Online learning with the Continuous Ranked Probability Score for ensemble forecasting

J. Thorey, V. Mallet, P. Baudin

Online learning with the CRPS for ensemble forecasting

J. Thorey[a,b], V. Mallet[b], P. Baudin[b]

[a] CEREA, joint research laboratory ENPC ParisTech – EDF R&D, Université Paris-Est, Marne-La-Vallée, France
[b] INRIA, Paris, France

Abstract

Ensemble forecasting resorts to multiple individual forecasts to produce a discrete probability distribution which accurately represents the uncertainties. Before every forecast, a weighted empirical distribution function is derived from the ensemble, so as to minimize the Continuous Ranked Probability Score (CRPS). We apply online learning techniques, which have previously been used for deterministic forecasting, and we adapt them for the minimization of the CRPS. The proposed method theoretically guarantees that the aggregated forecast competes, in terms of CRPS, against the best weighted empirical distribution function with weights constant in time. This is illustrated on synthetic data. Besides, our study improves the knowledge of the CRPS expectation for model mixtures. We generalize results on the bias of the CRPS computed with ensemble forecasts, and propose a new scheme to achieve fair CRPS minimization, without any assumption on the distributions.

Introduction

The minimization of the CRPS is a common way to drive probabilistic forecasts (Gneiting *et al.*, 2005; Junk *et al.*, 2015). From diagnostic tools to modeling techniques, Gneiting and Katzfuss (2014) review the state of the art of probabilistic forecasting. Using several forecasts, based on various models and perturbed input data, is a common way to produce probabilistic forecasts (Leutbecher and Palmer, 2008). The roots of this framework known as ensemble forecasting is reviewed by Lewis (2005). Ensemble of forecasts is the raw material of the techniques proposed in this paper.

Sequential aggregation targets optimal combinations, as thoroughly introduced in the monograph Cesa-Bianchi and Lugosi (2006). These techniques, also known under the scope of online learning, come with attractive theoretical guarantees of performance. Stoltz (2010) and Mallet *et al.* (2009); Mallet (2010)

summarized and tested these techniques on forecasts of respectively electricity consumption and ozone concentrations. Usually focused on scalar forecasting, sequential aggregation was applied to the Brier score and the quantile score by respectively Vovk and Zhdanov (2009) and Biau and Patra (2011). In this paper, we use sequential aggregation in order to target the best CRPS, with theoretical guarantees that do not require any assumptions on the forecast or observation distributions. In this sense, our method is a non-parametric post-processing method. Our techniques generate weights for each ensemble member so as to produce a linear opinion pool, also known as model mixture (Genest and McConway, 1990; Clemen and Winkler, 1999). Ranjan and Gneiting (2010) provide mathematical grounds on these combinations. Our technique was first designed to work with an ensemble of scalar forecasts. Still, it can be applied when a parameterized distribution is associated to each forecast.

In Section 1, we describe the mathematical background on the CRPS. We provide contributions related to ensemble forecasting with discrete Cumulative Distribution Functions (CDFs). Our contributions are mainly generalizations of existing results to the case of combinations of forecasts with unequal weights, in a probabilistic framework. We also provide a framework to work with classes of members, compatible with fair probabilistic evaluations. In Section 2, we detail online learning techniques, with adaptation for probabilistic ensemble forecasting based on the CRPS. In Section 3, we illustrate the notions of Section 1 with numerical experiments, and we demonstrate our algorithms with numerical examples. We summarize several useful identities involving CDFs in Appendix A.

1 Mathematical background

1.1 Bibliographical remarks

The evaluation of probabilistic forecasts is a long range discussion going on since Winkler and Murphy (1968); Savage (1971); see Dawid (2008) for a detailed bibliographical analysis, and more recently Gneiting and Raftery (2007) and Candille and Talagrand (2005) for detailed analyses. The Brier score was introduced by Brier (1950); Good (1952) to evaluate probabilistic forecasts for a given threshold and a binary observation. The Continuous Ranked Probability Score (CRPS) can be viewed as a continuous version of the Brier score (Epstein, 1969; Murphy, 1971) for any threshold.

1.2 The Continuous Ranked Probability Score (CRPS)

We want to forecast a scalar quantity y called the verification and we suppose that y admits an underlying distribution that is described by the CDF F. The CRPS is considered as a realization of a random variable, and it is defined as

$$\mathrm{CRPS}(\mathrm{G}, y) = \int \left(\mathrm{G} - \mathrm{H}_y\right)^2 , \tag{1}$$

where G is a CDF that is chosen by the forecaster in order to predict F, H is the unit (or Heaviside) step function, and $\mathrm{H}_y(x)$ indicates a centered Heaviside function $\mathrm{H}(x-y)$. The CRPS is negatively oriented, meaning that lower scores imply better performance. Gneiting and Raftery (2007) show that the CRPS may also be written as

$$\mathrm{CRPS}(\mathrm{G}, y) = \mathrm{E}(|\mathrm{X} - y|) - \frac{1}{2}\,\mathrm{E}(|\mathrm{X} - \mathrm{X}'|)\,, \tag{2}$$

where E is the expectation, and both X and X' are two random variables drawn from G. A decomposition of the average CRPS was introduced by Hersbach (2000). The decomposition of scores into divergence and uncertainty terms is explained in Bröcker (2009). The average CRPS is decomposed as follows:

$$\int \mathrm{CRPS}(\mathrm{G}, y)\mathrm{dF}(y) = \int (\mathrm{G} - \mathrm{F})^2 + \int \mathrm{F}(1 - \mathrm{F})\,, \tag{3}$$

where y is integrated over the values defined by F (using Equation 29 of Appendix A). The CRPS is a strictly proper score, which means that it is minimized on average if and only if the forecaster's choice G is equal to F. This is a straightforward observation from Equation 3.

The strict propriety of the CRPS can be compared to the non-strict propriety of the square loss (Bröcker and Smith, 2007), which reads

$$(y - \mathrm{E}(\mathrm{X}))^2 = \left(\int \mathrm{G} - \mathrm{H}_y\right)^2\,, \tag{4}$$

according to Equation 32 of Appendix A. We see that minimizing the square loss and minimizing the CRPS (Equation 1) are rather different objectives, due to the location of the square function inside or outside the integral expression. The CRPS objective is more demanding, because in this case the integration is applied to a positive function.

1.3 The ensemble CRPS

In the case of ensemble forecasting, the forecaster relies on an ensemble of M members x_m, $m \in \{1, \ldots, M\}$, to construct a CDF. The empirical CDF $\mathrm{G}^{\mathcal{E}}$ of the ensemble is a step function with jumps of heights u_m (called weights) at the members values x_m. Thus we write $\mathrm{G}^{\mathcal{E}}(x) = \sum_{m=1}^{M} u_m \mathrm{H}(x - x_m)$. In order to satisfy the definition of a CDF, the weights u_m should be nonnegative and sum to one, so that they produce a convex combination. Such weight vectors define the simplex $\mathcal{P}_M$ of $\mathbb{R}^M$. The weights u_m are to be optimized in order to minimize the CRPS.

The computation of the integral of Equation 1 is easy on step functions $\mathrm{G}^{\mathcal{E}}$.

When the CDF is a step function, we refer to the score as the ensemble CRPS:

$$\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) = \sum_{m=1}^{M} u_m |x_m - y| - \frac{1}{2} \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| \,. \tag{5}$$

The derivation of Equation 1 is detailed in Appendix B.

Without further information, the members are assumed to be i.i.d., thus the forecaster may arguably choose all weights equal to $1/M$. By definition, a scoring rule depending on the verification y and i.i.d. members x_m is fair if the average score is minimized when the members and the verification are sampled from the same distribution. Ferro *et al.* (2008) show that the ensemble CRPS is unfair due to the finite size of the ensemble. In the next section, we generalize this result to the case of unequal weights, with non identically distributed members.

The bias of the score is an important topic in our optimization framework. Indeed if our objective function is instrinsically biased, then the resulting probabilistic forecast cannot be calibrated.

