Unrestricted and Controlled Identification of Loss Functions: Possibility and Impossibility Results*

Robert P. Lieli† Maxwell B. Stinchcombe‡

January 23, 2018

Abstract

Some classes of loss functions seem difficult to identify from forecaster behavior: the property that the conditional mean is the unrestricted optimal forecast characterizes the Bregman class; the property that the α-quantile is the unrestricted optimal forecast characterizes the generalized α-piecewise linear (α-GPL) loss functions. 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 if 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 consistent for the same statistical functional as ℓ is a tiny subset of this class in a precise mathematical sense.

Keywords: point forecasts, nonparametric identification of loss functions, Bregman loss functions, GPL loss functions.

---

*An early version of this paper carried the title "GPL and Bregman Losses: Identified or Not?". We thank Andrew Patton for useful comments and for bringing Bregman losses to our attention. All errors are our responsibility.

†Department of Economics, Central European University, Budapest. Email: lielir@ceu.edu.

‡Department of Economics, University of Texas at Austin. Email: max.stinchcombe@gmail.com

1 Introduction

1.1 Overview

Under the assumption that point forecasts are constructed 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 always the optimal forecast under a given loss function, then that loss function must belong to the Bregman class. This class of loss functions goes back to Bregman (1967) and Savage (1971), and more recently, Banerjee et al. (2005), Gneiting (2011a) and Patton (2011, 2016) have used Bregman loss functions to make various points about the construction and evaluation of point forecasts.

There are other loss functions for which the optimal point forecast is given by a well-known statistical functional other than the mean. In 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 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, Komunjer 2005, Gneiting 2011b, Lieli and Stinchcombe 2013 for identification issues).

The Bregman class and the α-GPL classes both appear to designate strict boundaries for the literature concerned, directly or indirectly, with recovering loss functions from forecasts, realizations, and relevant covariates (e.g. Capistran 2008, Elliott, Komunjer, and Timmermann 2005, 2008, Patton and Timmermann 2007 in economics; Fissler and Ziegel 2016, Gneiting 2011b, Steinwart et al. 2014 in statistics). 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. It is also unclear whether or not this sort of identification problem is widespread.

We show that forecasters characterized by different Bregman loss functions can optimally make different forecasts when charged with making 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. As far as the question of how prevalent identification problems may be, from Lieli and Stinchcombe (2013; henceforth, LS) we know that the class of loss functions for which identification is not possible even under forecast restrictions is tiny in a precise mathematical sense. Here we show that within the class of smooth, convex loss functions, unrestricted forecasts identify each such loss up to a tiny (but non-trivial) equivalence class.

1.2 Detailed Contributions

This paper employs and extends the identification theory in LS to distinguish between various degrees of observational equivalence, and to point out that complete observational equivalence is quite rare. Throughout, we work with loss functions for which the unique best prediction in the face of certainty that a specific y will happen is to predict that y. This rules out e.g. constant loss functions, a thoroughly trivial case, and it rules out forecasters with an active desire to mislead.

We show the following.

(i) Different Bregman losses are distinguishable if off-support forecasts are excluded from consideration. This happens 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 a replacement. Thus, Bregman loss functions are at least potentially distinguishable, say, in controlled, experimental settings.

(ii) By contrast, 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) The "observational-equivalence-even-under-restrictions" property of 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 potentially identified when off-support forecasts are ruled out, and a "tiny" non-generic class of losses that are not.[1] We then make the theory in LS more accessible by stating simple conditions, not appearing in *ibid.*, that can be used to ascertain whether a loss function belongs to the generic potentially identified class.

(iv) 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 consistent for the same statistical functional as ℓ makes up at most a tiny subset of this class. This results limits the scope of Osband's principle (Gneiting 2011a, Steinwart et al. 2014), which implies that the tiny equivalence class of loss functions defined by the optimal forecast rule under ℓ contains materially different loss functions (rather than just scalar multiples).

1.3 Related Literature

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

The contributions explained in points (i) through (iv) above fit into the context of this literature in the following way. Items (i) and (ii) make the novel point that loss functions

[1] Here "large" and "tiny" have precise mathematical meanings; in particular, "tiny" is the analog of "Lebesgue measure zero" in a finite dimensional setting.

