Évaluation et construction des prévisions probabilistes : Score et calibration dans un cadre dynamique

Thibault Modeste

THÈSE de DOCTORAT DE L'UNIVERSITÉ CLAUDE BERNARD LYON 1

Ecole doctorale :
Informatique et Mathématiques de Lyon ED 512
Discipline: Mathématiques

Soutenue publiquement le 22 juin 2023 par

Thibault Modeste

Évaluation et construction des prévisions probabilistes : score et calibration dans un cadre dynamique

---

devant le jury composé de :

M. Clément Dombry	Université de Franche-Comté	Directeur de thèse
Mme Anne-Laure Fougères	Université Claude Bernard Lyon 1	Directrice de thèse
M. Ivan Kojadinovic	Université de Pau et des Pays de l'Adour	Examinateur
M. Guillaume Lecué	CREST-ENSAE	Examinateur
Mme Véronique Maume-Deschamps	Université Claude Bernard Lyon 1	Présidente
M. Bertrand Michel	École Centrale Nantes	Rapporteur
Mme Johanna Ziegel	University of Bern	Rapportrice

Institut Camille Jordan
Laboratoire de recherche en mathématiques Lyon/Saint-Étienne

Évaluation et construction des prévisions probabilistes : score et calibration dans un cadre dynamique

Thibault Modeste

Thèse de doctorat

Université Claude Bernard Lyon 1

Remerciements

Tout d'abord, je tiens à remercier mes deux encadrants Anne-Laure et Clément. Les réunions hebdomadaires, et les moments passés autour d'un café ont été d'une aide précieuse tout au long de ma thèse. Les discussions et vos relectures ont permis une amélioration non négligeable des différents articles contenus dans ce manuscrit. Merci d'avoir supporté pendant trois ans mes légers caprices de rédaction et mon entêtement à aller contre vos conseils.

Un grand merci à Johanna Ziegel et Bertrand Michel d'avoir accepté de rapporter cette thèse. I would particularly like to thank Johanna for the discussion we had in Aussois at the start of my second year of PhD, which led me to the study of RKHS and MMDs. Merci aussi aux différents membres constituant ce jury de thèse, Véronique Maume-Deschamps, Ivan Kojadinovic et Guillaume Lecué pour le temps que vous m'accordez pour la présentation de mes travaux.

Je souhaite également exprimer ma reconnaissance envers mes professeurs de mathématiques, qui ont eu un impact significatif, bien que distant, sur cette thèse. Je dis donc merci à Matthieu Solnon de m'avoir obligé à rédiger correctement mes copies durant mes années prépas. Merci à Bruno Arsac d'être un modèle de pédagogie et de bienveillance. J'espère pouvoir transmettre mes connaissances de manière similaire à la vôtre.

Je voudrais aussi remercier tous mes amis de longue date ainsi que ceux rencontrés récemment aux laboratoires de l'ICJ. Tout d'abord, comme promis, merci infiniment Valentin pour les nombreux morceaux de pain que tu m'as offert durant cette thèse. Ces cadeaux quotidiens ont permis à de nombreuses reprises d'éloigner de quelques quarts d'heure mes fringales. En second lieu, je remercie Sébastoche, nous avons partagé ensemble les nombreuses interrogations que constituent l'après thèse. Nous avons aussi partagé des bons moments que ça soit à la montagne ou à la plage avec le (très) grand Clément. Ces week-ends ont toujours été très ressourçants, merci Clément de les avoir organisés à chaque fois. Au passage, je remercie tous les membres du bureau B219, Mouksit et Lorenzo, de m'avoir accepté pour ces nombreuses pauses. Je remercie aussi pour tous les bons moments, mes co-bureaux passés, Martin, Sasha et Yanni, ainsi que les actuels, Solal, Yvon, Paul, Ri, Eduardo, et les derniers venus, Grégoire et William. D'autres bureaux m'ont accueilli de manière répétée, je souhaite donc remercier Annette, Baptiste et Luca pour n'avoir jamais fermé leurs portes à clé. À Guillaume et Lucas, je suis triste de vous quitter au bout d'un an, ces discussions passionnées autour de *One Piece* me manqueront. Je n'oublie pas les autres doctorants, Antoine, Bérénice, Dimitri, Gauthier, Jules, Jorge, Léa, Louna, Marianne, Matthieu, Oskar, Simon, Thibault, Wissam avec qui j'ai pu partager des discutions amusantes ou mathématiques. Merci aussi à Régis et son calendrier mathématiques qui a rendu nos pauses déjeuners plus studieuses en cette dernière année. Merci aux doctorants rencontrés brièvement en conférence avec qui j'ai quand même pu échanger des discussions ou des fous rires de qualité, Edmond, Jean-Baptiste, Nicolas J., Nicolas L., Paul, Romain. Myriam et Narimène , je vous l'avais promis il y a longtemps des théorèmes à vos noms. Voilà c'est chose faite, je vous dédie le Théroème 5.30 et la Proposition 5.33. Petite dédicace à mes *pélos* sûrs Karim, Mehdi et Guillaume. Je vous promets de ne pas abandonner la ville de Lyon. Charlie, nous nous sommes plus quittés depuis notre rencontre au fond de la salle 305. Je te remercie pour toute l'attention et l'écoute que tu me donnes.

Merci Maman, merci Papa, de m'avoir accompagné durant cette thèse. Les déjeuners dominicaux en votre compagnie m'ont permis de me reposer, de me relaxer et donc d'avancer dans les meilleures conditions pendant trois ans. Je remercie aussi Guillaume qui a accepté de prendre des congés pour me voir soutenir.

Pour finir, merci à Alix de me supporter une semaine par mois durant ces magnifiques moments de "télétravail". J'espère sincèrement que la fin de cette thèse marquera mon arrivée dans le Sud-Ouest, me permettant ainsi de te rejoindre pour de nouvelles balades pyrénéennes.

Table des matières

Introduction et contributions

Cette thèse peut paraître au premier abord un peu décousue, quatre chapitres sur quatre domaines différents, voire très différents. Essayons de rendre tout ceci cohérent. On se place dans un cadre de prévision probabiliste. Un évènement Y n'est plus prédit par une seule valeur y mais par plusieurs. Cette forme de prévision permet de prendre en compte l'incertitude inhérente à la prédiction. Si on se place d'un point de vue météorologique, chaque valeur provient d'un scénario possible. De plus, chaque valeur est associée à une probabilité d'apparition ou à une densité probabiliste. Une prévision probabiliste prend donc la forme d'une mesure de probabilité. Cette mesure dépend elle même de nos observations passées provenant de phénomène aléatoire. Ces mesures sont donc aussi aléatoires. Formellement, une prévision probabiliste F d'un phénomène $Y \in \mathcal{Y}$ est une variable aléatoire à valeurs dans $\mathcal{P}(\mathcal{Y})$, l'ensemble des mesures de probabilités. De manière plus générale, ces prévisions sont produites à partir d'information. Mathématiquement, l'information est souvent interprétée comme la sous tribu engendrée par une variable aléatoire. Dans la suite, une tribu $\mathcal{F}$ symbolisera l'information d'un prévisionniste F. Une prévision particulière qui est étudiée dans cette thèse est la prévision dite idéale. Cette prévision représente la distribution conditionnelle F^* de Y sachant l'information $\mathcal{F}$. Intuitivement, c'est la prévision qui prédit le mieux le phénomène Y en connaissant l'information $\mathcal{F}$.

Dans ce cadre de prévision probabiliste, plusieurs questions naturelles viennent à l'esprit. Comment comparer le travail de plusieurs prévisionnistes ? Est-ce qu'il existe des procédures permettant de déterminer si l'on prédit comme F^* ? Est-ce que se comparer à cette prévision idéale est pertinent ? Comment construire des estimateurs de la prévision F^* ayant de bonnes garanties théoriques ?

Le Chapitre 2 s'intéresse aux scores, un objet permettant de comparer des prévisions. On montre dans ce chapitre que cet objet permet de rendre la prévision idéale optimale dans un certain sens qui sera détaillé dans la suite. Être ou non la prévision idéale n'est pas le seul critère pouvant décrire une prévision. Par exemple, l'estimation en moyenne des marginales, i.e. pour tout $y \in \mathbb{R}$, l'espérance de la variable $F(y)$ est la probabilité $\mathbb{P}(Y \leq y)$. Cela signifie que la prévision F estime bien *en moyenne* la fonction de répartition de Y. Un résultat bien connu est que la prévision idéale vérifie cette propriété. Plusieurs autres définitions ont été introduites dans la littérature. Le Chapitre 1, introductif, détaillera les différentes propriétés ainsi que leurs liens. Dans cette thèse, on s'intéresse essentiellement à la distribution conditionnelle. Le but du Chapitre 3 est de donner une manière effective pour déterminer si une prévision est idéale ou pas en proposant différent tests. Le Chapitre 4 détaille une manière d'approcher cette prévision idéale. On montrera que la méthode à moyenne pondérée est consistante dans un certains sens en prolongeant le résultat classique sur la régression de l'espérance conditionnelle (Stone, 1982). On finira cette thèse avec un Chapitre 5 sur l'élaboration de nouvelles distances sur l'espace des mesures $\mathcal{M}(\mathcal{Y})$, lié partiellement au Chapitre 2.

1 Fonction de Score

1.1 Présentation

Un score (Gneiting and Raftery, 2007) est un objet permettant de comparer une prévision probabiliste avec une réalisation du phénomène étudié. On peut voir le score comme une pseudo distance entre deux ensembles de natures différentes. On note $S(F, y)$ cette comparaison où F est une mesure de probabilité et y une observation. Cette distance va servir à la validation de prévisions. Pour qu'un score soit pertinent, il doit vérifier la condition suivante, une prévision juste doit minimiser le score en moyenne.

Définition 0.1. *Un score S est dit propre sur $\mathcal{L} \subset \mathcal{P}(\mathcal{Y})$ si pour tout $F, G \in \mathcal{L}$,*

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

existe et si

$$\overline{S}(G, G) \leq \overline{S}(F, G).$$

Le score est dit strictement propre si l'égalité vraie si et seulement si $F = G$.

Pour un score propre, on définit sa divergence comme

$$\mathbf{div}_S(F, G) = \overline{S}(F, G) - \overline{S}(G, G) \geq 0.$$

Par exemple, un des scores le plus utilisés en pratique est le *Continuous Ranked Probability Score*, noté CRPS (Epstein, 1969a; Hersbach, 2000; Bröcker, 2012). Ce score est strictement propre sur l'ensemble des mesures de probabilités avec un moment d'ordre 1. Pour $F \in \mathcal{P}_1(\mathbb{R})$ et $y \in \mathbb{R}$, on définit

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

où on associe une mesure avec sa fonction de répartition. Ce score possède une autre écriture plus probabiliste,

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

Un rapide calcul montre que la divergence du CRPS est la distance $\mathrm{L}^2(\mathbb{R})$ des fonctions de répartitions. Ce score appartient à une famille beaucoup plus large basée sur l'écriture (1), appelée Kernel Score. En prenant une application ρ, on définit le score

$$S_\rho(F, y) = \mathbb{E}_{X \sim F}[\rho(X, y)] - \frac{1}{2}\mathbb{E}_{X, X' \sim F \otimes F}[\rho(X, X')]. \tag{2}$$

Il existe des conditions, basées sur l'inégalité de Hoeffding, sur ρ pour que le score associé soit propre.

1.2 Résultats

Dans Holzmann and Eulert (2014), les auteurs montrent qu'un score strictement propre est un objet pertinent pour discriminer la prévision parfaite. Sachant une information $\mathcal{F}$, l'espérance conditionnelle est minimisée uniquement par la prévision idéale F^* de Y

$$\mathbb{E}[S(F, Y) - S(F^*, Y) \mid \mathcal{F}] = \mathbf{div}_S(F, F^*) \geq 0, \tag{3}$$

où F est une prévision utilisant l'information $\mathcal{F}$. Nous montrons dans le Chapitre 2 que ce résultat se généralise à un cadre dynamique. Prenons un phénomène temporel $(Y_n)_{n \in \mathbb{N}}$ et un temps d'horizon T, i.e. à l'instant n, nous voulons prédire Y_{n+T}. Nous avons le résultat suivant

Théorème 0.2. *Soit une suite de prévisions $(F_n)_{n\in\mathbb{N}}$ utilisant l'information $(\mathcal{F}_n)_{n\in\mathbb{N}}$ et soit $(F^*_{n,T})_{n\in\mathbb{N}}$ les prévisions idéales sachant cette informations. Pour un score S propre, on a sous* ***certaines conditions techniques***

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) \geq 0\ . \tag{4}$$

De plus,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) = 0 \quad a.s. \tag{5}$$

si et seulement si

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

C'est à dire que la prévision idéale minimise, presque sûrement, asymptotiquement la moyenne empirique. Ce résultat est à mettre en parallèle avec le résultat précédent, où la minimisation ne se faisait qu'en espérance. Ce théorème est obtenu sans hypothèse de stationnarité. La stationnarité rend le résultat trivial car elle impliquerait la convergence de la moyenne empirique vers

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

d'après l'équation (3). Notre résultat justifie l'utilisation de la théorie des scores pour comparer plusieurs prévisions au delà des cadres classiques indépendants et identiquement distribués (i.i.d.) ou dépendants stationnaires. Pour être complètement exploitable, il serait intéressant de l'accompagner d'un test de comparaison mais à notre connaissance, de tels tests sont disponibles uniquement dans le cadre stationnaire.

2 Calibration et prévision probabiliste

2.1 Présentation

La calibration est une autre manière d'évaluer la qualité d'une prédiction. C'est un ensemble de propriétés plus ou moins fortes décrivant une bonne prévision (Gneiting et al., 2007; Gneiting and Ranjan, 2013; Tsyplakov, 2013; Gneiting and Resin, 2021). La section 1 du prochain chapitre détaillera un peu plus la théorie. On se concentre ici sur la propriété étudiée dans le Chapitre 3, la calibration idéale, i.e. la prévision F est égale à la loi conditionnelle F^*. Le but de l'article est de proposer une famille de tests pour déterminer, dans un cadre stationnaire, si une suite de prévisions $(F_n)_{n\in\mathbb{N}}$ est idéale ou non. Dans Strähl and Ziegel (2017), les auteurs construisent des tests pour la notion de cross-calibration. Lorsque plusieurs prévisionnistes $(F^{(1)}_n, \dots, F^{(k)}_n)_{n\in\mathbb{N}}$ s'affrontent pour prédire $(Y_{n+1})_{n\in\mathbb{N}}$, la prévision $(F^{(1)}_n)_{n\in\mathbb{N}}$ est dite cross-calibrée si

$$F^{(1)}_n = \mathcal{L}(Y_{n+1} \mid F^{(1)}_n, \dots, F^{(k)}_n, Y_j,\ \text{pour } j \leq n).$$

Cela signifie que le prévisionniste $F^{(1)}_n$ utilise de manière optimale son information et celles des autres aussi. L'idée est d'écrire cette propriété à partir d'une transformation très utilisée, dans le domaine de la calibration, la *Probability Integral Transform* (PIT) (David and Johnson, 1948). Dans le cas continu, il s'agit de l'évaluation $F^{(1)}_n(Y_{n+1})$. La formule générale est donnée dans le premier chapitre. Lorsqu'une prévision est cross-calibrée alors la PIT est uniformément distribuée sur $[0,1]$ et indépendante de l'information des autres prévisions. Leurs tests s'intéressent donc à ces deux critères.

Un autre article (Bröcker, 2022) introduit un test pour la calibration idéale dans le cas particulier où Y est binaire. C'est le cas lorsque l'on étudie si un évènement se passera ou non, par exemple

savoir s'il va pleuvoir ou non. Dans le cas binaire, une prévision peut être vu comme une variable aléatoire p sur $[0,1]$, car

$$F = (1-p)\delta_0 + p\delta_1.$$

La calibration idéale peut donc s'énoncer de la manière suivante

$$\mathbb{P}(Y = 1 \mid p) = p.$$

L'idée de l'article est d'écrire cette propriété sous forme de processus empirique. Prenons une suite stationnaire $(Y_{n+1}, p_n)_{n\in\mathbb{Z}}$ où p_n est la prévision pour l'évènement Y_{n+1}. L'auteur suppose, en plus de la stationnarité, le caractère Markovien de la suite et même un peu plus,

$$\mathcal{L}(Y_{n+1} \mid p_k, Y_k, \text{ pour } k \leq n) = \mathcal{L}(Y_{n+1} \mid p_n).$$

Sous ces conditions le processus suivant

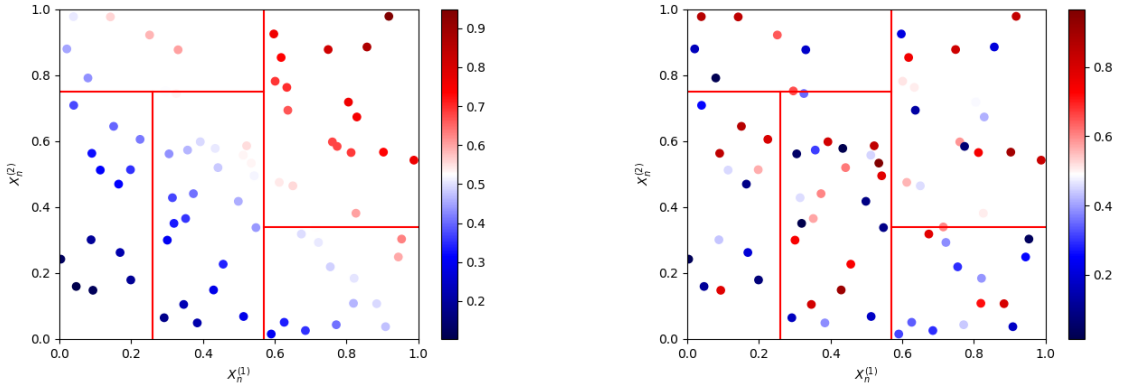
$$\mathbb{U}_n(t) = \frac{1}{n}\sum_{k=1}^{n}(Y_{k+1} - p_k)\mathbb{1}_{\{p_k \leq t\}}, \text{ avec } t \in [0,1],$$

possède de bonnes propriétés asymptotiques qui conduisent à un test de calibration.

2.2 Résultats

Pour nos tests, nous utilisons les deux idées précédentes simultanément : on se sert de la caractérisation de la calibration idéale par les PIT et nous testons si les PIT vérifient les bonnes propriétés en la faisant intervenir dans un processus empirique. Pour être plus précis, pour des prévisions $(F_n)_{n\in\mathbb{N}}$ d'un phénomène $(Y_{n+1})_{n\in\mathbb{N}}$, on note la PIT $Z_n := F_n(Y_{n+1})$. On se place dans le cadre où l'information $(\mathcal{F}_n)_{n\in\mathbb{N}}$ est engendrée par une suite $(X_n)_{n\in\mathbb{N}}$ à valeur dans $\mathbb{R}^d$ et Y_n est X_n mesurable. Ce cadre signifie que notre information provient d'observations et la condition de mesurabilité implique que l'on observe les réalisations passées du phénomène que l'on souhaite prédire. Dans ce cas-là, une suite de prévisions $(F_n)_{n\in\mathbb{N}}$ est idéalement calibrée si et seulement si la suite des PIT $(Z_n)_{n\in\mathbb{N}}$ est uniformément distribuée sur $[0,1]$ et que pour chaque $n \in \mathbb{N}$, la variable Z_n est indépendante de $(X_0, \ldots, X_n)$. Nos tests vont avoir la même méthode que ceux de l'article Strähl and Ziegel (2017), vérifier la distribution des PIT et l'indépendance avec l'information. Pour procéder à ces deux vérifications simultanément, nous allons utiliser des arbres de régression (CART Algorithm).

L'idée est de séparer le jeu de données $(X_n, Z_n)_{n\in\mathbb{N}}$ en plusieurs sous-groupes et étudier le comportement des PIT dans chaque sous-groupe. Ce découpage a pour but de créer des régions où les PIT sont le plus homogènes possibles ce qui contredit l'indépendance entre les variables $(X_n)_{n\in\mathbb{N}}$ et $(Z_n)_{n\in\mathbb{N}}$. Dans l'exemple ci-dessous, l'information $(X_n)_{n\in\mathbb{N}}$ est bidimensionnelle. Ainsi chaque point du plan est une réalisation de la variable explicative et la couleur représente la valeur de la PIT associée. Dans la première figure, il n'y a pas indépendance entre l'information et les PIT. On peut donc créer des régions où la PIT semble quasiment constante. Alors que sur la deuxième figure, sur chaque sous-groupe, les sous-échantillons semblent bien uniformément répartis. Ce fait suggère que les PIT sont bien indépendantes de l'information.



Concrètement, au premier découpage, le jeu de données est séparé en deux de manière rectangulaire. Cette découpe résulte de la minimisation de la quantité [1]

$$n(A)\mathrm{Var}(A) + n(A^c)\mathrm{Var}(A^c),$$

où $n(A)$ est le nombre de $(X_n)_{n\in\mathbb{N}}$ dans la région A et $\mathrm{Var}(A)$ est la variance des PIT de la région A. Si l'on minimise ce critère, on favorise l'apparition de zones où les PIT sont constantes. Si elles sont indépendantes des variables explicatives, il sera difficile de créer de telles régions. Puis on itère cette minimisation sur chaque sous-groupe créé jusqu'à la profondeur souhaitée. Ainsi, nous obtenons une variance pondérée du jeu de données définie par

$$\sum_{A\in R} n(A)\mathrm{Var}(A), \tag{6}$$

où R est l'ensemble des régions créées. Si cette quantité est *trop petite*, alors on estime que les prévisions $(F_n)_{n\in\mathbb{N}}$ ne sont pas idéalement calibrées. Pour définir le *trop petit*, on simule des variables booststrap $(Z_n^\star)_{n\in\mathbb{N}}$ uniformément distribuées sur $[0,1]$ et indépendantes des variables explicatives. On effectue la même procédure sur $(X_n, Z_n^\star)_{n\in\mathbb{N}}$ et on compare la vraie quantité (6) avec les quantités booststrappées. La variance pondérée intervenant dans notre test peut s'écrire comme une fonctionnelle du processus suivant

$$\mathbb{F}_n(y,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Z_i\le y\}} - y\right)\left(\mathbb{1}_{\{X_i\le t\}} - G(t)\right), \text{ pour } G \text{ la fonction de répartition des } (X_n)_{n\in\mathbb{N}}.$$

Cette écriture permet d'avoir des résultats théoriques asymptotiques de nos tests et la justification théorique du bootstrap. Tout ceci est détaillé dans le Chapitre 3.

3 Estimation de la loi conditionnelle

Après deux chapitres s'intéressant à la vérification qu'une prévision donnée soit idéale ou non, on s'intéresse à la construction effective de prévisions dans un cadre théorique offrant des garanties de convergence pour les grands échantillons. Maintenant la variable d'intérêt Y n'est plus simplement univariée mais vectorielle de dimension $k \ge 1$. Nos variables explicatives seront encore dans $\mathbb{R}^d$. Le but de ce chapitre est d'estimer la prévision idéale $F^* = \mathcal{L}(Y \mid X)$. On la notera dans cette section F_X^*. Plusieurs approches existent pour l'estimer. La plus simple est de se placer dans un cadre paramétrique. C'est-à-dire supposer des conditions sur la forme de la distribution. Un exemple basique est le cadre Gaussien. Déterminer la loi Y sachant X revient à déterminer les covariances entre les coordonnées des vecteurs X et Y. Nous allons ici considérer un cadre non paramétrique pour faire le moins d'hypothèses possibles. La méthode que l'on va étudier

1. Ce n'est pas exactement cette formule. Tous les détails exacts seront donnés dans le chapitre concerné.

est l'estimation par moyenne pondérée. Avec un échantillon $(X_i, Y_i)_{1\leq i\leq n}$, la distribution de Y sachant $X = x$ est une combinaison linéaire d'atomes δ_{Y_i},

$$\sum_{i=1}^{n} W_i(x; X_1, \dots, X_n)\ \delta_{Y_i}.$$

Intuitivement, si x est éloigné de X_i alors le poids devant le dirac δ_{Y_i} doit être faible. Nous détaillons le choix des poids dans la prochaine sous-section. Avant ceci, cette technique n'est pas la seule manière non paramétrique d'estimer la distribution conditionnelle. On peut, par exemple, parler des travaux récents de Henzi et al. (2021b); Mösching and Dümbgen (2020) dans le cas où Y est à valeurs réelles. Les auteurs supposent que la distribution conditionnelle est "conditionnellement croissante" (*isotonic* en anglais). Pour un ordre $\preceq$ sur $\mathbb{R}^d$, la distribution est conditionnellement croissante si $x \preceq x'$ implique $F_x \leq_{st} F_{x'}$, où $\leq_{st}$ désigne l'ordre stochastique des fonctions de répartition. On a $F \leq_{st} G$ si on peut trouver $U, V \sim F, G$ tel que $U \leq V$. Cette définition est équivalente à pour tout $x \in \mathbb{R}$, $F(x) \geq G(x)$.

3.1 Estimation par moyenne pondérée

L'estimation par moyenne pondérée estime F^*_X de la manière suivante

$$\hat{F}_{n,X} = \sum_{i=1}^{n} W_i(X; X_1, \dots, X_n)\ \delta_{Y_i},$$

où les poids $W_{n,i}$ sont positifs et de somme égale à 1. Donnons deux exemples de famille de poids.