1.4 Bias of the ensemble CRPS with underlying mixture model

We consider that the members x_m are independent samples from the CDFs G_m, and that y is fixed. The purpose of this section is to compare the score obtained with the step function $\mathrm{G}^{\mathcal{E}}$ averaged according to the CDFs G_m and the score obtained with the mixture model described by the average CDF $\mathrm{G} = \sum u_m \mathrm{G}_m$.

Taking the expectation with respect to the members x_m, we show that

$$\begin{aligned} \mathrm{E}(\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y)) = \int \mathrm{H}_y - 2 \sum_{m=1}^{M} u_m \mathrm{G}_m \mathrm{H}_y \\ + \sum_{m \neq k}^{M} u_m u_k \mathrm{G}_m \mathrm{G}_k + \sum_{m=1}^{M} u_m^2 \mathrm{G}_m \,, \end{aligned} \tag{6}$$

using Equation 29. The trick is that $\mathrm{H}^2(x - x_m) = \mathrm{H}(x - x_m)$, thus the average CRPS does not include G_m^2 terms but G_m terms instead. We conclude by introducing the terms $\sum_{m=1}^{M} u_m \mathrm{G}_m$ and $\sum_{m=1}^{M} u_m^2 \mathrm{G}_m^2$ in conjunction with

Equation 30, 34 and 35:

$$\begin{aligned}\mathrm{E}(\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y)) &= \mathrm{E}(|\mathrm{X} - y|) - \frac{1}{2}\mathrm{E}(|\mathrm{X} - \mathrm{X}'|) \\ &\quad + \frac{1}{2}\sum_{m=1}^{M} u_m^2\, \mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|) \qquad (7)\\ &= \mathrm{CRPS}(\mathrm{G}, y) + \frac{1}{2}\sum_{m=1}^{M} u_m^2\, \mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|)\,, \qquad (8)\end{aligned}$$

where X and X_m are random variables with CDFs G and G_m respectively. In the expectation of the ensemble CRPS, the diagonal terms $u_m^2\,\mathrm{E}\left(|\mathrm{X}_m - \mathrm{X}'_m|\right)$ are missing, because the spread of each member is assumed to be null. The absence of the diagonal terms is the cause of the bias of the ensemble CRPS.

As a consequence, the minimization of the ensemble CRPS should not be targeted because the solution of this optimization problem is not the underlying CDF of the verification. There is no contradiction with the strict propriety of the CRPS because, for the ensemble CRPS, the solution is only searched in a subspace made of step functions.

In the case of equal weights with i.i.d. members, Fricker *et al.* (2013) detailed in their Appendix why minimizing the ensemble CRPS is misleading as stated above. Ferro *et al.* (2008) exhibit a fair adjusted CRPS score, which includes correction terms to counteract the bias:

$$\begin{aligned}\mathrm{CRPS}_a(\mathrm{G}^{\mathcal{E}}, y) &= \frac{1}{M}\sum_{m=1}^{M}|x_m - y| - \frac{1}{2}\sum_{m,k=1}^{M}\frac{|x_m - x_k|}{M(M-1)} \qquad (9)\\ &= \mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) - \frac{1}{2M}\sum_{m,k=1}^{M}\frac{|x_m - x_k|}{M(M-1)}\,. \qquad (10)\end{aligned}$$

We see that rather than being a new score, the adjusted ensemble CRPS is a better estimation of the original CRPS, where the underlying distributions of the members are taken into account. In Equation 9, the dispersion of the ensemble $\mathrm{E}(|\mathrm{X} - \mathrm{X}'|)$ is estimated by $\sum_{m,k=1}^{M}|x_m - x_k|/(M(M-1))$. In other terms, the bias terms $u_m^2\,\mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|)$ of Equation 7 are taking into account in Equation 10 as $\mathrm{E}(|\mathrm{X}-\mathrm{X}'|)/M^2$, by considering that $\mathrm{E}(|\mathrm{X}_m - \mathrm{X}'_m|) = \mathrm{E}(|\mathrm{X}-\mathrm{X}'|)$.

1.5 Mixture model described by classes of members

We propose in this section a framework compatible with both ensemble forecasting and unbiased scores. In a standard model mixture design, a forecaster will assign weights to known parametric distributions (Raftery *et al.*, 2005; Grimit *et al.*, 2006). We do not want to make assumptions on distributions, thus we use a standard ensemble forecasting framework, where the members are usually assumed to be sampled from unknown CDFs. The goal of this section is to show

that despite the finite size of the ensemble, it is possible to use the CRPS by counteracting the discretization-induced bias. This framework is close to what is introduced in Fraley *et al.* (2010), however this previous work focused on Bayesian Model Averaging (BMA), and did not include considerations on the CRPS.

We assume that ensemble members are grouped into classes within which members are i.i.d. In this new setting, a class C has a weight $\mathcal{W}_C$ uniformly distributed among its members. The weight $u_m = \mathcal{W}_C/M_C$ is assigned to the mth member of the ensemble, assuming that it belongs to class C and that class C has M_C members. As an example, classes may be defined according to the rank of the members. Assuming that 10 members are available, two classes may be built by assigning the 5 members with the lowest values to the first class and the remaining members to the second class.

We introduce the CRPS using the classes. We call this score the class CRPS, and denote it

$$\begin{aligned}\mathrm{CRPS}_{\mathfrak{C}}(\mathrm{G}^{\mathfrak{C}}, y) = &\sum_{C\in\mathfrak{C}} \mathcal{W}_C \widehat{\mathrm{E}}(|\mathrm{X}_C - y|) \\ &- \frac{1}{2}\sum_{C,D\in\mathfrak{C}} \mathcal{W}_C\mathcal{W}_D \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|)\,. \end{aligned} \tag{11}$$

The terms of the class CRPS are detailed below.

For the class C, with M_C members x_c^C associated to the random variables X_C and X'_C, we have

$$\widehat{\mathrm{E}}(|\mathrm{X}_C - y|) = \sum_{c=1}^{M_C} |x_c^C - y|/M_C\,, \tag{12}$$

$$\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|) = \sum_{c=1}^{M_C}\sum_{d=1}^{M_D} |x_c^C - x_d^D|/(M_C M_D)\,, \tag{13}$$

where class D is different from class C, and

$$\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|) = \sum_{c,c'=1}^{M_C} |x_c^C - x_{c'}^C|/(M_C(M_C-1))\,. \tag{14}$$

This last quantity can be seen as the dispersion associated to the i.i.d. members of class C. Note the bias correction of $\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|)$ with the factor $M_C(M_C-1)$.

Now we show how the ensemble CRPS and the class CRPS are related. Summing among classes (which belong to the partition $\mathfrak{C}$ of the set of the members), we have

$$\sum_{C\in\mathfrak{C}} \mathcal{W}_C \sum_{c=1}^{M_C} |x_c^C - y|/M_C = \sum_{m=1}^{M} u_m |x_m - y|\,. \tag{15}$$

Then we sum inter- and intra-class dispersions to link them to inter-member

differences $|x_m - x_k|$. The key point is that inter-member differences for i.i.d. members are comprised in intra-class dispersions. We note that

$$\mathcal{W}_C^2 \sum_{c,c'=1}^{M_C} \frac{|x_c^C - x_{c'}^C|}{M_C(M_C-1)} = \frac{M_C}{M_C-1} \sum_{c,c'=1}^{M_C} \left(\frac{\mathcal{W}_C}{M_C}\right)^2 |x_c^C - x_{c'}^C|\,, \tag{16}$$

and $M_C/(M_C-1) = 1 + 1/(M_C-1)$ to obtain

$$\begin{aligned}
\sum_{C,D\in\mathfrak{C}} \mathcal{W}_C\mathcal{W}_D\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|) &= \sum_{C\neq D\in\mathfrak{C}} \mathcal{W}_C\mathcal{W}_D\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_D|) \\
&\quad + \sum_{C\in\mathfrak{C}} \mathcal{W}_C^2\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|) \\
&= \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + \sum_{C\in\mathfrak{C}} \frac{1}{M_C-1} \sum_{c,c'=1}^{M_C} \frac{\mathcal{W}_C^2}{M_C^2} |x_c^C - x_{c'}^C| \\
&= \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + \sum_{C\in\mathfrak{C}} \frac{\mathcal{W}_C^2}{M_C} \widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|) \\
&= \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + \sum_{m=1}^{M} u_m^2 \widehat{\mathrm{E}}(|\mathrm{X}_{C_m} - \mathrm{X}'_{C_m}|)\,,
\end{aligned}$$

where C_m is the class in which x_m falls. To obtain the last equation, consider that $\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|)$ is counted M_C times.