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. Nevertheless, this restriction still does not identify different α-GPL losses. As noted under item (iii), we provide general — and practically convenient — conditions for checking whether the on-support restriction can distinguish between losses consistent for the same 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 representing 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 f 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 Why we Care

There are at least two reasons to care about distinctions between loss functions that yield the same unrestricted optimal forecasts. The first concerns behavior in settings different from the idealized forecasting problem. The second concerns the interpretation and robustness of estimated loss function parameters reported in the literature.

Even when loss functions yield the same unrestricted optimal forecast, there are settings in which such loss functions lead to different optimal forecaster behavior. This point is well illustrated by Patton (2016), who shows that different Bregman (or α-GPL) loss functions rank competing forecasts the same way only if the forecasts come from correctly specified models for the conditional mean (or the conditional α-quantile), the information sets are nested, and 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 an influential paper, Elliott et al. (2005) propose a method for estimating loss functions in a parametric setting. In their application to budget deficit forecasts, they estimate the α parameter associated with asymmetric absolute loss, where $\alpha \in (0, 1)$ gives the quantile of the predictive distribution that is reported as the optimal point forecast. (A small α means

that the marginal cost of a positive error is small.) As shown in their Table 2, the IMF budget forecast for France gives an estimated value of α very close to 0.5 over the sample period, suggesting that the median is the optimal forecast and the forecaster dislikes a larger positive or larger negative error equally.

However, quantiles as optimal forecasts can also be represented by other types of α-GPL losses, for example, the two-parameter family used by Patton (2016), where in addition to α, there is a shape parameter $b > 0$. One could fix some value of b, and apply the Elliott et al. (2005) method to re-estimate α in the empirical example mentioned above. If the reported forecast is indeed the conditional median, one would still expect to obtain an estimated value of α close to 0.5. Nevertheless, depending on the value of b, the shape of the corresponding loss function can vary wildly, for example, it can exhibit strong asymmetry in either direction; see Figure 4 in Patton (2016). Clearly, there is a need then to distinguish between various α-GPL losses, which, as suggested by our results, is a tall order.[2]

1.5 Organization

The paper proceeds as follows. In Section 2 we define the Bregman and the α-GPL loss functions and review the LS identification theory. Section 3 proves our results making it easier to apply LS, treats the two classes under consideration, and states our theorem on the generic identifiability, by unrestricted forecasts, for smooth convex loss functions. Section 4 provides further discussion and points out directions for future research. Short, simple proofs that provide insight into our results are given in the main text; the proof of the identification result for smooth convex loss functions is in the Appendix.

2 Formal setup and the LS identification theory

The question studied by LS is the following: given the point forecasts published by an expected loss minimizing forecaster, and the distributions used in their construction, is it

[2]Should the re-estimated value of α differ from 0.5 and be overly sensitive to b, the interpretation of the results becomes even more tenuous. A plausible explanation would be that the linear model for the point forecast posited by Elliott et al. (2005) is misspecified for the median.

possible to nonparametrically identify the forecaster's loss function? This setup is a best-case scenario in that the econometrician is assumed to know the predicitive 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 expected loss,

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

The compact set $F_t \subseteq D$ is the set of allowable forecasts and Br is, mnemonically, "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 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 is y for any *nbcc* loss function. This restriction is satisfied by most commonly used losses and can also be motivated by deriving loss functions from an underlying decision problem as in Granger and Machina (2006).

We make 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.[3]

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

Forecasters with unrestrictedly equivalent loss functions forecast in the same way when there are no controls on the set of allowable forecasts,[4] while forecasters with completely equivalent loss functions forecast in the same way even when they are restricted to make forecasts in the support of the distribution of Y_{t+1}. Loss functions in $\mathcal{C}_{nbcc}$ that are scalar multiples of each other are, trivially, completely forecast equivalent. The central question of this paper is whether Bregman losses and α-GPL losses are non-trivial examples of complete forecast equivalence.