- **Poids à noyaux :** Ces poids sont définis de la manière suivante

$$W_i(x; X_1, \dots, X_n) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)}, \quad \text{avec } K \text{ une fonction positive.}$$

  L'idée est de prendre $h_n \to 0$ avec K s'annulant à l'infini. Ainsi, $F_{X=x}$ est estimée à partir des Y_i dont le X_i associé est proche de x.

- **$\kappa-$plus proches voisins :** Le poids est directement relié à la notion de x proche de X_i

$$W_i(x; X_1, \dots, X_n) = \begin{cases} \frac{1}{\kappa_n} & \text{si } X_i \text{ fait partie des } \kappa_n\text{-plus proches voisins de } x \\ 0 & \text{sinon} \end{cases}.$$

  Pour que ces poids estiment bien, il faut que $\kappa_n \to +\infty$, pour considérer assez d'information, **mais pas trop vite**, pour éviter de prendre des X_i trop éloignés de x.

Une utilisation historique de l'estimation par moyenne pondérée est celle de l'espérance conditionnelle, $r(X) = \mathbb{E}[Y \mid X]$. Cette fonction de régression s'estime par l'estimateur des moyennes pondérées

$$\hat{r}_n(X) = \sum_{i=1}^{n} W_i(X; X_1, \dots, X_n)\ Y_i.$$

Le résultat de consistance bien connu présent dans Stone (1982) est

Théorème 0.3. *Sous **certaines conditions** sur les poids, on a pour tout $p \geq 1$ et couple (X, Y) vérifiant $\mathbb{E}[\|Y\|^p] < \infty$,*

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \longrightarrow 0 \quad \textit{pour } n \to +\infty.$$

3.2 Consistance de l'estimateur

Avec les mêmes conditions que le théorème de Stone, nous avons la consistance de l'estimateur de la loi conditionnelle. La consistance ici sera à comprendre au sens de convergent pour la distance de Wasserstein. Pour n grand, l'estimateur $\hat{F}_{n,X}$ sera proche pour cette distance, en moyenne, de la quantité estimée.

Théorème 0.4. *Pour $p \geq 1$ et (X, Y) vérifiant $\mathbb{E}[\|Y\|^p] < \infty$,*

$$\mathbb{E}\big[W_p^p(\hat{F}_{n,X}, F_X^*)\big] \longrightarrow 0 \quad \textit{pour } n \to +\infty.$$

Remarque 0.5. D'autres consistances sont utilisées pour l'estimation de la distribution conditionnelle. Pour n'en citer qu'une, dans Stute (1986) la consistance de l'estimateur est au sens de la distance de Kolmogorov, i.e. pour tout x

$$\sup_{y \in \mathbb{R}^d} |\hat{F}_{n,x}(y) - F_x^*(y)| \to 0.$$

Ce résultat a des applications directes sur l'estimation de fonctionnelles conditionnelles. La plus simple de ces fonctionnelles est le quantile. On cherche à estimer $F_X^{*-1}(\alpha)$ le quantile conditionnel de Y sachant X d'ordre α. Comme on possède un estimateur de la distribution conditionnelle, on va l'utiliser pour estimer cette valeur en considérant $\hat{F}_{n,X}^{-1}(\alpha)$.

Corollaire 0.6. *Pour $\alpha \in (0, 1)$, on a la consistance suivante*

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{*-1}(\alpha) \quad \textit{dans } \mathrm{L}^p.$$

4 Distances sur l'espace des mesures

Finissons cette présentation par un chapitre un peu à part dans cette thèse. Cette partie est née après l'étude des Kernel Scores (2) et de leurs divergences. Comme la divergence est une manière de distinguer les mesures, on peut la voir comme une sorte de "distance". Avec des guillemets car la divergence ne vérifie pas forcément les axiomes d'une distance. Mais on peut tout de même se poser la question sur la signification de "la suite $(\mathbf{div}_S(F_n, F))_{n\in\mathbb{N}}$ tend vers 0" où $(F_n)_{n\in\mathbb{N}}$ est une suite de mesures, déterministes, et F une mesure.

4.1 RKHS

Cette question était en fait déjà présente dans la littérature mais avec un autre vocabulaire, celui des *Reproducting Kernel Hilbert Space* (RKHS) et du *Maximum Mean Discrepancy* (MMD). Ce lien entre les Kernels Scores et les RKHS est présent dans Sejdinovic et al. (2013) et sera rappelé dans le Chapitre 2. L'idée clé de cette théorie est de représenter un espace $\mathcal{Y}$ dans un espace de Hilbert $\mathcal{H}$. Le prochain chapitre s'intéresse à cette théorie et à la technique associée, le *kernel trick*. On donnera donc plus de détails dans quelques pages. Pour l'instant, on introduit seulement quelques notions de base. Dès que l'on a une application k symétrique et définie positive appelé *noyau*, i.e. pour $x_1, \ldots, x_n \in \mathcal{Y}$,

$$\sum_{1 \leq i,j \leq n} a_i a_j k(x_i, x_j) \geq 0,$$

on peut plonger l'espace $\mathcal{Y}$ dans un espace de Hilbert $\mathcal{H}$ grâce à une application K. De plus, cette fonction K est en relation avec le noyau de la manière suivante

$$k(x, y) = \langle K(x), K(y) \rangle_{\mathcal{H}}.$$

Sans rentrer dans les détails, avec un noyau k, on peut aussi prolonger le plongement K à l'espace des mesures $\mathcal{M}(\mathcal{Y})$ dans ce même espace de Hilbert. Cette extension est appelée *Kernel Mean Embedding* (KME). Elle permet de comparer des mesures μ et ν en comparant leurs représentants

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

Cette pseudo-distance, appelée *Maximum Mean Discrepancy* (MMD), possède une réécriture permettant de la manipuler explicitement

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

4.2 MMD et distances usuelles

Convergence en loi Revenons à la question de la convergence au sens de la divergence, ou du MMD. Que signifie $d_k(\mu_n,\mu)\to 0$? La première notion de convergence pour des mesures est la convergence en loi, ou faible. Lorsque la convergence du MMD est équivalente à la convergence en loi, on dira que le MMD métrise la convergence en loi. Le résultat le plus complet dans la littérature se trouve dans Simon-Gabriel et al. (2021). Il prolonge des travaux de l'article Sriperumbudur (2016) en affaiblissant les conditions sur l'espace $\mathcal{Y}$ pour qu'un MMD métrise la convergence en loi.

Théorème 0.7. *Soit k un noyau borné, alors sous **certaines conditions** sur k, le MMD métrise la convergence en loi si et seulement si le KME K est injectif.*

Les conditions sur le noyau concernent sa régularité et son comportement à l'infini. Ce résultat sera plus explicité dans le prochain chapitre.

Comparaison de distances Une autre propriété étudiée est le contrôle d'une distance d. En pratique, cette distance d peut être une distance classique métrisant la convergence en loi (distance de Prokorhov, distance de Fortet-Mourier) ou les distances de Wasserstein. Ce contrôle se fait partiellement dans le sens

$$\forall\varepsilon>0,\ \exists C>0,\ \forall\mu,\nu\in\mathcal{T}\subset\mathcal{M}(\mathcal{Y}),\ d(\mu,\nu)\leq C d_k(\mu,\nu)+\varepsilon. \tag{7}$$

Des résultats en ce sens sont présentés dans Sriperumbudur et al. (2010) et Vayer and Gribonval (2021).

4.3 Métrisation plus forte

Le Chapitre 5 donne des résultats analogues dans le cas où $\mathcal{Y}=\mathbb{R}^d$ muni de sa topologie classique.

Convergence pour les distances de Wasserstein Concernant le lien entre la convergence pour le MMD et les distances de Wasserstein, aucun résultat n'était présent dans la littérature. Les noyaux bornés, souvent utilisés en pratique, ne sont pas assez puissants pour métriser des convergences plus fortes que la convergence en loi. Une famille classique de noyaux bornés est les noyaux invariants par translation, c'est à dire de la forme $k(x,y)=k(x-y,0)$. Le théorème de Böchner permet de caractériser ces noyaux comme des fonctions caractéristiques de mesures. Si k est noyau invariant par translation alors il existe une mesure positive symétrique Λ tel que pour $x,y\in\mathbb{R}^d$,

$$k(x,y) = \int_{\mathbb{R}^d} \mathrm{e}^{i(x-y)\cdot\xi}\ \Lambda(\mathrm{d}\xi).$$

Un noyau peut être vu d'après le théorème de réalisation de Kolmogorov comme la fonction de covariance d'un processus, i.e. il existe un processus Gaussien $(B(x))_{x\in\mathbb{R}^d}$ vérifiant

$$k(x,y) = \mathrm{Cov}(B(x),B(y)).$$

Dans le cas où k est invariant par translation alors le processus associé est invariant par translation, $(B(x+h))_{x\in\mathbb{R}^d}$ et $(B(x))_{x\in\mathbb{R}^d}$ ont la même distribution. On va s'intéresser à une famille plus large de processus, les processus à accroissement stationnaires,

$$(B(x)-B(x_0))_{x\in\mathbb{R}^d} \overset{d}{=} (B(x+h)-B(x_0+h))_{x\in\mathbb{R}^d}.$$

D'après Yaglom and Silverman (1962), ces processus possèdent une fonction de covariance de la forme

$$k(x,y)=\int_{\mathbb{R}^d}\left(1-\mathrm{e}^{ix\cdot\xi}\right)\left(1-\mathrm{e}^{-iy\cdot\xi}\right)\ \Lambda(\mathrm{d}\xi)+x^T\Sigma y,$$

où Σ est une matrice définie positive et Λ est une mesure vérifiant une certaine propriété d'intégrabilité. Le noyau k n'est plus invariant par translation mais c'est à présent son MMD qui l'est, pour $\tau_h(x)=x+h$ la translation, le MMD vérifie

$$d_k(\mu\circ\tau_h^{-1},\nu\circ\tau_h^{-1})=d_k(\mu,\nu).$$

Les Energy Kernels forment une famille de noyaux possèdant un MMD invariant par translation. Ils sont définis pour $\alpha\in(0,1)$ comme

$$k_\alpha(x,y)=\|x\|^{2\alpha}+\|y\|^{2\alpha}-\|x-y\|^{2\alpha}. \tag{8}$$

Ils correspondent aux covariances des mouvements Brownien fractionnaires. On montrera que ces MMD sont liés aux distances de Wasserstein W_α pour $\alpha\in(0,1)$.

Théorème 0.8. *Pour $\alpha\in(0,1)$,*

i) $W_\alpha(\mu_n,\mu)\to 0$ *implique* $d_{k_\alpha}(\mu_n,\mu)\to 0$.

ii) $d_{k_\alpha}(\mu_n,\mu)\to 0$ *implique* $W_\beta(\mu_n,\mu)\to 0$ *pour* $\beta<\alpha$.

Malheureusement, les MMD invariants par translation ne sont pas assez forts pour métriser des distances de Wasserstein pour $\alpha\geq 1$. Une légère modification de ces noyaux permet de régler ce problème. On définit alors

$$k(x,y)=\int_{\mathbb{R}^d}\left(1-\mathrm{e}^{ix\cdot\xi}\right)\left(1-\mathrm{e}^{-iy\cdot\xi}\right)\ \Lambda(\mathrm{d}\xi)+(|x|^\alpha)^T\Sigma|y|^\alpha,$$

où $|x|^\alpha=(|x_1|^\alpha,\dots,|x_d|^\alpha)$. Le défaut de ce changement est que l'on perd l'invariance par translation et le lien avec les processus à accroissements stationnaires. Ces noyaux ne correspondent pas à des processus particuliers.

Théorème 0.9. *Pour un noyau k de la forme précédente, son MMD d_k métrise la distance de Wasserstein W_α si et seulement si* $\mathrm{supp}(\Lambda)=\mathbb{R}^d$ *et* $\ker\Sigma\cap(\mathbb{R}^+)^d=\{0\}$.

Comparaison de distances Les Energy kernels n'ont pas permis de métriser la distance de Wasserstein W_1 mais ils peuvent tout de même la contrôler partiellement. La preuve de ce contrôle se base sur l'écriture duale de la distance W_1 avec les fonctions lipschitziennes

$$W_1(\mu,\nu)=\sup\left\{\int_{\mathbb{R}^d}\varphi\,\mathrm{d}(\mu-\nu)\ :\ \varphi:\mathbb{R}^d\to\mathbb{R}\ \ \text{1-Lipschitz}\right\}.$$

Le contrôle ne se fera que sur un sous-ensemble des mesures de probabilités. On dit qu'un ensemble $\mathcal{T}$ est uniformément intégrable si

$$\forall\varepsilon>0,\ \exists K\subset\mathbb{R}^d \text{ compact},\ \forall\mu\in\mathcal{T},\ \int_{K^c}\|x\|\ \mu(\mathrm{d}x)\leq\varepsilon.$$

Ainsi sur un sous-ensemble de mesures de probabilité vérifiant cette propriété, les MMD des Energy kernels contrôlent partiellement la distance W_1 dans le sens de l'Équation (7).

Chapitre 1

Présentation d'outils clefs de la thèse

Résumé : Ce chapitre introductif présente deux théories utilisées dans cette thèse. Il ne contient aucun résultat original et peut ne pas être lu pour la lecture des chapitres concernés. L'idée de cette partie est de détailler quelques objets rencontrés durant ma thèse qui m'ont plu. La première section s'intéresse à la calibration (Chapitre 3). On donnera les définitions qui seront réintroduites dans le chapitre concerné et on présentera les preuves des relations des différentes définitions. Le but est de réunir dans un même document toutes les preuves concernant la calibration. La deuxième section sera en relation avec la théorie des *Reproducting Kernel Hilbert Space* (RKHS). Elle détaillera l'astuce du noyau (*kernel trick* en anglais) dans le cas du *Support Vector Machine* (SVM). Personnellement, cette théorie est l'une des théories les plus jolies que j'ai dû étudier durant mon parcours de mathématicien. Cette section est à mettre en relation avec l'article Modeste and Dombry (2022) constituant le Chapitre 5.

1 Théorie de la Calibration

1.1 Introduction

Nous redonnons le cadre *d'espace prédictif* introduit dans Gneiting and Ranjan (2013).

Définition 1.1. *Soit* $(\Omega, \mathcal{A}, \mathbb{P})$ *un espace probabilisé. On appelle espace prédictif associé aux informations* $(\mathcal{F}_1, \dots, \mathcal{F}_k)$ *un* $(k+2)$*-uplets de la forme* $(F_1, \dots, F_k, Y, V)$ *vérifiant*

1. Y *est une variable aléatoire réelle ;*
2. *pour* $1 \leq i \leq k$, $F_i \colon \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ *est un noyau qui est* $\mathcal{F}_i$*-mesurable lorsqu'on le regarde comme une mesure aléatoire* $F_i \colon \Omega \to \mathcal{P}(\mathbb{R})$ *;*
3. V *est une variable aléatoire indépendante de* $(F_1, \dots, F_k, Y)$ *ayant pour distribution* $\mathrm{Unif}([0,1])$.

Un espace prédictif modélise un problème de prédiction. La variable Y joue le rôle du phénomène à prédire et les mesures aléatoires $(F_1, \dots, F_k)$ sont les prévisionnistes en compétition. La variable V est une variable ad hoc qui servira dans la suite.

Remarque 1.2. Dans certains exemples, on mentionnera une variable explicative X à la place des tribus $\mathcal{F}$. L'information alors considérée sera la tribu engendrée par X.

Exemple 1.3. Prenons $k = 1$. Imaginons que l'on observe un phénomène suivant la loi $\mathcal{N}(m, 1)$ dont la moyenne m est inconnue. On considère n observations indépendantes $X \in \mathbb{R}^n$. On pose $F \sim \mathcal{N}(\overline{X}, 1)$, i.e. on prédit en remplaçant l'inconnue m par la moyenne empirique de nos observations. La variable aléatoire Y sera le prochain résultat du phénomène et V une variable uniforme indépendante. Le triplet (F, Y, V) est alors un espace prédictif.

Plusieurs propriétés peuvent être recherchées dans une prévision. Celle au cœur des Chapitres 2 et 3 a été introduite dans Tsyplakov (2011).

Définition 1.4. *Soit $\mathcal{F}$ une sous tribu, un prévisionniste F est dit idéalement calibré pour une information $\mathcal{F}$ si*

$$F = \mathcal{L}(Y \mid \mathcal{F}).$$

Dans le cas particulier où $\mathcal{F}$ est la tribu engendrée par F, la prédiction est dite auto-calibrée.

Dans le cas où le prévisionniste ne possède pas d'information, la tribu $\mathcal{F}$ est donc la tribu triviale, la prévision idéale est déterministe et il s'agit de la distribution de Y. Le Chapitre 4 s'intéresse à estimer cette loi conditionnelle dans le cas où la tribu $\mathcal{F}$ est engendrée par une variable X. Un prédicteur idéalement calibré utilise parfaitement l'information dont il dispose. Malheureusement en pratique, la calibration idéale est un but quasiment impossible à atteindre.

Exemple 1.5. Prenons $Y = Z_1 + Z_2 + \varepsilon$ où $(Z_1\, Z_2, \varepsilon)$ sont trois variables indépendantes de même loi $\mathcal{N}(0,1)$. On définit les trois prévisions suivantes

$$F_0 = \mathcal{N}(0,3),\ F_1 = \mathcal{N}(Z_1, 2) \text{ et } F_2 = \mathcal{N}(Z_1 + Z_2, 1).$$

On peut vérifier que les prédictions (F_0, F_1, F_2) sont idéalement calibrées respectivement par rapport à la tribu triviale, la tribu engendrée par Z_1 puis (Z_1, Z_2). Le prévisionniste F_0 ne possède comme information que la loi Y, il ne peut pas prédire de manière optimale. Tandis que les deux autres prévisionnistes ont connaissance de phénomènes Z_1 ou Z_2 qui ont des conséquences sur Y. On remarque dans cet exemple le résultat classique que l'espérance de la variance conditionnelle décroit avec la taille de l'information,

$$\mathbb{E}[\mathrm{Var}(Y \mid \mathcal{F}')] \leq \mathbb{E}[\mathrm{Var}(Y \mid \mathcal{F})],$$

quand $\mathcal{F} \subset \mathcal{F}'$ deux tribus.

On peut se demander à quoi sert la variable uniforme V dans la définition d'espace prédictif. Son introduction est nécessaire pour définir la transformation suivante. Par abus de notation, on associera une mesure de probabilité avec sa fonction de répartition.

Définition 1.6. *Pour un espace prédictif (F, Y, V), on définit la Probability Integral Transform (PIT) de F selon Y comme*

$$\begin{aligned} Z_F^Y &= F\left(]-\infty, Y[\right) + VF\left(\{Y\}\right) = F(Y^-) + V\left(F(Y) - F(Y^-)\right) \\ &= (1-V)F(Y^-) + VF(Y). \end{aligned}$$

Pour G une mesure, il s'agit de la généralisation de la transformation $G(Y)$. Un résultat classique dans le cas où la mesure G possède une fonction de répartition continue est que $G(Y)$ suit une loi uniforme sur $[0,1]$ si et seulement si G est la loi de la variable Y. Ce résultat reste vrai pour la PIT précédemment introduite. Une généralisation intéressante pour le cas de mesure aléatoire, énoncée dans la proposition suivante, sera l'élément clé du Chapitre 3.

Proposition 1.7. *Soit F un prédicteur $\mathcal{F}$ mesurable, on a l'équivalence entre*

1. *$F = \mathcal{L}(Y \mid \mathcal{F})$;*
2. *la PIT Z_F^Y est uniforme sur $[0,1]$ et indépendante de la tribu $\mathcal{F}$.*

Ainsi pour tester si un prédicteur est idéalement calibré, il suffit de regarder le comportement de sa PIT avec son information.

Remarque 1.8. On montrera dans le Chapitre 3 que cette transformation est mesurable. La preuve se base sur la mesurabilité de l'évaluation par un ouvert. On peut donc manipuler la PIT sans problème.

La prochaine définition est assez proche de celle de la calibration idéale. C'est une propriété, introduite dans Strähl and Ziegel (2017), qui permet de comparer un prévisionniste à plusieurs autres.

Définition 1.9. *On dit que F_1 est **cross calibrée** par rapport à $F_2, \dots, F_k$ si*

$$\mathcal{L}(Z_{F_1}^Y \mid \mathcal{F}_1, \dots, \mathcal{F}_k, \sigma(Y)) = \text{Unif}([0,1]),$$

où $(\mathcal{F}_i)_{1\leq i\leq k}$ sont les tribus associées aux prédicteurs.

Cette définition peut se reécrire en disant que F_1 est cross-calibrée si sa PIT est uniforme sur $[0,1]$ et indépendante de la sous tribu engendrée par $(\mathcal{F}_1, \dots, \mathcal{F}_k, \sigma(Y))$. La Proposition 1.7 donne une interprétation assez claire de cette définition. Un prédicteur F_1 est cross-calibré par rapport à $F_2, \dots, F_k$ si ce prédicteur prédit parfaitement avec son information et que l'ajout des informations des autres prédicteurs ne l'aide pas. Tout comme la calibration idéale, la cross-calibration est une propriété trop contraignante en pratique. Le point commun est le conditionnement, notion rendant la définition assez forte. Donnons maintenant des définitions de calibration beaucoup plus faibles. Pour créer facilement de telles notions de calibration, on ne va plus considérer l'information de nos prédicteurs. Une calibration est dite non conditionnelle dans ce cas, et conditionnelle lorsqu'elle prend en compte le conditionnement. Ces définitions se trouvent dans Gneiting and Ranjan (2013)

Définition 1.10. *Soit (F, G, Y, V) un tuplet prédictif, on dit que*

1. *la prédiction F est marginalement calibrée si $\mathbb{E}[F(y)] = \mathbb{P}(Y \leq y)$ pour tout y ;*
2. *la prédiction F est calibrée en probabilité si Z_F^Y est uniformément distribué sur $([0,1])$;*
3. *la prédiction F est sur- (resp. sous-) dispersée si la variance de Z_F^Y est strictement inférieure (resp. supérieure) à 1/12. Dans le cas égale à 1/12, on dit que la prédiction est dispersée de manière neutre ;*
4. *une prédiction F est plus dispersée qu'une prédiction G si la variance de Z_F^Y est plus petite que Z_G^Y.*

La calibration en probabilité peut être vu comme la calibration idéale pour la tribu triviale par la Proposition 1.7. Le $1/12$ est la variance de $\text{Unif}([0,1])$. Si Z_F^Y a une plus petite variance alors Y ne va pas avoir tendance à aller dans la queue supérieure de F, c'est pour cela que l'on dit que F est sur dispersée (par rapport aux données). Ces calibrations non conditionnelles ne sont pas assez riches pour évaluer la qualité d'une prédiction car elles ne s'intéressent pas à la manière dont les prévisionnistes utilisent leurs données, informations. Récemment, des calibrations conditionnelles plus faibles que la calibration idéale ont été introduites (Pohle, 2020). L'évaluation ne se fera pas sur la distribution entière, trop exigeant, mais sur certaines fonctionnelles de la prédiction, comme par exemple les quantiles ou la moyenne.

Définition 1.11. *Soit $\mathcal{T}$ une fonctionnelle définie sur un sous-ensemble des mesures de probabilité. Une prévision F est dite $\mathcal{T}$-calibrée si*

$$\mathcal{T}(\mathcal{L}(Y \mid \mathcal{T}(F))) = \mathcal{T}(F).$$

La fonctionnelle peut être à valeur dans des ensembles assez quelconques, dans $\mathbb{R}$, dans $\mathbb{R}^{\mathbb{R}}$ ou même dans l'ensemble des intervalles de $\mathbb{R}$. Pour certaines fonctionnelles, la $\mathcal{T}$-calibration peut être vue comme une version conditionnelle des calibrations définies dans la Définition 1.10. Par exemple, si la fonctionnelle $\mathcal{T}$ est l'ensemble des marginales,

$$\mathcal{T}(F) = (F(y))_{y\in\mathbb{R}}. \tag{1.1}$$

On peut donc voir la $\mathcal{T}-$calibration, pour cette fonctionnelle, comme la version conditionnelle de la calibration marginale. Nous énoncerons dans la partie suivante les différentes implications entre ces définitions, et dans le cas contraire des contre-exemples.

1.2 Lien entre les définitions

Commençons par les liens entre les différentes calibrations non conditionnelles. Cette première proposition découle directement de la définition de la calibration en probabilité. Nous donnons la preuve uniquement pour bien rappeler les différentes définitions.

Proposition 1.12 (Gneiting and Ranjan (2013)). *Une prédiction qui est calibrée en probabilité est dispersée de manière neutre.*

Démonstration. Si la prédiction F est calibrée en probabilité alors $Z_F^Y \sim \text{Unif}([0,1])$. Donc la variance de la PIT est bien $1/12$. Ainsi la prédiction est bien dispersée de manière neutre. □

Par contre, il n'existe pas de lien entre la calibration marginale et en probabilité. Ces contre-exemples sont basés sur des prévisions classiques que l'on rencontrera lors du Chapitre 3.