Compared to the ensemble CRPS, the class CRPS admits M additional terms corresponding to the dispersion of each member and resulting from the classes definition:

$$\mathrm{CRPS}_{\mathfrak{C}}(\mathrm{G}^{\mathfrak{C}}, y) = \mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) - \frac{1}{2} \sum_{m=1}^{M} u_m^2\, \mathrm{E}(|\mathrm{X}_{C_m} - \mathrm{X}'_{C_m}|)\,. \tag{17}$$

In the case of a single class, the class CRPS is equal to the adjusted ensemble CRPS described in Section 1.4.

The i.i.d. assumption on the members can be seen as too strong. The exchangeability of the members is a relaxation of the i.i.d. assumption. By definition, the joint distribution function of exchangeable members is invariant under permutation of the arguments, thus the members are indistinguishable. We refer the reader to Ferro (2014) for an analysis of fair scoring rules with the exchangeability assumption. In a few words, the user must investigate the (generally unknown) dependence structure and taylor the appropriate scoring rule accordingly. The simple case of pairwise uncorrelated members is however tractable. For the ensemble CRPS, the case of pairwise uncorrelated members is in practice equivalent to the case of i.i.d. members, because the terms $|x_m - x_k|$ rely on pairwise correlations only. In the same way for the class CRPS, the assumption of pairwise uncorrelated members within each class and independent

members between classes leads to similar results than i.i.d. members. Under the more general assumption of exchangeable members within each class, the definition of $\widehat{\mathrm{E}}(|\mathrm{X}_C - \mathrm{X}'_C|)$ should take into account the dependence between members.

Also note that these assumptions are only needed to counter the bias in the ensemble CRPS. Our aggregation methods still remain applicable without such correction. The theoretical bounds described in the next section do not rely on any stochastic assumption on the prediction data and the verifications. The assumptions of i.i.d. members and the use of the class CRPS should only guide the choice of a loss function.

2 Online learning methods

2.1 Theoretical background

Up to this section, a single time t was considered. Now we introduce online learning techniques. In this setting, the forecaster receives prediction data $\boldsymbol{\mathcal{D}}_t$ and wishes to produce the best prediction of y_t. In our case, prediction data are ensemble members and the algorithm gives a rule to compute the weights $u_{m,t}$ before each forecast time t. This rule takes into account only past information, and is therefore called the update rule. The goal of a given online learning algorithm is to provide the best possible weights according to a chosen loss function, for example the ensemble CRPS

$$\ell_t^{CRPS_\mathcal{E}}(\boldsymbol{u}) = \int \left(\sum_{m=1}^{M} u_m \mathrm{H}_{m,t} - \mathrm{H}_{y_t} \right)^2 , \tag{18}$$

written above for time t. The notation $\ell_t(\boldsymbol{u})$ emphasizes the importance of the weights, as opposed to the ensemble members and the verifications which are assumed to be given to the forecaster.

In practice, the algorithm reads

Initialization: $\boldsymbol{u}_1$;

For each time index $t = 1, 2, ..., T$

1. get prediction data $\boldsymbol{\mathcal{D}}_t$,
2. compute the forecaster's choice with $\boldsymbol{\mathcal{D}}_t$ and $\boldsymbol{u}_t$,
3. get the verification y_t and compute $\boldsymbol{u}_{t+1}$, based on the update rule.

The initial weight vector $\boldsymbol{u}_1$ is arbitrarily set, e.g., to $[1/M, \ldots, 1/M]^\top$.

The performance of an update rule comes with theoretical guarantee, where the forecaster's results are assessed against a reference, which is usually the best forecast with weights constant in time, called the oracle. An important aspect of these theoretical guarantees is that they come without any stochastic

assumption on the prediction data and the verifications. In this paper, the theoretical guarantees are regret bounds of the form

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}_M} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leq o(T)\,, \tag{19}$$

where ℓ_t is assumed to be bounded. The bound of ℓ_t can be arbitrarily small or large, so that this restriction is compatible with essentially all real world applications. Averaging the losses in time (i.e., dividing by T) shows that an algorithm giving the weights $\boldsymbol{u}_t$ is guaranteed to perform at least as well as any mixture model with weights constant in time and based on the same prediction data. This includes any individual forecast and any subset ensemble with uniform weights.

We now consider two algorithms: the online ridge regression and the exponentiated gradient method (EG). We introduce these methods in a general framework, and we show how the methods can be applied to the case of the CRPS. For the algorithm run with ensemble CRPS, a weight is explicitly given to each member. The quantities $|x_{m,t}-y_t|$ and $|x_{m,t}-x_{k,t}|$ are explicitly used in the minimization process. For the algorithm run with class CRPS, equal weights are given to all the members within a class. The weights $\mathcal{W}_{C,t}$ are computed using the terms $\widehat{\mathrm{E}}(|\mathrm{X}_{C,t}-y_t|)$ and $\widehat{\mathrm{E}}(|\mathrm{X}_{C,t}-\mathrm{X}_{D,t}|)$. Combining parameterized distributions is also possible with online learning techniques. It necessitates to compute the quantities $\mathrm{E}(|\mathrm{X}_{m,t}-y_t|)$ and $\mathrm{E}(|\mathrm{X}_{m,t}-\mathrm{X}_{k,t}|)$. These quantities are tractable from the CDFs using Equation 30. They are computed in Grimit *et al.* (2006) for a Gaussian mixture distribution.

2.2 Ridge regression

The approach of the ridge regression can be directly expressed in terms of minimization. The update rule for time $t+1$ and based on the loss ℓ is

$$\boldsymbol{u}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^M} \lambda \boldsymbol{w}^\top \boldsymbol{w} + \sum_{t'=1}^{t} \ell_{t'}(\boldsymbol{w})\,. \tag{20}$$

The regularization term with parameter $\lambda \geq 0$ controls the 2-norm of the weight vector. It is possible to add discount factors in the sum of the past losses, in order to give more importance to recent timesteps. At first sight, the ridge regression does not constrain the weights to be positive or sum to one. In practice, for the CRPS, we observed that these constraints are approximately satisfied after a spin up period. Other regularization terms of the form $\lambda(\boldsymbol{w}-\boldsymbol{u}_1)^\top(\boldsymbol{w}-\boldsymbol{u}_1)$ may also be used with arbitrary reference vector $\boldsymbol{u}_1 \in \mathcal{P}_M$. The reader interested in recent advances in online regularized regression is addressed to Orabona *et al.* (2015).