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 in 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 hard task (we will return to this problem in Section 3.3). The question that naturally arises then is whether there are additional conditions on the forecaster's environment that could make forecasters with such losses distinguishable. As we will explain

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

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

below, the restriction involved in the definition of complete forecast equivalence *almost always* does the job, in a precise technical sense.[5]

Furthermore, the concept of complete forecast equivalence is not entirely devoid of practical relevance, even in observational settings. For example, consider a variable $Y \in \{-1, 0, 1\}$, indicating whether a given stock will underperform, match, or outperform the market over some period. A financial analyst could report a continuous forecast from the $[-1, 1]$ interval, or 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 a continuous forecast.

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 it can eventually be distinguished from every other element of the class, i.e. completely recovered, by observing forecasts produced in a sufficiently diverse set of environments. This diversity of environments includes, of necessity, 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 pair of 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, $\mathcal{B}$, of 'bad' loss functions 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.[6] The definitions

[5] As pointed out by a referee, one could go beyond the complete forecast equivalence concept used here, and require that forecasters 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.

[6] The adjective "tiny" refers to a class that is not only small in the topological sense of Baire category,

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 for some distribution p satisfying $p(F) = 1$ and $p(\{y_i\}) = 0$ for some $y_i \in F$, $Br(p \mid F, \ell) = F$.*

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 in 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 failure of identification. Definition 4 and Definition 5 are both rather abstract; in Section 3.2 below we will provide 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 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 first show that this is plausible by way of restricted forecast examples. We then give 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 explicitly appear in LS. We end this section with our result characterizing the identifying power of unrestricted forecasts for smooth, strictly convex loss functions.

but also "shy," which is an infinite dimensional version of being a Lebesgue null set. For interpretations of shyness, see Stinchcombe (2001). Appendix B gives formal definitions as well as further references.

3.1 Plausibility

Patton (2016) considers the parametric family of Bregman losses

$$\ell(\hat{y}, y; a) = \frac{2}{a^2}(e^{ay} - e^{a\hat{y}}) - \frac{2}{a}e^{a\hat{y}}(y - \hat{y}), \quad a \neq 0. \tag{4}$$

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, for all $a \neq 0$, the unrestricted optimal forecast is $p(1) = E\,Y$ for all values of a, where $p(1)$ is the (conditional) probability that $Y = 1$. As this holds for all $a \neq 0$, identifying the forecaster's loss function by their unrestricted forecast is not possible.

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

$$p(1) > c_a = \frac{\ell(1, 0; a)}{\ell(1, 0; a) + \ell(0, 1; a)} = \frac{1}{1 - e^{-a}} - \frac{1}{a}; \tag{5}$$

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, given sufficient variation in $p(1)$, the different members of this class *can* be identified in this *controlled* environment.

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.[7] In stark contrast with 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, 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

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

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

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) has either no effect or eliminates the same off-support quantiles for any ψ. Hence, no information about ψ is revealed.

3.2 Controlled Identification Results

We begin with preliminaries, then give and prove results making it easier to apply the controlled identification in LS, and then apply the results to the two classes under consideration.

3.2.1 Preliminaries

If $F = \{y_1, y_2, y_3\}$ is a three point subset of D then testing for a boundary problem at F involves setting either $p(y_1) = 0$, $p(y_2) = 0$, or $p(y_3) = 0$ (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 conditions in (6), (7), and (8) in the following.

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 any one of the conditions below is satisfied:*

$$g_1(y_1, y_2, y_3) := \ell(y_2, y_3)\ell(y_3, y_2) - \ell(y_2, y_3)\ell(y_1, y_2) - \ell(y_1, y_3)\ell(y_3, y_2) = 0 \quad (6)$$

$$g_2(y_1, y_2, y_3) := \ell(y_1, y_3)\ell(y_3, y_1) - \ell(y_1, y_3)\ell(y_2, y_1) - \ell(y_2, y_3)\ell(y_3, y_1) = 0 \quad (7)$$

$$g_3(y_1, y_2, y_3) := \ell(y_1, y_2)\ell(y_2, y_1) - \ell(y_1, y_2)\ell(y_3, y_1) - \ell(y_3, y_2)\ell(y_2, y_1) = 0. \quad (8)$$

