamma(\cdot)$ is continuous over $\mathcal{D}^2_{\text{conv}}$; i.e., if a loss $\ell^\dagger$ is close to some given ℓ in Sobolev metric, then the statistical functional $\gamma(\ell^\dagger) = \hat{y}^*_{\ell^\dagger}(\cdot)$ is close to $\gamma(\ell) = \hat{y}^*_\ell(\cdot)$ in sup distance. The proof relies on the $\ell_{\hat{y},\hat{y}}$ of any $\ell \in \mathcal{D}^2_{\text{conv}}$ being uniformly above zero on the compact set $D \times D$.[12]

A.2. Main argument

To prove Proposition 5, we need to show that $\mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$ is 1-shy and Baire small. To make the notation cleaner, set $\mathcal{B}^2(\ell) := \mathcal{B}(\ell) \cap \mathcal{D}^2_{\text{conv}}$. This is the set of loss functions in $\mathcal{D}^2_{\text{conv}}$ that are consistent for the same functional as $\ell \in \mathcal{D}^2_{\text{conv}}$.

$\mathcal{B}^2(\ell)$ **is** 1-**shy**. Fix $\ell \in \mathcal{D}^2_{\text{conv}}$. We first show that for sufficiently small values of $|r| > 0$, both

$$\begin{aligned} &\ell^{\dagger}(\hat{y}, y) := \ell(\hat{y}, y) \cdot (1 + r\hat{y}) \quad \text{and} \\ &\ell^{\ddagger}(\hat{y}, y) := \ell(\hat{y}, y) \cdot (1 - r\hat{y}) \end{aligned} \tag{25}$$

belong to $\mathcal{D}^2_{\text{conv}}$. For any $|r| < 1$, the functions $\ell^{\dagger}$ and $\ell^{\ddagger}$ belong to $\mathcal{C}_{nbcc}$ and are twice continuously differentiable. By direct calculation, $\ell^{\dagger}_{\hat{y}\hat{y}} = \ell_{\hat{y}\hat{y}} + r(\ell + \ell_{\hat{y}} + \hat{y}\ell_{\hat{y}\hat{y}})$, where $\ell_{\hat{y}}$ can be negative. Since $\ell(\cdot,\cdot)$ is twice continuously differentiable over a set that contains the compact domain $D \times D$, the functions ℓ, $\ell_{\hat{y}}$ and $\ell_{\hat{y}\hat{y}}$ are all bounded over $D \times D$, and $\ell_{\hat{y}\hat{y}} > 0$ is bounded away from zero on $D \times D$ by assumption. This implies that $\ell^{\dagger}_{\hat{y}\hat{y}} > 0$ for values of $|r|$ that are sufficiently close to zero. The treatment of $\ell^{\ddagger}$ is parallel. For what follows, fix some $r \neq 0$ satisfying these conditions.

Pick an arbitrary $\ell^{\circ} \in \mathcal{D}^2_{\text{conv}}$ and consider the set of $\beta \in [0, 1]$ such that $\beta\ell^{\dagger} + (1-\beta)\ell^{\ddagger} \in \mathcal{B}^2(\ell) + \ell^{\circ}$. From Definition 6, it is sufficient to show that there is at most a single β in this set. Pick an arbitrary full support p and let $\hat{y}^* = \hat{y}^*_{\ell}(p)$ be the unique solution to the first-order condition for ℓ, i.e.,

$$\int \ell_{\hat{y}}(\hat{y}^*, y)\, dp(y) = 0. \tag{26}$$

If $\beta\ell^{\dagger} + (1-\beta)\ell^{\ddagger}$ belongs to $\mathcal{B}^2(\ell) + \ell^{\circ}$, then the unique optimal forecast under the loss function $-\ell^{\circ} + \beta\ell^{\dagger} + (1-\beta)\ell^{\ddagger}$ is also $\hat{y}^*$. In other words, replacing ℓ with $-\ell^{\circ} + \beta\ell^{\dagger} + (1-\beta)\ell^{\ddagger}$ in Eq. (26) must leave the equality intact, i.e.,

$$\begin{aligned} &-\int \ell^{\circ}_{\hat{y}}(\hat{y}^*, y)\, dp(y) + \beta \int \ell^{\dagger}_{\hat{y}}(\hat{y}^*, y)\, dp(y) \\ &+(1-\beta) \int \ell^{\ddagger}_{\hat{y}}(\hat{y}^*, y)\, dp(y) = 0. \end{aligned} \tag{27}$$

