Modeling and scoring dynamic probabilistic forecasts

Thibault Modeste, Clément Dombry, Anne-Laure Fougères

Modeling and scoring dynamic probabilistic forecasts

Thibault Modeste [*] Clément Dombry [†] Anne-Laure Fougères [‡]

Abstract

Probabilistic forecasts play a major role in many applications where forecast is needed together with an assessment of its uncertainty. Verification of probabilistic forecasts has become increasingly important and mostly relies on two sets of tools: scoring rules and calibration diagnostics. Proper scoring rules assign forecasts numerical scores such that the correct forecast achieves a minimal expected score. Calibration theory aims at verifying that the observations and the forecasts are consistent.

In practice, using a probabilistic forecast commonly involves a sequential decision making process where the environment evolves over time. In this article, we propose a mathematical framework for dynamic probabilistic forecasts. The forecasts take therein the form of stochastic processes adapted to a filtration that encodes the available information. Under minimal assumptions, we show that proper scoring rules can still be used in this dynamic framework to discriminate the ideal forecast - more precisely, we prove that the long term average score is close to minimum if and only if the forecasts are close to ideal. This result provides theoretical guarantees for several methods used in practice. Some connections are also done in terms of Wasserstein distance.

Keywords: dynamic probabilistic forecast; scoring rules; RKHS; MMD.

[*]CNRS / Université de Pau et des Pays de l'Adour / E2S UPPA Laboratoire de mathématiques et applications IPRA, UMR 5142 B.P. 1155, 64013 Pau Cedex, France. E-mail: thibault.modeste@univ-pau.fr

[†]Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, CNRS UMR 6623, F-25000 Besançon, France. E-mail: clement.dombry@univ-fcomte.fr

[‡]Universite Claude Bernard Lyon 1, ICJ UMR5208, CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Jean Monnet, 69622 Villeurbanne, France. E-mail: fougeres@math.univ-lyon1.fr

Contents

1 Introduction

In a wide range of applications, probabilistic forecasts (Dawid, 1984; Gneiting et al., 2007) have become an essential tool, as recently illustrated for example in hydrology (Tiberi-Wadier et al., 2021), health (Henzi et al., 2021), demography (Raftery and Ševčíková, 2021), or meteorology (Vannitsem et al., 2021). In such contexts, verification is of particular importance, and is based on two sets of tools: calibration diagnostics, and scoring rules. Calibration theory aims at verifying that the observations and the forecasts are consistent. See e.g. Tsyplakov (2011), Strähl and Ziegel (2017), or Taillardat et al. (2022, Appendix A) for formal definitions. Scoring rules are used for evaluating the quality of a forecast and to compare different forecasts, and proper scoring rules assign forecasts numerical scores such that the correct forecast achieves a minimal expected score.

Most of the phenomena considered in applications have a dynamical nature, see among others Holzmann and Eulert (2014) in a risk management scoring framework, or Bröcker and Ben Bouallègue (2020) for verification of ensemble weather forecasts. To meet these needs in terms of assumptions, the stationary setting is the most popular one, as is required in the papers cited above. Such hypotheses happen however to be too restrictive in most situations, and there is a real practical interest to have a flexible mathematical framework for probabilistic forecasts in a dynamical context. From this perspective, Strähl and Ziegel (2017) proposed a framework allowing for quite general serial dependence as well as a definition of calibration dedicated to this setting. Our work is in the same vein, and consists of proposing to consider as dynamic probabilistic forecasts some stochastic processes adapted to a filtration that encodes the available information. We show in such a general framework that proper scoring rules can still be used to discriminate the ideal forecast.

More precisely, we introduce in Section 2 a general model – called Model 1 – which only requires assumptions of measurability with respect to a σ-field gathering the available information. Various examples are also given to illustrate the defined structure. In Section 3, Theorem 1 shows that even under weak assumptions including Model 1, the long term averaged score is still almost surely minimized by the ideal forecast; it states additionally that the long term averaged score of a dynamic forecast is asymptotically equivalent to the long term averaged score of the ideal forecast if and only if the average divergence between the two forecasts tends to 0. Similar results can be easily obtained with stationarity assumptions. This theorem therefore justifies the common use of averaging under more general assumptions than the stationary one. Section 4 describes several applications of this theorem. First, it justifies the use of scoring rules to compare forecasters. Then, in some particular cases of Energy Scores, it is also stated that the average divergence between two forecasts tends to 0 if and only if the average Wasserstein distance between two forecasts tends to 0. A final application is to guarantee the estimation of certain parameters using scoring rules. Finally, we give a classic example of statistical post-processing to illustrate the result. Section 5 contains all the proofs, and Appendices A and B give some results on the regularity of scoring rules and the link between scoring rules with the RKHS.

2 Dynamic probabilistic forecasts

2.1 Mathematical models

In this section, we propose a simple mathematical framework for dynamic probabilistic forecasts. Let $\mathcal{U}$ be a Polish space considered as the *universe*. Let $\mathcal{Y}$ be a second Polish space and $f : \mathcal{U} \to \mathcal{Y}$ be a measurable application considered as the *observable*, that is to say the quantity we are interested in and we want to forecast. The space of Borel probability measures on $\mathcal{Y}$ is denoted by $\mathcal{P}(\mathcal{Y})$ and seen as the *space of predictive distributions*. It is equipped with the σ-algebra generated by the applications $\pi \in \mathcal{P}(\mathcal{Y}) \mapsto \pi(B) \in \mathbb{R}$, $B \subset \mathcal{Y}$ Borel.

Model 1. On an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider:

- a sequence $(U_n)_{n\in\mathbb{N}}$ of $\mathcal{U}$-valued random variables;
- the sequence $Y_n = f(U_n)$, $n \in \mathbb{N}$;
- a sub-filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of the natural filtration $(\mathcal{G}_n)_{n\in\mathbb{N}}$ associated with $(U_n)_{n\in\mathbb{N}}$, i.e. $\mathcal{F}_n \subset \mathcal{G}_n = \sigma(U_k; k \leq n)$ for all $n \in \mathbb{N}$;
- a sequence $(F_n)_{n\in\mathbb{N}}$ of $\mathcal{P}(\mathcal{Y})$-valued random variables adapted to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, i.e. F_n is $\mathcal{F}_n$-measurable for all $n \in \mathbb{N}$.

The sequence $(U_n)_{n\in\mathbb{N}}$ represents the evolution of the environment over time and $(Y_n)_{n\in\mathbb{N}}$ the quantity of interest. The sub-filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ represents the *information* and we call the sequence $(F_n)_{n\in\mathbb{N}}$ a *dynamic probabilistic forecast*. The forecast is understood relatively to a *lead time* $T \geq 1$, meaning that the forecaster produces at time n a forecast F_n for the future value Y_{n+T} and this predictive distribution F_n is built in view of the limited information encoded in $\mathcal{F}_n$ only.
In a context of meteorological forecasts, the space $\mathcal{U}$ may represent e.g. the different possible states of the atmosphere, and U_n its state at time n. If the quantity of interest is the temperature at some location, we may take $\mathcal{Y} = \mathbb{R}$ and Y_n is the temperature at time n. The available information $\mathcal{F}_n$ may be a record of temperature, pressure, precipitation at several locations up to time n. Using this information, the forecast for the future temperature Y_{n+T} is given by the predictive distribution F_n with lead time T.

For the forecaster, the Holy Grail is the so-called *ideal forecast* that we define below.

Definition 1. *The ideal forecast with respect to the filtration* $(\mathcal{F}_n)_{n\in\mathbb{N}}$ *and with lead time* $T \geq 1$ *is the random sequence* $(F^*_{n,T})_{n\in\mathbb{N}}$ *defined by*

$$F^*_{n,T} = \mathcal{L}(Y_{n+T} \mid \mathcal{F}_n) \quad a.s., \quad n \in \mathbb{N}.$$

Clearly, the ideal forecast $(F^*_{n,T})_{n\in\mathbb{N}}$ is adapted to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. The conditional distribution $F^*_{n,T}$ is considered as the best possible predictive distribution for Y_{n+T} given the information $\mathcal{F}_n$. We refer to Gneiting and Ranjan (2013) and Tsyplakov (2013) for a general discussion on ideal forecasts.

For the purpose of asymptotics, we also consider a stronger model assuming stationarity. The model is very similar to Model 1, but we assume additionally strict stationarity and that the time index is $n \in \mathbb{Z}$. We also assume that the information is steeming from auxiliary observations of the form $X_n = h(U_n)$, with $h : \mathcal{U} \to \mathcal{X}$ being a measurable mapping between Polish spaces.

Model 2. On an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider:

- a strictly stationary sequence $(U_n)_{n\in\mathbb{Z}}$ of $\mathcal{U}$-valued random variables;
- the sequences $Y_n = f(U_n)$ and $X_n = h(U_n)$, $n \in \mathbb{Z}$;
- the sub-filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ associated to $(X_n)_{n\in\mathbb{Z}}$, i.e. $\mathcal{F}_n = \sigma(X_k; k \leq n)$;

- a sequence $(F_n)_{n\in\mathbb{Z}}$ of $\mathcal{P}(\mathcal{Y})$-valued random variables adapted to $(\mathcal{F}_n)_{n\in\mathbb{Z}}$.