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

We show that the second case, $p(y_2) = 0$, is equivalent to (7), 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

$$\ell(y_1, y_3)p(y_3) = \ell(y_3, y_1)p(y_1), \tag{9}$$

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

$$\ell(y_1, y_3)p(y_3) = \ell(y_2, y_1)p(y_1) + \ell(y_2, y_3)p(y_3). \tag{10}$$

Using $p(y_1) + p(y_3) = 1$, one can solve for $p(y_1)$ and $p(y_3)$ using equation (9) and substitute the resulting expressions into (10). This yields equation (7). □

The following uses Proposition 1 to construct an easier-to-check sufficient condition for a loss function to belong to the identified set, $\mathcal{G}$.

Proposition 2. *Let D_0^3 denote the set of triples with distinct coordinates in the interior of D^3. 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, a special case of Theorem 1 in Ponomarev (1987), states 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 $f : O \to \mathbb{R}$ 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 $\mathbb{R}$.*

3.2.2 Controlled Identification is Possible for Bregman Loss Functions

We can now formally show 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 $\mathbb{R}^3$, 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

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

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

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

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

$$\frac{\partial}{\partial y_2} g_2(y_1, y_2, y_3) = 0 \Leftrightarrow \ell(y_1, y_3)y_1 - [\ell(y_1, y_3) + \ell(y_3, y_1)]y_2 + \ell(y_3, y_1)y_3 = 0. \tag{11}$$

Let $h(y_1, y_2, y_3) = \ell(y_1, y_3)y_1 - [\ell(y_1, y_3) + \ell(y_3, y_1)]y_2 + \ell(y_3, y_1)y_3$. By the equivalency stated in (11), 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 treated with 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.

3.2.3 Controlled Identification is Not Possible for α-GPL Loss Functions

The next proposition states that GPL losses have a boundary problem at *any* three point set, which of course means that they cannot be 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, hence $\ell \notin \mathcal{G}$.*

Proof: Pick a point in D_0^3. Letting, say, $y_1 < y_2 < y_3$, the defintion 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

$$[\psi(y_3) - \psi(y_1)]^2 + [\psi(y_1) - \psi(y_3)][\psi(y_2) - \psi(y_1)] + [\psi(y_2) - \psi(y_3)][\psi(y_3) - \psi(y_1)] = 0.$$

As ψ is strictly increasing, $\psi(y_3) > \psi(y_1)$, so the equation further simplifies 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 will set some g_j identically to zero. □

3.3 Unrestricted Identification for Smooth Convex Loss Functions

The fact that Bregman losses are potentially identified shows that restricted forecasts (controlled environments) can reveal strictly more information about loss functions than unrestricted forecasts. We will now focus attention to 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.

In particular, let $\mathcal{D}^2_{\text{conv}} \subset \mathcal{C}_{nbcc}$ be defined as the set of loss functions in $\mathcal{C}_{nbcc}$ that are twice continuously differentiable on some open subset of $\mathbb{R} \times \mathbb{R}$ containing $D \times D$ and satisfy $\partial^2 \ell(\hat{y}, y)/\partial \hat{y}^2 > 0$ at all $(\hat{y}, y)$ pairs in $D \times D$. For each loss function $\ell \in \mathcal{D}^2_{\text{conv}}$, we define $\mathcal{B}^2(\ell) = \{\ell^\dagger \in \mathcal{D}^2_{\text{conv}} : Br\left(\ell^\dagger \mid D, p\right) = Br\left(\ell \mid D, p\right) \text{for all } p\}$. This is the set of loss functions in $\mathcal{D}^2_{\text{conv}}$ that are consistent for the same statistical functional as ℓ. We first observe that $\mathcal{B}^2(\ell)$ does not just contain scalar multiples of ℓ; in general, $\mathcal{B}^2(\ell)$ is a non-trivial, infinite dimensional equivalence class akin to a Bregman class or a GPL class.