For a given experiment length T, for any vector $\boldsymbol{u} \in \mathcal{P}_M$, and if the CRPS

Method	Gradient loss
Ensemble CRPS	$\widetilde{\ell}_{m,t} = \lvert x_{m,t} - y_t \rvert - \sum_{k=1}^{M} u_{k,t} \lvert x_{m,t} - x_{k,t} \rvert + y_t - \sum_{k=1}^{M} u_{k,t} x_{k,t}$
Class CRPS	$\widetilde{\ell}_{C,t} = \widehat{\mathrm{E}}(\lvert \mathrm{X}_{C,t} - y_t \rvert) - \sum_{D \in \mathfrak{C}} \mathcal{W}_{D,t} \widehat{\mathrm{E}}(\lvert \mathrm{X}_{C,t} - \mathrm{X}_{D,t} \rvert) + y_t - \widehat{\mathrm{E}}(\mathrm{X}_t)$
CRPS for general mixture models	$\widetilde{\ell}_{m,t} = \mathrm{E}(\lvert \mathrm{X}_{m,t} - y_t \rvert) - \sum_{k=1}^{M} u_{k,t} \, \mathrm{E}(\lvert \mathrm{X}_{m,t} - \mathrm{X}_{k,t} \rvert) + y_t - \mathrm{E}(\mathrm{X}_t)$

Table 1: Formulae of the loss gradients. Equations from Appendix A are used for the simplifications. The terms of the form $y_t - \mathrm{E}(\mathrm{X}_t)$ do not impact the computation of the weights for EG, because they are independent of the member m or the class C.

losses $\ell_t(\boldsymbol{u}_t)$ are bounded, we have:

$$\mathcal{R}_T(\boldsymbol{u}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \leq \mathcal{O}\left(\ln T\right) , \tag{21}$$

so that the so-called regret $\mathcal{R}_T(\boldsymbol{u})$ is sublinear.

The Appendix C details technical aspects, such as the proof for the bound 21, as well as guidelines to compute the weights. The ridge regression applied to the square loss $(\mathrm{E}(\mathrm{X}) - y)^2$ gives a similar regret bound in terms of square losses. Our proof for the CRPS is inspired from the proof of the regret bound for the square loss, concisely described by Cesa-Bianchi and Lugosi (2006). We were helped by the work of Mallet *et al.* (2007), who demonstrated the case of multiple verification locations (also called stations) for the square loss. Our work transposes the results for the square loss with multiple locations to multiple Brier score with different thresholds, and to the CRPS.

2.3 Exponentiated gradient

Let the learning rate η be strictly positive, EG follows a multiplicative update rule of the form:

$$u_{m,t+1} = \frac{u_{m,t} \exp\left(-\eta \widetilde{\ell}_{m,t}\right)}{\sum_{m'=1}^{M} u_{m',t} \exp\left(-\eta \widetilde{\ell}_{m',t}\right)} , \tag{22}$$

where

$$\widetilde{\ell}_{m,t} = \frac{\partial \ell_t}{\partial u_m}(\boldsymbol{u}) . \tag{23}$$

This update relates to Bayesian inference (Catoni, 2004; Audibert *et al.*, 2009). The algorithm EG admits a formulation in terms of cost function minimization, where the regularization function is the entropy function, also known as the Kullback-Leibler divergence (Kivinen and Warmuth, 1997). The EG algorithm reads:

$$\boldsymbol{u}_{t+1} = \underset{\boldsymbol{w}\in\mathcal{P}_M}{\operatorname{argmin}} \sum_{m=1}^{M} w_m \ln(\frac{w_m}{u_{m,t}}) + \eta w_m \widetilde{\ell}_{m,t} \,. \tag{24}$$

Examples of loss gradients are provided in Table 1. The loss gradient $\widetilde{\ell}_{m,t}$ of the CRPS has two main terms: (i) $\mathrm{E}(|\mathrm{X}_{m,t} - y_t|)$ accounting for the distance between the verification and the mth random variable $\mathrm{X}_{m,t}$, and (ii) the weighted sum of $\mathrm{E}(|\mathrm{X}_{m,t} - \mathrm{X}_{k,t}|)$ accounting for distances between $\mathrm{X}_{m,t}$ and the $\mathrm{X}_{k,t}$. The first term controls a deviation from the median of the underlying distribution of the verifications, and the second term controls the dispersion of the mixture model. On average (on the verifications), the loss gradients are null if the verifications are correctly described by the forecaster's CDF.

The advantage of using the loss gradients is described (at least) in Devaine *et al.* (2013). In a few words, using the loss gradients makes the algorithm competitive against the best convex combination with constant weights, whereas simply using the loss $\ell_{m,t} = \mathrm{E}(|\mathrm{X}_{m,t} - y_t|) - 0.5\,\mathrm{E}(|\mathrm{X}_{m,t} - \mathrm{X}'_{m,t}|)$ would make the algorithm compete only against the best member. We insist on the fact that using the loss gradients provides the terms $\mathrm{E}(|\mathrm{X}_{m,t} - \mathrm{X}_{k,t}|)$ which are critical for the control of the ensemble spread.

The theoretical guarantee for EG states that, if the loss function ℓ is convex with respect to $\boldsymbol{u}$ and admits a subgradient, and if the losses $\widetilde{\ell}_{m,t}$ are bounded within a constant interval $[-a, a]$, then we have:

$$\sup \left[\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \inf_{\boldsymbol{u}\in\mathcal{P}_M} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \right] \leq \frac{\ln M}{\eta} + \eta \frac{a^2}{2} T \,, \tag{25}$$

where the supremum is taken for all possible values of the members $x_{m,t}$ and the verifications y_t, and η is the learning rate (Devaine *et al.*, 2013). For optimized values of η proportional to $1/\sqrt{T}$, the regret is sublinear. The theoretical guarantee of Equation 25 is verified for the square loss and for the CRPS.

3 Numerical example

3.1 Simple model

We use the simple model described in Bröcker (2012). The model is supposed to mimic local temperatures. We chose this model because the uncertainty terms are known, consequently we can easily draw conclusions from numerical tests.

We built the verifications y_t from the exact time series

$$a_t = (A sin(\pi\omega_1 t) + B sin(\pi\omega_2 t))^2 \,, \tag{26}$$

Table 2: Parameters of the numerical experiment.

s_1	s_2	A	B	ω_1	ω_2	T
0.3	0.3	1.68	0.336	1/365.25	1/11	730

combined with multiplicative and additive perturbation terms:

$$y_t \sim a_t(1 + s_1\mathcal{N}(0,1)) + s_2\mathcal{N}(0,1)\,. \tag{27}$$

Each term $\mathcal{N}(0,1)$ represents an independent Gaussian noise with zero mean and a variance of one. The perturbation terms are sampled independently at each timestep. The parameters are summarized in Table 2.

The members are sampled as

$$x_{m,t} \sim a_t(1 + s_1\mathcal{N}(0,d_{ens})) + s_2\mathcal{N}(0,d_{ens}) \tag{28}$$

analogously to the verification distribution, but the standard deviation d_{ens} describing the perturbations terms may differ from its optimal value (i.e., 1). The parameter d_{ens} is also referred to as the dispersion parameter.

3.2 Experiments without online learning

In this first experiment, ensembles are built for different values of the dispersion parameter d_{ens}. The members are drawn independently, and the weights of the members are taken constant and all equal to $1/M$. As expected, the adjusted ensemble CRPS gets the lowest value when the ensemble shows the correct spread (i.e., for $d_{ens} = 1$), see Figure 1. On the contrary, the best (non adjusted) ensemble CRPS is obtained for under-dispersed ensembles $d_{ens} < 1$. The shift of the ensemble CRPS minimum from the ideal location $d_{ens} = 1$ is larger for ensembles of small size, because the bias of the ensemble CRPS is proportional to $1/M$. This is a direct illustration of the bias due to the limited size of the ensemble explained in Section 1.4.

3.3 Experiments with weight updates

Now we test online learning techniques and more specifically their ability to discriminate between members. We build an ensemble of $M = 10$ members, that is composed of two subensembles, or classes, of equal size. The first subensemble is defined by the same distribution than the verifications. The second subensemble follows a distribution controlled by d_{ens}. If $d_{ens} = 1$, then the whole ensemble is correctly dispersed. In other words, half of the members follow the correct distribution, while the second half can follow a different distribution.