The sequence $(Y_n)_{n\in\mathbb{Z}}$ again corresponds to the quantity of interest that we want to forecast, whereas $(X_n)_{n\in\mathbb{Z}}$ gathers the observations that are available, generating the information encoded by the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$. Clearly, both sequences $(X_n)_{n\in\mathbb{Z}}$ and $(Y_n)_{n\in\mathbb{Z}}$ are strictly stationary, and we will mostly consider stationary forecasts as defined below.

Definition 2. *The dynamic probabilistic forecast $(F_n)_{n\in\mathbb{Z}}$ is called stationary if the sequence $(U_n, F_n)_{n\in\mathbb{Z}}$ is strictly stationary.*

Under Model 2, a stationary forecast takes the form

$$F_n = \Phi(X_n, X_{n-1}, X_{n-2}, \cdots), \quad n \in \mathbb{Z},$$

for some measurable mapping $\Phi\colon \mathcal{X}^{\mathbb{N}} \to \mathcal{P}(\mathcal{Y})$. The mapping Φ is seen as the forecast algorithm that produces the predictive distribution given the past observations. See Lemma 1 in Section 5.1.
One can also show that the ideal forecast in Model 2 is stationary and takes the form

$$F^*_{n,T} = \mathcal{L}(Y_{n+T} \mid X_n, X_{n-1}, X_{n-2}, \cdots) = \Phi^*_T(X_n, X_{n-1}, X_{n-2}, \cdots) \quad a.s.,$$

with $\Phi^*_T\colon \mathcal{X}^{\mathbb{N}} \to \mathcal{P}(\mathcal{Y})$ and $n \in \mathbb{Z}$. See Lemma 2 in Section 5.1.

2.2 Examples

Several examples are discussed in this section to illustrate the framework of dynamic probabilistic forecasts introduced previously.

Example 1. As a simple example of Model 1, consider the Gaussian autoregressive model of order 1 (Brockwell and Davis (1991), Chapter 3) defined by the initial value $U_0 = 0$ and the recursive relation

$$U_{n+1} = \alpha U_n + \varepsilon_{n+1}, \quad n \in \mathbb{N}, \tag{1}$$

where $\alpha \in \mathbb{R}$ and $(\varepsilon_n)_{n\geq 1}$ is an i.i.d. centered Gaussian sequence with variance $\sigma^2 > 0$. We assume that $\mathcal{U} = \mathcal{Y} = \mathbb{R}$ and that the quantity of interest is $Y_n = U_n$.

Consider first the trivial case where no information is available, i.e. $\mathcal{F}_n = \{\emptyset, \mathcal{U}\}$ is the trivial σ-field for all $n \in \mathbb{N}$. Then the ideal forecast with lead time $T \geq 1$ is

$$F^*_{n,T} = \mathcal{L}(U_{n+T} \mid \mathcal{F}_n) = \mathcal{L}(U_{n+T}), \quad n \in \mathbb{N},$$

because the conditional distribution with respect to the trivial σ-field reduces to the marginal distribution. Simple computations show that

$$U_{n+T} = \sum_{i=1}^{n+T} \alpha^{n+T-i} \varepsilon_i$$

whence we deduce

$$F^*_{n,T} = \mathcal{N}\left(0, \frac{1-\alpha^{2(n+T)}}{1-\alpha^2}\sigma^2\right), \quad n \in \mathbb{N}.$$

We next discuss the opposite case of complete information where $\mathcal{F}_n = \sigma(U_k, k \leq n)$ for all $n \in \mathbb{N}$. The ideal forecast with lead time $T = 1$ is then

$$F^*_{n,1} = \mathcal{L}(U_{n+1} \mid U_0, \dots U_n) = \mathcal{N}(\alpha U_n, \sigma^2), \quad n \in \mathbb{N}.$$

For a general lead time $T \geq 1$, the relation

$$U_{n+T} = \alpha^T U_n + \sum_{i=1}^{T} \alpha^{T-i}\varepsilon_{n+i}$$

implies that the ideal forecast is given by

$$F^*_{n,T} = \mathcal{N}\left(\alpha^T U_n, \frac{1-\alpha^{2T}}{1-\alpha^2}\sigma^2\right), \quad n \in \mathbb{N}. \tag{2}$$

Note that the variances are smaller than in the case with no information which corresponds to the general fact that exploiting information reduces forecast uncertainty and leads to sharper predictive distributions.
If the parameters α, σ^2 are unknown, the ideal forecast is not accessible but the forecaster may naturally provide sequential parameter estimates based on the observation record. Maximum likelihood estimation (Brockwell and Davis (1991), Chapter 8.7) yields

$$\hat{\alpha}_n = \frac{\sum_{i=1}^n U_{i-1}U_i}{\sum_{i=1}^n U_{i-1}^2} \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(U_i - \hat{\alpha}_n U_{i-1}\right)^2, \quad n \geq 1.$$

In view of Equation (2), the plug-in method suggests the dynamic probabilistic forecast with lead time T,

$$F_{n,T} = \mathcal{N}\left(\hat{\alpha}_n^T U_n, \frac{1-\hat{\alpha}_n^{2T}}{1-\hat{\alpha}_n^2}\hat{\sigma}_n^2\right), \quad n \geq 1, \tag{3}$$

which is adapted with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$, i.e. accessible in view of the available observation record.

A simple illustration of Model 2 can be obtained when $\alpha \in (-1, 1)$. Let $(\varepsilon_n)_{n\in\mathbb{Z}}$ be an i.i.d. sequence with distribution $\mathcal{N}(0, \sigma^2)$, and consider the infinite moving average

$$U_n = \sum_{i\geq 0} \alpha^i \varepsilon_{n-i}, \quad n \in \mathbb{Z}.$$

The sequence $(U_n)_{n\in\mathbb{Z}}$ is strictly stationary and satisfies the auto-regressive property (1). The marginal distribution $\mathcal{N}(0, \sigma^2/(1-\alpha^2))$ corresponds to the ideal forecast in absence of information. To produce a stationary forecast, one may consider maximum likelihood estimation based on the last p observations, where p is an integer in $[1, n]$, i.e.

$$\hat{\alpha}_n = \frac{\sum_{i=0}^{p-1} U_{n-i-1}U_{n-i}}{\sum_{i=0}^{p-1} U_{n-i}^2} \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{p}\sum_{i=0}^{p-1}\left(U_{n-i} - \hat{\alpha}_n U_{n-i-1}\right)^2.$$

Clearly, the sequence of parameter estimates $(\hat{\alpha}_n, \hat{\sigma}_n^2)_{n\in\mathbb{Z}}$ is strictly stationary and $(F_{n,T})_{n\in\mathbb{Z}}$ defined by Equation (3) is a stationary forecast.

Example 2. An extension of the previous model can be built by incorporating an additive measurement error. To do this, consider the sequence

$$\begin{cases} U_{n+1}^{(1)} &= \alpha U_n^{(1)} + \varepsilon_{n+1}, \\ U_{n+1}^{(2)} &= U_{n+1}^{(1)} + \delta_{n+1}\ , \end{cases}$$

where the two sequences of innovation $(\varepsilon_n)_{n\geq 1}$ and noise $(\delta_n)_{n\geq 1}$ are assumed i.i.d. and with respective distribution $\mathcal{N}(0,\sigma^2)$ and $\mathcal{N}(0,\tau^2)$. Assume that the quantity of interest is the first component $Y_n = U_n^{(1)}$ and is observed with a measurement error δ_n, so that the observation available is the second component $X_n = U_n^{(2)}$. Here, one thus has $\mathcal{U} = \mathbb{R}^2$ and $\mathcal{Y} = \mathbb{R}$, and the information is given by the natural filtration associated with (X_n), i.e. $\mathcal{F}_n = \sigma(X_k, k \leq n)$.
The sequence $(U_n)_{n\in\mathbb{N}} = (U_n^{(1)}, U_n^{(2)})_{n\in\mathbb{N}}$ is a bivariate Gaussian vector as soon as the initial value U_0 is assumed to be bivariate Gaussian. When $\alpha \in (-1,1)$, it is strictly stationary for a suitable choice of the initial distribution.

Example 3. An example in a spatio-temporal setting can be constructed starting from a vectorial AR(1). Let $A \in \mathcal{M}_d(\mathbb{R})$ be a real square matrix, U_0 be a d-variate random vector and $(\varepsilon_n)_{n\in\mathbb{N}}$ an i.i.d. sequence of d-variate Gaussian vectors with distribution $\mathcal{N}_d(0,\Sigma)$. Consider for $n \in \mathbb{N}$

$$U_{n+1} = AU_n + \varepsilon_n.$$

We write

$$U_n = (X_{1,n}, \ldots, X_{d,n})^T.$$

The quantity of interest is the first coordinate, ie $Y_n = X_{1,n}$. For $I \subset \{1,\ldots,d\}$, we define

$$\mathcal{F}_{I,n} = \sigma\left(X_{i,k}, i \in I, k \leq n\right).$$

Example 4. As discussed above, the three examples considered can be made stationary by choosing a specific initialisation U_0. Let us finally illustrate a case of Model 1 that can not be converted into Model 2; let $(t_n)_{n\in\mathbb{N}}$ and $(s_n)_{n\in\mathbb{N}}$ be two sequences, and assume that $(s_n)_{n\in\mathbb{N}}$ is periodic. It typically represents seasonal variability. The sequence $(t_n)_{n\in\mathbb{N}}$ is a general trend that can represent a global warming in the context of climate change. For a given r.v. U_0, one can define for all $n \in \mathbb{N}$ the sequence

$$U_{n+1} = t_n + s_n + \alpha U_n + \varepsilon_{n+1}\ ,$$

which fulfills Model 1's general assumptions but not Model 2's.