The basis of this observation is what Gneiting (2011a) refers to as *Osband's principle* (after Osband 1985). Given an initial loss function, $\ell(\hat{y}, y) \in \mathcal{D}^2_{\text{conv}}$, the idea is to generate unrestrictedly forecast equivalent losses via the integral

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

where $w(t) > 0$ is a continuously differentiable weight function. Direct computation shows that if $w'(t) > 0$, then $\ell^\dagger$ also belongs to $\mathcal{D}^2_{\text{conv}}$, hence the first order conditions (FOCs) uniquely determine the unrestricted optimal forecast. The next step is to show that the FOCs for ℓ and $\ell^\dagger$ are the same. Interchanging the order of integration yields

$$\int \ell^\dagger(\hat{y}, y)dp(y) = \int_a^{\hat{y}} \left[\int \ell_{\hat{y}}(t, y)dp(y) \right] w(t)dt.$$

Passing the derivative through the integral sign yields, for all $\hat{y} \in (a, b)$,

$$\frac{d}{d\hat{y}} \int \ell^\dagger(\hat{y}, y)dp(y) = w(\hat{y}) \int \ell_{\hat{y}}(\hat{y}, y)dp(y) = w(\hat{y}) \frac{d}{d\hat{y}} \int \ell(\hat{y}, y)dp(y). \tag{12}$$

Equation (12) shows that $Br\left(p \mid D, \ell^\dagger\right)$ satisfies the same first order condition as $Br\left(p \mid D, \ell\right)$ so that ℓ and $\ell^\dagger$ are unrestrictedly forecast equivalent.[8]

By varying the weight function $w(t)$, one can then generate an entire class of forecast equivalent loss functions to ℓ. For example, starting from 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 actually be generated this way.

Our contribution to this theory is to show that Osband's principle notwithstanding, the set $\mathcal{B}^2(\ell)$ is still a tiny subset of $\mathcal{D}^2_{\text{conv}}$ in the same technical sense discussed in Section 2, i.e., $\mathcal{B}^2(\ell)$ is shy and Baire small. This implies that unrestricted forecasts still have a lot of identifying power in that they distinguish any given loss function ℓ from a generic or large subset of $\mathcal{D}^2_{\text{conv}}$, given by $\mathcal{G}^2(\ell) := \mathcal{D}^2_{\text{conv}} \setminus \mathcal{B}^2(\ell)$. Stated more formally:

Proposition 5. *For each $\ell \in \mathcal{D}^2_{\text{conv}}$, the set $\mathcal{B}^2(\ell)$ is shy and Baire small. That is, for each $\ell \in \mathcal{D}^2_{\text{conv}}$ the set $\mathcal{G}^2(\ell)$ 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

1. The proof of Proposition 5 is technical and is given in the Appendix in a series of steps.

[8] We thank an anonymous referee for this.

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

3. Together with Osband's principle, Proposition 5 offers novel insight 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 uncountably many tiny equivalence classes. Given any loss in $\mathcal{D}^2_{\text{conv}}$, "almost all" other loss functions give rise to a different statistical functional as the unrestricted optimal forecast, save for the losses in the same equivalence class.

4. While Proposition 5 theoretically limits the extent of the identification problem one faces in trying to recover loss functions from observing unrestricted forecasts (along with predictive distributions), the set $\mathcal{B}^2(\ell)$ is still generally large enough to cause substantial ambiguity in practice. For example, as noted in the introduction, different Bregman losses may rank different conditional mean forecasts differently in case of non-overlapping information sets, model misspecification, or estimation error.

5. The result stated here is stronger than the identification results in LS in that we do not need control over the sets in which the forecasters choose, and weaker in two senses. First, the generic set $\mathcal{G}^2(\ell)$ depends on ℓ, and we do not know if there exists a single generic $\mathcal{E} \subset \mathcal{D}^2_{\text{conv}}$ with the property that each $\ell \in \mathcal{E}$ is distinguishable from every other member of $\mathcal{E}$ (except possibly scalar multiples) based on unrestricted forecasts. 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, and that contains none of the α-GPL loss functions.

6. We conjecture that the result could also be strengthened in another direction. There may exist a single set of distributions $\mathcal{P}$, independent of ℓ and smaller than $\Delta(D)$, such that $\mathcal{P}$ alone is capable of distinguishing each ℓ from all members of $\mathcal{G}^2(\ell)$. We do not know how to construct such a collection $\mathcal{P}$ or how large the smallest such collection needs to be.