An example of the temporal evolution of the weights is given in Figure 2. We used the algorithm EG ($\eta = 0.05$) with the gradients of the ensemble CRPS. At the middle of the experiment, we swap the dispersion parameters of the

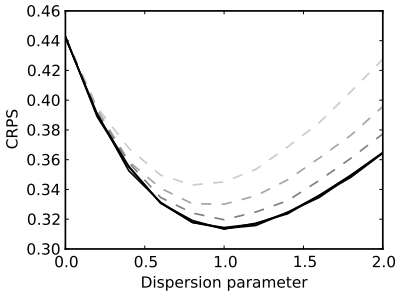

Figure 1: Ensemble CRPS (dotted gray) and adjusted ensemble CRPS (solid black), for ensembles of various sizes (10, 20, 50) from light gray to dark gray. The dispersion parameter (x-axis) is d_{ens}. The scores are averaged over nearly 200 years of data (73,000 timesteps). The (solid black) lines of the adjusted CRPS are approximately at the same location for all ensemble sizes.

members. Correct members become incorrect members and conversely. The members with incorrect dispersion parameter ($d_{ens} = 1.5$) see their weights decrease on average. After the swap, the weights of the newly incorrect members also decrease on average. The impact of the learning rate is shown in Figure 3, where a larger value $\eta = 0.2$ leads to a faster evolution of the weights. Note the difference of scales between Figures 2 and 3.

Now we show the average weight of the second subensemble parameterized by d_{ens} for different learning algorithms. Here we did not include a change of the dispersion parameter at mid-experiment. The first subensemble therefore remains the correct one all the time. The discrimination procedure tests whether the algorithm makes a difference between the subensembles and whether the incorrect members are given lower weights than the correct members.

We show the importance of the CRPS gradients in EG for probabilistic forecasting. We show in Figure 4 the average weights of EG using: (i) $\ell_{m,t} = |x_{m,t} - y_t|$, using the CRPS without the gradients; or (ii) $\check{\ell}_{m,t} = 2(\boldsymbol{u}_t^\top \boldsymbol{x}_t - y_t)x_{m,t}$, using the gradients of the square loss $(\boldsymbol{u}_t^\top \boldsymbol{x}_t - y_t)^2$, instead of the CRPS gradients $\widetilde{\ell}_{m,t}$. We see that in either case, the members with the lowest dispersion parameter are the most weighted. The members with the correct distribution receive the highest weights when the incorrect members are overdispersed ($d_{ens} > 1$). Formulation (i) and (ii) do not tend to forecast the distribution of the verifications, but only the mean or the median of the distribution of the verifications. These formulations are therefore not suited for probabilistic forecasting, as opposed to the CRPS gradients (see below). Note that we can rewrite $\check{\ell}_{m,t} = (x_{m,t} - y_t)^2 - (x_{m,t} - \boldsymbol{u}_t^\top \boldsymbol{x}_t)^2$ plus terms independent of m. Thus using the gradients (or equivalently trying to get the best combination) is a diversification strategy compared to simply using $(x_{m,t} - y_t)^2$.

Using the same representation, the algorithms EG and the ridge regression are tested with the ensemble CRPS and the class CRPS, see Figure 5. The

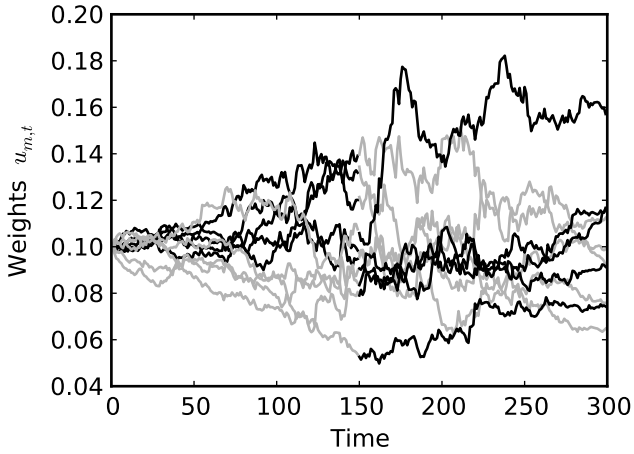

Figure 2: Temporal evolution of the weights $u_{m,t}$, with learning rate $\eta = 0.05$. The weights of members with correct dispersion are in black, and the weights of members with the incorrect dispersion $d_{ens} = 1.5$ are in gray.

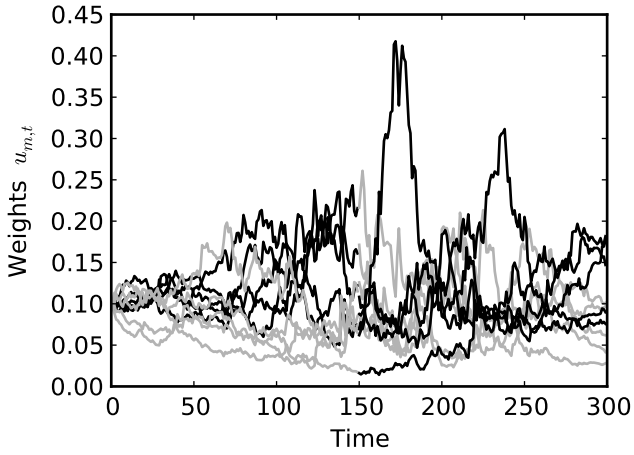

Figure 3: Temporal evolution of the weights $u_{m,t}$, with learning rate $\eta = 0.2$. The weights of members with correct dispersion are in black, and the weights of members with the incorrect dispersion $d_{ens} = 1.5$ are in gray.

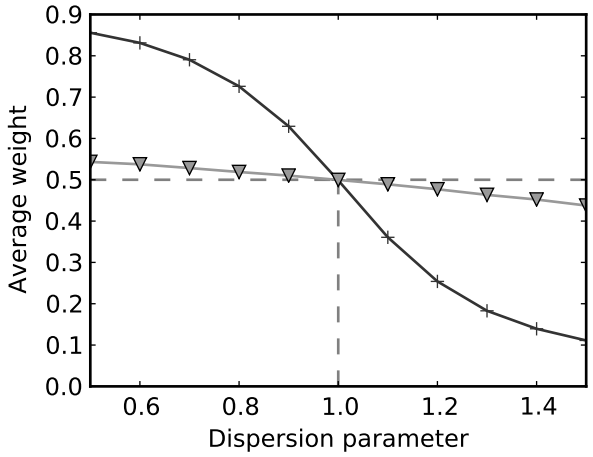

Figure 4: Average cumulated weights of the members with (possibly) incorrect dispersion parameter d_{ens} (x-axis). The black crosses indicate that the CRPS of each member are used in EG (i). The light gray triangles indicate that the square loss gradients are used in EG (ii). This figure shows that not using the CRPS gradients favors the less dispersed members, even though they do not show the correct dispersion. The experiment of roughly ten years is repeated 200 times for each dispersion parameter. We used the learning parameters $\eta = 0.05$ and $\lambda = 0.5$.

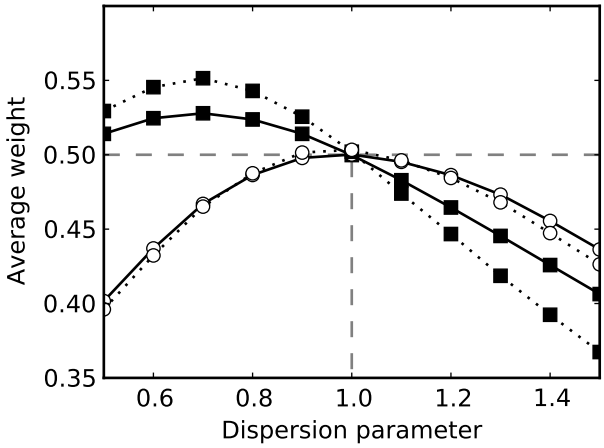

Figure 5: Average cumulated weights of the members with (possibly) incorrect dispersion parameter. Both learning algorithms based on the CRPS are tested: EG (solid line) and ridge (dotted lined). The white circles indicate that the algorithm is run for class CRPS (equal weights within the class) and the black squares indicate that the weights are computed explicitly for each member.

algorithms based on the class CRPS show correct discrimination: whatever the dispersion parameter of the wrongly dispersed members, the class with incorrect dispersion shows smaller weights on average. The sum of the weights attributed to the incorrect members stays below 0.5 (equal weights between the two subensembles). On the contrary, the algorithms based on the ensemble CRPS does not give a correct discrimination. When the dispersion parameter d_{ens} is close to 0.70, the underdispersed members receive larger weights than the correct members. We see that the minimization of the ensemble CRPS is misleading for an ensemble of small size. We interpret these results as direct consequences from the bias of the ensemble CRPS described in Section 1.4.