However, using the definitions of $\ell^{\dagger}$ and $\ell^{\ddagger}$ and taking Eq. (26) into account gives

$$\begin{aligned} &\int \ell^{\dagger}_{\hat{y}}(\hat{y}^*, y)\, dp(y) = r \int \ell(\hat{y}^*, y)\, dp(y) \text{ and} \\ &\int \ell^{\ddagger}_{\hat{y}}(\hat{y}^*, y)\, dp(y) = -r \int \ell(\hat{y}^*, y)\, dp(y). \end{aligned} \tag{28}$$

Substituting Eq. (28) into Eq. (27) yields

$$\begin{aligned} &-\int \ell^{\circ}_{\hat{y}}(\hat{y}^*, y)\, dp(y) - r \int \ell(\hat{y}^*, y)\, dp(y) \\ &+2r\beta \int \ell(\hat{y}^*, y)\, dp(y) = 0. \end{aligned}$$

The last equation is linear in β with a strictly positive or negative slope. Therefore, it has at most one solution in $[0, 1]$.

$\mathcal{B}^2(\ell)$ **is small in the sense of Baire.** $\mathcal{B}^2(\ell)$ is closed because it is the inverse image of a point under the continuous mapping $\gamma(\cdot)$. To show that the interior of $\mathcal{B}^2(\ell)$ is empty, pick an arbitrary $\ell' \in \mathcal{B}^2(\ell)$. It is sufficient to show that there are points (losses) in $\mathcal{D}^2_{\text{conv}}$ that are arbitrarily close to ℓ' but are not in $\mathcal{B}^2(\ell)$. Consider the point $\ell' + \delta\ell^{\dagger}$, where $\delta > 0$ and $\ell^{\dagger}$ is defined in Eq. (25) with some sufficiently small $r > 0$. Because $\mathcal{D}^2_{\text{conv}}$ is a convex cone, this loss also belongs to $\mathcal{D}^2_{\text{conv}}$. For any full support p, the optimal forecast $\hat{y}^*_{\ell'}(p)$ is in (0,1) and is equal to $\hat{y}^*_{\ell}(p)$. However, it is not hard to see that the optimum forecast for $\ell' + \delta\ell^{\dagger}$ is strictly smaller than $\hat{y}^*_{\ell'}(p)$ for any $\delta > 0$, and hence, $\ell' + \delta\ell^{\dagger} \notin \mathcal{B}^2(\ell)$. Nevertheless, $\ell' + \delta\ell^{\dagger}$ converges to ℓ' as $\delta \downarrow 0$. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2018.11.007.

References

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

Banerjee, A., Guo, X., & Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions of Information Theory*, *51*, 2664–2669.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, *7*, 200–217.

Capistran, C. (2008). Bias in Federal Reserve inflation forecasts: is the Federal Reserve irrational or just cautious? *Journal of Monetary Economics*, *55*, 1415–1427.

Corbae, D., Stinchcombe, M. B., & Zeman, J. (2009). *An introduction to mathematical analysis for economic theory and econometrics*. Princeton, NJ: Princeton University Press.

Ehm, W., Gneiting, T., Jordan, A., & Krueger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B.*, *78*, 505–562.