4 Conclusion

The main result in 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"—though it may require a tremendous amount of variation in the conditional distributions faced by the forecaster, as well as the set of allowable forecasts, eventually any loss function in $\mathcal{G}$ can be nonparametrically identified up to scale. While observational data on forecasts is unlikely to incorporate sufficient variation for this result to be practically applicable, preference recovery in, say, experimental settings 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 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 to distinguish between all possible loss functions, albeit these counterexamples are part of a "tiny," measure zero set within $\mathcal{C}_{nbcc}$. Proposition 2 gives a novel, and applicable, method for checking whether a given loss function belongs to the "good" class of loss functions.

Motivated by these results, we studied identification by unrestricted forecasts more generally. If ℓ is a twice continuously differentiable, convex loss function, then unrestricted forecasts distinguish it from a large set, $\mathcal{G}^2(\ell)$, of such losses. A number of open questions remain: (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 multiples)? (ii) Can a single class of distributions smaller than $\Delta(D)$ distinguish each ℓ from every loss in $\mathcal{G}^2(\ell)$? How large does this collection need to be? (iii) Can results be extended to classes more general than $\mathcal{D}^2_{\text{conv}}$?

Appendix: Proof of Proposition 5

Part A provides preliminary results, part B discusses the relevant definition of smallness/genericity, and part C 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 only concerned 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 will use both the $\ell_{\hat{y}}(\hat{y}, y)$ and $\frac{\partial}{\partial \hat{y}}\ell(\hat{y}, y)$ style notation to denote partial derivatives of ℓ.

A. 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 $p \in \Delta(D)$,

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

(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 putting positive mass on $(0, 1)$, the interior of D, the solution $\hat{y}^*_\ell(p)$ satisfies $0 < \hat{y}^*_\ell(p) < 1$.

Proof: (i) Immediate. (ii) Follows from the dominated convergence theorem by $\frac{\partial^j}{\partial \hat{y}^j}\ell(\hat{y}, y)$ being continuous and hence bounded over $D \times D$. (iii) The function $F_p(\hat{y}) := \int_D \ell(\hat{y}, y)\, dp(y)$ 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 theorem of the maximum and (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 $\frac{d}{d\hat{y}}F_p(\hat{y}) = \frac{\partial}{\partial \hat{y}}\ell(\hat{y}, 0)p(0) + \int_{(0,1)} \frac{\partial}{\partial \hat{y}}\ell(\hat{y}, y)\, dp(y) + \frac{\partial}{\partial \hat{y}}\ell(\hat{y}, 1)p(1)$. □

B. Small and large subsets of $\mathcal{D}^2_{\text{conv}}$

In order to define 'small' and 'large' (generic), we must specify the metric we use on $\mathcal{D}^2_{\text{conv}}$. Given a bi-index $\alpha = (\alpha_1, \alpha_2) \in \{0, 1, 2\}^2$, with $|\alpha|$ defined as $\alpha_1 + \alpha_2$, the α-derivative of a function $f : \mathbb{R}^2 \to \mathbb{R}$ at a point (x°_1, x°_2) is

$$f^{(\alpha)}(x^\circ_1, x^\circ_2) = \frac{\partial^{|\alpha|} f(x^\circ_1, x^\circ_2)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}. \tag{13}$$

We measure the distance between twice continuously differentiable functions on $D \times D$ with the (Sobolev) metric

$$d(f, g) = \sum_{|\alpha| \le 2} \left(\max_{(x_1, x_2) \in D \times D} |f^{(\alpha)}(x_1, x_2) - g^{(\alpha)}(x_1, x_2)| \right). \tag{14}$$

Intuitively, if f and g are close to each other in Sobolev metric, then in addition to $|f(x) - g(x)|$ being uniformly small, the distance between various derivatives (here up to order two) is also uniformly small.