Conclusion

We introduced new tools for probabilistic forecasting using an ensemble of forecasts. Our algorithms use online learning techniques to produce forecast combinations that tend to minimize the CRPS. In the long run, they guarantee that the performance of the weighted ensemble is at least as good as the performance of the best weighted ensemble with weights constant in time. This theoretical guarantee holds without any assumptions on the distributions of the forecasts and verifications. In this sense, our method is a non-parametric post-processing method.

A new framework using classes of members is introduced in order to coun-

teract the bias in the ensemble CRPS. With this framework and the proposed algorithms, numerical tests showed that our online learning techniques tend to give higher weights to the forecasts with the same distribution as the verifications.

The algorithms should now be tested against real data, in order to assess their potential in operational applications against Bayesian model averaging (BMA) or other post-processing techniques. The work of the forecaster is then to obtain numerous forecasts to combine. The methods do not require any assumptions on the forecasts to be applied (bias, spread, or any other stochastic or deterministic assumptions). However, some good practices may be applied to improve the overall performance. For example, the forecasts can be altered before their inclusion in the ensemble, or additional forecasts may be derived from the raw ensemble. Also, it is recommended to draw ensembles with enough spread, so that they encompass the verifications. We argue that for most applications, the use of a multimodel ensemble combined with several post-processing techniques is an efficient way to obtain an ensemble to be calibrated with our algorithms. From a meteorological point of view, new members can be added to the ensemble by using nearby grid-points or time-shifted forecasts. This approach may be particularly efficient to account for the ability of a forecasting system to predict an event, but at the wrong time or location.

On theoretical side, a next step could be the inclusion of the uncertainty in the verifications. Also, other non local strictly proper scoring rules could serve as loss function.

Acknowledgements

The EDF company (Électricité de France) and ANRT (Association nationale de la recherche et de la technologie) are sincerely acknowledged for providing thesis fellowship and supporting this work. We also express our warm thanks to Gilles Stoltz (CNRS) and Yannig Goude (EDF) for their remarks and involvement.

Appendix A Identities implying CDFs

Let the random variable Z be described by the probability density function K' and the CDF K. We have for any real number x:

$$\mathrm{E}(\mathrm{H}(x-\mathrm{Z})) = \int \mathrm{K}'(\mathrm{Z})\mathrm{H}(x-\mathrm{Z})\mathrm{dZ} = \mathrm{K}(x)\,, \tag{29}$$

or equivalently $\mathrm{E}(\mathrm{H}_\mathrm{Z}) = \mathrm{K}$. The demonstration of the strict propriety of the CRPS uses this property for the integration over the CDF of the verifications.

Let X and Z be two random variables described respectively by the CDFs G and K. We have:

$$\mathrm{E}(|\mathrm{X}-\mathrm{Z}|) = \int \mathrm{G}(1-\mathrm{K}) + \mathrm{K}(1-\mathrm{G})\,. \tag{30}$$

For $\mathrm{G} = \mathrm{K}$, the above quantity is the Gini mean difference, which is thoroughly introduced in the monograph of Yitzhaki and Schechtman (2012).

The product GK of CDFs is itself the CDF of the random variable $\max(\mathrm{X}, \mathrm{Z})$. This can be used to explain simply Equation 30, using:

$$2\max(a, b) = |a - b| + a + b\,, \tag{31}$$

for any $(a, b) \in \mathbb{R}^2$, and

$$\mathrm{E}(\mathrm{Z}) = \int_{-\infty}^{+\infty} (\mathrm{H}(x) - \mathrm{K}(x))\mathrm{d}x\,. \tag{32}$$

Let $\mathrm{G} = \sum_{i=1}^{I} u_i \mathrm{G}_i$ and $\mathrm{K} = \sum_{j=1}^{J} w_j \mathrm{K}_j$ be CDFs of mixture models with respectively I and J components, i.e., the G_i and K_j are CDFs, and the weight vectors $\boldsymbol{u}$ and $\boldsymbol{w}$ respectively belong to the simplexes $\mathcal{P}_I$ and $\mathcal{P}_J$. Let X, X_i, Z and Z_j be random variables respectively following G, G_i, K, and K_j. We have

$$\mathrm{E}(|\mathrm{X} - \mathrm{Z}|) = \sum_{i=1}^{I}\sum_{j=1}^{J} u_i w_j\, \mathrm{E}(|\mathrm{X}_i - \mathrm{Z}_j|)\,, \tag{33}$$

based on Equation 30. Indeed,

$$\begin{aligned}
&\sum_{i=1}^{I}\sum_{j=1}^{J} u_i w_j \int (\mathrm{G}_i(1 - \mathrm{K}_j) + \mathrm{K}_j(1 - \mathrm{G}_i)) \\
&= \int \sum_{i=1}^{I} u_i \mathrm{G}_i(1 - \sum_{j=1}^{J} w_j \mathrm{K}_j) + \sum_{j=1}^{J} w_j \mathrm{K}_j(1 - \sum_{i=1}^{I} u_i \mathrm{G}_i) \\
&= \mathrm{E}(|\mathrm{X} - \mathrm{Z}|)\,,
\end{aligned}$$

because the weights w_i and u_j respectively sum to one, and using the linearity of integration.

It is straightforward to use Equation 33 to show to that

$$\mathrm{E}\left(|\mathrm{X} - y|\right) = \sum_{i=1}^{M} u_i\, \mathrm{E}\left(|\mathrm{X}_i - y|\right)\,, \tag{34}$$

and that

$$\mathrm{E}\left(|\mathrm{X} - \mathrm{X}'|\right) = \sum_{i=1}^{M} u_i\, \mathrm{E}\left(|\mathrm{X}_i - \mathrm{X}|\right) = \sum_{i,j=1}^{M} u_i u_j\, \mathrm{E}\left(|\mathrm{X}_i - \mathrm{X}'_j|\right)\,, \tag{35}$$

with X' and X'_j being random variables respectively described by G and G_j.

Appendix B Computation of the ensemble CRPS

We have

$$\begin{aligned}
\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) &= \int \left(\sum_{m,k=1}^{M} u_m u_k \mathrm{H}(x - x_m)\mathrm{H}(x - x_k) \right. \\
&\quad \left. -2 \sum_{m=1}^{M} u_m \mathrm{H}(x - x_m)\mathrm{H}(x - y) + \mathrm{H}(x - y) \right) \mathrm{d}x \\
&= \sum_{m,k=1}^{M} u_m u_k (\Gamma - \max(x_m, x_k)) \\
&\quad - 2 \sum_{m=1}^{M} u_m (\Gamma - \max(x_m, y)) + \Gamma - y \\
&= - \sum_{m,k=1}^{M} u_m u_k \max(x_m, x_k) + 2 \sum_{m=1}^{M} u_m \max(x_m, y) - y \,,
\end{aligned}$$

where Γ is the upper bound of the integral. Because the weights sum to one, we get the last simplification.

We rewrite the above expression using Equation 31:

$$\begin{aligned}
\mathrm{CRPS}(\mathrm{G}^{\mathcal{E}}, y) &= -\frac{1}{2} \left(\sum_{m,k=1}^{M} u_m u_k |x_m - x_k| + 2 \sum_{m=1}^{M} u_m x_m \right) \\
&\quad + \sum_{m=1}^{M} u_m |x_m - y| + \sum_{m=1}^{M} u_m (x_m + y) - y \\
&= \sum_{m=1}^{M} u_m |x_m - y| - \frac{1}{2} \sum_{m,k=1}^{M} u_m u_k |x_m - x_k| \,, \qquad (36)
\end{aligned}$$

because the weights u_m sum to one. We highlight the fact that the diagonal terms $u_m^2 |x_m - x_m|$ are null, so that the double sum of Equation 36 is computed for $m \neq k$.