3 Scoring rules for dynamic probabilistic forecast

3.1 Background on scoring rules

Scoring rules (Gneiting and Raftery, 2007) provide a major tool for forecast validation. A scoring rule compares forecasts and realizations and assigns a numerical score assessing

the forecast quality. Proper scoring rules have the property that the correct forecast minimizes the expected score. They are commonly used to compare different forecast methods and the forecast with the lowest score is preferred.
First we recall some basic definitions. Let $\mathcal{L}$ be a subset of $\mathcal{P}(\mathcal{Y})$ and d be a distance on this space, as for example the Wasserstein distance, see Section 4.2 for a definition. A scoring rule on $\mathcal{L}$ is a *measurable* real valued function $S\colon \mathcal{L} \times \mathcal{Y} \to \mathbb{R}$, and $S(F, y)$ is the quantity assigned to the probabilistic forecast F when the outcome y occurs.

Remark 1. Let specify here that we add the assumption of measurability in the definition of the scoring rule S, as done in Holzmann and Eulert (2014), in order to lighten the presentation. Some sufficient conditions to get this measurability are provided in Appendix A.

Definition 3. *The score S is said to be proper on $\mathcal{L}$ if for all $F, G \in \mathcal{L}$, the integral*

$$\overline{S}(F, G) = \int_{\mathcal{Y}} S(F, y) \, \mathrm{d}G(y)$$

is well-defined and if the following inequality holds

$$\overline{S}(G, G) \leq \overline{S}(F, G), \quad \text{for all } F, G \in \mathcal{L}.$$

The scoring rule is said to be strictly proper if the equality holds above if and only if $F = G$.

The quantity $\overline{S}(F, G)$ is the average score when the forecast is F and the observations have distribution G. For a proper scoring rule, the minimum of $F \mapsto \overline{S}(F, G)$ is achieved when $F = G$. The divergence associated with a proper scoring rule S is the non-negative function

$$\mathbf{div}_S(F, G) = \overline{S}(F, G) - \overline{S}(G, G).$$

For a strictly proper scoring rule, the divergence vanishes if and only if $F = G$, so that the divergence can be seen as a pseudo-distance between F and G (the symmetry or triangle inequality may not be satisfied).

Example 5. For real-valued observations, the most important and widely used scoring rule is the Continuous Ranked Probability Score, shortly noted CRPS (Epstein (1969); Hersbach (2000); Bröcker (2012)). The CRPS is a strictly proper scoring rule on the class $\mathcal{P}_1(\mathbb{R})$ of probability measures on $\mathbb{R}$ having a finite first moment. It is defined for $F \in \mathcal{P}_1(\mathbb{R})$ and $y \in \mathbb{R}$ by

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(x) - \mathbb{1}_{\{y \leq x\}})^2 \mathrm{d}x,$$

where the probability measure F is identified with its cumulative distribution function. An alternative representation is

$$\mathrm{CRPS}(F, y) = \mathbb{E}[|X - y|] - \frac{1}{2}\mathbb{E}[|X - X'|] \tag{4}$$

where X, X' are independent random variables with distribution F. Several other decompositions are available, see e.g. Taillardat et al. (2022) and references therein.

Example 6. A generalization of the CRPS can be obtained from (4) by introducing the so-called energy kernel, defined for $\alpha > 0$ and $\beta \in (0, \infty]$ by

$$\rho_{\alpha,\beta}(x, y) = \|x - y\|_\beta^\alpha, \quad x, y \in \mathbb{R}^d, \tag{5}$$

with $\|x\|_\beta = (\sum_{i=1}^d |x_i|^\beta)^{1/\beta}$ and $\|x\|_\infty = \max_{1 \le i \le d} |x_i|$. More precisely, consider the Energy Score, defined as

$$S_{\rho_{\alpha,\beta}}(F, y) = \int_{\mathbb{R}^d} \rho_{\alpha,\beta}(x, y) \, \mathrm{d}F(x) - \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} \rho_{\alpha,\beta}(x, x') F(\mathrm{d}x) F(\mathrm{d}x').$$

The case $d = 1$ and $\alpha = 1$ boils down to Equation (4) and hence to the CRPS.
It can be shown that the Energy Score is a proper scoring rule in the following cases:

- $d = 1$ and $\alpha \in (0, 2]$;
- $d \ge 2$, $\beta \in (0, 2]$ and $\alpha \in (0, \beta]$;
- $d = 2$, $\beta \in (2, +\infty]$ and $\alpha \in (0, 2]$.

See e.g. Schoenberg (1938) (case $\beta = 2$), Koldobsky (1992) and Zastavnyi (1993). Additionally, some interesting results can be established on the divergence of the Energy Score when $\beta = 2$ and $\alpha \in (0, 2)$, see Section 4.2.

Example 7. Another score that is widely used in practice is the logarithmic score, introduced in Good (1952). It is defined for measures dominated by the Lebesgue measure with non vanishing density functions. Consider such a measure F, denote by f its density, and define for $y \in \mathbb{R}$ the score as

$$S(F, y) = -\log f(y).$$

This scoring rule is strictly proper, and its divergence is the well-known Kullback–Leibler divergence,

$$\mathbf{div}_S(F, G) = \int_{\mathbb{R}} \log\left(\frac{g(y)}{f(y)}\right) g(y) \, \mathrm{d}y.$$

3.2 Asymptotic results for evaluating dynamic forecasts with proper scoring rules

The aim of this section is to discuss the scoring rule in a dynamic probabilistic forecasts framework, as defined in Model 1. It is established in particular that the ideal forecast minimizes the averaged score in different ways. Note that Model 1 provides a sequential model and, for simplicity, we first look at a single step.
Consider a random forecast F on $(\Omega, \mathcal{F}, \mathbb{P})$, which is measurable with respect to a sub-σ-algebra $\mathcal{F}_0 \subset \mathcal{F}$, and consider an observation Y. The ideal forecast is thus defined as $F^* = \mathcal{L}(Y \mid \mathcal{F}_0)$. A natural consequence of the definition of proper scoring rule is that the ideal forecast minimizes the expected score. Here, the expectation is taken with respect to both the observation and the forecast randomness. The following proposition is an important result in the non dynamic framework, and can be found in Holzmann and Eulert (2014) (Theorem 3).

Proposition 1 (Holzmann and Eulert (2014)). *Let S be a proper scoring rule. Then a.s.*

$$E[S(F,Y) - S(F^*,Y) \mid \mathcal{F}_0] = \mathbf{div}_S(F,F^*) \geq 0\ ,$$

and this implies $\mathbb{E}[S(F,Y)] \geq \mathbb{E}[S(F^*,Y)]$. Moreover, if the score is strictly proper, then equality holds if and only if $F = F^*$ a.s..

As a consequence, in the sequential framework defined by Model 1, the forecaster has to predict at each time n the ideal forecast $F^*_{n,T}$ in order to minimize its expected score. The main result of this section is a stronger optimality property in the sense of almost sure convergence. It states that the ideal forecast minimizes the long-term score *almost surely*. In some sense, expectation is replaced by a temporal average but it should be stressed that this is not straightforward because we do not assume any stationary condition.

Theorem 1. *Let $(F_n)_{n\in\mathbb{N}}$ be a probabilistic dynamic forecast as defined in Model 1, measurable with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Let $(F^*_{n,T})_{n\in\mathbb{N}}$ be the ideal forecast with lead time $T \geq 1$. Let S be a proper scoring rule with associated divergence $\mathbf{div}_S$. We assume that, for $k = 1,\dots,T$,*

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(\delta_i^k\right)^2 \mid \mathcal{F}_{i+T-k}\right] = O(n), \quad a.s.. \tag{6}$$

where $\delta_i^k = \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T+1-k}] - \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T-k}]$ and $\Delta_i = S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right)$. Then a.s., the following inequality holds

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) \geq 0\ . \tag{7}$$

Moreover, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) = 0 \quad a.s. \tag{8}$$

if and only if

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{div}_S(F_i, F^*_{i,T}) \to 0 \quad a.s.\ .$$

The condition (6) means for almost every $\omega \in \Omega$,

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(\delta_i^k\right)^2 \mid \mathcal{F}_{i+T-k}\right](\omega) = O(n).$$

The implicit constant in $O(n)$ depends to ω. Equation (7) states that the ideal forecast minimizes the long term averaged score almost surely – here the average is temporal and for a fixed realization, in opposition to the expected score considered in Proposition 1. The vanishing limit (8) means that the long term averaged score of the dynamic forecast $(F_n)_{n\in\mathbb{N}}$ is equal to the one of the ideal forecast $(F^*_{n,T})_{n\in\mathbb{N}}$, stating that both predictions are

equally good in this sense. This is characterized by an asymptotically negligible average divergence between the two sequences.
The proof of Theorem 1 is based on the strong law of large numbers for square integrable martingales and does not assume any stationarity condition. The technical condition (6) is required for the martingale convergence theorem. See Section 5.2 for more details.