Proposition 1.13 (Gneiting et al. (2007)). *Il existe une prédiction F marginalement calibrée mais pas en probabilité. Et inversement, il existe F calibrée en probabilité mais pas marginalement.*

Démonstration. Prenons $\mu, \varepsilon \sim \mathcal{N}(0,1)$ deux variables Gaussiennes indépendantes. Posons $Y = \mu + \varepsilon$. La variable Y est une Gaussienne centrée de variance 2. Considérons les deux prévisions suivantes qui seront

- $F_1 = \mathcal{N}(-\mu, 1)$: marginalement calibrée mais pas en probabilité ;
- $F_2 = \frac{1}{2}\mathcal{N}(\mu, 1) + \frac{1}{2}\mathcal{N}(\mu + \tau, 1)$ avec $\tau = \pm 1$ avec probabilité $1/2$: calibrée en probabilité mais pas marginalement.

On a $Z_{F_1}^Y = \phi(Y - (-\mu))$ où ϕ est la fonction de répartition d'une loi Gaussienne centrée réduite. Or $Y + \mu$ n'est pas réduit, donc $Z_{F_1}^Y$ n'est pas uniforme sur $[0,1]$, et ainsi F_1 n'est pas calibrée en probabilité. Soit $y \in \mathbb{R}$, notons f la densité de $\mathcal{N}(0,1)$,

$$\begin{aligned}
\mathbb{E}[F_1(y)] = \mathbb{E}[\phi(y+\mu)] &= \int_{-\infty}^{+\infty} \phi(y+x) f(x) \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \phi(y-x) f(x) \mathrm{d}x, \text{ par symétrie} \\
&= \int_{-\infty}^{+\infty} f(x) \int_{-\infty}^{y} f(t-x)\ \mathrm{d}t \mathrm{d}x \\
&= \int_{-\infty}^{y} \int_{-\infty}^{+\infty} f(t-x) f(x)\ \mathrm{d}x \mathrm{d}t \\
&= \int_{-\infty}^{y} f \star f(t)\ \mathrm{d}t = \mathbb{P}(Y \leq y),
\end{aligned}$$

car la densité $f \star f$ est celle d'une Gaussienne centrée de variance 2 donc F_1 est bien marginalement calibrée. De manière très similaire, on montre que

$$\mathbb{E}[F_2(y)] = \frac{1}{2}\mathbb{P}(Y \leq y) + \frac{1}{4}\mathbb{P}(Y \leq y+1) + \frac{1}{4}\mathbb{P}(Y \leq y-1) \neq \mathbb{P}(Y \leq y).$$

Ainsi F_2 n'est pas marginalement calibrée. Pour la calibration en probabilité, introduisons la fonction

$$\Psi(x) = \frac{1}{2}\phi(x) + \frac{1}{2}\phi(x+1).$$

On a $Z_{F_2}^Y = \frac{1}{2}\phi(\varepsilon) + \frac{1}{2}\phi(\varepsilon - \tau)$. Pour g une fonction continue, on a en conditionnant selon τ

$$\begin{aligned}
\mathbb{E}[g(Z_{F_2}^Y] &= \frac{1}{2}\int_{-\infty}^{+\infty} g\left(\frac{1}{2}\phi(x) + \frac{1}{2}\phi(x-1)\right) f(x)\ \mathrm{d}x + \frac{1}{2}\int_{-\infty}^{+\infty} g\left(\frac{1}{2}\phi(x) + \frac{1}{2}\phi(x+1)\right) f(x)\ \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} g(\Psi(x)) \frac{f(x+1) + f(x)}{2}\ \mathrm{d}x = \int_0^1 g(u)\ \mathrm{d}u,
\end{aligned}$$

par un changement de variable. Donc la PIT est uniformément distribuée sur $[0,1]$. Ainsi la prévision F_2 est calibrée en probabilité. □

Le prochain résultat énonce que les premières calibrations conditionnelles (idéal, auto et cross) sont plus fortes que les non conditionnelles.

Théorème 1.14 (Gneiting and Ranjan (2013))**.** *Une prédiction F idéale par rapport à une tribu $\mathcal{F}$ quelconque est marginalement calibrée et calibrée en probabilité.*

Démonstration. Si F est idéal alors la prévision est calibrée en probabilité par la Proposition 1.7. Et pour $y \in \mathbb{R}$, par la définition de la loi conditionnelle,

$$F(y) = \mathbb{P}(Y \leq y \mid \mathcal{F}),$$

on voit bien que la prévision est marginalement calibrée. □

Une réciproque assez amusante est énoncée dans Gneiting and Ranjan (2013, Théorème 2.11) pour le cas binaire. Nous allons proposer ici une preuve différente.

Théorème 1.15. *Soit F une prédiction binaire, i.e. une mesure de probabilité sur $\{0,1\}$. On a équivalence entre F auto-calibrée et calibrée en probabilité.*

Démonstration. Si F est auto-calibrée alors F est idéale pour $\sigma(F)$ et donc par le théoèrme précédent, la prédiction est calibrée en probabilité. La réciproque est le point difficile de la preuve. Dans le cas où F est binaire, la prédiction s'écrit

$$F(y) = (1-p)\mathbb{1}_{\{y\geq 0\}} + p\mathbb{1}_{\{y\geq 1\}}, \quad \text{pour } y \in \mathbb{R},$$

où p est une variable aléatoire sur $[0,1]$ de distribution μ. Cette variable contient toute l'information de F donc $\sigma(F) = \sigma(p)$. Dans ce cas, la PIT s'écrit

$$Z_F^Y = V(1-p)\mathbb{1}_{\{Y=0\}} + (1-p+Vp)\mathbb{1}_{\{Y=1\}}.$$

Comme $\mathbb{P}(Y=1 \mid p)$ est $\sigma(p)$ mesurable, on peut l'écrire $h(p)$ avec h une fonction mesurable. Pour montrer l'auto-calibration, il faut donc montrer que $h = \mathrm{Id}$ $\mu-$pp. Si μ a des atomes en 0 ou 1, il est direct de voir que $h(0) = 0$ et $h(1) = 1$. Car sinon Z_F^Y aurait lui même un atome en 1 ou 0. Premièrement, déterminons la distribution conditionnelle de Z_F^Y par rapport à p. Pour cela, prenons une fonction positive bornée f, le théorème de Fubini pour les noyaux conditionnels donne

$$\int_0^1 f(z)\,\mathrm{d}z = \int_0^1 \int_0^1 f(v(1-x))(1-h(x)) + f(1-x+vx)h(x)\,\mathrm{d}v\mu(\mathrm{d}x),$$

car V est indépendante de (p,Y) et Z_F^Y est uniforme sur $[0,1]$. Pour $x \in (0,1)$, on a par un changement de variable

$$\int_0^1 f(v(1-x))(1-h(x)) + f(1-x+vx)h(x)\,\mathrm{d}v = \int_0^{1-x} f(z)\frac{1-h(x)}{1-x}\,\mathrm{d}z + \int_{1-x}^1 f(z)\frac{h(x)}{x}\,\mathrm{d}z.$$

Donc Z_F^Y est encore à densité dont la densité est

$$\begin{cases} \frac{1-h(x)}{1-x} & , z \in [0,1-x) \\ \frac{1-h(x)}{1-x} + \frac{h(x)}{x} & , z = 1-x \\ \frac{h(x)}{x} & , z \in (1-x,1] \end{cases}.$$

Comme la distribution est à densité, on peut décider que dans le cas $z = 1-x$, il n'y ait que le premier terme de la somme. Et donc par le théorème de Fubini-Tonelli

$$\begin{aligned}\int_0^1 f(z)\,\mathrm{d}z &= \int_0^1 \left(\int_{[0,1-x]} f(z)\frac{1-h(x)}{1-x}\,\mathrm{d}z + \int_{(1-x,1]} f(z)\frac{h(x)}{x}\,\mathrm{d}z \right) \mu(\mathrm{d}x) \\ &= \int_0^1 f(z) \left(\int_{[0,1-z]} \frac{1-h(x)}{1-x}\,\mu(\mathrm{d}x) + \int_{(1-z,1]} \frac{h(x)}{x}\,\mu(\mathrm{d}x) \right) \mathrm{d}z.\end{aligned}$$

Donc pour tout $z \in [0,1]$, $\lambda-$pp on a

$$\int_{[0,1-z]} \frac{1-h(x)}{1-x}\,\mu(\mathrm{d}x) + \int_{(1-z,1]} \frac{h(x)}{x}\,\mu(\mathrm{d}x) = 1.$$

Ce qui revient à pour tout $z \in [0,1]$, $\lambda-$pp

$$\int_{[0,z]} \frac{1-h(x)}{1-x}\,\mu(\mathrm{d}x) + \int_{(z,1]} \frac{h(x)}{x}\,\mu(\mathrm{d}x) = 1.$$

En faisant tendre z vers 0 par densité dans un ensemble de mesure plein, on a

$$1 = \mu(0) + \int_{(0,1]} \frac{h(x)}{x}\,\mu(\mathrm{d}x).$$

Ainsi pour tout $z \in [0,1]$, $\lambda-$pp

$$\int_{[0,z]} \frac{1-h(x)}{1-x}\,\mu(\mathrm{d}x) = \int_{[0,z]} \frac{h(x)}{x}\mathbb{1}_{\{x\neq 0\}} + \mathbb{1}_{\{x=0\}}\,\mu(\mathrm{d}x).$$

Et donc par densité, on a $\mu-$pp l'égalité pour $x \in [0,1)$,

$$\frac{1-h(x)}{1-x} = \frac{h(x)}{x}.$$

En développant, on retombe bien sur $h(x) = x$ $\mu-$pp. Donc F est bien auto-calibrée, i.e.

$$F = \mathcal{L}(Y \mid F).$$

□

La calibration conditionnelle par rapport à la fonctionnelle des marginales (1.1) est plus forte que la calibration marginale. En effet, pour $y \in \mathbb{R}$, on a

$$\mathbb{P}(Y \leq y \mid \mathcal{T}(F)) = F(y).$$

On clôture cette partie avec un fait amusant.

Proposition 1.16 (Gneiting and Resin (2021))**.** *L'auto-calibration n'implique pas forcément la $\mathcal{T}-$calibration.*

Démonstration. Prenons comme fonctionnelle la variance avec le cadre de la Proposition 1.13. Donc soit $\mu, \varepsilon \sim \mathcal{N}(0,1)$ et $Y = \mu + \varepsilon$. La prévision $F = \mathcal{N}(\mu, 1)$ est auto-calibrée mais comme $\mathrm{Var}(F)$ est constante égale à 1,

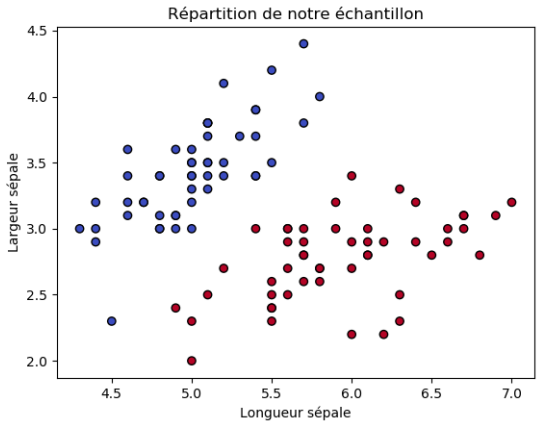
$$\mathrm{Var}(Y \mid \mathrm{Var}(F)) = \mathrm{Var}(Y) = 2.$$

Donc la prévision F n'est pas Var$-$calibrée. □

2 Reproducing Kernel Hilbert Space

2.1 Motivation : séparation par un hyperplan

Prenons comme point de départ le jeu de données classique des *Iris de Fisher*[1]. Ce jeu est constitué d'un échantillon de 150 iris de 3 espèces différentes. Il y a 50 représentants par espèce. Pour notre cas, on va simplifier le problème et considérer seulement deux espèces. Pour chaque individu de l'échantillon, on mesure 4 caractéristiques, la longueur et la largeur des sépales et des pétales. Pour formaliser ceci, on a 100 couples de variables $(X_i, Y_i)_{1\leq i\leq 100}$ à valeurs dans $\mathbb{R}^4 \times \{0, 1\}$ représentants la valeur des 4 caractéristiques ainsi que l'espèce associée. Le but est de déterminer une relation entre ces 4 caractéristiques pour pouvoir prédire l'espèce. Pour des raisons de visualisation, on ne va prendre en compte que 2 de ces caractéristiques.



On voit bien que les deux espèces sont bien séparées par une droite. On va donc essayer de déterminer l'équation de cette dernière. Ainsi, si l'on ramasse un nouvel iris, on pourra *potentiellement* déterminer son espèce juste en mesurant ces pétales. En effet, sa position par rapport à la droite nous indiquera si l'iris est de l'espèce 0 ou de l'espèce 1. La première étape est d'écrire ce problème mathématiquement et de savoir s'il existe une manière de séparer les points. L'idée principale est de changer les labels $(Y_i)_{1\leq i\leq 100}$. Au lieu d'être dans l'ensemble $\{0, 1\}$, on sera dans $\{-1, 1\}$. Ce changement permet de relier le signe du label avec son rapport à l'hyperplan (ou droite en dimension 1) séparateur. En effet, un hyperplan affine s'écrit

$$H = \{x \in \mathbb{R}^d \mid \langle \mathrm{w}, x\rangle + b = 0\},$$

avec $\mathrm{w}, b \in \mathbb{R}^d \times \mathbb{R}$. À noter que pour un hyperplan H, le couple (w, b) le définissant n'est pas unique car $(2\mathrm{w}, 2b)$, par exemple, définit le même hyperplan. Un vecteur $x \in \mathbb{R}^d$ est dit au dessus (resp. en dessous) de l'hyperplan H si $\langle \mathrm{w}, x\rangle + b > 0$ (resp. < 0). Ainsi, on peut séparer un jeu de données $(X_i, Y_i)_{1\leq i\leq 100}$ si l'on peut trouver w, b tel que pour tout $i \in \{1, \ldots, n\}$,

$$Y_i \times (\langle \mathrm{w}, X_i\rangle + b) > 0.$$

L'existence d'un tel couple (w, b) peut se réécrire avec le théorème de séparation des convexes fermés de la manière suivante :

Théorème 1.17. *Un jeu de données $(X_i, Y_i)_{1\leq i\leq n}$ est séparable si et seulement si*

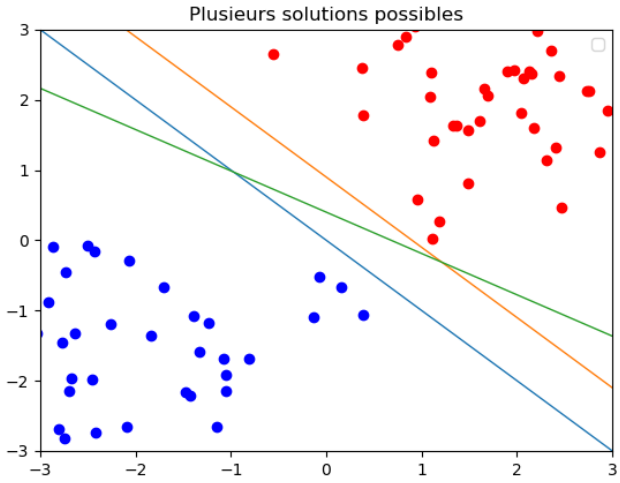
$$\mathrm{Conv}(\{X_i \mid Y_i = -1\}) \cap \mathrm{Conv}(\{X_i \mid Y_i = +1\}) = \emptyset,$$

où $\mathrm{Conv}(A)$ est le plus petit convexe contenant l'ensemble A.

1. https://fr.wikipedia.org/wiki/Iris_de_Fisher

Malheureusement, ce théorème ne nous aide pas vraiment à savoir si l'on peut séparer ou non un jeu de données. En effet, la vérification informatique de la séparation de convexe n'est pas forcément évidente. De plus, la preuve de ce résultat ne nous donne pas une équation d'un hyperplan séparateur. L'algorithme *Perceptron Learning Algortihm* (PLA) permet de trouver un tel hyperplan de manière itérative (Rosenblatt (1958)). Pour la suite, on se place dans le cas où les points sont bien séparables par un hyperplan.

Dans ce cas, il peut exister une infinité d'hyperplans séparant notre jeu de données.



L'idée intuitive est de considérer l'hyperplan qui sépare le mieux notre jeu de données. C'est le but de la fin de cette section, introduire un critère de *bonne séparabilité* et donner une manière de trouver la solution optimale. Pour choisir parmi tous ces hyperplans, un critère naturel peut être la distance minimale entre le jeu de données et l'hyperplan séparateur. Ainsi, plus cette distance minimale est grande, plus l'hyperplan sépare les points en étant éloigné simultanément des deux espèces. On appelle *marge* d'un hyperplan cette distance. Cette marge s'écrit grâce à la formule de la distance entre un point et un hyperplan comme

$$M(\mathrm{w}, b) = \min_{i \in \{1,\dots,n\}} \frac{|\langle \mathrm{w}, X_i \rangle + b|}{\|\mathrm{w}\|_2}$$

Le but est donc de maximiser cette marge sous la contrainte

$$\forall i \in \{1, \dots, n\},\ Y_i \times (\langle \mathrm{w}, X_i \rangle + b) \geq 0.$$

On a déjà mentionné le fait que si (w, b) est solution alors pour $\lambda \in \mathbb{R}$ le couple $(\lambda \mathrm{w}, \lambda b)$ est aussi solution car ces vecteurs définissent le même hyperplan. Ainsi avec cette renormalisation, on peut supposer que

$$\min_{i \in \{1,\dots,n\}} |\langle \mathrm{w}, X_i \rangle + b| = 1.$$

Et donc le problème de maximisation sous contrainte s'écrit de la manière suivante

$$\begin{cases} \max\limits_{\mathrm{w} \neq 0, b} \frac{1}{\|\mathrm{w}\|_2} \\ \text{sous la contrainte } \forall i \in \{1, \dots, n\},\ Y_i \times (\langle \mathrm{w}, X_i \rangle + b) \geq 1 \end{cases}.$$

Avant de continuer, montrons que le maximum est atteint. Cela est possible s'il existe i, j tel que $Y_i Y_j = -1$, i.e. tous les individus n'ont pas le même label. Soit $(\mathrm{w}_n, b_n)_{n \in \mathbb{N}}$ vérifiant les contraintes et tel que

$$f(\mathrm{w}_n, b_n) \to \sup f,$$

avec $f(\mathrm{w}, b) = 1/\|\mathrm{w}\|_2$. Si le sup est infini, alors $\|\mathrm{w}_n\| \to 0$. Mais ceci est impossible car sinon b serait du signe de Y_i et Y_j. De plus, la suite $(\mathrm{w}_n)_{n\in\mathbb{N}}$ est bornée car le $\sup f > 0$, on peut donc extraire une sous suite convergence vers un vecteur $\mathrm{w} \neq 0$. Comme Y_i et Y_j sont de signes opposés, la suite $(b_n)_{n\in\mathbb{N}}$ est majorée et minorée. Ainsi à une sous suite près, elle converge vers un réel b. Le couple (w, b) vérifie encore les contraintes et par la continuité de f, le maximum est bien atteint. Une écriture équivalente mais plus régulière de cette optimisation est

$$\begin{cases} \min\limits_{\mathrm{w},b} \frac{1}{2}\|\mathrm{w}\|_2^2 \\ \text{sous la contrainte } \forall i \in \{1,\dots,n\},\ 1 - Y_i \times (\langle \mathrm{w}, X_i\rangle + b) \leq 0 \end{cases}.$$

Nous avons vu que le cas $\mathrm{w} = 0$ est évité par les contraintes, on peut donc l'ajouter dans la recherche du minimum. Pour résoudre ce problème d'optimisation, nous allons le réécrire de manière duale grâce au théorème des extréma liés. Nous venons de voir que le problème admettait une solution. Le Lagrangien de cette optimisation est

$$\mathcal{L}(\mathrm{w}, b, \alpha_1, \dots, \alpha_n) = \frac{1}{2}\|\mathrm{w}\|^2 + \sum_{i=1}^{n} \alpha_i (1 - Y_i \times (\langle \mathrm{w}, X_i\rangle + b)).$$

En différenciant le Lagrangien, on obtient une condition nécessaire sur (w, b),

$$\begin{cases} \nabla_{\mathrm{w}}\ \mathcal{L}(\mathrm{w}, b, \alpha) = \mathrm{w} - \sum_{i=1}^{n} \alpha_i Y_i X_i = 0 \\ \partial_b\ \mathcal{L}(\mathrm{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i Y_i = 0 \end{cases}.$$

Donc

$$\begin{cases} \sum_{i=1}^{n} \alpha_i Y_i b = 0 \\ \mathrm{w} = \sum_{i=1}^{n} \alpha_i Y_i X_i \end{cases}, \tag{1.2}$$

en injectant ceci dans la forme duale, l'optimisation s'écrit

$$\begin{cases} \max\limits_{\alpha_1,\dots,\alpha_n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j\rangle \\ \text{sous la contrainte } \alpha_i \geq 0, \\ \text{et } \sum_{i=1}^{n} \alpha_i Y_i = 0 \end{cases}. \tag{1.3}$$

Une fois le $n-$uplet α trouvé, on connait w grâce à la formule (1.2). Pour trouver b, on utilise le fait que pour $i \in \{1,\dots,n\}$, si la contrainte i est active, i.e. $\alpha_i > 0$, alors forcément $Y_i \times (\langle \mathrm{w}, X_i\rangle + b) = 1$. À présent lorsque l'on observe un nouvel individu X, on prédit son label par

$$\text{Signe}\left(\sum_{i=1}^{n} \alpha_i Y_i \langle X_i, X\rangle + b\right), \tag{1.4}$$

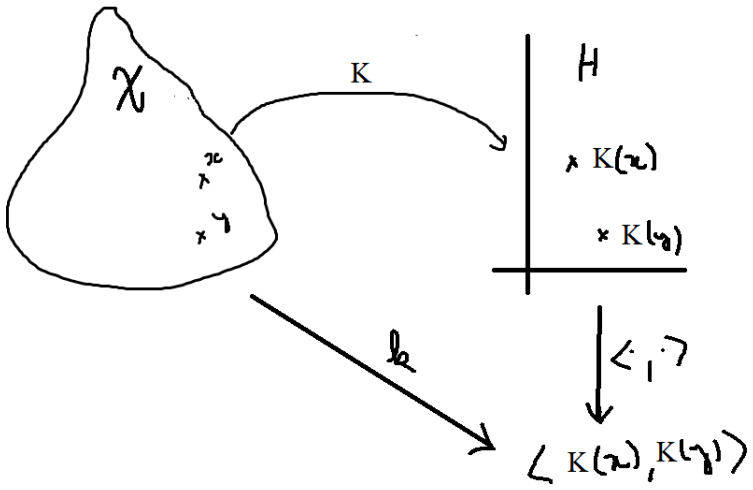
où b s'écrit en fonction du jeu de données $(X_i, Y_i)_{i=1}^n$ et du $n-$uplet α.
On peut faire la remarque que cette solution ne dépend que du produit scalaire entre nos points du jeu de données.

2.2 Définition

Approche 1

On vient de donner une manière de séparer des points lorsque ces derniers sont dans $\mathbb{R}^d$. On a remarqué que la formulation analytique ne dépend que du produit scalaire entre ces points. L'idée derrière la théorie des Noyaux Reproduisants se base sur cette remarque. On va vouloir séparer des points qui ne sont pas forcément dans $\mathbb{R}^d$ ou un espace vectoriel. Pour faire cette généralisation, on va plonger un espace topologique $\mathcal{X}$ assez quelconque dans un espace de Hilbert

$\mathcal{H}$, i.e. représenter chaque point $x \in \mathcal{X}$ par un vecteur $K(x) \in \mathcal{H}$. L'utilité est de pouvoir utiliser la régularité des espaces de Hilbert pour résoudre des problèmes classiques de statistique (SVM, ACP, ...).



En fait, la remarque sur la solution de la séparation d'un hyperplan en terme de produit scalaire s'applique à plusieurs problèmes. Ainsi, il n'est pas nécessaire de plonger entièrement notre espace $\mathcal{X}$, mais simplement de définir le produit scalaire entre les différents points $x, y \in \mathcal{X}$. Cette tache est plus simple. Au lieu de devoir construire entièrement un espace de Hilbert $\mathcal{H}$ (et son produit scalaire $\langle \cdot, \cdot \rangle$) ainsi que le plongement K, il suffit juste de définir la composition $x, y \mapsto \langle K(x), K(y) \rangle$. Cette composition sera notée k et on l'appellera *noyau*. Ce noyau k doit vérifier quelques propriétés. En effet, il doit être symétrique mais aussi défini positif, i.e.

$$\forall n \geq 1,\ \forall x_1, \ldots, x_n \in \mathcal{X},\ \forall a_1, \ldots, a_n \in \mathbb{R},\ \sum_{1 \leq i,j \leq n} a_i a_j k(x_i, x_j) \geq 0, \tag{1.5}$$

car k doit pouvoir représenter la composition d'un produit scalaire et d'une application K, donc

$$\sum_{1 \leq i,j \leq n} a_i a_j k(x_i, x_j) = \langle \sum_{i=1}^{n} a_i K(x_i), \sum_{i=1}^{n} a_i K(x_i) \rangle \geq 0.$$

On appellera alors dans la suite *noyau*, toute application à valeurs réelles de $\mathcal{X} \times \mathcal{X}$ qui est symétrique et définie positive. À noter que k n'est pas nécessairement un produit scalaire, il n'a pas à être bilinéaire. En fait, considérer un tel noyau symétrique et défini positif suffit à construire un espace de Hilbert $\mathcal{H}$ et une application K.