The calculus of this section can also be written with expectations and random variables using the content of Appendix A.

Appendix C Regret bound of the ridge regression with the CRPS

This section is written for general model mixtures $\mathrm{G}m, t$ and for general CDF F_t for the verifications. For simplicity, we assume that the integrals of the CRPS can be computed on an interval $[\gamma, \Gamma]$ of limited size. All the considered CDFs

hit 0 at γ and 1 at Γ, which formalizes the assumption of bounded values for the members and the verifications. Thus the considered CDFs verify $\int G_{m,t} \leq \Gamma - \gamma$.

This appendix is structured as follows: (i) we exhibit an update rule between $\boldsymbol{u}_{t+1}$ and $\boldsymbol{u}_t$; (ii) we bound the regret against a constant vector $\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$ by the regret against the best a posteriori vector $\sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_{t+1})$; (iii) we provide an interpretable regret bound by using the update rule and the convexity of ℓ_t.

The CRPS has a quadratic form

$$\ell_t(\boldsymbol{u}) = \boldsymbol{u}^\top \left(\int \mathbf{G}_t \mathbf{G}_t^\top \right) \boldsymbol{u} - 2\boldsymbol{u}^\top \int \mathrm{F}_t \mathbf{G}_t + \int \mathrm{F}_t^2 \,, \tag{37}$$

where $\mathrm{F}_t(x) = \mathrm{H}(x - y_t)$ and $\mathbf{G}_t(x)$ is the vector of the CDFs $\mathrm{G}_{m,t}(x)$.

The cost function $J_t(\boldsymbol{u}) = \lambda \boldsymbol{u}^\top \boldsymbol{u} + \sum_{t'=1}^{t} \ell_{t'}(\boldsymbol{u})$ is written in a quadratic matricial form with:

$$J_t(\boldsymbol{u}) = \boldsymbol{u}^\top \mathbf{A}_{t+1} \boldsymbol{u} - 2\boldsymbol{u}^\top \boldsymbol{b}_{t+1} + \sum_{t'=1}^{t} \int \mathrm{F}_{t'}^2 \,, \tag{38}$$

where the vector $\boldsymbol{b}_t$ is defined by:

$$\boldsymbol{b}_t = \sum_{t'=1}^{t-1} \int \mathrm{F}_{t'} \mathbf{G}_{t'} \,, \tag{39}$$

and the matrix $\mathbf{A}_t$ of size $M \times M$ is symmetrical positive-definite:

$$\mathbf{A}_t = \lambda \mathbf{I}_M + \sum_{t'=1}^{t-1} \int \mathbf{G}_{t'} \mathbf{G}_{t'}^\top \,, \tag{40}$$

with $\mathbf{I}_M$ the identity matrix. The matrix $\mathbf{A}_t$ admits an inverse which is also symmetrical positive-definite. Note the trivial recurrence relation $J_{t+1} = \ell_{t+1} + J_t$.

The weight $\boldsymbol{u}_{t+1}$ is by definition the minimizer of J_t. Simple derivation gives the equality $\mathbf{A}_t \boldsymbol{u}_t = \boldsymbol{b}_t$. In practice, the weights are found via matrix inversion. Besides, a recurrence relation can be obtained. We successively deduce:

$$\begin{aligned}
\mathbf{A}_{t+1} \boldsymbol{u}_{t+1} = \boldsymbol{b}_{t+1} &= \boldsymbol{b}_t + \int \mathrm{F}_t \mathbf{G}_t \,, \\
&= \mathbf{A}_t \boldsymbol{u}_t + \int \mathrm{F}_t \mathbf{G}_t \,, \\
&= \left(\mathbf{A}_{t+1} - \int \mathbf{G}_t \mathbf{G}_t^\top \right) \boldsymbol{u}_t + \int \mathrm{F}_t \mathbf{G}_t \,.
\end{aligned}$$

The recurrence relation holds for any quadratic definition of the loss ℓ, and is

expressed as:

$$\begin{aligned}\boldsymbol{u}_{t+1} - \boldsymbol{u}_t &= \mathbf{A}_{t+1}^{-1} \int \left(\mathrm{F}_t - \boldsymbol{u}_t^\top \mathbf{G}_t\right) \mathbf{G}_t \\ &= -\frac{1}{2} \mathbf{A}_{t+1}^{-1} \nabla \ell_t(\boldsymbol{u}_t)\,. \end{aligned} \tag{41}$$

Demonstration of the regret bound
We iteratively use the fact that $\boldsymbol{u}_{t+1}$ is the minimizer of J_t to get

$$\begin{aligned} J_T(\boldsymbol{u}) &\geq J_T(\boldsymbol{u}_{T+1}) = \ell_T(\boldsymbol{u}_{T+1}) + J_{T-1}(\boldsymbol{u}_{T+1}) \\ &\geq \ell_T(\boldsymbol{u}_{T+1}) + J_{T-1}(\boldsymbol{u}_T) \\ &\geq \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_{t+1}) + \lambda \boldsymbol{u}_1^\top \boldsymbol{u}_1\,. \end{aligned} \tag{42}$$

The nonnegativity of $\lambda \boldsymbol{u}_1^\top \boldsymbol{u}_1$ gives:

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{u}) \geq \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_{t+1}) - \lambda \boldsymbol{u}^\top \boldsymbol{u}\,. \tag{43}$$

Thus the regret can be bounded:

$$\begin{aligned} \mathcal{R}_T(\boldsymbol{u}) &= \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \\ &\leq \lambda \boldsymbol{u}^\top \boldsymbol{u} + \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}_{t+1}) \\ &\leq \lambda \boldsymbol{u}^\top \boldsymbol{u} + \sum_{t=1}^{T} \left(\nabla \ell_t(\boldsymbol{u}_t)\right)^\top (\boldsymbol{u}_t - \boldsymbol{u}_{t+1}) \\ &= \lambda \boldsymbol{u}^\top \boldsymbol{u} + \frac{1}{2} \sum_{t=1}^{T} \left(\nabla \ell_t(\boldsymbol{u}_t)\right)^\top \mathbf{A}_{t+1}^{-1} \nabla \ell_t(\boldsymbol{u}_t)\,, \end{aligned} \tag{44}$$

where we have used Equation 43, the convexity of the functions ℓ_t and Equation 41. At this point of the demonstration, one may have the feeling that a logarithm bound can be obtained, because the matrix $\mathbf{A}_t$ is a sum of t matrices, and because the logarithmic function is the primitive of the inverse function.

We define $\mathbf{Q}_t = \mathbf{A}_{t+1}^{-1/2} \mathbf{G}_t$ and $s_t = (\boldsymbol{u}_t^\top \mathbf{G}_t - \mathrm{F}_t)$, so that the symmetry of $\mathbf{A}_{t+1}^{-1/2}$ gives

$$\frac{1}{2} \left(\nabla \ell_t(\boldsymbol{u}_t)\right)^\top \mathbf{A}_{t+1}^{-1} \nabla \ell_t(\boldsymbol{u}_t) = 2 \left(\int s_t \mathbf{Q}_t\right)^\top \left(\int s_t \mathbf{Q}_t\right)\,.$$

The inequality of Cauchy-Schwartz gives

$$\begin{aligned}\left(\int s_t \mathbf{Q}_t\right)^\top \left(\int s_t \mathbf{Q}_t\right) &= \sum_{m=1}^{M} \left[\left(\int s_t \mathbf{Q}_t\right)_m\right]^2 \\ &\leq \sum_{m=1}^{M} \int s_t^2 \int [(\mathbf{Q}_t)_m]^2 \\ &= \int s_t^2 \left(\int \mathbf{Q}_t^\top \mathbf{Q}_t\right) \\ &= \ell_t(\boldsymbol{u}_t) \left(\int \mathbf{G}_t^\top \mathbf{A}_{t+1}^{-1} \mathbf{G}_t\right) . \end{aligned} \tag{45}$$