We say that a subset $\mathcal{B}$ of a topologically complete, convex set of functions (such as $\mathcal{D}^2_{\text{conv}}$) is **tiny** if it is small in the sense of Baire and shy. A subset is **large** if its complement is tiny. A relatively simple sufficient condition for a set to be shy is 1-shyness.[9]

Definition 6. *For M a topologically complete metric space, $B \subset M$ is **small in the sense of Baire** if it is closed and has empty interior or if it is the countable union of closed sets with empty interior. For C a convex, topologically complete set of functions, $B \subset C$ is 1-shy if $\mu(B + \ell^\circ) = 0$ for all $\ell^\circ \in C$, where μ is the uniform distribution on some 1-dimensional line segment in C.*

We want to apply these definitions to $M = \mathcal{D}^2_{\text{conv}}$ and $C = \mathcal{D}^2_{\text{conv}}$, endowed with the Sobolev metric. A one-dimensional line segment in $\mathcal{D}^2_{\text{conv}}$ is parameterized as $L = \{\beta\ell^\dagger + (1-\beta)\ell^\ddagger : 0 \leq \beta \leq 1\}$, where $\ell^\dagger$ and $\ell^\ddagger$ are distinct elements of $\mathcal{D}^2_{\text{conv}}$. Intuitively, 1-shyness of $B \subset \mathcal{D}^2_{\text{conv}}$ means that the intersection between L and any given translation of B is either empty or corresponds to a set of β's with Lebsegue measure zero. More formally, we require $\lambda\{\beta \in [0,1] : \beta\ell^\dagger + (1-\beta)\ell^\ddagger \in B + \ell^\circ\} = 0$, where λ is the Lebsegue measure.

C. Existence and genericity of $\mathcal{G}^2(\ell)$

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 second derivative of any $\ell \in \mathcal{D}^2_{\text{conv}}$ being uniformly strictly above zero on $D \times D$.[10]

[9] A metric space is topologically complete if it has an equivalent metric in which it is complete. With the given metric, $\mathcal{D}^2_{\text{conv}}$ is topologically complete. The idea of 1-shyness, though not referred to as such, appears in Hunt, Sauer and Yorke (1992). The general treatment of shy subsets of topologically complete convex subsets of metric vector spaces is given in Anderson and Zame (2001). LS call tiny sets 'totally small' and the complements of tiny sets 'totally large'.

[10] 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

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

$\mathcal{B}^2(\ell)$ **is 1-shy**. We first show that for $|r| > 0$ sufficiently small, both

$$\ell^\dagger(\hat{y}, y) := \ell(\hat{y}, y) \cdot (1 + r\hat{y}) \text{ and } \ell^\ddagger(\hat{y}, y) := \ell(\hat{y}, y) \cdot (1 - r\hat{y}) \tag{15}$$

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 $|r|$ sufficiently close to zero. The treatment of $\ell^\ddagger$ is parallel. For the following, fix some $r \neq 0$ satisfying these conditions.

Pick arbitrary $\ell^\circ \in \mathcal{D}^2_{\text{conv}}$ and consider the set of $\beta \in [0, 1]$ such that $\{\beta\ell^\dagger + (1-\beta)\ell^\ddagger\} \in \mathcal{B}(\ell) + \ell^\circ$. 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.,

$$\int \ell_{\hat{y}}(\hat{y}^*, y)\, dp(y) = 0. \tag{16}$$

If $\beta\ell^\dagger + (1-\beta)\ell^\ddagger$ belongs to $\mathcal{B}(\ell) + \ell^\circ$, 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 equation (16) must leave the equality intact, i.e.,

$$-\int \ell^\circ_{\hat{y}}(\hat{y}^*, y)dp(y) + \beta \int \ell^\dagger_{\hat{y}}(\hat{y}^*, y)dp(y) + (1-\beta)\int \ell^\ddagger_{\hat{y}}(\hat{y}^*, y)dp(y) = 0. \tag{17}$$

However, using the definition of $\ell^\dagger$ and $\ell^\ddagger$, and taking (16) into account gives

$$\int \ell^\dagger_{\hat{y}}(\hat{y}^*, y)dp(y) = r\int \ell(\hat{y}^*, y)dp(y) \text{ and } \int \ell^\ddagger_{\hat{y}}(\hat{y}^*, y)dp(y) = -r\int \ell(\hat{y}^*, y)dp(y). \tag{18}$$