Théorème 1.18 (Moore–Aronszajn)**.** *Soit $\mathcal{X}$ un espace topologique et soit k un noyau. Alors il existe un espace de Hilbert $\mathcal{H}$ et une application $K \colon \mathcal{X} \to \mathcal{H}$ vérifiant pour $x, y \in \mathcal{X}$,*

$$k(x, y) = \langle K(x), K(y) \rangle.$$

Une conséquence de ce résultat est qu'un noyau vérifie l'inégalité de Cauchy-Schwarz, i.e.

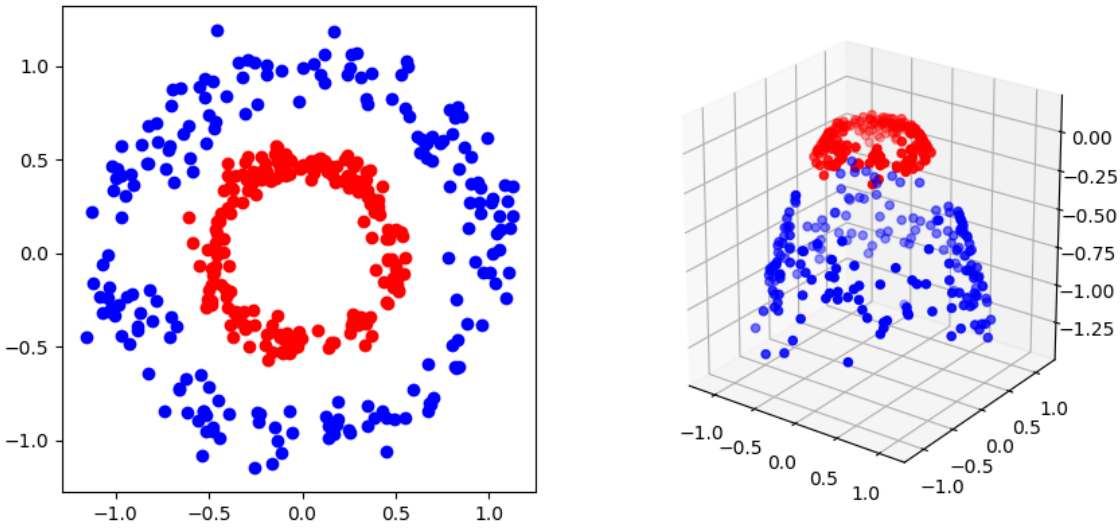
$$\forall x, y \in \mathcal{X},\ k(x, y) \leq \sqrt{k(x, x)} \sqrt{k(y, y)}. \tag{1.6}$$

Cette méthode de plongement est nommée *kernel trick* (astuce du noyau) dans la littérature (Aizerman, 1964). Plusieurs familles sont classiquement utilisées en pratique.

Exemple 1.19. Notre jeu de données peut être dans $\mathbb{R}^d$ mais pas linéairement séparable. Dans ce cas, l'astuce du noyau permet de modifier la géométrie de $\mathbb{R}^d$. Une famille souvent utilisée est la famille des noyaux gaussiens (*Radial basis function kernel*) qui s'écrit

$$k(x, y) \exp(-\gamma \|x - y\|^2),$$

avec $\gamma > 0$. Ces noyaux permettent par exemple de séparer un jeu de données radial.



Exemple 1.20. Pour $\alpha \in (0, 1)$, on définit le noyau sur $\mathbb{R}^d$,

$$k(x, y) = \|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x - y\|^{2\alpha}. \tag{1.7}$$

Cette famille de noyaux s'appelle les Energy Kernels. Ces noyaux sont assez bien étudiés en statistique et sont connectés aux α-distances de corrélation pour des tests d'indépendance (Székely and Rizzo, 2009, Section 4).

Approche 2

La théorie des noyaux reproduisants possède une autre approche plus constructive. Dans certains domaines, la structure de l'espace de Hilbert associé au noyau k est importante. C'est le cas dans le Chapitre 5. Considérons un espace quelconque $\mathcal{X}$ et notons $\mathcal{F}(\mathcal{X}, \mathbb{R})$ l'espace des fonctions à valeurs réelles sur $\mathcal{X}$. Un espace de Hilbert $\mathcal{H}$ inclus dans $\mathcal{F}(\mathcal{X}, \mathbb{R})$ est appelé *Reproducting Kernel Hilbert Space* (RKHS) si pour tout $x \in \mathcal{X}$, l'application évaluation $f \mapsto f(x)$ est continue sur $\mathcal{H}$. Cette continuité permet d'après le théorème de représentation de Riesz de représenter cette application par un vecteur $K(x) \in \mathcal{H}$, i.e.

$$\forall f \in \mathcal{H},\ f(x) = \langle K(x), f \rangle.$$

On définit alors le noyau k associé au RKHS $\mathcal{H}$ pour $x, y \in \mathcal{X}$,

$$k(x, y) = \langle K(x), K(y) \rangle.$$

Cette application k est bien symétrique et défini positif. De plus, par le lien entre $K(y)$ et l'évaluation, on obtient la relation $K(x) = k(x, \cdot)$. Cette approche est équivalente à la précédente.

En effet, dans le Théorème 1.18, l'espace de Hilbert construit dans la preuve est bien un RKHS. Présentons les grandes lignes de cette construction. Considérons le sous-espace vectoriel $\mathcal{H}_0$ de $\mathcal{F}(\mathcal{X}, \mathbb{R})$,

$$\mathcal{H}_0 := \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \mid n \geq 1,\ x_1, \dots, x_n \in \mathcal{X},\ \alpha_1, \dots, \alpha_n \in \mathbb{R} \right\}.$$

Pour $x \in \mathcal{X}$, le point sera représenté par le vecteur $K(x) := k(x, \cdot)$. On munit cet espace vectoriel de la forme bilinéaire suivante pour $f = \sum_{i=1}^{n} \alpha_i K(x_i)$ et $g = \sum_{j=1}^{m} \beta_j K(y_j)$,

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j).$$

On peut remarquer que le vecteur $K(x)$ est relié à l'évaluation,

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i)(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = \langle f, K(x) \rangle,$$

donc l'évaluation est bien continue, et même uniformément continue, sur $\mathcal{H}_0$.

Proposition 1.21. *Cette application a bien un sens et est un produit scalaire.*

Démonstration. Soit deux écritures de g,

$$g = \sum_{j=1}^{m} \beta_j K(y_j) = \sum_{j=1}^{m} \delta_j K(y_j).$$

On a la suite d'égalité suivante

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) &= \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \beta_j K(y_j)(x_i) \\ &= \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{m} \delta_j K(y_j)(x_i) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \delta_j k(x_i, x_j). \end{aligned}$$

Donc la forme bilinéaire est bien définie. Montrons maintenant qu'il s'agit d'un produit scalaire. La symétrie et la positivité proviennent des propriétés du noyau k. Soit f tel que $\langle f, f \rangle = 0$, on a pour $x \in \mathcal{X}$ et $t \in \mathbb{R}$,

$$\langle f + tK(x), f + tK(x) \rangle = \langle f, f \rangle + 2t\langle f, K(x) \rangle + t^2 k(x, x) \geq 0.$$

Comme ce polynôme est toujours positif, son discriminant est négatif ou nul, d'où

$$f(x) = \langle f, K(x) \rangle = 0.$$

Donc f est le vecteur nul. □

La construction de $\mathcal{H}$ se base sur la complétion de cet espace. Ainsi, dans la Section 2.4, on considérera cette approche.

2.3 Exemple d'applications

SVM et RKHS

Restons dans le cadre d'une classification à deux classes, mais prenons un jeu de données qui ne vit pas dans un espace vectoriel mais un ensemble $\mathcal{X}$ quelconque. Séparer linéairement notre jeu de données n'a plus vraiment de sens. Pour régler ce problème, on envoie ce jeu dans un espace de Hilbert $\mathcal{H}$ grâce à un noyau k. Ainsi, on applique la même optimisation présentée précédemment en utilisant non pas le produit scalaire entre les points x, y mais $k(x, y)$. L'optimisation duale (1.3) s'écrit dans ce cas

$$\begin{cases} \max\limits_{\alpha_1,\dots,\alpha_n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j Y_iY_j k(X_i, X_j) \\ \text{sous la contrainte } \alpha_i \geq 0, \\ \text{et } \sum_{i=1}^n \alpha_i Y_i = 0 \end{cases} . \tag{1.8}$$

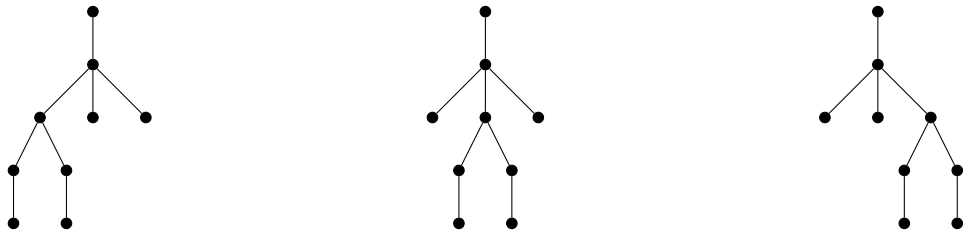
Et comme auparavant l'hyperplan séparateur, ou plutôt le produit scalaire avec l'hyperplan, s'écrira

$$k(\mathrm{w}, \cdot) = \sum_{i=1}^n \alpha_i Y_i k(X_i, \cdot),$$

avec α un $n-$uplet solution. En pratique, on possède une famille de noyaux $(k_\theta)_{\theta\in\Theta}$, et un jeu de données $(X_i, Y_i)_{1\leq i\leq n}$ que l'on coupe en deux, l'un pour la phase d'*entraînement* et l'autre pour la phase de *test*. La partie training permet de trouver $k(\mathrm{w}_\theta, \cdot)$ et b_θ pour chaque $\theta \in \Theta$. Puis on prend le θ faisant le plus de bonnes prédictions sur la partie testing.

Applications exotiques

On appelle arbre enraciné T un graphe orienté connexe sans cycle possédant une unique racine, i.e. un sommet $r(T)$ représentant le sommet initial. On note $\mathcal{T}$ l'ensemble des arbres enracinés. On utilise la convention des arbres non-ordonnés, c'est à dire que l'ordre des enfants est sans importance. Pour un arbre T, on note $h(T)$ sa hauteur qui est la distance entre la racine et le noeud le plus éloigné. Par exemple, ces trois arbres sont identiques dans notre cas et de hauteur 4.



Cette structure de données apparaît naturellement dans plusieurs domaines assez variés, structure de l'ARN en biologie (Le et al., 1989), structure des polymères en chimie (Martín-Delgado et al., 2002), structure de pages web en informatique (Costa et al., 2004). La difficulté est de définir un noyau sur l'ensemble $\mathcal{T}$. Un arbre étant un objet avec une forme singulière, il n'est pas forcément évident de vérifier le caractère défini positif d'une application. Nous donnons donc un exemple de famille de noyaux assez large.

Subtree kernel : Le subtree kernel a été introduit dans Smola and Vishwanathan (2002), l'idée est de comparer deux arbres par rapport à leurs nombres de sous arbres en commun. Pour $T_1, T_2 \in \mathcal{T}$, on définit le noyau

$$k(T_1, T_2) = \sum_{\tau\in\mathcal{T}} w_\tau \kappa(N_\tau(T_1), N_\tau(T_2)),$$

où w_τ est un poids (positif) associé à l'arbre τ, $N_\tau(T_1)$ est le nombre de sous-arbres identiques de T_1 à τ et κ est un noyau sur $\mathbb{N}, \mathbb{Z}$ ou $\mathbb{R}$. Une condition classique sur κ est que si τ n'est pas présent dans l'un des arbres alors le noyau κ s'annule, i.e. pour $n \in \mathbb{N}$, $\kappa(0, n) = \kappa(n, 0) = 0$. Dans Azaïs and Ingels (2020), les auteurs comparent plusieurs poids différents dont le poids exponentiel, $\mathrm{w}_\tau = \lambda^{h(\tau)}$ pour $\lambda \in [0, 1]$. À noter que $\lambda \in [0, 1]$ est naturel car si un sous-arbre τ est dans T_1 et T_2 alors tous les sous arbres de τ aussi. Donc $\lambda \leq 1$ permet de contrer la croissance exponentielle du nombre de sous-arbres. Ils comparent ces différents noyaux sur plusieurs problèmes dont un consistant à déterminer la langue d'un article Wikipédia à partir de sa structure.

2.4 Pour les mesures

On se place à présent dans un espace métrique $(\mathcal{X}, d)$ et l'on note par $\mathcal{B}$ sa tribu borélienne. On note $\mathcal{M}$ (resp. $\mathcal{P}$) l'ensemble des mesures finies signées (resp. mesures de probabilité). Pour $\mu \in \mathcal{M}$, on peut la décomposer, d'après le théorème de Jordan, en $\mu = \mu^+ - \mu^-$ où μ^+ (resp. μ^-) est la partie positive (resp. négative) de la mesure μ. On considère alors $|\mu| = \mu^+ + \mu^-$ sa variation totale.

Construction

Prenons un noyau k et considérons le RKHS $\mathcal{H}_k$ associé. Le but maintenant est de plonger l'espace des mesures finies signées $\mathcal{M}$. Dans la section 2.2, nous avions utilisé l'évaluation $f(x) = \langle K(x), f\rangle$ pour pouvoir représenter via le Théorème de Riesz un point x. La forme linéaire canonique pouvant représenter une mesure est l'intégrale. Avant cela, il va falloir s'assurer que les fonctions dans le RKHS soient mesurables. C'est le cas dès que le noyau k est mesurable.

Lemme 1.22. *Soit k un noyau et $\mathcal{H}_k$ son RKHS associé. Si pour tout $x \in \mathcal{X}$, l'application $k(x, \cdot)$ est mesurable alors toutes fonctions $f \in \mathcal{H}_k$ l'est.*

Démonstration. On rappelle que pour $x \in \mathcal{X}$, $K(x) = k(x, \cdot)$. Soit $f \in \mathcal{H}_k$, nous allons montrer que f est la limite d'une combinaison linéaire de $(k(x, \cdot))_{x\in\mathcal{X}}$. Pour montrer ceci, il suffit d'avoir la densité de $\mathbb{K} = \mathrm{Vect}(k(x, \cdot))$. Comme $\mathcal{H}_k$ est un espace de Hilbert, c'est équivalent à $\mathbb{K}^\perp = \{0\}$. Donc prenons g dans l'orthogonal de $\mathbb{K}$, ainsi pour tout $x \in \mathcal{X}$,

$$g(x) = \langle K(x), g\rangle = 0.$$

Ce qui montre bien que $g = 0$. Comme $\mathbb{K}$ est dense, soit $(f_n)_{n\in\mathbb{N}}$ une suite de $\mathbb{K}$ tendant vers f pour la norme $\|\cdot\|_{\mathcal{H}}$. Or la convergence pour une norme hilbertienne implique la convergence faible, ainsi pour $x \in \mathcal{X}$,

$$f_n(x) = \langle K(x), f_n\rangle \to \langle K(x), f\rangle = f(x).$$

Comme les $(K(x))_{x\in\mathcal{X}}$ sont mesurables, la suite $(f_n)_{n\in\mathbb{N}}$ aussi et donc f aussi car c'est la limite simple de fonction mesurable. □

Pour pouvoir appliquer le théorème de Riesz, il faut avoir la continuité de l'intégration, ce qui n'est pas toujours le cas. On va donc devoir se restreindre à un sous-ensemble de $\mathcal{M}$. Dans la littérature, il existe deux restrictions. Mais on ne sait pas si ces restrictions sont équivalentes.

Proposition 1.23. *Pour k un noyau, on définit*

$$\mathcal{M}_k := \left\{\mu \in \mathcal{M} \mid \int_{\mathcal{X}} \sqrt{k(x, x)}\ |\mu|(\mathrm{d}x) < +\infty\right\} \text{ et } \mathcal{M}^k := \left\{\mu \in \mathcal{M} \mid \mathcal{H}_k \subset \mathcal{L}^1(\mathrm{d}\mu)\right\}.$$

Sur ces deux ensembles, l'intégrale est continue et on a l'inclusion suivante,

$$\mathcal{M}_k \subset \mathcal{M}^k.$$

Démonstration.

- Soit $\mu \in \mathcal{M}_k$, on a pour $x \in \mathcal{X}$,

$$|f(x)| = |\langle K(x), f\rangle| \leq \|f\|\|K(x)\| = \|f\|\sqrt{k(x,x)}.$$

Donc par inégalité triangulaire, $f \in \mathcal{L}^1(\mathrm{d}\mu)$. De plus,

$$\int f(x)\ \mu(\mathrm{d}x) \leq \|f\| \int \sqrt{k(x,x}\ |\mu|(\mathrm{d}x),$$

d'où la continuité de l'intégrale.

- Soit $\mu \in \mathcal{M}^k$, on définit la projection identité

$$\mathrm{Id}\colon (\mathcal{H}_k, \|\cdot\|_k) \to (\mathrm{L}^1(\mathrm{d}\mu), \|\cdot\|_1).$$

On souhaite montrer la continuité de cette application. Comme l'intégrale sur $\mathrm{L}^1(\mathrm{d}\mu)$ est continue, on aura la conclusion. Pour montrer la continuité de Id, nous allons utiliser la caractérisation du graphe fermé des applications continues dans les espaces de Banach. Soit $(f_n)_{n\in\mathbb{N}}$ une suite tendant vers $f \in \mathcal{H}_k$ tel que $\mathrm{Id}(f_n) \to g$ pour $\|\cdot\|_1$. Il faut montrer que $f = g$ $\mu-$pp. Or la convergence dans $\mathcal{H}_k$ implique la convergence simple et la convergence dans $\mathrm{L}^1(\mathrm{d}\mu)$ implique la convergence $\mu-$ pp à une sous suite près, ainsi par unicité de la limite, $f = g$ $\mu-$pp.

□

Dans la suite, on se limitera à l'ensemble $\mathcal{M}_k$. Cet ensemble a l'avantage d'être plus explicite. En effet, l'ensemble $\mathcal{H}_k$ est construit en partie par complétion donc il est difficile de déterminer exactement $\mathcal{M}^k$ juste à partir du noyau k. Ainsi comme la forme linéaire intégrale est continue, on peut représenter une mesure, i.e. il existe $K(\mu) \in \mathcal{H}_k$ tel que

$$\forall f \in \mathcal{H}_k,\ \langle K(\mu), f\rangle = \int_{\mathcal{X}} f(x)\ \mu(\mathrm{d}x).$$

Cette application $K\colon \mathcal{M}_k \to \mathcal{H}_k$ est appelée *Kernel Mean Embedding* (KME). On peut noter que pour le dirac δ_x, l'intégrale correspond à l'évaluation au point x, on peut donc voir le KME comme le prolongement du plongement $\mathcal{X} \to \mathcal{H}_k$. Ce plongement de l'espace $\mathcal{M}_k$ permet de comparer d'une nouvelle manière des mesures. Prenons $\mu, \nu \in \mathcal{M}_k$, pour comparer ces deux mesures, on va considérer la distance de leurs représentants, i.e. définir

$$d_k(\mu, \nu) = \|K(\mu) - K(\nu)\|_{\mathcal{H}_k}.$$

Cette application d_k, appelée *Maximum Mean Embedding* (MMD), est une distance si et seulement si le KME K est injectif. Dans le cas où le MMD définit bien une distance, la question naturelle est de comparer cette nouvelle distance aux topologies usuelles sur l'espace $\mathcal{P}$. La première topologie avec laquelle on comparera est celle de la convergence en loi, correspondant à la topologie faible. Nous verrons, sous quelques hypothèses, que l'article Simon-Gabriel et al. (2021) a caractérisé les noyaux bornés dont le MMD associé métrise la convergence en loi. Nous terminerons cette section avec des réponses préliminaires pour la métrisation de la distance de Wasserstein. L'étude complète sera effectuée dans le chapitre associé de cette thèse.

Topologie induite

Convergence en loi : On dira qu'un noyau métrise la convergence en loi (voir Billingsley (1999)) si $\mathcal{P} \subset \mathcal{M}_k$ et si pour toute suite de mesures de probabilité $(\mu_n)_{n\in\mathbb{N}}$ et $\mu \in \mathcal{P}$, la suite converge en loi vers μ si et seulement si $d_k(\mu_n, \mu) \to 0$. La première remarque que l'on peut faire vient de Sriperumbudur et al. (2010, Proposition 2).

Lemme 1.24. *Soit k un noyau, alors $\mathcal{P} \subset \mathcal{M}_k$ si et seulement si le noyau k est borné. De plus, si k est borné alors $\mathcal{M}_k$ est l'ensemble des mesures signées.*

Démonstration. Si le noyau k est borné alors naturellement $\mathcal{P} \subset \mathcal{M}_k$. Maintenant supposons k non borné. Par l'inégalité de Cauchy Schwarz, pour $x, y \in \mathcal{X}$,

$$|k(x,y)|^2 \leq k(x,x)k(y,y).$$

Donc un noyau k n'est pas borné si et seulement si $\sup k(x,x) = +\infty$. Ainsi on peut construire $(x_n)_{n\in\mathbb{N}}$ une suite dans $\mathcal{X}$ vérifiant

$$k(x_n, x_n) \geq n^2,$$

pour $n \in \mathbb{N}$. Alors la mesure $\mu = \sum_{n=0}^{+\infty} \frac{6}{\pi^2(n+1)^2}\delta_{x_n}$ est une mesure de probabilité et

$$\int_{\mathcal{X}} \sqrt{k(x,x)}\ \mu(\mathrm{d}x) \geq \sum_{n=0}^{+\infty} \frac{6n}{\pi^2(n+1)^2} = +\infty,$$

donc $\mu \notin \mathcal{M}_k$. □

L'autre propriété intéressante que l'on peut avoir sur le noyau k est

Lemme 1.25. *Soit k un noyau métrisant la convergence en loi alors le noyau k est continu. De plus, si le noyau k est continu alors $\mathcal{H}_k$ est inclus dans l'espace des fonctions continues.*

Démonstration. Comme $(\mathcal{X}, d)$ est un espace métrique, la continuité est équivalente à la continuité séquentielle. Soit $x, y \in \mathcal{X}$ et soit $(x_n, y_n)_{n\in\mathbb{N}}$ une suite tendant vers (x, y). On a $(\delta_{x_n})_{n\in\mathbb{N}}$ (resp. $(\delta_{y_n})_{n\in\mathbb{N}}$) converge en loi vers δ_x (resp. δ_y). Donc $K(\delta_{x_n}) \xrightarrow{\mathcal{H}} K(\delta_x)$ et $K(\delta_{y_n}) \xrightarrow{\mathcal{H}} K(\delta_y)$, ainsi

$$k(x_n, y_n) = \langle K(\delta_{x_n}, K(\delta_{y_n})\rangle \to \langle K(\delta_x), K(\delta_y)\rangle = k(x,y).$$

□

Dans l'article Sriperumbudur (2016, Theorem 3.2), l'auteur a donné une condition suffisante pour qu'un noyau métrise la convergence en loi mais sous des conditions trop fortes sur l'espace $\mathcal{X}$. En effet, il supposait que l'ensemble $\mathcal{X}$ était un espace polonais localement compact. Cette hypothèse empêchait de considérer, par le Lemme de la boule unité compacte, des espaces de Banach de dimension infinie. Simon-Gabriel et al. (2021) ont alors généralisé le résultat obtenu en diminuant les hypothèses. Si l'on note par $\mathcal{C}_0(\mathcal{X}, \mathbb{R})$ l'ensemble des fonctions continues tendant vers zéro à l'infinie, i.e.

$$\forall \varepsilon > 0,\ \exists D \subset \mathcal{X} \text{ compact},\ \forall x \notin D,\ |f(x)| < \varepsilon.$$

Théorème 1.26. *Soit k un noyau continu borné tel que le RKHS associé $\mathcal{H}_k$ soit inclus dans $\mathcal{C}_0$, alors on a équivalence entre*

i) le MMD d_k métrise la convergence en loi ;

ii) le KME K est injectif sur $\mathcal{M}$ (sur $\mathcal{P}$ si $\mathcal{X}$ est compact).

Il faut noter que le cas compact et non compact diffèrent. On a besoin d'une condition plus forte dans le cas non compact. On demande au MMD d_k de séparer non pas uniquement les mesures de probabilités mais toutes les mesures signées finies. De plus, l'hypothèse $\mathcal{H}_k \subset \mathcal{C}_0$ n'est restrictive que dans le cas non compact. En effet, si $\mathcal{X}$ est compact alors $\mathcal{C}_0$ est l'ensemble des fonctions continues et par le Lemme 1.25, le RKHS $\mathcal{H}_k$ est bien inclus dans l'espace des fonctions continues. Pour éclaircir ce résultat, on peut donner une condition nécessaire et suffisante sur cette hypothèse dans le cas général.