Remark 2. In the case of the Energy Score with $\alpha \leq 1$ and $1 \leq \beta$ (see Example 6), a simple sufficient condition for condition (6) to hold is

$$\sup_i m(F_i) + m\left(F^*_{i,T}\right) < +\infty \quad a.s.,$$

where $m(F)$ is the first moment of a probability measure F. In other words, the probabilistic forecast and the ideal forecast have uniformly bounded first moments. See Proposition 3 in the Appendix A for more details.

We briefly comment the stationary case and state simple results for Model 2. Note that stationarity combined with ergodicity allows to use the *ergodic theorem* and greatly simplifies the proof.

Corollary 1. *In the framework of Model 2 with the ergodicity condition, let $(F_n)_{n\in\mathbb{Z}}$ be a stationary dynamic forecast measurable with respect to $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ and $(F^*_{n,T})_{n\in\mathbb{Z}}$ be the ideal forecast with lead time $T \geq 1$. Let S be a proper scoring rule. Then,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) = \mathbb{E}\left[S(F_0, Y_T)\right] \geq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S\left(F^*_{i,T}, Y_{i+T}\right) = \mathbb{E}\left[S(F^*_{0,T}, Y_T)\right],$$

where the limits are meant almost surely. Moreover, if S is strictly proper, equality holds if and only if $(F_n)_{n\in\mathbb{Z}} = (F^*_{n,T})_{n\in\mathbb{Z}}$ a.s.

4 Application of this asymptotic optimality

4.1 Comparison of forecasters

4.1.1 Classical framework

The first application of this result is a commentary on the use of scoring rules to compare forecasters. In practice, a scoring rule S is used to compare two forecasters $(F_n)_{n\in\mathbb{N}}$ and $(G_n)_{n\in\mathbb{N}}$. If the next quantity is negative, then the F forecaster is considered better, and vice versa if the quantity is positive,

$$\frac{1}{n}\sum_{i=1}^{n} S(F_i, Y_{i+T}) - \frac{1}{n}\sum_{i=1}^{n} S(G_i, Y_{i+T}).$$

Theorem 1 provides a theoretical guarantee if one of the forecasters is ideal. Indeed, if $(F_n)_{n\in\mathbb{N}}$ is an ideal forecaster with respect to information $(\mathcal{F}_n)_{n\in\mathbb{N}}$, then assymptotically, this rule will designate forecaster $(F_n)_{n\in\mathbb{N}}$ better than forecaster $(G_n)_{n\in\mathbb{N}}$.

4.1.2 Comparison of ideal forecasters

Theorem 1 has an interesting application dealing with partial information. Assume that two experts have access to different information and that the first expert is better informed. This is formalized by two filtrations $(\mathcal{F}_n^1)_{n\in\mathbb{N}}$ and $(\mathcal{F}_n^2)_{n\in\mathbb{N}}$ with $\mathcal{F}_n^2 \subset \mathcal{F}_n^1$ for all $n \in \mathbb{N}$. The best possible forecast for each expert is the ideal forecast with respect to the available information, noted $F_{n,T}^{*,j} = \mathcal{L}(Y_{n+T} \mid \mathcal{F}_n^j)$, $j = 1, 2$. Theorem 1 yields

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(F_{i,T}^{*,2}, Y_{i+T}\right) - S\left(F_{i,T}^{*,1}, Y_{i+T}\right) \geq 0 \quad a.s.,$$

meaning that the extra information possessed by the first expert allows him to reach a lower averaged score in the long term.

4.1.3 Forecasting and cross-calibration

The notion of cross-calibration could be seen as a generalization of the previous example (Strähl and Ziegel, 2017). Assume J different experts produce dynamic forecasts $(F_n^j)_{n\in\mathbb{N}}$ with respect to different filtrations $(\mathcal{F}_n^j)_{n\in\mathbb{N}}$, $1 \leq j \leq J$. Assume that the information $(\mathcal{F}_n^j)_{n\in\mathbb{N}}$ is private but the forecasts $(F_n^j)_{n\in\mathbb{N}}$ are public. Then the information encoded in the filtration

$$\mathcal{F}_n = \sigma(F_i^j; i \leq n, 1 \leq j \leq J), \quad n \in \mathbb{N},$$

is publicly accessible. Note that $\mathcal{F}_n$ does not necessarily contain $\mathcal{F}_n^j$ but nevertheless F_n^j is measurable with respect to $\mathcal{F}_n$. Considering the ideal forecast $F_{n,T}^* = \mathcal{L}(Y_{n+T} \mid \mathcal{F}_n)$, Theorem 1 yields, for all $1 \leq j \leq J$,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(F_i^j, Y_{i+T}\right) - S\left(F_{i,T}^*, Y_{i+T}\right) \geq 0 \quad a.s. .$$

This means that the public forecasts can be used to produce a new forecast that outperforms the J experts in terms of averaged score.

4.2 Interpretation for Energy Score

The purpose of this subsection is to provide a more explicit interpretation of the divergence result (8) in the case of the Energy Scores defined via (5). We assume $\beta = 2$ so that $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$. The $p-$Wasserstein space on $\mathbb{R}^d$ consists in the set $\mathcal{P}_p(\mathbb{R}^d)$ of Borel probability measures F on $\mathbb{R}^d$ with finite $p-$moment, i.e.

$$\mathcal{P}_p(\mathbb{R}^d) = \{F \in \mathcal{P}(\mathbb{R}^d) : \int \|x\|_2^p \, \mathrm{d}F(x) < \infty\}.$$

It follows from Hölder's inequality that $\mathcal{P}_p(\mathbb{R}^d) \subset \mathcal{P}_1(\mathbb{R}^d)$. In the case of $p = 1$, it is endowed with the Kantorovich-Rubinstein distance

$$W_1(F_1, F_2) = \sup_{\mathrm{Lip}(\phi)\leq 1} \left| \int \varphi(x) \, \mathrm{d}F_1(x) - \int \varphi(x) \, \mathrm{d}F_2(x) \right|,$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz function and $\mathrm{Lip}(\phi) = \sup_{x \neq y} |\phi(x) - \phi(y)|/\|x - y\|_2$. Fore more details on Wasserstein spaces, we refer to Villani (2009) Chapter 6.

Theorem 2. *Let $\alpha \in (0, 2)$ and $(F_n)_{n\in\mathbb{N}}$ and $(G_n)_{n\in\mathbb{N}}$ be sequences in $\mathcal{P}_{max(1,\alpha)}(\mathbb{R}^d)$ and assume the sequences are uniformly integrable, i.e.*

$$\forall \varepsilon > 0,\ \exists K \subset \mathbb{R}^d \text{ compact},\ \forall n \in \mathbb{N},\ \int_{K^c} \|x\|\ \mathrm{d}(F_n + G_n)(x) < \varepsilon.$$

*Let $\mathbf{div}_S$ be the divergence of the **Energy Score** with $\alpha \in (0, 2)$ and $\beta = 2$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{div}_S(F_i, G_i) \to 0 \quad \text{if and only if} \quad \frac{1}{n}\sum_{i=1}^{n} W_1(F_i, G_i) \to 0.$$

Consequently, Theorem 1 can be rewritten as follows: assuming condition (6) together with uniform integrability, we have the equivalence when the scoring rule S is an Energy Score,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) = 0 \iff \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} W_1(F_i, F^*_{i,T}) = 0.$$

This means that the dynamic forecast $(F_n)_{n\in\mathbb{N}}$ achieves asymptotically the minimal averaged score if and only if it is closed to the ideal forecast $(F^*_{n,T})_{n\in\mathbb{N}}$ in Wasserstein distance.

4.3 Convergence of argmin forecasters

The last result of this section gives a theoretical guarantee on the estimation of the ideal forecast with scoring rules. They can be used for statistical post-processing to improve forecasts by removing certain biases and reducing under-dispersion. Let us state a framework with more conditions to have this guarantee.

(A1) **Parametric model :** for $i \in \mathbb{N}$, the ideal forecast belongs to a parametric family of distributions

$$F^*_{i,T} = f_{\theta^*}(X_i),$$

where $\theta^* \in \mathbb{R}^k$ (not depends to i);

(A2) **Previous framework :** for all $\theta \in \mathbb{R}^k$, the dynamical forecaster $(f_\theta(X_n))_{n\in\mathbb{N}}$ satisfies Condition (6);

(A3) **Lipschitz condition :** the scoring rules S is lispchitz continuous for W_1, i.e. there exists $C > 0$ such that for $F, G \in \mathcal{L}$ and $y \in \mathbb{R}$

$$|S(F, y) - S(G, y)| \leq C W_1(F, G);$$

(A4) **Identifiable model :** there exists $C > 0$ and $p > 0$, such that for all $\theta, \eta \in \mathbb{R}^k$ and $x \in \mathbb{R}^d$,

$$W_1(f_\theta(x), f_\eta(x)) \leq C(1 + x)^p \|\theta - \eta\|;$$

(A5) **Dominated condition :** the empirical mean

$$\frac{1}{n}\sum_{i=1}^{n}|1+X_i|,$$

is dominated, i.e. there exists a measurable function φ such that

$$\forall n\in\mathbb{N},\ \frac{1}{n}\sum_{i=1}^{n}|1+X_i|\leq\varphi,\ a.s.$$

The condition (A1) does not mean that the quantity of interest $(Y_n)_{n\in\mathbb{N}}$ is stationary but only its dependence with explicative covariates $(X_n)_{n\in\mathbb{N}}$. The condition (A2) makes it possible to use the results set out in the previous sections. In Appendix A, examples of scoring rules verifying condition (A3) are given, including the CRPS. A reproach of these conditions is that we assume that the true ideal forecaster is included in our parametric model. An alternative way is to define a parameter $\theta^{(*)}$ such that the forecast $f_{\theta^{(*)}}$ is as close as possible to the ideal forecast in the sense of divergence. But in a non-stationary framework, it is not possible to define such a minimising parameter.