Substituting (18) into (17) yields

$$-\int \ell^\circ_{\hat{y}}(\hat{y}^*, y)dp(y) - r\int \ell(\hat{y}^*, y)dp(y) + 2r\beta\int \ell(\hat{y}^*, y)dp(y) = 0.$$

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

$\epsilon > 0$ such that $\frac{\partial^2}{\partial \hat{y}^2}\ell_n(\cdot, \cdot) > \epsilon$ for all n large enough. Then $|\hat{y}^*_n(p) - \hat{y}^*(p)|$ is bounded from above by

$$\tfrac{1}{\epsilon}\big|\textstyle\int \frac{\partial}{\partial \hat{y}}\ell_n(\hat{y}^*, y)dp(y)\big| = \tfrac{1}{\epsilon}\big|\textstyle\int (\frac{\partial}{\partial \hat{y}}\ell_n(\hat{y}^*, y) - \frac{\partial}{\partial \hat{y}}\ell(\hat{y}^*, y))dp(y)\big| < \tfrac{1}{\epsilon}\sup_y \big|\tfrac{\partial}{\partial \hat{y}}\ell_n(\hat{y}^*, y) - \tfrac{\partial}{\partial \hat{y}}\ell(\hat{y}^*, y)\big|.$$

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.

$\mathcal{B}^2(\ell)$ is small in the sense of Baire. $\mathcal{B}(\ell)$ is closed because it is the inverse image of a point under the continuous mapping $\gamma(\cdot)$. To show that the interior of $\mathcal{B}(\ell)$ is empty, pick arbitrary $\ell' \in \mathcal{B}(\ell)$. 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 $\mathcal{B}(\ell)$. Consider the point $\ell' + \delta\ell^\dagger$ where $\delta > 0$ and $\ell^\dagger$ is defined in (15) with some $r > 0$ sufficiently small. 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 $\delta > 0$, hence, $\ell' + \delta\ell^\dagger \notin \mathcal{B}(\ell)$. Nevertheless, as $\delta \downarrow 0$, $\ell' + \delta\ell^\dagger$ converges to ℓ'. □

References

Anderson, R. M. and W. R. Zame (2001): "Genericity with Infinitely Many Parameters." *Advances in Theoretical Economics*, 1, pp. 1-62.

Banerjee, A. X. Guo and H. Wang (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.

Ehm, W., T. Gneiting, A. Jordan and F. Krueger (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., I. Komunjer and A. Timmermann (2005): "Estimation and Testing of Forecast Rationality under Flexible Loss." *Review of Economic Studies*, 72, pp. 1107-1125.

Elliott, G. I. Komunjer and A. Timmermann (2008): "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss." *Journal of European Economic Association*, 6, pp. 122-157.

Fissler, T., and J. F. Ziegel (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, pp. 746-762.

Gneiting, T. (2011b): "Quantiles as Optimal Point Forecasts." *International Journal of Forecasting*, 27, pp. 197-207.

Granger, C.W.J. and M. Machina (2006): "Forecasting and Decision Theory." In G. Elliott, C.W.J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting.* Elsevier.

Komunjer, I. (2005): "Quasi Maximum-Likelihood Estimation for Conditional Quantiles." *Journal of Econometrics*, 128, pp. 137-164.

Lieli, R. P. and M. B. Stinchcombe (2013): "On the Recoverability of Forecasters' Preferences." *Econometric Theory*, 29, pp. 517-544.

Osband, K. H. (1985): "Providing Incentives for Better Cost Forecasting." Unpublished Ph.D. thesis, University of California, Berkeley.

Patton, A. J. and A. Timmermann (2007): "Testing Forecast Optimality under Unknown Loss." *Journal of the American Statistical Association*, 102, pp. 1172-1184.

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, Duke University.

Ponomarev, S. P. (1987): "Submersions and Preimages of Sets of Measure Zero". *Siberian Mathematical Journal*, 28, pp. 153-163.

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

Savage, L. J. (1971): "Elicitation of Personal Probabilities and Expectations." *Journal of the American Statistical Association*, 66, pp. 783-801.

Steinwart, I., C. Pasin, R. C. Williamson, S. Zhang (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 American Mathematical Society*, 129, pp. 451-7.