Lemme 1.27. *Soit k un noyau, on a l'équivalence entre*

i) pour tout $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}_k$ et $\sup k(x,x) < +\infty$;

ii) $\mathcal{H}_k \subset \mathcal{C}_0$.

Nous n'allons pas donner toute la preuve de ce théorème. En effet, le sens direct se base sur deux lemmes un peu techniques (Lemme 9 et 10 de Simon-Gabriel et al. (2021)). On va plutôt s'attarder sur la réciproque qui est amusante dans un certains sens. En effet, cette implication dit que dès que le MMD d_k est une distance sur $\mathcal{M}$ alors automatiquement cette distance métrise la convergence en loi. Pour voir rapidement les raisons de ce fait, on peut reformuler l'expression de la norme du KME, pour $\mu \in \mathcal{M}$,

$$\|K(\mu)\|_k = \sup_{\|f\|_k=1} \langle K(\mu), f\rangle = \sup_{\|f\|_k=1} \int_{\mathcal{X}} f(x)\ \mu(\mathrm{d}x).$$

Ainsi, l'injectivité du KME K va impliquer la densité, pour la norme infinie, du RKHS $\mathcal{H}_k$ dans $\mathcal{C}_0$. Si ce n'était pas le cas, nous pourrions, d'après le théorème de Hahn-Banach, trouver deux formes linéaires continues φ_1, φ_2 sur $\mathcal{C}_0$ distinctes mais identiques sur $\mathcal{H}_k$. Or le théorème de représentation de Riesz-markov relie les formes linéaires continues sur $\mathcal{C}_0$ et $\mathcal{M}$, i.e. on peut trouver $\mu_1, \mu_2 \in \mathcal{M}$ distinctes tel que

$$\forall f \in \mathcal{C}_0,\ \varphi_1(f) = \int_{\mathcal{X}} f(x)\ \mu_1(\mathrm{d}x) \quad \text{et} \quad \varphi_2(f) = \int_{\mathcal{X}} f(x)\ \mu_2(\mathrm{d}x).$$

Comme φ_1 et φ_2 coïncident sur $\mathcal{H}_k$, on obtient que $K(\mu_1) = K(\mu_2)$, ce qui contredit l'injectivité de K. La suite de la preuve est donc de montrer que la densité de $\mathcal{H}_k$ est suffisante pour avoir la métrisation de la convergence en loi.

Distance de Wasserstein : Métriser la convergence en loi c'est bien, mais peut-on espérer faire mieux ? Une distance très utilisée en pratique plus forte que la convergence en loi, aussi appelée convergence faible, est la distance de Wasserstein. Ces distances sont utilisées dans de nombreux domaines tels que le transport optimal (Villani, 2009), l'algorithme d'apprentissage (Frogner et al., 2015; Arjovsky et al., 2017), la recherche d'images (Rubner et al., 2000)... Pour $\alpha > 0$, on définit l'espace des mesures de probabilité ayant un moment d'ordre α,

$$\mathcal{P}_\alpha(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P} \mid \int_{\mathbb{R}^d} \|x\|^\alpha\ \mu(\mathrm{d}x) < +\infty \right\}.$$

Pour $\mu, \nu \in \mathcal{P}_\alpha$, on pose $\Gamma(\mu,\nu)$ l'ensemble des couplages de μ et ν, c'est à dire les mesures γ sur $\mathbb{R}^d \times \mathbb{R}^d$ tel que

$$\gamma(A \times \mathbb{R}^d) = \mu(A) \quad \text{et} \quad \gamma(\mathbb{R}^d \times A) = \nu(A),$$

pour tout borélien $A \subset \mathbb{R}^d$. La distance de Wasserstein d'ordre α est définie, pour $\alpha \geq 1$, par

$$W_\alpha(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \|x-y\|^\alpha\, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/\alpha}, \quad \mu, \nu \in \mathcal{P}_\alpha.$$

Pour $\alpha \in (0,1)$, elle est définie comme

$$W_\alpha(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \|x-y\|^\alpha\, \gamma(\mathrm{d}x, \mathrm{d}y).$$

Cette famille est rangée de manière croissante, i.e. la convergence pour W_α implique la convergence pour W_β lorsque $\beta < \alpha$. De plus, ces distances sont plus fortes que la convergence en loi. Pour ces distances, la classification des noyaux les métrisant était très peu abordée dans la littérature. Le Chapitre 5 s'intéresse à cette question. Le but de ce chapitre, basé sur l'article

Modeste and Dombry (2022), est d'exhiber une famille de noyaux métrisant les distances de Wasserstein d'ordre $\alpha > 1$. Rappelons la famille des Energy kernels pour $\alpha \in (0,1)$ et $x, y \in \mathbb{R}^d$,

$$k_\alpha(x,y) = \|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x-y\|^{2\alpha}.$$

Nous montrerons aussi dans le chapitre que

Théorème 1.28. *Soit $\alpha \in (0,1)$ et $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_\alpha$.*

i) $W_\alpha(\mu_n, \mu) \to 0$ *implique* $d_{k_\alpha}(\mu_n, \mu) \to 0$.

ii) $d_{k_\alpha}(\mu_n, \mu) \to 0$ *implique* $W_\beta(\mu_n, \mu) \to 0$ *pour tout* $\beta < \alpha$.

Ce résultat peut être interprété comme le fait que le MMD de k_α est une distance intermédiaire entre W_α et W_β pour $\beta < \alpha$. Malheureusement, ces noyaux ne sont pas assez forts pour métriser les distances de Wasserstein pour $\alpha \geq 1$ notamment W_1 ou W_2. Nous introduirons donc une nouvelle famille de noyaux pouvant métriser ces distances.

Chapitre 2

Modeling and scoring dynamic probabilistic forecasts

Ce premier chapitre est constitué du préprint Modeste et al. (2023), soumis à *Electronic Journal of Statistics*. Nous montrons que le score propre est un objet pertinent pour discriminer la loi idéale.

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. Some connections are also done in terms of Wasserstein distance, and links are given between scoring rules and reproducing kernels.

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., 2021a), 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 2.13 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. 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. Finally, Section 4 focuses on the family of kernel scores, for which sufficient conditions on the kernel are discussed that guarantee a proper scoring rule, and the links with the theory of reproducing kernels is discussed. Section 5 contains all the proofs, and an appendix gives some results on the regularity of scoring rules.

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 2.1. *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 2.2. *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 2.23 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 2.24 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 2.3. 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

$$U_{n+1} = \alpha U_n + \varepsilon_{n+1}, \quad n \in \mathbb{N}, \tag{2.1}$$

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

$$F^*_{n,T} = \mathcal{N}\left(\alpha^T U_n, \frac{1-\alpha^{2T}}{1-\alpha^2}\sigma^2\right), \quad n \in \mathbb{N}. \tag{2.2}$$

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 (2.2), the plug-in method suggests the dynamic probabilistic forecast with lead time T,

$$F_{n,T} = \mathcal{N}\left(\hat{\alpha}_n^T U_n, \frac{1-\hat{\alpha}_n^{2T}}{1-\hat{\alpha}_n^2}\hat{\sigma}_n^2\right), \quad n \geq 1, \tag{2.3}$$

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 (2.1). 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 (2.3) is a stationary forecast.

Example 2.4. 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 2.5. 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 2.6. 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 3.3 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 2.7. 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 1.

Definition 2.8. *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

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

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 2.9. For real-valued observations, the most important and widely used scoring rule is the Continuous Ranked Probability Score, shortly noted CRPS (Epstein (1969a); 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

$$\mathrm{CRPS}(F, y) = \mathbb{E}[|X - y|] - \frac{1}{2}\mathbb{E}[|X - X'|] \tag{2.4}$$

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 2.10. A generalization of the CRPS can be obtained from (2.4) by introducing the so-called energy kernel, defined for $\alpha > 0$ and $\beta \in (0, \infty]$ by

$$\rho_{\alpha,\beta}(x, y) = \|x - y\|_\beta^\alpha, \quad x, y \in \mathbb{R}^d, \tag{2.5}$$

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 (2.4) 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), as well as the discussion in Section 4. Additionally, some interesting results can be established on the divergence of the energy score when $\beta = 2$ and $\alpha \in (0, 2)$, see Section 3.3.

Example 2.11. 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 2.12. *Let S be a proper scoring rule. Then a.s.*

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

and this implies $\mathbb{E}[S(F, Y)] \ge \mathbb{E}[S(F^*, Y)]$. 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 2.13. *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 $k = 1, \dots, T$,*

$$\sum_{i=1}^{n} \mathbb{E}\left[\left(\delta_i^k\right)^2 \mid \mathcal{F}_{i+T-k}\right] = O(n), \quad a.s.. \tag{2.6}$$

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

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) \geq 0\,. \tag{2.7}$$

Moreover, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} S\left(F_i, Y_{i+T}\right) - S\left(F^*_{i,T}, Y_{i+T}\right) = 0 \quad a.s. \tag{2.8}$$

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 (2.6) 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 (2.7) 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 2.12. The vanishing limit (2.8) 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 2.13 is based on the strong law of large numbers for square integrable martingales and does not assume any stationarity condition. The technical condition (2.6) is required for the martingale convergence theorem. See Section 5.2 for more details.

Remark 2.14. In the case of the Energy Score with $\alpha \leq 1$ and $1 \leq \beta$ (see Example 2.10), a simple sufficient condition for condition (2.6) 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 A.5 in the Appendix 1 for more details.

Example 2.15. Theorem 2.13 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 2.13 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.

Example 2.16. 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 2.13 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.

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 2.17. *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.

3.3 Links between energy scores and Wasserstein distance

The purpose of this section is to provide a more explicit interpretation of the divergence condition (2.8) in the case of the energy scores defined via (2.5). 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 2.18. *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 2.13 can be rewritten as follows : assuming condition (2.6) 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 Kernel Score

4.1 Presentation

The energy score defined in Example 2.10 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

$$\forall y \in \mathcal{Y},\ \rho(y, y) = 0. \tag{2.9}$$

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

$$\sum_{1\leq i,j\leq n} a_i a_j \rho(y_i, y_j) \leq 0\ . \tag{2.10}$$

Note that a kernel ρ satisfying Assumptions (2.9) and (2.10) 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

$$\mathcal{L}_\rho := \left\{ F \in \mathcal{P}(\mathcal{Y}) \mid \exists y_0 \in \mathcal{Y},\ \int_{\mathcal{Y}} \rho(y, y_0)\ \mathrm{d}F(y) < +\infty \right\}, \tag{2.11}$$

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 2.19. The well-defined aspect of these integrals is justified in Remark 21 of Sejdinovic et al. (2013). Note that conditions (2.9) and (2.10) 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 (2.9) and (2.10) 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 (2.11),

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

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.

4.2 Scoring rules and Reproducing Kernels

This review section aims at explaining in the proof of Theorem 2.18, the links between this article and the results used in the article Modeste and Dombry (2022). 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.

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

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 2.20 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.

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

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

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

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

$$\mathcal{L}_k := \left\{ F \in \mathcal{P}(\mathcal{Y}) \mid \int \sqrt{k(y,y)}\ \mathrm{d}F(y) < +\infty \right\}. \tag{2.12}$$

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 2.20. *[(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 (2.9) and (2.10), then the kernel defined by*

$$\forall x, y \in \mathcal{Y},\ k(x,y) = \rho(y_0, x) + \rho(y_0, y) - \rho(x, y) \tag{2.13}$$

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 γ_k satisfy

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

Example 2.21. 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 γ_k is defined for *strictly* more probability measures.

Remark 2.22. 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 (2.12) and $y \in \mathbb{R}^d$. If the kernel k is associated with the conditionally negative kernel ρ by Equation (2.13), 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.

5 Proofs

5.1 Proofs of Section 2

Lemma 2.23. *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 2.24. *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), or Appendix 2 for more details. 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 2.13. 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 2.12 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 (2.6) because $\langle M^k \rangle_n = O(n)$. As a conclusion, one gets also in ii) that $M^k_n/n \to 0$ as $n \to \infty$.

□

Our proof of Theorem 2.18 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 $F, G \in \mathcal{P}^\alpha(\mathbb{R}^d)$,

$$\mathbf{div}_S(F,G) = \frac{1}{C(d,\alpha)} \int_{\mathbb{R}^d} \frac{|\hat{F}(t) - \hat{G}(t)|^2}{\|t\|_2^{d+\alpha}} \, \mathrm{d}t, \tag{2.14}$$

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 Subsection 4.2.

Proof of the Theorem 2.18. 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 (2.14)

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

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 2.25. *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 $t \in \mathbb{R}^+ \mapsto t^{d-1}\phi(t) \in \mathrm{L}^1(\mathbb{R})$.*

□

Chapitre 3

Testing ideal calibration for sequential predictions

L'objectif de ce chapitre est de présenter des tests permettant de déterminer si une prédiction est idéale, notion introduite dans le chapitre précédent, ou pas. La rédaction d'un article est en cours de finalisation sur la base de ce chapitre et donnera lieu prochainement à soumission.

Abstract : Forecasts and their evaluation are major tasks in statistics. In real applications, forecasts often take the form of a dynamic process evolving over time and this sequential point of view must be taken into account. A strategy for forecast evaluation is calibration theory based on the *Probability Integral Transform*. The idea is to check the conformity between the forecast and the observation. Here, ideal forecasts are characterized by conditional calibration and we present some new tests based on regression trees.

1 Introduction

Forecasting is an important task in statistics with many applications such as in meteorology (Vannitsem et al., 2021), hydrology (Tiberi-Wadier et al., 2021), health (Henzi et al., 2021a), energy (Hong et al., 2016). To take account for uncertainties or statistical errors, a prediction is not just a single value. Several approaches exist to deal with these uncertainties, including interval prediction and probabilistic prediction. We will focus on the latter approach. To estimate the trend of the future, predictions then take the form of probability measures, i.e. the forecaster predicts several, even infinitely many, possible scenarios with different probabilities of success. This is the idea introduced by Epstein (1969b) where the process studied is governed by deterministic laws but the initial conditions are unknown.

Forecast verification is crucial in order to improve a forecasting method or compare different forecast strategies. Assessing forecast accuracy is done by comparing the predictive distributions and the actual observations which are real valued. But since these two quantities are not intrinsically comparable, assessing in this case is a difficult task. Two main methods exist, scoring rules (Gneiting and Raftery, 2007) and calibration theory (Gneiting et al., 2007). The idea of the first approach is to create a pseudo distance between the observations and the predictions, while the second one is in some way more qualitative. We can find an early introduction of the notion of calibration in Dawid (1984) and Diebold et al. (1998). The reliability of a probabilistic forecast is defined as its skill to be conform with the actual observation.

The evaluation methods depend on the assumptions about these measures. In some cases, probability measure is discrete and uniform over its atoms. This type of prediction is called ensemble forecast. It is a common approach in meteorology with the *General Circulation Models* (GCMs) where each atom represents a different scenario starting from a different initial point. Several diagnostic tools have been introduced (Bröcker, 2009; Weigel, 2011), the rank histogram is one

of them and one of the most studied (Anderson, 1996; Talagrand et al., 1997). The measures can also be assumed with a density, as is the case in financial risk management (Diebold et al., 1998).

In the general case, the fundamental notion of *Probability Integral Transform* (PIT, David and Johnson (1948)) is introduced, which plays a crucial role in calibration theory. More recently, Tsyplakov (2013) describes the use of the PIT as a diagnostic tool for calibration. The review by Gneiting and Katzfuss (2014) provides a nice discussion of these notions. Several types of calibrations have been introduced over time. Their definitions and links will be given in Section 2. For this article, we are interested in a particular type of calibration, the ideal calibration. A prediction of a phenomenon Y is said to be ideally calibrated with respect to an information $\mathcal{F}$ if the latter prediction is the conditional distribution of Y for $\mathcal{F}$. That is to say that the forecaster predicts the phenomenon perfectly with respect to the known information. Few results exist in the literature for this type of calibration. This notion of ideal calibration is closely related to the cross-calibration defined by Strähl and Ziegel (2017), which aims at predicting perfectly with the knowledge of other forecasters' information. In this context, the authors propose two tests based on the study of PIT.

In Bröcker (2022), the author also focuses on ideal calibration but in the specific case of binary events. This sub-case of probabilistic prediction is called *probability forecasting*. The concept of his test is to write the calibration in terms of an empirical process involving the observations of the phenomenon and its forecasters. Under some conditions, this empirical process has good asymptotic behavior under the ideal calibration hypothesis. We will proceed in an analogous way.

Section 2 sets the definitions and details the classical framework, the PIT and the different notions of calibration. This section ends with Corollary 3.6 that is a key result from which we will be able in Section 3 to write the ideal calibration in terms of an empirical process. Under certain conditions, this process will converge to a Gaussian limit process. Since this limit has an unknown distribution, we will justify the bootstrap to approximate it. Section 4 introduces three new tests based on regression trees (CART algorithm). The purpose of this section is to show that this regression is rewritten as the functional of the empirical process that involves the PITs. The performance of these three tests are then numerically investigated in Section 5 in an autoregressive model framework. We conclude this paper with Section 6 where we give some limitations of ideal calibration et our tests. Then we present a recent notion of weaker calibration and discuss an adaptation perspective for our tests. All the proofs are relegated to Section 7.

2 Calibration theory for the validation of dynamic forecast

2.1 Probability Integral Transform

Before introducing the different notions of calibration for probabilistic forecasts, the *Probability Integral Transform* (PIT) is defined in the case where the probability measures are not random.

Definition 3.1. *Let F be a deterministic CDF, Y be a random variable and $V \sim \mathrm{Unif}([0,1])$ independent of the variable Y. We define the Probability Integral Transform (PIT) as*

$$Z_F^Y = VF(Y^-) + (1-V)F(Y).$$

It is the generalization of the classical PIT $F(Y)$ when F is continuous. It is well known that in the continuous case, $F(Y)$ is uniformly distributed on $[0,1]$ if and only if F is the CDF of Y. It is still true for the generalization. We do not know a reference showing the first implication of the following lemma, so we prove it in Section 7.

Lemma 3.2. *Let F be a CDF and Y be a random variable. Let Z_F^Y denote the associated PIT. The two statements are equivalent*

1. $Z_F^Y \sim \text{Unif}([0,1])$;
2. F *is the CDF of* Y.

2.2 Prediction Space

We describe the classical framework of calibration developed in Gneiting and Ranjan (2013) and Strähl and Ziegel (2017). Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space and let $\mathcal{F}$ be a sub σ-algebra. The σ-algebra $\mathcal{F}$ will represent the information we have. It will be sometimes assumed that this information comes from observations $X\colon \Omega \to \mathbb{R}^d$. We denote by $\mathcal{M}_1(\mathbb{R})$ the space of probability measures on $\mathbb{R}$. This set is endowed with its Borel σ-algebra. A real valued random forecast for a σ-algebra $\mathcal{F}$ is a measurable map F from $(\Omega, \mathcal{G})$ to $\mathcal{M}_1(\mathbb{R})$ which is $\mathcal{F}$ measurable. This means that the prediction F uses only the information contained in the σ-algebra $\mathcal{F}$. With a slight abuse of notation, we will often identify a probability measure and its CDF. We also consider the variable of interest Y. In our case, this variable will be real valued. It represents the quantity that we try to predict. To complete this model, we introduce an ad hoc variable $V \sim \text{Unif}([0,1])$ independent of the information $\mathcal{F}$, and thus of X, and of the variable of interest Y. This variable will be used to define the PIT for random forecasters.

We will also need this framework in a sequential format. Let $(\Omega, \mathcal{G}, \mathbb{P})$ be still a probability space and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration, i.e. for $k \leq n$, $\mathcal{F}_k \subset \mathcal{F}_n$. This condition means that no information is forgotten in the process. We will also suppose sometimes that this filtration will be generated by a sequence $(X_n)_{n\in\mathbb{N}}$, then

$$\mathcal{F}_n = o(X_0, \dots, X_n), \text{ for } n \in \mathbb{N}.$$

In the sequential framework, a sequence of random forecasts $(F_n)_{n\in\mathbb{N}}$ will be maps from $(\Omega, \mathcal{G})$ to $\mathcal{M}_1(\mathbb{R})$ which will be measurable for the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. The sequence of variables of interest will be denoted $(Y_n)_{n\in\mathbb{N}}$ and we introduce a lead time $T \geq 1$. At the time n, the forecaster F_n will aim to predict the variable Y_{n+T}. We will see that the case $T = 1$ is easier than the case $T \geq 2$. This fact has already been noted in the literature, see Knüppel (2015) for a discussion about it. Finally, we consider an i.i.d. sequence $(V_n)_{n\in\mathbb{N}}$ of random variables uniformly distributed on $[0,1]$ independent of the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and the sequence $(Y_n)_{n\in\mathbb{N}}$.

2.3 Calibration of a probabilistic forecast

In the following, we will consider the PIT of a random forecast. The justification of the measurability is provided in Proposition 3.22. A random forecast F is said to be *probabilistically calibrated* for Y if the PIT is uniformly distributed. Given a σ-algebra $\mathcal{F}$, there exists a special forecast F called *ideal forecast* defined as

$$F = \mathcal{L}(Y \mid \mathcal{F}).$$

This forecast is said to be *ideally calibrated*. The case where F is ideal with respect to $\sigma(F)$ is called *auto-calibration*. Gneiting and Ranjan (2013, Theorem 2.8) show that ideal calibration relative to any σ-algebra implies probabilistic calibration. The converse is true only for binary outcome and auto calibration (see Gneiting and Ranjan (2013, Theorem 2.11) for the converse and Gneiting and Resin (2021, Example 2.4) for a counter example in the general case). Another notion of calibration can be defined from the PIT which was introduced by Mitchell and Wallis (2011) in the case where F is continuous. We recall that the variable V in the definition of the PIT is independent of $\mathcal{F}$ and Y.

Definition 3.3. *Let F be a random forecast and Y be a random variable. The forecast is completely calibrated with respect to a σ-algebra $\mathcal{F}$ if $Z_F^Y \sim \text{Unif}([0,1])$ and Z_F^Y is independent of $\mathcal{F}$.*

It is clear that *complete calibration* for any σ-algebra implies probabilistic calibration. We find a similar notion in Strähl and Ziegel (2017) with *cross-calibration*. We will discuss the link between their article and ours in Section 6.1. This link is clearer when we rewrite the definition as follows

$$\mathcal{L}\left(Z_F^Y \mid \mathcal{F}\right) = \text{Unif}([0,1]). \tag{3.1}$$

It is easy to check the equivalence between Definition 3.3 and (3.1). The link between complete and ideal calibration was partially found by Diebold et al. (1998). Indeed, in the first Proposition of their Part 3, they show under a $\mathcal{C}^2$ condition on the forecast, an ideally calibrated forecast is completely calibrated. We find also a part of the first implication in Gneiting and Ranjan (2013) with their Theorem 2.9. In Strähl and Ziegel (2017), the equivalence between these two notions was established (Proposition 2.11). We propose a different proof based on the Fubini theorem for the conditional distribution.

Proposition 3.4. *Let F be a $\mathcal{F}$-measurable random forecast and Y be a random variable. The two following statements are equivalent*

1. *F is ideally calibrated relative to $\mathcal{F}$*
2. *The forecast is completely calibrated relative to $\mathcal{F}$.*

In the following, we will use the designation ideally calibrated for historical reason but we will use the characterization of the complete calibration. These definitions naturally extend to a dynamic framework.

Definition 3.5. *Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of forecasts and $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration. A dynamic forecaster $(F_n)_{n\in\mathbb{N}}$ is said to be $*$-calibrated with $* \in \{$auto, ideally, completely$\}$ if for each $n \in \mathbb{N}$, the forecast F_n is $*$-calibrated with respect to $\mathcal{F}_n$.*

Proposition 3.4 thus has an immediate rewriting for dynamic forecasts. For a lead time $T \geq 1$ and $n \in \mathbb{N}$, we denote by $Z_n := Z_{F_n}^{Y_{n+T}}$ the PIT of Y_{n+T} by F_n.

Corollary 3.6. *Let $T \geq 1$ be a lead time. We consider a dynamic forecast $(F_n)_{n\in\mathbb{N}}$ measurable with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. We have equivalence between*

1. *The dynamic forecast $(F_n)_{n\in\mathbb{N}}$ is completely calibrated relatively to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*
2. *The dynamic forecast $(F_n)_{n\in\mathbb{N}}$ is ideally calibrated relatively to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.*

Moreover if $(\mathcal{F}_n)_{n\in\mathbb{N}}$ contains the filtration endowed by $(Y_n)_{n\in\mathbb{N}}$, the sequence $(Z_n)_{n\in\mathbb{N}}$ is $(T-1)$-dependent.

The assumption that $(\mathcal{F}_n)_{n\in\mathbb{N}}$ contains the filtration endowed by $(Y_n)_{n\in\mathbb{N}}$ is realistic in many applications. It means that we used at least the information coming from the past observations.Under calibration, in the case where $T = 1$, the sequence of PIT is i.i.d. For a larger lead time $T \geq 2$, it is more difficult to determine the law of $(Z_n)_{n\in\mathbb{N}}$. Even with a stationary assumption, we do not know the behaviour of the marginal $(Z_0, \ldots, Z_{T-1})$. It is for this reason that the following section will restrict itself to the case $T = 1$.