For a scoring rule S, let us consider the following estimator

$$\hat{\theta}_n=\underset{\theta\in\mathbb{R}^k}{\operatorname{argmin}}\frac{1}{n}\sum_{i=1}^{n}S(f_\theta(X_i),Y_{i+T}).$$

There is no reason why this estimate should be accurate, i.e. $\hat{\theta}_n=\theta^*$. Indeed, the optimality result of Theorem 1 is only asymptotic. In this parametric model, the *best* way to predict is to use this estimator to approximate the ideal forecast. Using the information $(X_1,\dots,X_n)$ and observations $(Y_2,\dots,Y_{n+1})$, the estimator $\hat{\theta}_n$ is computed and Y_{n+2} is predicted by $f_{\hat{\theta}_n}(X_{n+1})$. The following theorem states the consistency guarantee for this method.

Theorem 3. *Under the previous assumption, if $(\hat{\theta}_n)_n$ is bounded then*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{div}_S(f_{\hat{\theta}_n}(X_i),f_{\theta^*}(X_i))\to 0.$$

Moreover, if S is an Energy Score with $\alpha\in(0,2)$ and the following set

$$\mathcal{T}=\{f_\theta(x)\mid x\in\mathbb{R}^d,\ \theta\in\mathbb{R}^k\},$$

is uniformly integrable, then

$$\frac{1}{n}\sum_{i=1}^{n}W_1(f_{\hat{\theta}_n}(X_i),f_{\theta^*}(X_i))\to 0.$$

As the model is not assumed to be stationary, almost surely convergence cannot be achieved on the sequence of divergences. We end this section by giving a classic example of the use of this method.

Example 8. The *Ensemble model output statistics* (EMOS) is one of the most widely used methods for correcting weather forecasts. It has been introduced in Gneiting et al. (2005), and improved on several occasions (Messner et al., 2017; Schulz and Lerch, 2022). EMOS was introduced as part of statistical post-processing to debias predictions obtained by numerical simulation (NWP). To predict meteorological data, e.g. temperature or rainfall, meteorological institutes produce several dozen equiprobable scenarios $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_k)$. These raw numerical data have a number of flaws that need to be corrected (Hamill and Colucci, 1997; Richardson, 2001). The simple idea is to assume that the predicted phenomenon with respect to the information $\mathbf{x}$ produced by the algorithms is of the following form

$$Y \mid \mathbf{x} \ \sim \ \mathcal{N}(a^* + b^* \overline{\mathbf{x}}, c^* + d^* \sigma(\mathbf{x})^2),$$

where $\overline{\mathbf{x}}, \sigma(\mathbf{x}$ is the mean and variance of the vector $\mathbf{x}$. With $(\mathbf{x}_n)_{n\in\mathbb{N}}$ the NWP outputs, the parameters (a^*, b^*, c^*, d^*) are estimated as follows

$$(\hat{a}_n, \hat{b}_n, \hat{c}_n, \hat{d}_n) = \operatorname*{argmin}_{a,b,c,d} \frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(\mathcal{N}(a + b\overline{\mathbf{x}_i}, c + d\sigma(\mathbf{x}_i)^2), Y_{i+1}).$$

Theorem 3 states that the estimation error, represented by the divergence, vanishes.

5 Proofs

5.1 Proofs of Section 2

Lemma 1. *Under Model 2, if $(F_n)_{n\in\mathbb{Z}}$ is stationary, there exists a map $\Phi\colon \mathcal{X}^{\mathbb{N}} \to \mathcal{P}(\mathcal{Y})$ such that*

$$F_n \overset{d}{=} \Phi(X_n, X_{n-1}, \ldots).$$

Proof. As $\mathcal{P}(\mathcal{Y})$ is a Polish Space, there exists for all $n \in \mathbb{Z}$, a map $\Phi_n\colon \mathcal{X}^{\mathbb{N}} \to \mathcal{P}(\mathcal{Y})$ such that

$$F_n = \Phi_n(X_n, X_{n-1}, \ldots),$$

because F_n is $\sigma(X_n, X_{n-1}, \ldots)$-measurable. Moreover, $(F_n)_{n\in\mathbb{Z}}$ is stationary, so that

$$\Phi_n(X_n, X_{n-1}, \ldots) = F_n \overset{d}{=} F_0 = \Phi_0(X_0, X_{-1}, \ldots).$$

But $(X_n)_{n\in\mathbb{Z}}$ is also stationary, and thus $\Phi_0(X_0, X_{-1}, \ldots) \overset{d}{=} \Phi_0(X_n, X_{n-1}, \ldots)$. Then we note $\Phi := \Phi_0$ and get the proof. □

Lemma 2. *Under Model 2, the ideal forecast with respect to the filtration $(\mathcal{F}_n)_{n\geq\mathbb{Z}}$ and with lead time $T \geq 1$ can be written as*

$$F^*_{n,T} = \Phi^*_T(X_n, X_{n-1}, \ldots) \ a.s.,$$

where $\Phi^*_T\colon \mathcal{X}^{\mathbb{N}} \to \mathcal{P}(\mathcal{Y})$ is a measurable map.

Proof. By definition, we have $F^*_{n,T} = \mathcal{L}(Y_{n+T}|\mathcal{F}_n) = F^{\mathcal{F}_n}_{Y_{n+T}}$ for all $n \in \mathbb{Z}$. Standard properties of conditional distributions then give that

$$F^*_{n,T} = F^{(X_n, X_{n-1}, \ldots)}_{Y_{n+T}}(X_n, X_{n-1}, \ldots).$$

See e.g. Kallenberg (1997, Chapter 5). Now, $(X_n)_{n\in\mathbb{Z}}$ and $(Y_n)_{n\in\mathbb{Z}}$ are strictly stationary (since $(U_n)_{n\in\mathbb{Z}}$ is so), and one thus gets for all $k \in \mathbb{Z}$,

$$\mathcal{L}(Y_{n+T} \mid X_n, X_{n-1}, \ldots) = \mathcal{L}(Y_{n+T+k} \mid X_{n+k}, X_{n+k-1}, \ldots).$$

This yields

$$\Phi^*_T := F^{(X_n, X_{n-1}, \ldots)}_{Y_{n+T}} = F^{(X_{n+k}, X_{n+k-1}, \ldots)}_{Y_{n+T+k}} ,$$

implying that $F^*_{n,T} = \Phi^*_T(X_n, X_{n-1}, \ldots)$. □

5.2 Proofs for Section 3

Proof of Theorem 1. Defining $\Delta_i = S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right)$, our goal is to prove that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Delta_i \geq 0 \, .$$

Applying Proposition 1 for all $i \geq 1$ yields $\mathbb{E}[\Delta_i \mid \mathcal{F}_i] = \mathbf{div}_S(F_i, F^*_{i,T}) \geq 0$. Now, since Δ_i is $\mathcal{F}_{i+T}$-measurable, let us introduce the following decomposition as a telescopic sum

$$\Delta_i - \mathbb{E}[\Delta_i \mid \mathcal{F}_i] = \sum_{k=1}^{T} \left(\mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T+1-k}] - \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T-k}]\right) \, .$$

Defining $\delta^k_i = \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T+1-k}] - \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T-k}]$ and $M^k_n = \sum_{i=1}^n \delta^k_i$ implies

$$\frac{1}{n}\sum_{i=1}^{n} \Delta_i = \frac{1}{n}\sum_{i=1}^{n} \mathbf{div}_S(F_i, F^*_{i,T}) + \frac{1}{n}\sum_{k=1}^{T} M^k_n \, .$$

The announced results therefore follow as soon as the second term of the right hand side in the equality above is shown to converge a.s. to 0. To see this, notice that for $1 \leq k \leq T$, the sequence $(M^k_n)_{n\in\mathbb{N}}$ is a martingale with respect to the filtration $(\mathcal{F}_{n+T+1-k})_{n\in\mathbb{N}}$ and its quadratic variation is defined by

$$\langle M^k \rangle_n = \sum_{i=1}^{n} \mathbb{E}\left[\left(\delta^k_i\right)^2 \mid \mathcal{F}_{i+T-k}\right], \quad n \in \mathbb{N}.$$

It is a nondecreasing process and we denote by $\langle M^k \rangle_\infty$ its almost sure limit in $[0, +\infty]$. The strong law of large numbers for square-integrable martingales, see e.g. (Hall and Heyde, 1980, Section 2.6), implies that:

i) on the event $\langle M^k \rangle_\infty < +\infty$, the martingale M^k_n converges to a finite limit as $n \to \infty$;

ii) on the event $\langle M^k \rangle_\infty = +\infty$, the ratio $M^k_n / \langle M^k \rangle_n$ converges to 0 as $n \to \infty$.