We continue with

$$\begin{aligned}\int \mathbf{G}_t^\top \mathbf{A}_{t+1}^{-1} \mathbf{G}_t &= \operatorname{Tr}\left(\int \mathbf{A}_{t+1}^{-1} \mathbf{G}_t \mathbf{G}_t^\top\right) \\ &= \operatorname{Tr}\left(\mathbf{A}_{t+1}^{-1} \int \mathbf{G}_t \mathbf{G}_t^\top\right) \\ &= \operatorname{Tr}\left(\mathbf{I}_M - \mathbf{A}_{t+1}^{-1} \mathbf{A}_t\right) \\ &\leq \ln \frac{\det \mathbf{A}_{t+1}}{\det \mathbf{A}_t} . \end{aligned} \tag{46}$$

The first equality holds with the linearity of the integration and because $\boldsymbol{z}_1^\top \mathbf{A} \boldsymbol{z}_2 = \operatorname{Tr}(\mathbf{A} \boldsymbol{z}_2 \boldsymbol{z}_1^\top)$ for any vectors $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and matrix $\mathbf{A}$. The inequality holds because $\mathbf{A}_{t+1}^{-1} \mathbf{A}_t$ is positive definite and $1 - 1/x \leq \ln x$ for any $x > 0$.

At this step of the proof, we have shown that:

$$\mathcal{R}_T(\boldsymbol{u}) \leq \lambda \boldsymbol{u}^\top \boldsymbol{u} + 2 \sum_{t=1}^{T} \ell_t(\boldsymbol{u}_t) \ln \frac{\det \mathbf{A}_{t+1}}{\det \mathbf{A}_t} . \tag{47}$$

We assume that the losses $\ell_t(\boldsymbol{u}_t)$ are bounded by $a > 0$. Then we easily reach:

$$\mathcal{R}_T(\boldsymbol{u}) \leq \lambda \boldsymbol{u}^\top \boldsymbol{u} + 2a \ln \frac{\det \mathbf{A}_{T+1}}{\lambda^M} . \tag{48}$$

The inequality of arithmetic and geometric means applied to the eigenvalues of $\mathbf{A}_{T+1}$ leads to the conclusion

$$\begin{aligned}\det(\mathbf{A}_{T+1}) &\leq \left(\frac{\operatorname{Tr} \mathbf{A}_{T+1}}{M}\right)^M = \left(\frac{M\lambda + \sum_{t=1}^{T} \operatorname{Tr} \int \mathbf{G}_t \mathbf{G}_t^\top}{M}\right)^M \\ &\leq \left(\frac{M\lambda + MT(\Gamma - \gamma)}{M}\right)^M , \end{aligned} \tag{49}$$

from which we conclude that

$$\begin{aligned}\mathcal{R}_T(\boldsymbol{u}) &\leq \lambda \boldsymbol{u}^\top \boldsymbol{u} + 2aM \ln\left(1 + \frac{T(\Gamma - \gamma)}{\lambda}\right) \\ &\leq \lambda \boldsymbol{u}^\top \boldsymbol{u} + \mathcal{O}(\ln T)\,. \end{aligned} \tag{50}$$

We logically compete against any constant vector $\boldsymbol{u}$ on the simplex so that

$$\sup_{\boldsymbol{u} \in \mathcal{P}_M} \mathcal{R}_T(\boldsymbol{u}) \leq \mathcal{O}(\ln T)\,. \tag{51}$$

□

References

Audibert JY, *et al.* 2009. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics* **37**(4): 1591–1646.

Biau G, Patra B. 2011. Sequential quantile prediction of time series. *Information Theory, IEEE Transactions on* **57**(3): 1664–1674.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1): 1–3.

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society* **135**(643): 1512–1519.

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, Smith LA. 2007. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting* **22**(2): 382–388.

Candille G, Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society* **131**(609): 2131–2150, doi:10.1256/qj.04.71, URL http://dx.doi.org/10.1256/qj.04.71.

Catoni O. 2004. Statistical learning theory and stochastic optimization, Lectures on probability theory and statistics, Ecole d'été de Probabilités de Saint-Flour XXXI–2001. *Lecture Notes in Mathematics* **1851**: 1–269.

Cesa-Bianchi N, Lugosi G. 2006. *Prediction, learning, and games*. Cambridge University Press.

Clemen RT, Winkler RL. 1999. Combining probability distributions from experts in risk analysis. *Risk analysis* **19**(2): 187–203.

Dawid A. 2008. Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST* **17**(2): 243–244, doi:10.1007/s11749-008-0118-6, URL http://dx.doi.org/10.1007/s11749-008-0118-6.

Devaine M, Gaillard P, Goude Y, Stoltz G. 2013. Forecasting electricity consumption by aggregating specialized experts. *Machine Learning* **90**(2): 231–260.

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

Ferro C. 2014. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society* **140**(683): 1917–1923.

Ferro CAT, Richardson DS, Weigel AP. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications* **15**(1): 19–24, doi:10.1002/met.45.

Fraley C, Raftery AE, Gneiting T. 2010. Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging. *Monthly Weather Review* **138**(1): 190–202.

Fricker TE, Ferro CAT, Stephenson DB. 2013. Three recommendations for evaluating climate predictions. *Meteorological Applications* **20**(2): 246–255, doi:10.1002/met.1409, URL http://dx.doi.org/10.1002/met.1409.

Genest C, McConway KJ. 1990. Allocating the weights in the linear opinion pool. *Journal of Forecasting* **9**(1): 53–73, doi:10.1002/for.3980090106, URL http://dx.doi.org/10.1002/for.3980090106.

Gneiting T, Katzfuss M. 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1**: 125–151.

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

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

Good IJ. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* : 107–114.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA. 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society* **132**(621C): 2925–2942.

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

Junk C, Delle Monache L, Alessandrini S. 2015. Analog-based ensemble model output statistics. *Monthly Weather Review* (2015).

Kivinen J, Warmuth MK. 1997. Exponentiated Gradient versus Gradient Descent for Linear Predictors. *Information and Computation* **132**(1): 1 – 63, doi:http://dx.doi.org/10.1006/inco.1996.2612.

Leutbecher M, Palmer TN. 2008. Ensemble forecasting. *Journal of Computational Physics* **227**(7): 3515–3539.

Lewis JM. 2005. Roots of ensemble forecasting. *Monthly weather review* **133**(7): 1865–1885.

Mallet V. 2010. Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research* **115**(D24303).

Mallet V, Mauricette B, Stoltz G. 2007. Description of sequential aggregation methods and their performances for ozone ensemble forecasting. Technical Report DMA-07-08, École normale supérieure de Paris.

Mallet V, Stoltz G, Mauricette B. 2009. Ozone ensemble forecast with machine learning algorithms. *jgr* **114**(D05307).

Murphy AH. 1971. A note on the ranked probability score. *Journal of Applied Meteorology and Climatology* **10**(2): 155–156.

Orabona F, Crammer K, Cesa-Bianchi N. 2015. A generalized online mirror descent with applications to classification and regression. *Machine Learning* **99**(3): 411–435.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**: 1,155–1,174.

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

Savage LJ. 1971. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**(336): 783–801.

Stoltz G. 2010. Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique* **151**(2): 66–106.

Vovk V, Zhdanov F. 2009. Prediction with expert advice for the brier game. *J. Mach. Learn. Res.* **10**: 2445–2471, URL http://dl.acm.org/citation.cfm?id=1577069.1755868.

Winkler RL, Murphy AH. 1968. "good" probability assessors. *Journal of applied Meteorology* **7**(5): 751–758.

Yitzhaki S, Schechtman E. 2012. *The gini methodology: A primer on a statistical methodology*, vol. 272. Springer Science & Business Media.