Recently, several notions of calibration were defined (Gneiting and Resin, 2021) which are weaker. We will discuss these notions in Section 6.

3 Empirical process and calibration

3.1 Assumptions and notation

The idea of this article is to write the calibration assumption in terms of some empirical processes as in Bröcker (2022). We will slightly modify the dynamic framework and assume that the sequences are indexed in $\mathbb{Z}$. Let us recall that, Z_n is the PIT of Y_{n+T} by F_n. The following assumptions are made :

(A1) **Model :** the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ is generated by a vectorial sequence $(X_n)_{n\in\mathbb{Z}}$, contains the filtration endowed by the quantity of interest $(Y_n)_{n\in\mathbb{Z}}$ and for $n \in \mathbb{Z}$

$$\mathcal{L}(Y_{n+T} \mid \mathcal{F}_n) = \mathcal{L}(Y_{n+T} \mid X_n);$$

(A2) **Stationarity :** the sequence $(Z_n, X_n)_{n\in\mathbb{Z}}$ is stationary ;

(A3) **Dependence :** For two $\sigma-$algebras $\mathcal{A}$ and $\mathcal{B}$, the $\alpha-$mixing coefficient is defined by

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A\in\mathcal{A}, B\in\mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

The $\sigma-$algebra are independent if and only if the coefficient is null, see Rosenblatt (1961) for more details on mixing dependent coefficients. We assume that there exists $\varepsilon > 0$, such that

$$\alpha(n) := \alpha(\sigma(\ldots, X_{-1}, X_0), \sigma(X_n, X_{n+1}, \ldots)) = O(n^{-2d-\varepsilon}).$$

The consequence of the Markovian Assumption (A1) will be discussed in Remark 3.9. This formulation of Assumption (A2) is mathematical. It is stated in this way to make fewer assumptions. In a less general framework, for example, if the triplet $(X_n, Y_{n+1}, F_n)_{n\in\mathbb{Z}}$ is stationary then Assumption (A2) is verified. This framework means that the predicted and observed phenomena are stationary as well as the way of forecasting. We want to test the following null hypothesis

$$(H_0) : \text{the sequence } (F_n)_{n\in\mathbb{Z}} \text{ is ideally calibrated relatively to } (\mathcal{F}_n)_{n\in\mathbb{Z}}.$$

This means that for each $n \in \mathbb{Z}$, $F_n = \mathcal{L}(Y_{n+T} \mid \mathcal{F}_n)$. With Assumption (A1) and Corollary 3.6, this null hypothesis can be rewritten

$$(H_0) : \text{for each } n \in \mathbb{Z},\ Z_n \sim \text{Unif}([0,1]) \text{ and } Z_n \text{ is independent of } X_n.$$

We may assume without loss of generality that the stationary sequence $(X_n)_{n\in\mathbb{Z}}$ takes its values in $[0,1]^d$. The common CDF of the X_n is denoted by F. We consider the empirical process

$$\mathbb{G}^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{Z_i\le y, X_i\le t\}} - yF(t)\right), \quad y \in [0,1], x \in [0,1]^d,$$

and denote by Γ the limit covariance function

$$\Gamma((y,t),(y',t')) = \sum_{i\in\mathbb{Z}} \text{Cov}\left(\mathbb{1}_{\{Z_0\le y, X_0\le t\}}, \mathbb{1}_{\{Z_i\le y', X_i\le t'\}}\right).$$

Here the symbol $\le$ denotes componentwise comparison of vectors so that $x \le t$ means that $x_i \le t_i$ for all $i = 1, \ldots, d$. The negation $x \not\le t$ means that $x_i > t_i$ for some $i = 1, \ldots, d$.
The following Proposition is a direct application of Theorem 10.2 in Dedecker et al. (2007).

Proposition 3.7. *Under Assumptions (A1)-(A3) and assuming the calibration null hypothesis (H_0), the empirical process converges in distribution*

$$\sqrt{n}\mathbb{G}^{(n)} \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty\left([0,1]\times[0,1]^d\right)$$

and the limit $\mathbb{G}$ is a centered Gaussian process with covariance function Γ.

3.2 Decomposition of the empirical process

Let us consider the decomposition of the empirical process

$$\mathbb{G}^{(n)}(y,t) = \mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) + y\mathbb{F}_3^{(n)}(y,t)$$

with

$$\mathbb{F}_1^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Z_i\le y\}} - y\right)F(t)$$

$$\mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Z_i\le y\}} - y\right)\left(\mathbb{1}_{\{X_i\le t\}} - F(t)\right)$$

$$\mathbb{F}_3^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\le t\}} - F(t)$$

This decomposition is motivated by the following simple yet interesting lemma.

Lemma 3.8. *The following properties hold true :*

1. *the process $\mathbb{F}_1^{(n)}$ is centered if and only if $Z_i \sim \mathrm{Unif}([0,1])$ for all i ;*
2. *the process $\mathbb{F}_2^{(n)}$ is centered if and only if Z_i and X_i are independent for all i ;*
3. *the process $\mathbb{F}_1^{(n)} + \mathbb{F}_2^{(n)}$ is centered if and only if $Z_i \sim \mathrm{Unif}([0,1])$ and Z_i and X_i are independent for all i.*

As a consequence, the ideal calibration assumption (H_0) holds if and only if $\mathbb{F}_1^{(n)}$ and $\mathbb{F}_2^{(n)}$ are both centered processes.

A short interpretation of this lemma is that the first term tests probabilistic calibration and the second term tests the independence of the PIT with the information. The last term only uses the information $(X_n)_{n\in\mathbb{Z}}$. Thus in general, to test ideal calibration, one should essentially use the first two terms $\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}$ and not the last one $\mathbb{F}_3^{(n)}$. Indeed, the sum encodes ideal calibration and has the advantage of being observable since it does not depend on the unknown CDF F, for $y,t \in [0,1]\times[0,1]^d$,

$$\mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Z_i\le y\}} - y\right)\mathbb{1}_{\{X_i\le t\}}.$$

Remark 3.9. Assumption (A1) on the Markovian character of the conditional distribution is used only to prove the last consequence of the previous lemma. Without it, the three points are still true.

We now investigate the asymptotic behavior of the empirical processes $(\mathbb{F}_i^{(n)})_{1\le i\le 3}$. We need to introduce these two limit covariances,

$$\Gamma_1((a,t),(b,s)) = \sum_{|i|\le T-1} F(t)F(s)\mathrm{Cov}\left(\mathbb{1}_{\{Z_0\le a\}}, \mathbb{1}_{\{Z_i\le b\}}\right),$$

$$\Gamma_2((a,t),(b,s)) = \sum_{|i|\le T-1} \mathrm{Cov}\left(\left(\mathbb{1}_{\{Z_0\le a\}} - a\right)\left(\mathbb{1}_{\{X_0\le t\}} - F(t)\right), \left(\mathbb{1}_{\{Z_i\le b\}} - b\right)\left(\mathbb{1}_{\{X_i\le s\}} - F(s)\right)\right)$$

The next theorem follows from Proposition 3.7 and the continuous mapping theorem.

Theorem 3.10. *Under Assumptions (A1)-(A3) and (H_0),*

$$\sqrt{n}\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \rightsquigarrow (\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3),$$

where $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3$ are Gaussian processes, with covariances respectively given by Γ_1 and Γ_2 for the two first processes. In the specific case where $T = 1$, these two processes are independent.

We remark that the covariance structure of $\mathbb{G}_1$ and $\mathbb{G}_2$ is simple because the covariance functions involve sums with $2T-1$ terms, which is a consequence of the $(T-1)$ dependence of the sequence of PITs $(Z_i)_{i\in\mathbb{Z}}$. In particular, when $T=1$, Γ_1 and Γ_2 involve only a single term. For the sake of brevity, we do not provide the full expression for the covariance function of $(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3)$.
For the following, we need also

$$\bar{\mathbb{G}}^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i \le y\}} \mathbb{1}_{\{X_i \not\le t\}} - y\bar{F}(t), \text{ where } \bar{F}(t) = 1 - F(t).$$

Similarly as before, we decompose this process as

$$\overline{\mathbb{G}}^{(n)}(y,t) = \overline{\mathbb{F}}_1^{(n)}(y,t) + \overline{\mathbb{F}}_2^{(n)}(y,t) + y\overline{\mathbb{F}}_3^{(n)}(y,t).$$

Note that for $i = 1, 2$,

$$\overline{\mathbb{F}}_i^{(n)}(y,t) = \mathbb{F}_i^{(n)}(y,\mathbf{1}) - \mathbb{F}_i^{(n)}(y,t), \ \forall y,t \in [0,1]\times[0,1]^d. \tag{3.2}$$

We get also a convergence result for this other empirical process.

Corollary 3.11. *Under Assumptions (A1)-(A3) and* (H_0),

$$\sqrt{n}\begin{pmatrix} \mathbb{F}_1^{(n)} & \overline{\mathbb{F}}_1^{(n)} \\ \mathbb{F}_2^{(n)} & \overline{\mathbb{F}}_2^{(n)} \\ \mathbb{F}_3^{(n)} & \overline{\mathbb{F}}_3^{(n)} \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 & \mathbb{G}(\cdot,\mathbf{1}) - \mathbb{G}_1 \\ \mathbb{G}_2 & -\mathbb{G}_2 \\ \mathbb{G}_3 & -\mathbb{G}_3 \end{pmatrix},$$

where $(\mathbb{G}_i)_{1\le i\le 3}$ *is the same as in Theorem 3.10.*

The tests we will propose in the next Section can be seen as functionals of these empirical processes and we will use the functional delta method to derive their asymptotic behaviour (van der Vaart and Wellner, 1996, Section 3.9).
As the ideal calibration is encoded by $(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)})$, it makes sense to ask that the asymptotics involve only the components $(\mathbb{G}_1, \mathbb{G}_2)$ as in the following theorem. Here $\ell^\infty = \ell^\infty([0,1]\times[0,1]^d)$.

Theorem 3.12. *Assume* $\Psi\colon \ell^\infty \times \ell^\infty \times \ell^\infty \to \ell^\infty$ *is differentiable at* $(0,0,0)$ *with*

$$\Psi(0,0,0) = 0 \text{ and } \partial_3\Psi(0,0,0) = 0.$$

Then under Assumptions (A1)-(A3) and the ideal calibration hypothesis (H_0),

$$\sqrt{n}\begin{pmatrix} \Psi\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \\ \Psi\left(\overline{\mathbb{F}}_1^{(n)}, \overline{\mathbb{F}}_2^{(n)}, \overline{\mathbb{F}}_3^{(n)}\right) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, 0) \\ \mathrm{d}_0\Psi(\mathbb{G}(\cdot,\mathbf{1}) - \mathbb{G}_1, -\mathbb{G}_2, 0) \end{pmatrix}.$$

It is too difficult to compute the exact law of these asymptotic empirical processes. Therefore we will approximate these limits by bootstrap. As already mentioned, the case $T=1$ is the simplest and we will focus on this case.

3.3 Bootstrap for the lead time $T = 1$

To approximate the limiting distribution arising in Theorem 3.12, we will use a form of bootstrap where we replace the sequence of PITs $(Z_n)_{n\in\mathbb{Z}}$ by a new uniform sequence $(Z_n^\star)_{n\in\mathbb{Z}}$ which is i.i.d. and independent of the information $(X_n)_{n\in\mathbb{Z}}$. Note that it is not obvious that the new empirical process has the same limit, because the sequences $(X_n, Z_n)_{n\in\mathbb{Z}}$ and $(X_n, Z_n^\star)_{n\in\mathbb{Z}}$ do not have the same distributions under (H_0). Indeed, (H_0) implies that Z_n is independent of the past $(X_1, \ldots, X_n)$ but in general, as Z_n depends on Y_{n+1}, there exists some dependency between

Z_n and X_{n+1}, and more generally between Z_n and the future observations. We introduce the usual notation to denote the bootstrap process

$$\mathbb{G}^{(n)\star}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i^\star \le y\}}\mathbb{1}_{\{X_i \le t\}} - yF(t).$$

and we construct the same process $(\mathbb{F}_i^{(n)\star}, \overline{\mathbb{F}}_i^{(n)\star})$ as previously, but with $Z_n^\star$ replacing Z_n.

Theorem 3.13. *Let $(Z_n^\star)_{n\in\mathbb{Z}}$ be an i.i.d. uniform random sequence, independent of $(Z_i)_{i\in\mathbb{Z}}$ and $(X_i)_{i\in\mathbb{Z}}$. Under the Assumptions (A1)-(A3) and the null hypothesis (H_0),*

$$\sqrt{n}\begin{pmatrix} \mathbb{F}_1^{(n)} & \overline{\mathbb{F}}_1^{(n)} \\ \mathbb{F}_2^{(n)} & \overline{\mathbb{F}}_2^{(n)} \\ \mathbb{F}_1^{(n)\star} & \overline{\mathbb{F}}_1^{(n)\star} \\ \mathbb{F}_2^{(n)\star} & \overline{\mathbb{F}}_2^{(n)\star} \\ \mathbb{F}_3^{(n)} - F & \overline{\mathbb{F}}_3^{(n)} - \overline{F} \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 & \mathbb{G}(\cdot,\mathbf{1}) - \mathbb{G}_1 \\ \mathbb{G}_2 & -\mathbb{G}_2 \\ \mathbb{G}_1^\star & \mathbb{G}^\star(\cdot,\mathbf{1}) - \mathbb{G}_1^\star \\ \mathbb{G}_2^\star & -\mathbb{G}_2^\star \\ \mathbb{G}_3 & -\mathbb{G}_3 \end{pmatrix},$$

where

$$\begin{pmatrix} \mathbb{G}_1 & \mathbb{G}(\cdot,\mathbf{1}) - \mathbb{G}_1 \\ \mathbb{G}_2 & -\mathbb{G}_2 \end{pmatrix} \overset{law}{=} \begin{pmatrix} \mathbb{G}_1^\star & \mathbb{G}^\star(\cdot,\mathbf{1}) - \mathbb{G}_1^\star \\ \mathbb{G}_2^\star & -\mathbb{G}_2^\star \end{pmatrix},$$

and these two vector processes are independent.

Remark 3.14. The term bootstrap is a slight abuses of language. In fact, resampling is done with a new sample.

This result states that the bootstrapped process has the same asymptotic behavior even if the two sequences $(Z_n, X_n)_{n\in\mathbb{Z}}$ and $(Z_n^\star, X_n)_{n\in\mathbb{Z}}$ do not have the same distribution. Then thanks to Theorem 3.12 and Theorem 3.13, we get the next result that justifies the use of bootstrap to adjust our tests. As the distributions of $(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3)$ and $(\mathbb{G}_1^\star, \mathbb{G}_2^\star, \mathbb{G}_3)$ are not the same, it is important to assume that the functional Ψ satisfies $\partial_3\Psi(0,0,0) = 0$.

Corollary 3.15. *Let $(Z_n^\star)_{n\in\mathbb{Z}}$ be an i.i.d. uniform random sequence, independent of $(Z_n)_{n\in\mathbb{Z}}$ and $(X_n)_{n\in\mathbb{Z}}$. Assume $\Psi\colon \ell^\infty \times \ell^\infty \times \ell^\infty \to \ell^\infty$ is differentiable at $(0,0,0)$ with*

$$\Psi(0,0,0) = 0 \text{ and } \partial_3\Psi(0,0,0) = 0,$$

then under Assumptions (A1)-(A3) and (H_0),

$$\sqrt{n}\begin{pmatrix} \Psi\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \\ \Psi\left(\mathbb{F}_1^{(n)\star}, \mathbb{F}_2^{(n)\star}, \mathbb{F}_3^{(n)}\right) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, 0) \\ \mathrm{d}_0\Psi(\mathbb{G}_1^\star, \mathbb{G}_2^\star, 0) \end{pmatrix}.$$

Moreover, the two limit processes have the same distribution and are independent.

In practice, the calibration of the test uses several independent sequences $(Z_n^\star)_{n\in\mathbb{Z}}$ to obtain a sample approximation of the limit distribution.

4 Testing for ideal calibration

4.1 Heuristic and strategy

The main idea driving our tests for ideal calibration relies on Corollary 3.6, stating that the forecast $(F_n)_{n\in\mathbb{Z}}$ is ideally calibrated if and only if

$$(H_0) : \text{for each } n \in \mathbb{Z},\ Z_n \sim \mathrm{Unif}([0,1]) \text{ and } Z_n \text{ is independent of } X_n.$$

We recall that this characterization was obtained thanks to the Markov assumption (A1), which implies that Z_n is independent from $\mathcal{F}_n$ if and only if it is independent from X_n. Note that even if the Markov assumption does not hold, our tests can still be used with a controlled level but will detect only if Z_n depends on X_n and will not be able to detect more subtle forms of non-calibration. However, it is always possible in theory to augment the dimension of the covariate space and test the dependency of Z_n from $(X_n, \dots, X_{n-m+1})$, that is consider memory of length $m \geq 1$. In practice, augmenting the dimension of the covariate space has a cost, both in terms of computational time and loss of power.

The proposed methodology for testing (H_0) relies on the simple observation that Z_n is uniformly distributed on $[0,1]$ and independent of X_n if and only if

$$\mathbb{E}\big[g(Z_n) \mid X_n \in A\big] = 0,$$

for all functions $g : [0,1] \to \mathbb{R}$ such that $\int_0^1 g(y)\, \mathrm{d}y = 0$ and measurable sets $A \subset [0,1]^p$. In practice, this integral can be estimated by

$$\frac{\frac{1}{n}\sum_{i=1}^n g(Z_i)\mathbb{1}_{\{X_i \in A\}}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}}} \tag{3.3}$$

which must be approximately zero under (H_0). In order to detect default of calibration, one needs to find a test function g and a region A where this mean significantly deviates from 0. We will consider several natural choices for the choice of g that are related to cumulative distribution functions, moments or histograms. The search for the region A is guided by the CART algorithm as explained in the next section.

4.2 General approach for tree based tests

The CART algorithm (Breiman et al., 1984) is a popular method from statistics and machine learning used for prediction, both in classification and regression. It produces simple predictors, called trees, in the sense that they can be represented graphically by a decision tree and predict finitely many different values on finitely many different subgroups of the population. More precisely, the procedure constructs a partition of the feature space $[0,1]^d$ into regions $A_1, \dots, A_K$, called leaves, and the tree function $T\colon [0,1]^d \to \mathbb{R}$ is constant on each leaf. In regression, the predicted value on A_k is simply the sample mean of the response variable in the sub-sample of individuals with features in A_k.

For our purpose, we will use the CART algorithm to predict the PIT Z_n, or rather a transformation of it $g(Z_n)$, as a function of the covariate X_n. Under (H_0), the covariate is uninformative to predict $g(Z_n)$ but, under the alternative, the CART algorithm should detect the dependency between $g(Z_n)$ and X_n. Interestingly, the CART algorithm is completely non parametric and assumes no model between $g(Z_n)$ and X_n. Furthermore, it is known to be quite robust in high dimension, i.e. its power is not too much hindered by the curse of dimensionality.

We next briefly describe the construction of the tree in our setting. Let $(X_i, g(Z_i))_{1 \leq i \leq n} \in [0,1]^d \times \mathbb{R}$ be the sample data. The construction of the partition $A_1, \dots, A_K$ of $[0,1]^d$ relies on recursive binary splitting, where a splitting rule is used repeatedly to form the partition. The first split forms the partition $[0,1]^d = A_1 \cup A_2$ in order to minimize the mean square error

$$\sum_{X_i \in A_1} \Big(g(Z_i) - \overline{g(A_1)}\Big)^2 + \sum_{X_i \in A_2} \Big(g(Z_i) - \overline{g(A_2)}\Big)^2, \tag{3.4}$$

where $\overline{g(A)}$ is the mean of the transformed PITs $g(Z_i)$ for $X_i \in A$. Not all possible partitions are used, but only the so-called admissible ones. An admissible partition is obtained by choosing a covariate index $j \in \{1, \dots, d\}$ and a threshold $u \in (0,1)$ and by letting

$$A_1 = \{x \in [0,1]^d \colon x_j \leq u\} \quad \text{and} \quad A_2 = \{x \in [0,1]^d \colon x_j > u\}.$$

Finding the admissible partitions that minimize the mean square error (3.4) can be done very efficiently, see Breiman et al. (1984) for more details. Using this splitting rule recursively on both A_1 and A_2 we then obtain 4 leaves, renamed $A_1, \dots, A_4$. Repeating the procedure d times, we obtain the tree with depth d with 2^d leaves. For our purpose, we will consider shallow trees with depth $d = 1, 2, 3$ only.

As a simple illustration, Figure 3.1 represents a data set in dimension $d = 2$ and the associated partition. The points $X_i \in [0,1]^2$ represent the covariates and the transformed PITs $g(Z_i)$ are represented by the color of the points. We can see a strong dependence since X_i and $g(Z_i)$ tend to be large at the same time. The CART algorithm detects this dependence and produces regions $A_1, \dots, A_5$ which are quite homogeneous.

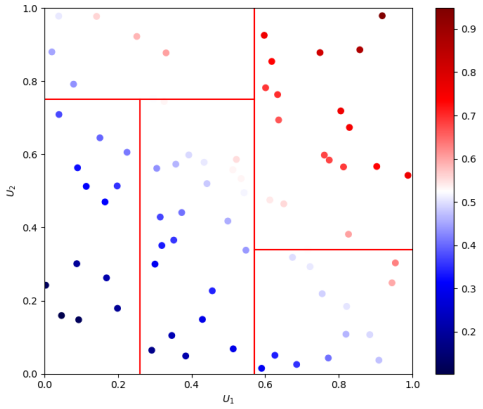


FIGURE 3.1: Created region by CART algorithm for the dummy case

König-Huygens Formula implies this rewriting of the splitting criterion (3.4)

$$\sum_{i=1}^{n} g(Z_i)^2 - \sum_{X_i \in A_1} \overline{g(A_1)}^2 - \sum_{X_i \in A_2}^{n} \overline{g(A_2)}^2. \tag{3.5}$$

So minimizing this variance quantity is equivalent to maximizing the following quantity

$$\sum_{X_i \in A_1} \overline{g(A_1)}^2 + \sum_{X_i \in A_2}^{n} \overline{g(A_2)}^2. \tag{3.6}$$

In its simplest version, our proposed test for ideal calibration relies on the choice of a test function $g : [0,1] \to \mathbb{R}$ such that $\int_0^1 g(u)\mathrm{d}u = 0$ and on the choice of a depth $d \geq 1$ (typically $d = 1, 2, 3$). The test statistic takes the form

$$\Delta = \sum_{k=1}^{2^d} \sum_{X_i \in A_k} \overline{g(A_k)}^2,$$

where the partition $(A_k)_{1 \leq k \leq 2^d}$ is the one associated to the tree with depth d. Under the null hypothesis (H_0), this statistic must be close to zero. To calibrate the test, we use the bootstrap (see Theorem 3.13) and we compute the mean square error $\Delta^\star$ obtained when fitting a regression tree with depth d to the sample $(X_i, g(Z_i^\star))_{1 \leq i \leq n}$, where $(Z_i^\star)_{1 \leq i \leq n}$ denotes an i.i.d. sample with uniform distribution on $[0,1]$ (independent of everything else). This statistic has a more explicit form coming from an optimization problem

$$\Delta = \max_{\substack{\text{admissible regions} \\ (B_1,\dots,B_{2^d})}} \sum_{k=1}^{2^d} \sum_{X_i \in B_k} \overline{g(B_k)}^2. \tag{3.7}$$

4.3 Specification of test functions

In practice, using only one test function seems limited and several of them should be used. We discuss different natural choices and also how to combine different test functions in our approach.

Test 1 : cumulative distribution function

The first test focuses on the CDF of the uniform distribution and considers the family of test functions

$$g_p(u) = \mathbb{1}_{\{u \le p\}} - p, \quad p \in (0,1).$$

This approach is closely related to the test based on Conditional Exceedance Probability (CEP) by Strähl and Ziegel (2017), see also section 5.1 for more details on this test. It is natural choice because, since the CDF characterize the PIT distribution, ideal calibration is equivalent to the fact that

$$\mathbb{E}[g_p(Z_n) \mid X_n \in A] = 0, \quad \text{for all } p \in (0,1) \text{ and } A \subset [0,1]^d \ .$$

In practice we consider finitely many values $p_1 < \ldots < p_K$, (for instance $p_1 = 0.1, \ldots, p_9 = 0.9$ in the simulation study) and fit K trees with depth d. The k-th tree uses the sample $(X_i, g_{p_k}(Z_i))_{1 \le i \le n}$ and the corresponding mean squared error is noted Δ_k. Similarly, for a bootstrap sample $(Z_i^\star)_{1 \le i \le n}$ we obtain the mean squared errors $\Delta_k^\star$, $1 \le k \le K$. We use here aggregation of the K errors, that is the test statistic is $\Delta = \sum_{i=1}^k \Delta_k$ that we compare with the bootstrap distribution of $\Delta = \sum_{i=1}^k \Delta_k^\star$.