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

[12] Take a sequence $\ell_n \to \ell$ and fix p. The optimal forecasts are $\hat{y}^*_n = \hat{y}^*_n(p)$ for ℓ_n and $\hat{y}^* = \hat{y}^*(p)$ for ℓ. Expanding the first-order condition $\int \frac{\partial}{\partial \hat{y}} \ell_n(\hat{y}^*_n, y)\, dp(y) = 0$ around $\hat{y}^*$ gives $\int \frac{\partial}{\partial \hat{y}} \ell_n(\hat{y}^*, y)\, dp(y) + (\hat{y}^*_n - \hat{y}^*) \int \frac{\partial^2}{\partial \hat{y}^2} \ell_n(\tilde{y}^*_n, y)\, dp(y) = 0$, where $\tilde{y}^*_n$ is between y^*_n and $\hat{y}^*$. As $\ell_n \to \ell$ in Sobolev metric, there exists some $\epsilon > 0$ such that $\frac{\partial^2}{\partial \hat{y}^2} \ell_n(\cdot,\cdot) > \epsilon$ for all large enough values of n. Then, $|\hat{y}^*_n(p) - \hat{y}^*(p)|$ is bounded from above by

$$\begin{aligned} \frac{1}{\epsilon}\Big| \int \frac{\partial}{\partial \hat{y}} \ell_n(\hat{y}^*, y)\, dp(y)\Big| &= \frac{1}{\epsilon}\Big| \int \Big(\frac{\partial}{\partial \hat{y}} \ell_n(\hat{y}^*, y) \\ &\quad - \frac{\partial}{\partial \hat{y}} \ell(\hat{y}^*, y)\Big)\, dp(y)\Big| \\ &< \frac{1}{\epsilon} \max_{\hat{y},y} \Big| \frac{\partial}{\partial \hat{y}} \ell_n(\hat{y}, y) \\ &\quad - \frac{\partial}{\partial \hat{y}} \ell(\hat{y}, y)\Big|. \end{aligned}$$

The last upper bound does not depend on p and converges to zero. Hence, $|\hat{y}^*_n(p) - \hat{y}^*(p)|$ converges to zero uniformly in p.

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

Fissler, T. (2017). *On higher order elicitability and some limit theorems on the Poisson and Wiener space* (Ph.D. thesis), University of Bern.

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

Gneiting, T. (2011a). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*, 746–762.

Gneiting, T. (2011b). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*, 197–207.

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

Hunt, B. R., Sauer, T., & Yorke, J. A. (1992). Prevalence: A translation-invariant "almost every" on infinite-dimensional spaces. *Bulletin of the American Mathematical Society*, *27*, 217–238.

Komunjer, I. (2005). Quasi maximum-likelihood estimation for conditional quantiles. *Journal of Econometrics*, *128*, 137–164.

Lieli, R. P., & Stinchcombe, M. B. (2013). On the recoverability of forecasters' preferences. *Econometric Theory*, *29*, 517–544.

Osband, K. H. (1985). *Providing incentives for better cost forecasting* (Ph.D. thesis), Berkeley: University of California, Unpublished.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*, 246–256.

Patton, A. J. (2016). Comparing possibly misspecified forecasts. Working paper, Department of Economics, Duke University.

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

Ponomarev, S. P. (1987). Submersions and preimages of sets of measure zero. *Siberian Mathematical Journal*, *28*, 153–163.

Ross, K. A. (2013). *Undergraduate texts in mathematics*, *Elementary Analysis*. Springer.

Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks*, *11*, 1263–1271.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Schmidt, P., & Katzfuss, M. (2018). Interpretation of point forecasts with unknown directive. Working paper, Heidelberg Institute for Theoretical Studies, Goethe University, Frankfurt.

Steinwart, I., Pasin, C., Williamson, R. C., & Zhang, S. (2014). Elicitation and identification of properties. *JMLR: Workshop and Conference Proceedings*, *35*, 1–45.

Stinchcombe, M. B. (2001). The gap between probability and prevalence: loneliness in vector spaces. *Proceedings of the Americal Mathematical Society*, *129*, 451–457.