The first case clearly implies that $M^k_n/n \to 0$ as $n \to \infty$. This also holds in the second case thanks to the Assumption (6) because $\langle M^k \rangle_n = O(n)$. As a conclusion, one gets also in ii) that $M^k_n/n \to 0$ as $n \to \infty$.

□

5.3 Proof of Section 4

Our proof of Theorem 2 rely on the inequality stated in Modeste and Dombry (2022, Proposition 3.9). This results is based on the following representation of the Energy Score divergence in the case $\alpha \in (0,2)$ and $\beta = 2$ due to Szekely (2003). For $F, G \in \mathcal{P}_\alpha(\mathbb{R}^d)$,

$$\mathbf{div}_S(F,G) = \frac{1}{C(d,\alpha)} \int_{\mathbb{R}^d} \frac{|\hat{F}(t) - \hat{G}(t)|^2}{\|t\|_2^{d+\alpha}} \, \mathrm{d}t, \tag{9}$$

with $C(d,\alpha) > 0$ and $\hat{F}$ (resp. $\hat{G}$) the characteristic function of F (resp. G) defined by $\hat{F}(t) = \int_{\mathbb{R}^d} e^{it\cdot x} \, \mathrm{d}F(x)$. Note that the subject of Modeste and Dombry (2022) is not conditionally negative kernels, but their Formula (14) shows that we consider the same object. The link between these two articles is detailed in Appendix B.

Proof of the Theorem 2. The direct implication is a consequence of this inequality present in Modeste and Dombry (2022, Proposition 3.9 and Formula (14))

$$\forall \varepsilon > 0, \ \forall n \in \mathbb{N}, \ \exists C > 0, \ W_1(F_n, G_n) \leq C\sqrt{\mathbf{div}_S(F_n, G_n)} + \varepsilon,$$

because the sequences are uniformly integrables. Moreover, the Cauchy-Schwarz inequality implies

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^n \sqrt{\mathbf{div}_S(F_j, G_j)} &= \sum_{j=1}^n \sqrt{1/n}\sqrt{\mathbf{div}_S(F_j, G_j)/n} \\
&\leq 1 \times \sqrt{\frac{1}{n}\sum_{j=1}^n \mathbf{div}_S(F_j, G_j)} \\
&\to 0.
\end{aligned}$$

Then for $\varepsilon > 0$, let $C > 0$ of the previous inequality. We deduce that

$$\limsup_{n\to+\infty} \frac{1}{n}\sum_{j=1}^n W_1(F_j, G_j) \leq C \limsup_{n\to+\infty} \frac{1}{n}\sum_{j=1}^n \sqrt{\mathbf{div}_S(F_j, G_j)} + \varepsilon = \varepsilon.$$

We now prove the converse implication and assume that the sequences $(F_n)_{n\in\mathbb{N}}$, $(G_n)_{n\in\mathbb{N}}$ are uniformly integrable and that $n^{-1}\sum_{i=1}^n W_1(F_i, G_i) \to 0$. Because, for all $t \in \mathbb{R}^d$, the functions $x \mapsto \cos(t\cdot x)$ and $x \mapsto \sin(t\cdot x)$ are $\|t\|$-Lipschitz continuous, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n \left|\hat{F}_j(t) - \hat{G}_j(t)\right| \\
&\leq \frac{1}{n}\sum_{j=1}^n \left| \int_{\mathbb{R}^d} \cos(t\cdot x) \ (F_j - G_j)(\mathrm{d}x)\right| + \left| \int_{\mathbb{R}^d} \sin(t\cdot x) \ (F_j - G_j)(\mathrm{d}x)\right| \\
&\leq \frac{2\|t\|}{n}\sum_{j=1}^n W_1(F_j, G_j) \to 0.
\end{aligned}$$

By the following inequality as a Fourier Transform is bounded by 1, we also have

$$\frac{1}{n}\sum_{j=1}^{n}|\hat{F}_j(t)-\hat{G}_j(t)|^2 \leq \frac{1}{n}\sum_{j=1}^{n}2|\hat{F}_j(t)-\hat{G}_j(t)| \to 0, \quad t\in\mathbb{R}^d.$$

Thanks to Equation (9)

$$\frac{1}{n}\sum_{j=1}^{n}\mathbf{div}_S(F_j,G_j) = \int_{\mathbb{R}^d}\frac{1}{nC(d,\alpha)}\sum_{j=1}^{n}\frac{|\hat{F}_j(t)-\hat{G}_j(t)|^2}{\|t\|^{d+\alpha-2}}\,\mathrm{d}t.$$

This is shown to converge to 0 by dominated convergence. Indeed,

$$h_n(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{|\hat{F}_j(t)-\hat{G}_j(t)|^2}{\|t\|^{d+\alpha-2}} \to 0, \quad \text{for all } t\in\mathbb{R}^d\setminus\{0\}.$$

Furthermore, the uniformly integrability of $(F_n)_{n\in\mathbb{N}}$, $(G_n)_{n\in\mathbb{N}}$ implies a first moment uniformly bounded by some constant $M>0$ so that the characteristic functions are M-Lipschitz continuous and

$$|h_n(t)| \leq \frac{M^2}{\|t\|^{d+\alpha-2}}\mathbb{1}_{\|t\|\leq 1} + \frac{2}{\|t\|^{d+\alpha}}\mathbb{1}_{\|t\|>1} \in \mathrm{L}^1(\mathbb{R}^d),$$

for $\alpha\in(0,2)$. The integrability of the dominant function comes from the following lemma (Olivier Garet, 2011, Theorem 4.12.9) based on the pushforward measure.

Lemma 3. *Let $\phi\colon\mathbb{R}\to\mathbb{R}$ be a measurable map and $\|\cdot\|$ be a norm on $\mathbb{R}^d$. The map $\phi\circ\|\cdot\|\in\mathrm{L}^1(\mathbb{R}^d)$ if and only if $t\in\mathbb{R}^+\mapsto t^{d-1}\phi(t)\in\mathrm{L}^1(\mathbb{R})$.*

□

The proof of Theorem 3 is based on this classical topological lemma.

Lemma 4. *Let $(x_n)_{n\in\mathbb{N}}$ be a real sequence. If for each extraction $(x_{n'})_{n\in\mathbb{N}}$ there is exists a second extraction $(x_{n''})_{n\in\mathbb{N}}$ such that*

$$x_{n''}\to 0,$$

then the sequence $(x_n)_{n\in\mathbb{N}}$ vanishes.

Proof. By contraposition. □

Proof of Theorem 3. Consider an extraction, as $(\hat{\theta}_{n'})_{n\in\mathbb{N}}$ is bounded, consider another extraction where $\tilde{\theta}$ is the sub-sequential limit. For the sake of clarty, we note n instead

of n''.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} S(f_{\hat{\theta}_n}(X_i), Y_{i+T}) - S(f_{\theta^*}(X_i), Y_{i+T}) &= \frac{1}{n}\sum_{i=1}^{n} S(f_{\hat{\theta}_n}(X_i), Y_{i+T}) - S(f_{\tilde{\theta}}(X_i), Y_{i+T}) \\ &+ S(f_{\tilde{\theta}}(X_i), Y_{i+T}) - S(f_{\theta^*}(X_i), Y_{i+T}) \\ &= \frac{1}{n}\sum_{i=1}^{n} S(f_{\hat{\theta}_n}(X_i), Y_{i+T}) - S(f_{\tilde{\theta}}(X_i), Y_{i+T}) \\ &+ \frac{1}{n}\sum_{i=1}^{n} S(f_{\tilde{\theta}}(X_i), Y_{i+T}) - S(f_{\theta^*}(X_i), Y_{i+T}).\end{aligned}$$

Moreover, with Assumptions (A3)-(A5)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} S(f_{\hat{\theta}_n}(X_i), Y_{i+T}) - S(f_{\tilde{\theta}}(X_i), Y_{i+T}) &\leq \frac{1}{n}\sum_{i=1}^{n} W_1(f_{\hat{\theta}_n}(X_i), f_{\tilde{\theta}}(X_i)) \\ &\leq \frac{\|\hat{\theta}_n - \tilde{\theta}\|}{n}\sum_{i=1}^{n}(1+X_i)^p \\ &\to 0\end{aligned}$$

The first term vanishes and the second is asymptotically positive by Theorem 1. As the sum is non positive, it means that

$$\frac{1}{n}\sum_{i=1}^{n} S(f_{\hat{\theta}_n}(X_i), Y_{i+T}) - S(f_{\theta^*}(X_i), Y_{i+T}) \to 0.$$

It implies that the empirical mean of the divergence vanishes with the previous lemma.

The second convergence is a direct application of Theorem 2. □

References

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Springer-Verlag.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA.

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1611–1617.

Bröcker, J. and Ben Bouallègue, Z. (2020). Stratified rank histograms for ensemble forecast verification under serial dependence. *Quarterly Journal of the Royal Meteorological Society*, 146(729):1976–1990.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer.

Dawid, A. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

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

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118.

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

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B*, 14(1):107–114.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

Hamill, T. M. and Colucci, S. J. (1997). Verification of eta-rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327.

Henzi, A., Kleger, G.-R., Hilty, M. P., Wendel Garcia, P. D., Ziegel, J. F., and for Switzerland, R.-.-I. I. (2021). Probabilistic analysis of covid-19 patients' individual length of stay in swiss intensive care units. *PloS one*, 16(2):e0247265.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Holzmann, H. and Eulert, M. (2014). The role of the information set for forecasting - with applications to risk management. *The Annals of Applied Statistics*, 8(1):595–621.