Test 2 : moments

The second test focuses on the moments of the uniform distribution and more precisely on the first four moments. In order to check whether the PITs are uniformly distributed, we want to verify that mean, variance, skewness and kurtosis match those of the uniform distribution. In a slightly different context of calibration of ensemble forecast, this approach was used by Jolliffe and Primo (2008).
Here we consider the orthogonal polynomials

$$g_0(u) = 1, \quad g_1(u) = u - \frac{1}{2}, \quad g_2(u) = \sqrt{12}\left(u - \frac{1}{2}\right)^2, \ldots$$

that are obtained by the Gram-Schmidt orthonormalisation procedure applied to family of polynomials $(u^k)_{0 \le k \le 4}$ in the Hilbert space $L^2([0,1])$. As we will see in Proposition 3.18, orthogonality offers the benefit to yield asymptotically independent tests.
Thanks to this asymptotic independence, we do not aggregate the four mean square errors but rather perform four independent tests with test function $g_1, \ldots, g_4$ respectively. This strategy here seems interesting because it offers some qualitative interpretation : if a deviation to uniformity is detected, we are able to see whether it is rather in mean, variance, skewness or kurtosis.

Test 3 : histogram and χ^2-test

The third test is related to the histogram and χ^2-test and is slightly different from the previous ones as we will see that it can be related to classification trees rather than regression trees.

For $L \geq 2$, we consider the histogram of the PITs $(Z_n)_{n\geq 1}$ based on L bins of equal size $[0, 1/L)$, $\ldots$, $[(L-1)/L, 2]$. Here we introduce the centered *vectorial* test function

$$g(u) = \left(\mathbb{1}_{[\frac{(l-1)}{L}, \frac{l}{L})}(u) - \frac{1}{L} \right)_{1\leq l\leq L}.$$

On a region $A \subset [0,1]^2$, the squared (euclidean) norm

$$\|\overline{g(A)}\|^2 = \sum_{l=1}^{L} \left(\frac{\sum_{i=1}^n \mathbb{1}_{\{X_i\in A, Z_i\in[\frac{(l-1)}{L}, \frac{l}{L})\}}}{\sum_{i=1}^n \mathbb{1}_{\{X_i\in A\}}} - \frac{1}{L} \right)^2$$

corresponds, up to a multiplicative constant, to the χ^2-distance obtained when testing the uniformity of the PIT in A with the χ^2 test with L classes. This quantity is related to the Gini criterion

$$G(A) = \sum_{l=1}^{L} \left(\frac{\sum_{i=1}^n \mathbb{1}_{\{X_i\in A, \lceil LZ_i\rceil = l}}{\sum_{i=1}^n \mathbb{1}_{\{X_i\in A\}}} \right)^2$$

by the relation $\|\overline{g(A)}\|^2 = G(A) - 1/L$. Here the L classes used for the computation of the Gini criterion are the bins number $l = 1, \ldots L$ and the class associated with a PIT Z_i is $\lceil LZ_i\rceil$. Hence the l-th term of the sum defining $G(A)$ corresponds to the estimated probability of the l-th class in A.

Interestingly, the Gini criterion is one of the homogeneity criteria used in the construction of classification trees. Their construction is similar as the one of regression trees and is again based on recursive binary splitting. But now a split consists in the search of the admissible partition $A_1 \cup A_2$ that maximizes the Gini criterion $G(A_1) + G(A_2)$ (instead of minimisation of the mean squared error in regression).

The methodology for our third test is the following. We fit a classification tree with depth d on the sample $(X_i, \lceil LZ_i\rceil)$ using the Gini criterion. Denoting by $(A_k)_{1\leq k\leq 2^d}$ the resulting partition, the test statistic is

$$\Delta = \sum_{k=1}^{2^d} G(A_k).$$

Up to constants, this is the sum, over the different leaves, of χ^2-distances obtained when testing the uniformity of the PIT in each leaf. Calibration of the test is again based on a bootstrap replication $\Delta^\star$ of the test statistic using the bootstrap sample $(X_i, \lceil LZ_i^\star\rceil)$.

4.4 Statistics Δ and empirical processes

In this section, we see how the CART algorithm, and especially the splitting criterion (3.4) and the statistic Δ, can be rewritten in terms of the empirical processes introduced in Section 3. Let us recall the form of Δ in term of optimization problem

$$\Delta = \max_{\substack{\text{admissible regions} \\ (B_1,\ldots,B_{2^d})}} \sum_{k=1}^{2^d} \sum_{X_i\in B_k} \overline{g(B_k)}^2.$$

We give a proof only for the first split. For this split, the shape of the region B_1, B_2 is $\{x \in [0,1]^d \mid x \leq t\}, \{x \in [0,1]^d \mid x \not\leq t\}$ for an admissible $t \in [0,1]^d$, then

$$\Delta = \max_{\text{admissible } t\in[0,1]^d} \sum_{X_i\leq t} \overline{g(\{x \leq t\})}^2 + \sum_{X_i\not\leq t} \overline{g(\{x \not\leq t\})}^2 \tag{3.8}$$

These two terms can be rewritten in the following way to make the empirical processes appear more naturally,

$$\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}} \left(\frac{\sum_{k=1}^{n} g(Z_k) \mathbb{1}_{\{X_k \leq t\}}}{\sum_{k=1}^{n} \mathbb{1}_{\{X_k \leq t\}}} \right)^2 + \sum_{i=1}^{n} \mathbb{1}_{\{X_i \not\leq t\}} \left(\frac{\sum_{k=1}^{n} g(Z_k) \mathbb{1}_{\{X_k \not\leq t\}}}{\sum_{k=1}^{n} \mathbb{1}_{\{X_k \not\leq\}}} \right)^2,$$

or in the same way

$$\left(\sqrt{n} \frac{\frac{1}{n} \sum_{k=1}^{n} g(Z_k) \mathbb{1}_{\{X_k \leq t\}}}{\sqrt{\frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k \leq t\}}}} \right)^2 + \left(\sqrt{n} \frac{\frac{1}{n} \sum_{k=1}^{n} g(Z_k) \mathbb{1}_{\{X_k \not\leq t\}}}{\sqrt{\frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k \not\leq t\}}}} \right)^2. \tag{3.9}$$

These are the sums that can be interpreted as functionals of our processes. Indeed, let us introduce the following integral notation. Let g be a piecewise continuously differentiable function, it can be decomposed as

$$g(y) = g_0(y) + \sum_{i=1}^{k} w_i \mathbb{1}_{\{y \leq \alpha_i\}},$$

where $g_0 \in \mathcal{C}^1$ and $w_i, \alpha_i \in \mathbb{R}$. For $f \in \ell^\infty$, we define its integral with respecto to $\mathrm{d}g(y)$ as

$$\int_{\mathbb{R}} f(y) \, \mathrm{d}g(y) := \int_{\mathbb{R}} f(y) g_0'(y) \, \mathrm{d}y - \sum_{i=1}^{k} w_i f(\alpha_i).$$

This integral is only a notation. It is not directly related to the Stieltjes integral because our definition allows to consider functions f, g with common points of discontinuities. The minus sign in the notation is natural because the function $\mathbb{1}_{\{\cdot \leq \alpha\}}$ has a negative jump. The following Lemma is useful. It is a kind of integration by parts.

Lemma 3.16. *Assume $g : [0,1] \to \mathbb{R}$ piecewise continuously differentiable. Then for each $t \in [0,1]^d$,*

$$\frac{1}{n} \sum_{i=1}^{n} \big(g(Z_i) - \int_0^1 g(u) \, \mathrm{d}u \big) \mathbb{1}_{\{X_i \leq t\}} = - \int_0^1 \left(\mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) \right) \mathrm{d}g(y).$$

Similarly

$$\frac{1}{n} \sum_{i=1}^{n} \big(g(Z_i) - \int_0^1 g(u) \, \mathrm{d}u \big) \mathbb{1}_{\{X_i \not\leq t\}} = - \int_0^1 \left(\overline{\mathbb{F}}_1^{(n)}(y,t) + \overline{\mathbb{F}}_2^{(n)}(y,t) \right) \mathrm{d}g(y).$$

This lemma links quantity (3.9) with empirical processes because $\int_0^1 g(u) \, \mathrm{d}u = 0$, because the quantity (3.9) can be rewritten

$$\left(\sqrt{n} \frac{\int_0^1 \mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) \, \mathrm{d}g(y)}{\sqrt{F(t) + \mathbb{F}_3^{(n)}(1,t)}} \right)^2 + \left(\sqrt{n} \frac{\int_0^1 \overline{\mathbb{F}}_1^{(n)}(y,t) + \overline{\mathbb{F}}_2^{(n)}(y,t) \, \mathrm{d}g(y)}{\sqrt{\bar{F}(t) + \overline{\mathbb{F}}_3^{(n)}(1,t)}} \right)^2 \tag{3.10}$$

This invites us to define the functional

$$\Psi_g^F(F_1, F_2, F_3; t) = \frac{\int_0^1 \big(F_1(y,t) + F_2(y,t) \big) \, \mathrm{d}g(y)}{\sqrt{F(t) + F_3(1,t)}}. \tag{3.11}$$

By combining (3.8) and (3.10),

$$\Delta = n \max_{\text{admissible } t \in C_\varepsilon} \Psi_g^F \left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}; t \right)^2 + \Psi_g^{\overline{F}} \left(\overline{\mathbb{F}}_1^{(n)}, \overline{\mathbb{F}}_2^{(n)}, \overline{\mathbb{F}}_3^{(n)}; t \right)^2. \tag{3.12}$$

We had already noticed that the sum $\mathbb{F}_1^{(n)}+\mathbb{F}_2^{(n)}$ was observable because it did not depend on the CDF F. We can make the same remark about this quantity. Of course, the process $\mathbb{F}_3^{(n)}$ and the application Ψ_g^F depend on F separately but this dependence is simplified in the denominator. To use the delta method on Ψ_g^F, the condition of differentiability must be verified. This is only possible on a restriction of this set because the square root is not derivable in 0, so we need to avoid regions where F or $\overline{F}$ is too close to 0. For $\varepsilon > 0$, Let us consider the restriction on $\ell^\infty(C_\varepsilon)$ with $C_\varepsilon \subset [0,1]^d$ and

$$\forall t \in C_\varepsilon,\ \varepsilon \leq F(t) \leq 1-\varepsilon. \tag{3.13}$$

Proposition 3.17. *Assume $g : [0,1] \to \mathbb{R}$ piecewise continuously differentiable. Let $\varepsilon \in (0,1)$, the functionals Ψ_g^F and $\Psi_g^{\overline{F}}$ are differentiable at $(0,0,0)$ when the variable t is restricted to a subset $C_\varepsilon \subset [0,1]^d$ satisfying (3.13) and*

$$\partial_3\Psi_g^F(0,0,0) = \partial_3\Psi_g^{\overline{F}}(0,0,0) = 0.$$

It remains to be seen that the application max is continuous in ℓ^∞. Then Δ converges to a certain distribution by the contiunous mapping theorem.

The Corollary 3.15 states that $\Delta^\star$ approximates the asymptotic distribution of Δ. Sometimes, the ideal calibration of a forecaster $(F_n)_{n\in\mathbb{Z}}$ will be tested with several functions g. When these functions are pairwise orthogonal, see Equation (3.14), the aggregation of the multiple tests will be exact.

Proposition 3.18. *Assume $g, f : [0,1] \to \mathbb{R}$ **centred** piecewise continuously differentiable. Let $\varepsilon \in (0,1)$, under Assumptions (A1)-(A3) and (H_0), if*

$$\int_0^1 f(u)g(u)\ \mathrm{d}u = 0, \tag{3.14}$$

then $\sqrt{n}\begin{pmatrix}\Psi_g^F\left(\mathbb{F}_1^{(n)},\mathbb{F}_2^{(n)},\mathbb{F}_3^{(n)}\right)\\ \Psi_g^{\overline{F}}\left(\overline{\mathbb{F}}_1^{(n)},\overline{\mathbb{F}}_2^{(n)},\overline{\mathbb{F}}_3^{(n)}\right)\end{pmatrix}$ *and* $\sqrt{n}\begin{pmatrix}\Psi_f^F\left(\mathbb{F}_1^{(n)},\mathbb{F}_2^{(n)},\mathbb{F}_3^{(n)}\right)\\ \Psi_f^{\overline{F}}\left(\overline{\mathbb{F}}_1^{(n)},\overline{\mathbb{F}}_2^{(n)},\overline{\mathbb{F}}_3^{(n)}\right)\end{pmatrix}$ *are asymptotically independent in $\ell^\infty(C_\varepsilon)$.*

5 Numerical illustrations

5.1 Presentation of algorithms

Regression and classification tree : We will use our three previously introduced tests. We limit the depth of our trees to two splits. The other arbitrarily fixed parameters are the partition of the interval [0,1] of Test 1, the number of polynomials of Test 2 and the number of classes of Test 3. For the first test, we choose as partition $\{0, 1/10, 2/10, \ldots, 1\}$. For the second test, we will take the polynomials of degree $\{1,2,3,4\}$. And finally, we will take $L = 7$ classes. The number of bootstrap replications $(Z_n^\star)_{n\in\mathbb{Z}}$ for adjusting the tests is $B = 600$. The reject rates is classically chosen at $\alpha = 0.05$.

Conditional Excedence Probability : The Strähl and Ziegel's test (Strähl and Ziegel, 2017, Section 6.1) is to write the characterization of the ideal calibration, i.e. the sequence of PIT $(Z_n)_{n\in\mathbb{N}}$ are uniformly distributed and independent of the information $(X_n)_{n\in\mathbb{N}}$, in terms of logistic regression. In our framework, this is written as follows for $n \geq 0$,

$$\mathrm{logit}(\mathbb{P}(Z_n \leq z \mid X_n)) = \beta_{0,z} + \sum_{i=1}^{d} \beta_{i,z} X_n^{(i)}, \tag{3.15}$$

with $z \in (0,1)$, $\text{logit}(z) = \log(z/(1-z))$ and $X_n^{(i)}$ is the i^{th} coordinate of X_n. Under (H_0), one has

$$\beta_{0,z} = \text{logit}(z) \text{ and } \beta_{i,z} = 0 \text{ for each } z \in (0,1) \text{ and } i \in \{1,\dots,d\}. \tag{3.16}$$

Then they run the regression for different values of z to test if the Equation (3.16) is true. To combine the different tests, they use the multiple test adjustment method of Cox and Lee (2008). For the exact details of the implementation, we refer to Strähl and Ziegel (2017, Section 6.1)

Remark 3.19. Other tests related to this problem have been introduced in the literature. For example, still in Strähl and Ziegel (2017) or Held et al. (2010), the authors present another tests that can be used in this context. They add a parametric hypothesis on the model. That's why, we have decided not to include them in the benchmark.

5.2 Autoregressive model

The first model used for the simulations will be the classical AR(1). It is presented as follows : for $n \geq 0$ and $\rho \in (-1,1), \alpha > 0$,

$$Y_{n+1} = \rho Y_n + \varepsilon_n, \quad Y_0 \sim \mathcal{N}(0, \alpha^2),$$

where $(\varepsilon_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of $\mathcal{N}(0,\sigma^2)$ with $\sigma > 0$. The number α is chosen such that the sequence $(Y_n)_{n\in\mathbb{N}}$ is a stationary process, i.e. $\alpha^2 = \sigma^2/(1-\rho^2)$. The real number ρ represents the time dependence of the phenomenon. If $\rho = 0$, then the sequence $(Y_n)_{n\in\mathbb{N}}$ is completely independent. In the simple case, we assume that $X_n = Y_n$ for $n \geq 0$ and the known information $\mathcal{F}_n$ is $\sigma(X_0,\dots,X_n)$. We recall that this means that we only know what we observe. At the time n, the ideal forecast F_n^* is $\mathcal{N}(\rho X_n, \sigma^2)$. We propose several alternatives. These alternatives have already been used in Gneiting and Ranjan (2013); Strähl and Ziegel (2017), except the last one

- **Climatological forecaster :** $F_n^{(2)} = \mathcal{N}(0,\alpha^2)$;
- **Unfocused forecaster :** $F_n^{(3)} = \frac{1}{2}\mathcal{N}(\rho X_n, \sigma^2) + \frac{1}{2}\mathcal{N}(\rho X_n + \tau_n, \sigma^2)$ with $\tau_n = \pm 1$ with probability $1/2$ independently of $(X_n, \varepsilon_n)_{n\in\mathbb{N}}$;
- **Sign-reversed forecaster :** $F_n^{(4)} = \mathcal{N}(-\rho X_n, \sigma^2)$;
- **Error observation forecaster :** $F_n^{(5)} = \mathcal{N}(\rho(X_n + \delta_n), \sigma^2)$, where $\delta_n \sim \mathcal{N}(0,1)$ independently of $(X_n, \varepsilon_n)_{n\in\mathbb{N}}$.

For the two first alternatives, the PITs of these forecasters are uniformly distributed (see Chapter 1 of this manuscript or Gneiting and Ranjan (2013)). We hope that our tests will focus on the absence of independence with information and not on the distribution of PITs. The number N is the number of realizations that we used in the test. Intuitively, if this number is large, then the tests perform better because we use more data.

The following results are based on 1000 replications. The level of the 4 tests is equal to 0.95. This is quite natural as the bootstraps used to adjust the rejection are well justified theoretically. The next tables estimate the test rejection rate for the various alternatives. We recall that the closer this rate is to 1, the better the test performs. All test parameters (depth, number of classes, level, ...) are defined in Section 5.1.

Alternative 1 : Climatological Forecaster

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	0.08	0.07	0.06	0.07
$(0.3, 50)$	**0.35**	0.23	0.17	0.27
$(0.5, 50)$	**0.81**	0.63	0.48	0.75
$(0.8, 50)$	**0.99**	**0.97**	**0.97**	**0.99**
$(0.1, 100)$	0.10	0.07	0.06	0.07
$(0.3, 100)$	**0.62**	0.46	0.31	0.54
$(0.5, 100)$	**0.97**	0.91	0.78	**0.98**
$(0.8, 100)$	1	1	1	1

Alternative 2 : Unfocused Forecaster

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	0.04	0.06	0.04	0.03
$(0.3, 50)$	0.05	0.05	0.05	0.03
$(0.5, 50)$	0.05	0.06	0.05	0.05
$(0.8, 50)$	0.06	0.05	0.05	0.05
$(0.1, 100)$	0.04	0.06	0.04	0.03
$(0.3, 100)$	0.05	0.05	0.05	0.03
$(0.5, 100)$	0.05	0.06	0.05	0.05
$(0.8, 100)$	0.06	0.06	0.05	0.05

Alternative 3 : Sign-reversed Forecaster

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	**0.17**	0.13	0.10	**0.14**
$(0.3, 50)$	**0.92**	0.81	0.63	**0.87**
$(0.5, 50)$	1	1	1	1
$(0.8, 50)$	1	1	1	1
$(0.1, 100)$	**0.32**	0.19	0.14	0.25
$(0.3, 100)$	**0.99**	**0.98**	0.90	**0.99**
$(0.5, 100)$	1	1	1	1
$(0.8, 100)$	1	1	1	1

Alternative 4 : Error observation Forecaster

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	0.05	0.05	0.05	0.05
$(0.3, 50)$	0.06	0.08	0.05	0.05
$(0.5, 50)$	0.08	**0.16**	0.09	0.10
$(0.8, 50)$	0.16	**0.46**	0.17	0.23
$(0.1, 100)$	0.05	0.05	0.05	0.05
$(0.3, 100)$	0.06	0.08	0.05	0.05
$(0.5, 100)$	0.08	**0.19**	0.09	0.10
$(0.8, 100)$	0.20	**0.65**	0.31	0.36

Several comments can be formulated on the basis of the table above :

- Alternative 2 is not detected by all tests. This fact has already been noted in Strähl and Ziegel (2017). This is still very surprising as the PITs are only uniformly distributed.
- Tests based on tree regression do the best or are comparable to CEP test. Unfortunately, there is not one of these tests that comes out better every time.
- Not surprisingly, the power of the tests increases with the sample size.
- Due to the shape of the alternatives, the greater the time dependency, the more powerful the tests are.

5.3 Non linear model

To finish this short simulation section, we create a toy framework to challenge the CEP test but not the Tree Regression based tests. Let $(\mu_n, \varepsilon_n, \delta_n)_{n\in\mathbb{N}}$ be a random i.i.d. sequence where the marginals are also independent and $\varepsilon_n \in \{-1, 1\}$ uniformly, $\mu_n \sim \mathcal{N}(0, 1)$ and $\delta_n \sim \mathcal{N}(0, 1)$. We define $Y_{n+1} = \varepsilon_n \mu_n + \delta_n$ and $\mathcal{F}_n = o(\mu_1, \ldots, \mu_n)$.

We consider the Climatological Forecaster $F = \mathcal{N}(0, 2)$. We recall that this forecaster is not ideally calibrated for the information $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Thus the sequence of PIT $(Z_F^{Y_{n+1}})_{n\in\mathbb{N}}$ is not simultaneously independent of the information and uniformly distributed. More precisely, the PITs will be uniformly distributed because F is the distribution of $(Y_n)_{n\in\mathbb{N}}$, and even the PITs will be i.i.d. The only difference with samples bootstraps $(Z_n^\star)_{n\in\mathbb{N}}$ will be their independence with information $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

We first compare our tests for different values of N. Naturally, we notice that the power increases with the size of the sample. You can quickly notice that you should avoid making too many splits in the Tree Regression. By increasing the depth d, you can decrease the power.

N=50	Test 1	Test 2	Test 3
$d=1$	0.14	**0.29**	0.14
$d=2$	0.08	**0.26**	0.11
$d=3$	0.08	0.20	0.11

N=100	Test 1	Test 2	Test 3
$d=1$	0.25	0.56	0.24
$d=2$	0.20	**0.64**	0.23
$d=3$	0.12	0.53	0.19

N=200	Test 1	Test 2	Test 3
$d=1$	0.44	**0.94**	0.63
$d=2$	0.42	**0.96**	0.65
$d=3$	0.29	**0.94**	0.5

We now compare ourselves to the CEP test in the case N=200 and for a depth $d = 2$ for our trees. We notice that the test introduced in this article performs better.

Test 1 $d=2$	Test 2 $d=2$	Test 3 $d=2$	Test CEP
0.44	**0.96**	0.63	0.13

6 Discussion

6.1 Testing cross-calibration

In the simulations, we compared ourselves to the test in Strähl and Ziegel (2017). The framework of this article is not exactly the same. So we had to adapt their test for our framework. The reverse is quite possible, meaning to take our tests to test the cross-calibration. First, let us recall the definitions of cross-calibration. Let us introduce the $\sigma-$algebra generated by the sequence $(Y_n)_{n\in\mathbb{N}}$, denoted by $(\mathcal{G}_n)_{n\in\mathbb{N}}$. Let $(F_{1,n})_{n\in\mathbb{N}}, \dots, (F_{k,n})_{n\in\mathbb{N}}$ be k dynamical forecasters. The dynamical forecaster $(F_{1,n})_{n\in\mathbb{N}}$ is said *cross-calibrated* with respect to $(F_{1,n})_{n\in\mathbb{N}}, \dots, (F_{k,n})_{n\in\mathbb{N}}$ if

$$\forall n \in \mathbb{N},\ Z_{F_{1,n}}^{Y_{n+1}} \sim \text{Unif}([0,1]) \text{ and } Z_{F_{1,n}}^{Y_{n+1}} \perp\!\!\!\perp (F_{1,n}, \dots, F_{k,n}, \mathcal{G}_n).$$

Proposition 3.4 states an equivalent definition

$$\forall n \in \mathbb{N},\ F_{1,n} = \mathcal{L}(Y_{n+1} \mid F_{1,n}, \dots, F_{k,n}, \mathcal{G}_n).$$

This means that the first forecaster uses the information he has perfectly and that the information of the other forecasters does not help him to predict better. Our algorithms based on regression trees can be adapted with very little modification to the study of cross-calibration. With the idea of the CEP test, the test of independence with a forecaster will be based on the quantiles of the forecast $F_{i,n}$. As it is impossible to test all quantiles, we check a weaker condition by considering a partition $(p_1, \dots, p_K)$ of $[0,1]$. The sequence of variables $(X_n)_{n\in\mathbb{N}}$ is defined as

$$\forall n \in \mathbb{N},\ X_n = (Y_n, F_{i,n}^{-1}(p_j) \text{ for } 1 \le i \le k,\ 1 \le j \le K).$$

The adaptation of our tests to the cross-calibration is to test independence with these variables $(X_n)_{n\in\mathbb{Z}}$.

6.2 Weaker calibration

In practice, with a meteorological point of view, it is impossible to predict exactly with the true conditional distribution. Using these tests with a real data set is unlikely to produce any interesting results. Indeed, they test a small null hypothesis and will therefore reject almost all predictions from a concrete case. We have nevertheless proposed several statistics to test whether a forecast is ideal or not. We think that the idea behind this family of tests could be useful. Moreover, the proof of the asymptotic behaviour of these tests is quite interesting since even under no independence assumption, we do not need to compute the co-variance matrix. So an advantage of these tests is to not compute the temporal correlation which could be a difficult task.

A recent kind of calibration, Gneiting and Resin (2021), is the $\mathcal{T}$-calibration where $\mathcal{T}$ is a functional on the probability measure, e.g. the median, the mean...