Kallenberg, O. (1997). *Foundations of modern Probability*. Springer.

Koldobsky, A. (1992). Schoenberg's problem on positive definite functions. *Algebra and Analysis*, 3.

Messner, J. W., Mayr, G. J., and Zeileis, A. (2017). Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145(1):137–147.

Modeste, T. and Dombry, C. (2022). Characterization of translation invariant MMD on $\mathbb{R}^d$ and connections with Wasserstein distances. Submitted.

Olivier Garet, A. K. (2011). *De l'intégration aux probabilités*. ellipses, 1 edition.

Raftery, A. E. and Ševčíková, H. (2021). Probabilistic population forecasting: Short to very long-term. *International Journal of Forecasting*.

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127(577):2473–2489.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536.

Schulz, B. and Lerch, S. (2022). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235 – 257.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5).

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In Hutter, M., Servedio, R. A., and Takimoto, E., editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg. Springer Berlin Heidelberg.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition.

Steinwart, I. and Ziegel, J. F. (2021). Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542.

Strähl, C. and Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electronic journal of statistics*, 11(1):608–639.

Szekely, G. J. (2003). E-statistics: The energy of statistical samples. Technical report, Bowling Green State University.

Taillardat, M., Fougères, A.-L., Naveau, P., and de Fondeville, R. (2022). Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting*.

Tiberi-Wadier, A.-L., Goutal, N., Ricci, S., Sergent, P., Taillardat, M., Bouttier, F., and Monteil, C. (2021). Strategies for hydrologic ensemble generation and calibration: On the merits of using model-based predictors. *Journal of Hydrology*, 599:126233.

Tsyplakov, A. (2011). Evaluating density forecasts: a comment. *Available at SSRN 1907799*.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: Proper scoring rules and moments. *SSRN Electronic Journal*.

Vannitsem, S., Bremnes, J., Demaeyer, J., Evans, G., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Bouallègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenković, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., and Ylhaisi, J. (2021). Statistical postprocessing for weather forecasts—review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):681–699.

Villani, C. (2009). *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Old and new.

Zastavnyi, V. (1993). Positive definite functions depending on the norm. *Russian Journal of Mathematical Physics*, 1:511–522.

A Measurability of the score

The assumption of measurability added in the definition is there for mathematical reasons. In this small Appendix, we will give some natural properties about a scoring rules involving measurability.

Definition 4. *An application S is said continuous for a distance δ on $\mathcal{L}$ if for each $y \in \mathcal{Y}$, the map $S(\cdot, y)$ is continuous for δ.*

Lemma 5. *Let $S\colon \mathcal{L} \times \mathcal{Y} \to \mathbb{R}$ be an application measurable in its second variable, if S is continuous for a metric δ such that $\mathcal{L}$ is separable then S is a score.*

Proof. Let $(F_0, F_1, F_2, \ldots)$ be a countable dense family of $\mathcal{L}$. We define for $n \in \mathbb{N}^*$

$$k_n(F) = \inf\{k \in \mathbb{N} \mid F \in B_\delta(F_k, 1/n)\},$$

where $B_\delta(F, \varepsilon)$ is the closed balled centered at point F. By density, this set is not empty. The map k_n is measurable cause

$$\{k_n(F) = k\} = B_\delta(F_k, 1/n) \setminus \bigcup_{i=0}^{k-1} B_\delta(F_i, 1/n).$$

We define

$$S_n(F,y) = \sum_{k=0}^{+\infty} \mathbb{1}_{\{k_n(F)=k\}} S(F_k, y).$$

This map is measurable by the measurability of k_n. By continuity of S and construction of k_n, we have

$$S_n(F,y) \to S(F,y), \text{ forall } F, y \in \mathcal{L} \times \mathcal{Y}.$$

Then S is limit of measurable maps, then S is measurable. □

Remark 3. If $(\mathcal{Y}, d)$ is separable, the space of probability measure $\mathcal{P}(\mathcal{Y})$ is still measurable for the Levy-Prokhorov metric which metrizes the weak convergence. This is still true when the set $\mathcal{P}_1(\mathbb{R}^d)$ is fitted with the Wasserstein distance W_1, introduced in Section 4.2. Moreover, a subset of a separable metric space remains separable.

In the Example 6, we introduced the family of Energy score. A natural extension is widely used in practice, the family of kernel score

$$S_\rho$$

We will show the continuity of these scores for the Wasserstein distance introduced in Section 4.2. We need another writing of this distance, for $F, G \in \mathcal{P}_1(\mathbb{R}^d)$,

$$W_1(F,G) = \inf_{X\sim F, Y\sim G} \mathbb{E}[\|X - Y\|], \tag{10}$$

where $X \sim F$ means that F is the distribution of the random variable X.

Proposition 2. *Let ρ be a Lipschitz continuous kernel, ie there exists $C > 0$,*

$$\forall x_1, x_2, y_1, y_2 \in \mathbb{R}^d, \ |\rho(x_1, y_1) - \rho(x_2, y_2)| \leq C(\|x_1 - x_2\| + \|y_1 - y_2\|),$$

then the score S_ρ is defined on $\mathcal{P}_1(\mathbb{R}^d)$ and is continuous for the Wasserstein distance W_1. More precisely for all $F, G \in \mathcal{P}_1(\mathbb{R}^d)$ and $y \in \mathbb{R}^d$,

$$|S_\rho(F,y) - S_\rho(G,y)| \leq 2CW_1(F,G).$$

Proof. Let $C > 0$ be a Lipschitz constant of ρ, then S_ρ is well defined on $\mathcal{P}_1(\mathbb{R}^d)$. Let $F, G \in \mathcal{P}_1(\mathbb{R}^d)$ and X, X', Z, Z' be four random variables associated with this probabilities and X (resp. Z) and X' (resp. Z') independents. For $y \in \mathbb{R}^d$, we have

$$\left| \int_{\mathbb{R}^d} \rho(x,y) \, \mathrm{d}F(x) - \int_{\mathbb{R}^d} \rho(z,y) \, \mathrm{d}G(z) \right| = |\mathbb{E}[\rho(X,y) - \rho(Z,y)]| \leq C\mathbb{E}[\|X - Z\|]$$

and

$$\left| \int_{\mathbb{R}^d \times \mathbb{R}^d} \rho(x,x') \, \mathrm{d}F(x)\mathrm{d}F(x') - \int_{\mathbb{R}^d \times \mathbb{R}^d} \rho(z,z') \, \mathrm{d}G(z)\mathrm{d}G(z') \right| = |\mathbb{E}[\rho(X,X') - \rho(Z,Z')]|$$
$$\leq C\mathbb{E}[\|X - Z\| + \|X' - Z'\|].$$

As we do not do assumption on the dependence between X (resp. X') and Z (resp. Z'), we conclude with the formulation (10) of the Wasserstein distance that

$$\left|\int_{\mathbb{R}^d} k(x,y)\ \mathrm{d}F(x) - \int_{\mathbb{R}^d} k(z,y)\ \mathrm{d}G(z)\right| \leq CW_1(F,G)$$
$$\left|\int_{\mathbb{R}^d\times\mathbb{R}^d} k(x,x')\ \mathrm{d}F(x)\mathrm{d}F(x') - \int_{\mathbb{R}^d\times\mathbb{R}^d} k(z,z')\ \mathrm{d}G(z)\mathrm{d}G(z')\right| \leq 2CW_1(F,G).$$

Then

$$|S_\rho(F,y) - S_\rho(G,y)| \leq 2CW_1(F,G).$$

□

Proposition 3. *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration and $F^*_{n,T}$ be the ideal forecast with respect to this filtration with a lead time $T \geq 1$. We consider also an admissible dynamic forecast $(F_n)_{n\in\mathbb{N}}$. If ρ is Lipschitz continuous and*

$$\sum_{i=1}^{n}\left(\int_{\mathbb{R}^d}\|y\|\ \mathrm{d}\left(F_i + F^*_{i,T}\right)(y)\right)^2 = O(n),$$

then one verifies the condition (6). This is true specially in the case where the moments are uniformly bounded

$$\sup_i \int_{\mathbb{R}^d}\|y\|\ \mathrm{d}\left(F_i + F^*_{i,T}\right)(y) < +\infty.$$

Proof. Let $k \in \{1,\dots,T\}$ and $C > 0$ be a Lipschitz constant of the kernel ρ, let's remember the definition of

$$\Delta_i = S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) \text{ and } \delta_i^k = \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T+1-k}] - \mathbb{E}[\Delta_i \mid \mathcal{F}_{i+T-k}], \text{ for } i \in \mathbb{N}$$

For $a, b \in \mathbb{R}, (a+b)^2 \leq 2(a^2+b^2)$, then with the Jensen Inequality

$$\mathbb{E}[(\delta_i^k)^2 \mid \mathcal{F}_{i+T-k}] \leq 4\mathbb{E}[\Delta_i^2 \mid \mathcal{F}_{i+T-k}].$$

The Proposition 2 rewrites this inequality in terms of Wasserstein distance

$$\mathbb{E}[(\delta_i^k)^2 \mid \mathcal{F}_{i+T-k}] \leq 16C^2\mathbb{E}[W_1^2\left(F_i, F^*_{i,T}\right) \mid \mathcal{F}_{i+T-k}] = 16C^2W_1^2\left(F_i, F^*_{i,T}\right)$$

Then by Triangular Inequality and Equation (10),

$$\begin{aligned} W_1^2\left(F_i, F^*_{i,T}\right) &\leq \left(\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-x'\|\ \mathrm{d}F_i \otimes F^*_{i,T}(x,x')\right)^2 \\ &\leq \left(\int_{\mathbb{R}^d}\|y\|\ \mathrm{d}\left(F_i + F^*_{i,T}\right)(y)\right)^2. \end{aligned}$$

So the condition (6) is checked. □

B Scoring rules and Reproducing Kernels

This Appendix aims at explaining the links between the elements of the proof of Theorem 2 and the results used in Modeste and Dombry (2022). Without being new, this gives a synthetic overview of Sejdinovic et al. (2013)'s results for self-contents of this article.

B.1 Presentation of Kernel Scores

The Energy Score defined in Example 6 is a specific case of the larger family of the so-called kernel scores. On the observation space $\mathcal{Y}$, consider a measurable kernel, $\rho\colon \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ and assume that

$$\forall y\in\mathcal{Y},\ \rho(y,y)=0. \tag{11}$$

Recall that ρ is said to be *conditionally negative definite* if it is symmetric and if for all $n\geq 2$, $(y_1,\dots,y_n)\in\mathcal{Y}^n$ and $(a_1,\dots,a_n)\in\mathbb{R}^n$ such that $\sum_{i=1}^n a_i=0$, it holds that

$$\sum_{1\leq i,j\leq n} a_i a_j \rho(y_i,y_j)\leq 0\ . \tag{12}$$

Note that a kernel ρ satisfying Assumptions (11) and (12) is necessarily non-negative. Indeed, for $x,y\in\mathcal{Y}$, these conditions yield, when $a_1=1$ and $a_2=-1$,

$$0\geq \rho(x,x)+\rho(y,y)-2\rho(x,y)=-2\rho(x,y).$$

The following subset of probability measures is then introduced for such kernels

$$\mathcal{L}_\rho:=\left\{F\in\mathcal{P}(\mathcal{Y})\mid \exists y_0\in\mathcal{Y},\ \int_{\mathcal{Y}}\rho(y,y_0)\ \mathrm{d}F(y)<+\infty\right\}, \tag{13}$$

and the score S_ρ associated with the kernel ρ is defined on $\mathcal{L}_\rho\times\mathcal{Y}$ by

$$S_\rho\colon (F,y)\mapsto \int_{\mathcal{Y}}\rho(x,y)\ \mathrm{d}F(x)-\frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Y}}\rho(x,x')F(\mathrm{d}x)F(\mathrm{d}x').$$

Remark 4. The well-defined aspect of these integrals is justified in Remark 21 of Sejdinovic et al. (2013). Note that conditions (11) and (12) are essential. We will propose a quick justification similar to the previous article but using the terminology from geostatistics. If the kernel ρ verifies conditions (11) and (12) then we can consider a Gaussian process $(B_y)_{y\in\mathcal{Y}}$ verifying

$$\forall x,y\in\mathcal{Y},\ \rho(x,y)=\mathrm{Var}(B_x-B_y).$$

The kernel ρ is then called the variogramm of the process $(B_y)_{y\in\mathcal{Y}}$. Let $F\in\mathcal{L}_\rho$ and $y_0\in\mathcal{Y}$ from definition (13),

$$\begin{aligned}\rho(x,y)&=\mathrm{Var}(B_x-B_y)=\mathrm{Var}(B_x-B_{y_0}+B_{y_0}-B_y)\\&=\mathrm{Var}(B_x-B_{y_0})+\mathrm{Var}(B_{y_0}-B_y)+2\mathrm{Cov}(B_x-B_{y_0},B_{y_0}-B_y)\\&\leq \rho(x,y_0)+\rho(y,y_0)+2\sqrt{\rho(x,y_0)}\sqrt{\rho(y,y_0)}.\end{aligned}$$

This concludes because $\mathrm{L}^1(\mathcal{Y},\mathrm{d}F)\subset\mathrm{L}^{1/2}(\mathcal{Y},\mathrm{d}F)$ as F is a finite measure. This also shows that $\overline{S}(F,G)$ is well defined for $F,G\in\mathcal{L}_\rho$.

The scoring rule S_ρ is proper as soon as the kernel ρ is continuous. This comes from Hoeffding's inequality, see Berg et al. (1984), section 7, Theorem 2.1. Note that it is not always simple to show that a kernel score is strictly proper. Sufficient conditions are discussed in Steinwart and Ziegel (2021), especially the case where $\mathcal{Y}$ is compact.

B.2 Reproducing kernel Hilbert Space

The scoring rules from a kernel can be compared to the Maximum Mean Discrepancy (MMD) of the Reproducing Kernel Hilbert Space (RKHS) theory. We will not give details of this theory but only the essence. We invite the curious reader to refer to Berlinet and Thomas-Agnan (2004), Smola et al. (2007) or Steinwart and Christmann (2008, Section 4). The idea of this theory is to embed any topological space $\mathcal{Y}$ into a Hilbert space $\mathcal{H}$, i.e. each point $y \in \mathcal{Y}$ is represented by a vector $K(y) \in \mathcal{H}$. To do this embedding, we define the scalar product between each point of the space $\mathcal{Y}$. This scalar product is represented by a kernel k, which is this time positive definite, i.e.

$$\forall n \geq 2, (y_1, \dots, y_n) \in \mathcal{Y}^n, \ (a_1, \dots, a_n) \in \mathbb{R}^n, \quad \sum_{1 \leq i,j \leq n} a_i a_j k(y_i, y_j) \geq 0$$

and not conditionally definite negative as previously. In the following, k will represent a positive definite kernel and ρ a conditionally negative definite kernel. We will see in Theorem 4 the links between these two properties. Returning to the RKHS, after embedding the space $\mathcal{Y}$ into a Hilbert space $\mathcal{H}$, one can embed a set $\mathcal{L} \subset \mathcal{P}(\mathcal{Y})$ into this same Hilbert. Thus each measure $F \in \mathcal{L}$ is represented by a vector $K(F)$. This idea allows then to compare two measures $F, G \in \mathcal{L}$ through the Hilbert space, i.e.

$$d_k(F, G) = \|K(F) - K(G)\|_{\mathcal{H}}.$$

The function d_k comparing these two measures is called the MMD. It has another more explicit form in terms of the kernel k,

$$d_k(F, G)^2 = \int_{\mathcal{Y}^2} k(x, y) \ \mathrm{d}(F - G) \otimes (F - G)(x, y).$$

Moreover, this integral is well defined for F, G in the set

$$\mathcal{L}_k := \left\{ F \in \mathcal{P}(\mathcal{Y}) \mid \int \sqrt{k(y, y)} \ \mathrm{d}F(y) < +\infty \right\}. \tag{14}$$

The links between kernel scores and this theory have already been explained in several articles (see (Sejdinovic et al., 2013, Section 4 and 5), Steinwart and Ziegel (2021)).

Theorem 4. *[(Berg et al., 1984, Section 3 Lemma 2.1.), (Sejdinovic et al., 2013, Proposition 20, Theorem 22 and Remark 23)] Let $y_0 \in \mathcal{Y}$, let ρ be a kernel verifying the Assumption (11) and (12), then the kernel defined by*

$$\forall x, y \in \mathcal{Y}, \ k(x, y) = \rho(y_0, x) + \rho(y_0, y) - \rho(x, y) \tag{15}$$

is positive definite. Moreover, this kernel is defined on a larger set of probability measures, i.e. $\mathcal{L}_\rho \subset \mathcal{L}_k$. The divergence $\mathbf{div}_{S_\rho}$ and the MMD d_k satisfy

$$\forall F, G \in \mathcal{L}_\rho, \ \mathbf{div}_{S_\rho}(F, G) = 2 d_k^2(F, G).$$

Example 9. The CRPS corresponds to the conditionally negative kernel $\rho(x, y) = |x - y|$ and is defined on $\mathcal{P}_1(\mathbb{R})$, the set of probability measures with a first absolute moment. An associated positive definite kernel is

$$k(x, y) = |x| + |y| - |x - y|.$$

Then $k(x, x) = 2|x|$ so that the MMD is defined on the space $\mathcal{P}_{1/2}(\mathbb{R})$ of probability measures with a half absolute moment. So we notice that the MMD d_k is defined for *strictly* more probability measures.

Remark 5. Following Dawid (2007), kernel scores can also be defined for positive definite kernel by

$$S_k(F, y) = \int_{\mathcal{Y}^2} k(x, x') \, \mathrm{d}(F - \delta_y) \otimes (F - \delta_y)(x, x'),$$

for $F \in \mathcal{L}_k$ defined in (14) and $y \in \mathbb{R}^d$. If the kernel k is associated with the conditionally negative kernel ρ by Equation (15), then the scoring rules S_ρ and S_k are defined on different distribution sets $\mathcal{L}_\rho \subset \mathcal{L}_k$ and

$$\forall F, G \in \mathcal{L}_\rho, \ \mathbf{div}_{S_\rho}(F, G) = \mathbf{div}_{S_k}(F, G).$$

The construction with positive definite kernels is more general since it is defined on a larger space.