Definition 3.20. *A random forecast F is $\mathcal{T}$-calibrated if*

$$\mathcal{T}(\mathcal{L}(Y \mid \mathcal{T}(F))) = \mathcal{T}(F) \text{ a.s.}$$

A functional $\mathcal{T}$ is said to be identifiable if there exists $V\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ such that $V(\cdot, y)$ is increasing, left-continuous for all $y \in \mathbb{R}$ and satisfying for each $F \in \mathcal{P}$,

$$\begin{cases} \forall x < \mathcal{T}(F),\ \int_{\mathbb{R}} V(x,y)\ F(\mathrm{d}y) < 0 \\ \forall x > \mathcal{T}(F),\ \int_{\mathbb{R}} V(x,y)\ F(\mathrm{d}y) > 0 \\ \int_{\mathbb{R}} V(\mathcal{T}(F),y)\ F(\mathrm{d}y) = 0 \end{cases}.$$

In a similar way to their proof of Theorem 2.10, the $\mathcal{T}$-calibration can be written in terms of a conditional expectation.

Proposition 3.21. *Let $\mathcal{T}$ be an identifiable functional, and let F be a random forecast. The forecast F is $\mathcal{T}$-calibrated if and only if*

$$\mathbb{E}[V(\mathcal{T}(F), Y) \mid \mathcal{T}(F)] = 0.$$

Proof.

$$\mathbb{E}[V(\mathcal{T}(F), Y) \mid \mathcal{T}(F)] = \int_{\mathbb{R}} V(\mathcal{T}(F), y)\ F_{\mathcal{T}}(\mathrm{d}y), \text{ where } F_{\mathcal{T}} = \mathcal{L}(Y \mid \mathcal{T}(F)).$$

Then it is equal to 0 if and only if $\mathcal{T}(F) = \mathcal{T}(F_{\mathcal{T}})$. □

This writing in terms of conditional expectation makes it possible to use the same method based on empirical process to analyse $\mathcal{T}-$calibration. More precisely, similarly as in Lemma 3.8, one can show that the following process is centered if and only if the dynamic forecast $(F_n)_{n\in\mathbb{Z}}$ is $\mathcal{T}$-calibrated

$$x \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{\mathcal{T}(F_i)\le x\}} V(\mathcal{T}(F_i), Y_i).$$

However, it is not clear that the results and techniques presented in this article are generalizable to this process because the covariance of the limit process is not easy to use.

7 Proofs

7.1 Proofs of Section 2

Proof of Lemma 3.2. The reciprocal implication is the Theorem 1 of Brockwell (2007). A rewriting of this result is for $z \in [0,1]$ and $Y \sim \mu$ with $\mu \in \mathcal{M}_1(\mathbb{R})$ and F its CDF,

$$\mathbb{P}\left(Z_F^Y \le z\right) = \mathbb{E}\left[\mathbb{1}_{\{Z_F^Y \le z\}}\right] = \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{v\mu(]-\infty,y[)+(1-v)\mu(]-\infty,y])\le z\}} \mathrm{d}\mu(y)\mathrm{d}v = z \tag{3.17}$$

Now, assume that the PIT is uniformly distributed. For all $x \in \mathbb{R}$, these inclusions hold

$$\begin{cases} \{Y \le x\} \subset \{Z_F^Y \le F(x)\} \\ \{Y > x\} \subset \{Z_F^Y \ge F(x)\} \end{cases},$$

because $Z_F^Y = VF(Y^-) + (1-V)F(Y)$ with F non decreasing. As $\mathbb{P}$ is also non decreasing,

$$\begin{cases} G(x) = \mathbb{P}\left(\{Y \le x\}\right) \le \mathbb{P}\left(\{Z_F^Y \le F(x)\}\right) = F(x) \\ 1 - G(x) = \mathbb{P}(\{Y > x\}) \le \mathbb{P}(\{Z_F^Y \ge F(x)\}) = 1 - F(x) \end{cases},$$

where G denotes the CDF of Y and $Z_F^Y \sim \text{Unif}([0,1])$. Then we get $F(x) = G(x)$.
Let $\mu, \nu \in \mathcal{M}_1(\mathbb{R})$, a rewriting of this result is

$$\left(\forall z \in [0,1],\ \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{v\mu(]-\infty,y[)+(1-v)\mu(]-\infty,y])\leq z\}} \mathrm{d}\nu(y)\mathrm{d}v = z\right) \Rightarrow \mu = \nu. \tag{3.18}$$

The set $[0,1]$ could be replaced by any dense subset. □

The following proposition justifies the measurability of PIT. This technical proposition is not necessary at first reading.

Proposition 3.22. *Let μ be a kernel, F be its associated random forecast and Y be a random vatiable. The PIT Z_F^Y is a random variable, i.e. it is a measurable map.*

Proof. Let us define the applications for $w, y \in \Omega \times Y$,

$$g_1(\omega, y) = \mu(\omega,]-\infty, y[) = F(\omega, y^-) \text{ and } g_2(\omega, y) = \mu(\omega,]-\infty, y]) = F(\omega, y).$$

For $y \in \mathbb{R}$, $g_1(\cdot, y)$ and $g_2(\cdot, y)$ are measurable by the definition of a kernel. Now for $\omega \in \Omega$ fixed, $g_1(\omega, \cdot)$ is left continuous and $g_2(\omega, \cdot)$ is right continuous. The aim of the proof is to show measurability for both variables simultaneously. For $n \in \mathbb{N}^*$, the maps $g_1^{(n)} \colon (\omega, y) \mapsto g_1\left(\omega, \frac{\lfloor ny \rfloor}{n}\right)$ and $g_2^{(n)} \colon (\omega, y) \mapsto g_2\left(\omega, \frac{\lceil ny \rceil}{n}\right)$ are measurable because they can be rewritten as follows, for $\omega, y \in \Omega \times \mathbb{R}$,

$$g_1^{(n)}(\omega, y) = \sum_{k \in \mathbb{Z}} \mathbb{1}_{\{\lfloor ny \rfloor = k\}} g_1\left(\omega, \frac{k}{n}\right).$$

As g_1 is left continuous, the sequence $(g_1^{(n)})_{n\in\mathbb{N}}$ converges point-wise to g_1, so g_1 is measurable. In the same way, g_2 is measurable. Then Z_F^Y is measurable by composition and sum. □

Proof of Proposition 3.4. For the direct implication, let $A \in \mathcal{F}$ and $z \in [0,1]$, as $F = \mathcal{L}(Y \mid \mathcal{F})$,

$$\mathbb{P}(A \cap \{Z_F^Y \leq z\}) = \int_A \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{vF(\omega,]-\infty,y[)+(1-v)F(\omega,]-\infty,y])\leq z\}} F(\omega, \mathrm{d}y)\mathrm{d}v\mathbb{P}(\mathrm{d}\omega),$$

by the Proposition A.10 and V is independent of $\mathcal{F}$. Then with the Equation (3.17) for $\omega \in A$ fixed,

$$\int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{uF(\omega,]-\infty,y[)+(1-u)F(\omega,]-\infty,y])\leq z\}} F(\omega, \mathrm{d}y)\mathrm{d}u = z$$

Hence $\mathbb{P}(A \cap \{Z_F^Y \leq z\}) = \mathbb{P}(A)z$, so $Z_F^Y \sim \text{Unif}([0,1])$ and is independent of $\mathcal{F}$.

For the reciprocal implication, the $\mathcal{F}-$measurability of F allows us to apply Proposition A.10. Let $z \in [0,1] \cap \mathbb{Q}$ and $A \in \mathcal{F}$, the independence and uniform distribution on $[0,1]$ imply,

$$\mathbb{E}\left[\mathbb{1}_A \mathbb{1}_{\{Z_F^Y \leq z\}}\right] - \mathbb{P}(A)z = \int_A \underbrace{\int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{vF(\omega,]-\infty,y[)+(1-v)F(\omega,]-\infty,y])\leq z\}} \mu(\omega, \mathrm{d}y)\mathrm{d}v - z}_{\mathcal{F}\text{-measurable}} \mathbb{P}(\mathrm{d}\omega),$$

where μ is $\mathcal{L}(Y \mid \mathcal{F})$. The $\mathcal{F}$-measurability of the integrand is a consequence of the Fubini Theorem. As this integral is null for $A \in \mathcal{F}$,

$$a.s.,\ \forall z \in [0,1] \cap \mathbb{Q}, \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{uF(\omega,]-\infty,y[)+(1-u)F(\omega,]-\infty,y])\leq z\}} \mu(\omega, \mathrm{d}y)\mathrm{d}u = z.$$

Then by the Equation (3.18),

$$a.s.,\ F = \mu = \mathcal{L}(Y \mid \mathcal{F}).$$

□

Proof of Corollary 3.6. It is a direct consequence of the Proposition 3.4. To get the last point, it suffices to remark that if $(\mathcal{F}_n)_{n\in\mathbb{N}}$ contains the filtration endowed by $(Y_n)_{n\in\mathbb{N}}$ then Z_n is $\sigma(\mathcal{F}_{n+T}, V_n)$ measurable, and Z_{n+T} is independent of $\mathcal{F}_{n+T}$ and V_n. □

7.2 Proofs of Section 3

Proof of Proposition 3.7. Under (H_0) with Assumptions (A1)-(A3), Theorem 10.2 of Dedecker et al. (2007) is applicable. In this book, the convergence is in the Skorokhod space $\mathcal{D}([0,1]^d)$. But the authors prove the tighness in $\ell^\infty([0,1]^d)$. Then by their Proposition 4.2, the convergence is also in ℓ^∞. □

Proof of Lemma 3.8. 1. This equivalence is the definition of $Z \sim \mathrm{Unif}([0,1])$;

2. Recall that $(Z_n, X_n)_{n\in\mathbb{N}}$ is stationary

$$\begin{aligned}\mathbb{F}_2^{(n)} \text{ is centred} &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathrm{Cov}\left(\mathbb{1}_{\{Z_1\le y\}}, \mathbb{1}_{\{X_1\le t\}}\right) = 0\\ &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathbb{P}(Z_1\le y, X_1\le t) = \mathbb{P}(Z_1\le y)\mathbb{P}(X_1\le t)\\ &\Leftrightarrow (Z_1,X_1) \text{ are independent.}\end{aligned}$$

3. For $y,t \in [0,1]\times[0,1]^d$,

$$\mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}_{\{Z_i\le y\}} - y\right)\mathbb{1}_{\{X_i\le t\}}.$$

The stationarity implies

$$\begin{aligned}\mathbb{F}_1^{(n)} + \mathbb{F}_2^{(n)} \text{ is centred} &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathbb{E}\left[\left(\mathbb{1}_{\{Z_1\le y\}} - y\right)\mathbb{1}_{\{X_1\le t\}}\right] = 0\\ &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathbb{P}(Z_1\le y, X_1\le t) = yF(t)\\ &\Leftrightarrow (Z_1,X_1) \text{ are independent and } Z_1\sim\mathrm{Unif}([0,1]).\end{aligned}$$

□

Proof of Theorem 3.10. As the evaluation is continuous, the mapping theorem yields

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\mathbb{F}_1^{(n)}\\ \mathbb{F}_2^{(n)}\\ \mathbb{F}_3^{(n)}\end{pmatrix} &= \begin{pmatrix}\Phi_1\\ \Phi_2\\ \Phi_3\end{pmatrix}\left(\sqrt{n}\mathbb{G}^{(n)}\right)\\ &\rightsquigarrow \begin{pmatrix}\Phi_1(\mathbb{G})\\ \Phi_2(\mathbb{G})\\ \Phi_3(\mathbb{G})\end{pmatrix} = \begin{pmatrix}\mathbb{G}_1\\ \mathbb{G}_2\\ \mathbb{G}_3\end{pmatrix},\end{aligned}$$

where

$$\begin{cases}\Phi_1(G)(y,t) = G(y,\mathbf{1})F(t)\\ \Phi_2(G)(y,t) = G(y,t) - G(y,\mathbf{1})F(t) - yG(1,t)\\ \Phi_3(G)(y,t) = G(1,t)\end{cases}.$$

The covariance functions of the two first processes are, for $a,b\in[0,1]$ and $s,t\in[0,1]^d$,

$$\begin{cases}\Gamma_1((a,t),(b,s)) = \sum_{i\in\mathbb{Z}}\mathrm{Cov}\left(F(t)\mathbb{1}_{\{Z_0\le a\}}, F(s)\mathbb{1}_{\{Z_i\le b\}}\right)\\ \Gamma_2((a,t),(b,s)) = \sum_{i\in\mathbb{Z}}\mathrm{Cov}\left(H_0(a,t), H_i(b,s)\right),\\ \text{where } H_i(a,t) = \left(\mathbb{1}_{\{Z_i\le a\}} - a\right)\left(\mathbb{1}_{\{X_i\le t\}} - F(t)\right)\end{cases}.$$

The first function is directly simplified because the sequence of PITs $(Z_i)_{i\in\mathbb{Z}}$ is $(T-1)$ dependent. However, this is not true for the sequence $(X_i,Z_i)_{i\in\mathbb{Z}}$. Nevertheless, the simplification is possible. Let $|i|\ge T$, then Z_i is independent of (X_0,Z_0,X_i),

$$\mathrm{Cov}\left(H_0(a,t), H_i(b,s)\right) = \mathbb{E}[\mathbb{1}_{\{Z_i\le b\}} - b]\mathbb{E}[\left(\mathbb{1}_{\{X_i\le s\}} - F(s)\right)H_0(a,t)] = 0.$$

In the case where $T = 1$, the sequence $(Z_i)_{i\in\mathbb{Z}}$ is independent. The independence of $\mathbb{G}_1$ and $\mathbb{G}_2$ is a consequence of the decorrelation between $\mathbb{F}_1^{(n)}$ and $\mathbb{F}_2^{(n)}$. For $i \neq j$, Z_i is independent of (X_i, Z_j) then

$$\mathrm{Cov}(H_i(a,t), \mathbb{1}_{\{Z_j\leq b\}} - b) = \mathbb{E}[\mathbb{1}_{\{Z_i\leq a\}} - a]\mathbb{E}[\left(\mathbb{1}_{\{X_i\leq t\}} - F(t)\right)\left(\mathbb{1}_{\{Z_j\leq b\}} - b\right)] = 0.$$

For $i = j$, the PIT Z_i is independent of X_i then

$$\mathrm{Cov}(H_i(a,t), \mathbb{1}_{\{Z_i\leq b\}} - b) = \mathbb{E}[\mathbb{1}_{\{X_i\leq t\}} - t]\mathbb{E}[\left(\mathbb{1}_{\{Z_i\leq a\}} - a\right)\left(\mathbb{1}_{\{Z_i\leq b\}} - b\right)] = 0.$$

□

Proof of Theorem 3.12. It is a direct consequence of the δ-method in van der Vaart and Wellner (1996, Theorem 3.9.4) and the fact that as $\partial_3\Psi(0,0,0) = 0$,

$$\mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3) = \mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, 0)$$

. □

Proof of Theorem 3.13. The random sequence $(Z_i, Z_i^\star, X_i)_{i\in\mathbb{Z}}$ still checks Assumptions (A1)-(A3) then this convergence is a consequence of Theorem 10.2 in Dedecker et al. (2007),

$$\sqrt{n}\begin{pmatrix}\mathbb{G}^{(n)}\\ \mathbb{G}^{(n)\star}\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathbb{G}\\ \mathbb{G}^\star\end{pmatrix}.$$

The application of the Mapping Theorem is done in the same way as in the proof of Theorem 3.10. The essential part of this theorem is the equality between the classical limit and the bootstrapped limit and the asymptotic independence. This part will be shown in several steps. As many of these steps are identical, we will not show them all. Moreover, since they are Gaussian vectors, the pairwise independence of the components implies the independence of the vectors.

1. $\mathbb{G}_1$ and $\mathbb{G}_1^\star$ have the same distribution ;
2. $\mathbb{G}_2$ and $\mathbb{G}_2^\star$ have the same distribution ;
3. $\mathbb{G}_1$ is independent of $\mathbb{G}_1^\star$;
4. $\mathbb{G}_2$ is independent of $\mathbb{G}_2^\star$;
5. $\mathbb{G}_2$ is independent of $\mathbb{G}_1^\star$;
6. $\mathbb{G}_1$ is independent of $\mathbb{G}_2^\star$.

Firstly, for $T = 1$, the PITs $(Z_i)_{i\in\mathbb{Z}}$ and the bootstrapped PITs $(Z_i^\star)_{i\in\mathbb{Z}}$ are independent and have the same distribution. Then it is the same for the processes $(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)})$ and their limits. So the Point 1 and 3 are shown. Point 2 comes from the fact that for the lead time $T = 1$, the covariance of $\mathbb{G}_2$ simplifies to

$$\Gamma_2((a,t),(b,s)) = \mathrm{Cov}\left(\left(\mathbb{1}_{\{Z_0\leq a\}} - a\right)\left(\mathbb{1}_{\{X_0\leq t\}} - F(t)\right), \left(\mathbb{1}_{\{Z_0\leq b\}} - b\right)\left(\mathbb{1}_{\{X_0\leq s\}} - F(s)\right)\right),$$

which is the same that $\mathbb{G}_2^\star$. The last three points are shown by studying the correlation of the processes. As the proofs are identical, we only detail Point 4. For $i, j \in \mathbb{Z}$,

$$\begin{aligned}
&\mathrm{Cov}\left(\left(\mathbb{1}_{\{Z_i\leq a\}} - a\right)\left(\mathbb{1}_{\{X_i\leq t\}} - F(t)\right), \left(\mathbb{1}_{\{Z_j^\star\leq b\}} - b\right)\left(\mathbb{1}_{\{X_j\leq s\}} - F(s)\right)\right)\\
&\quad= \mathbb{E}\left[\left(\left(\mathbb{1}_{\{Z_i\leq a\}} - a\right)\left(\mathbb{1}_{\{X_i\leq t\}} - F(t)\right)\left(\mathbb{1}_{\{Z_j^\star\leq b\}} - b\right)\left(\mathbb{1}_{\{X_j\leq s\}} - F(s)\right)\right)\right]\\
&\quad= \mathbb{E}[\mathbb{1}_{\{Z_j^\star\leq b\}} - b]\mathbb{E}\left[\left(\left(\mathbb{1}_{\{Z_i\leq a\}} - a\right)\left(\mathbb{1}_{\{X_i\leq t\}} - F(t)\right)\left(\mathbb{1}_{\{X_j\leq s\}} - F(s)\right)\right)\right]\\
&\quad= 0
\end{aligned}$$

□

Proof of Lemma 3.16. Let us recall that for $y, t \in [0,1] \times [0,1]^d$,

$$\mathbb{F}^{(n)}(y,t) := \mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{Z_i \le y\}} - y\right) \mathbb{1}_{\{X_i \le t\}}.$$

Let us develop the integral

$$\begin{aligned}
\int_0^1 \mathbb{F}^{(n)}(y,t)\ \mathrm{d}g(y) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}} \left(\int_0^1 g_0'(y)(\mathbb{1}_{\{Z_i \le y\}} - y)\ \mathrm{d}y - \sum_{j=1}^{k} w_j(\mathbb{1}_{\{Z_i \le \alpha_j\}} - \alpha_j) \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}} \left(g_0(1) - g_0(Z_i) - [g_0(y)y]_{y=0}^1 + \int_0^1 g_0(y)\ \mathrm{d}y - \sum_{j=1}^{k} w_j(\mathbb{1}_{\{Z_i \le \alpha_j\}} - \alpha_j) \right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}} \left(\int_0^1 g(u)\ \mathrm{d}u - g(Z_i) \right).
\end{aligned}$$

□

Proof of Proposition 3.17. Let $h_1, h_2, h_3 \in \ell^\infty$, the function Ψ_g^F is null at $(0,0,0)$,

$$\begin{aligned}
\Psi_g^F(h_1^{(n)}, h_2^{(n)}, h_3^{(n)}; t) &= -\frac{\int_0^1 h_1(y,t) + h_2(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t) + h_3(1,t)}} \\
&= -\frac{\int_0^1 h_1(y,t) + h_2(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t)}} \times \left(1 - \frac{h_3(1,t)}{2F(t)} + o(\|h_3\|/F)\right).
\end{aligned}$$

The Taylor expansion is uniform in t because $\varepsilon \le F(t) \le 1 - \varepsilon$. We therefore define the linear application

$$H(h_1, h_2, h_3; t) = -\frac{\int_0^1 h_1(y,t) + h_2(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t)}} \tag{3.19}$$

The continuity of this linear application is a direct consequence of the following inequality

$$\left| \int_0^1 h_1(y,t) + h_2(y,t)\ \mathrm{d}g(y) \right| \le \|h_1 + h_2\|_\infty \left(\|g_0'\|_\infty + \sum_{j=1}^{k} |w_j| \right). \tag{3.20}$$

This inequality also proves that there exists $C > 0$ such that

$$\left\| \Psi_g^F(h_1^{(n)}, h_2^{(n)}, h_3^{(n)}; \cdot) + \frac{\int_0^1 h_1(y,\cdot) + h_2(y,\cdot)\ \mathrm{d}g(y)}{\sqrt{F(\cdot)}} \right\|_\infty \le C(\|h_1\| + \|h_2\|)\|h_3\|,$$

and this bound is $o(\|h_1\| + \|h_2\| + \|h_3\|)$. This concludes that Ψ_g^F is differantiable at $(0,0,0)$ and $\mathrm{d}_0\Psi_g^F$ is the linear application H. □

Proof of Proposition 3.18. Let f, g be centred piecewise continuously differentiable. For sake of simplification, we assume that they are just continuously differentiable. Let us prove that for all $t, s \in C_\varepsilon$,

$$-\int_0^1 (\mathbb{G}_1(y,t) + \mathbb{G}_2(y,t))\ \mathrm{d}g(y) \perp\!\!\!\perp -\int_0^1 (\mathbb{G}_1(y,s) + \mathbb{G}_2(y,s))\ \mathrm{d}f(y).$$

By the Mapping theorem as the integral is continuous and Lemma 3.16,

$$\begin{aligned}\int_0^1 -g'(y)\left(\mathbb{G}_1(y,t)+\mathbb{G}_2(y,t)\right)\,\mathrm{d}y &= \lim\sqrt{n}\int_0^1 -g'(y)\left(\mathbb{F}_1^{(n)}(y,t)+\mathbb{F}_2^{(n)}(y,t)\right)\,\mathrm{d}y\\ &= \lim\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n g(Z_i)\mathbb{1}_{\{X_i\le t\}}\right).\end{aligned}$$

By a dependent Central Limit Theorem,

$$\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n g(Z_i)\mathbb{1}_{\{X_i\le t\}}\\ \frac{1}{n}\sum_{i=1}^n f(Z_i)\mathbb{1}_{\{X_i\le s\}}\end{pmatrix}\rightsquigarrow\mathcal{N}\left(\mathbf{0},\Sigma\right),$$

where

$$2\Sigma_{1,2}=\sum_{i\in\mathbb{Z}}\mathrm{cov}\left(g(Z_0)\mathbb{1}_{\{X_0\le t\}},f(Z_i)\mathbb{1}_{\{X_i\le s\}}\right)$$

If $i\neq 0$ by the Assumption (A1) and the Corollary 3.6, Z_i is independent of (Z_0,X_0,X_i), for $i>0$, or Z_0 is independent of (Z_i,X_i,X_0), for $i<0$, then

$$\mathrm{cov}\left([g(Z_0)-m(g)]\mathbb{1}_{\{X_0\le t\}},[f(Z_i)-m(f)]\mathbb{1}_{\{X_i\le s\}}\right)=0,$$

and for $i=0$,

$$\begin{aligned}\mathrm{cov}\left(g(Z_0)\mathbb{1}_{\{X_0\le t\}},f(Z_0)\mathbb{1}_{\{X_0\le s\}}\right) &= \mathbb{E}\left(g(Z_0)f(Z_0)\mathbb{1}_{\{X_0\le t\}}\right)\\ &= F(t)\times\mathbb{E}\left(f(Z_0)g(Z_0)\right)\\ &= F(t)\int_0^1 f(u)g(u)\,\mathrm{d}u=0.\end{aligned}$$

As the limit is Gaussian, this decorrelation implies the independence of the marginals.

□

Chapitre 4

Stone's theorem for distributional regression in Wasserstein distance

Ce chapitre a donné lieu au préprint Dombry et al. (2023), soumis dans *Bernoulli*. Nous allons nous intéresser à l'estimation de cette fameuse prédiction idéale, prédiction centrale dans les deux derniers chapitres.

Abstract : We extend the celebrated Stone's theorem to the framework of distributional regression. More precisely, we prove that weighted empirical distributions with local probability weights satisfying the conditions of Stone's theorem provide universally consistent estimates of the conditional distributions, where the error is measured by the Wasserstein distance of order $p \geq 1$. Furthermore, for $p = 1$, we determine the minimax rates of convergence on specific classes of distributions. We finally provide some applications of these results, including the estimation of conditional tail expectation or probability weighted moments.

1 Introduction

Forecasting is a major task from statistics and often of crucial importance for decision making. In the simple case when the quantity of interest is univariate and quantitative, point forecasting often takes the form of regression where one aims at estimating the conditional mean (or the conditional quantile) of the response variable Y given the available information encoded in a vector of covariates X. A point forecast is only a rough summary statistic and should at least be accompanied with an assessment of uncertainty (e.g. standard deviation or a confidence interval). Alternatively, probabilistic forecasting and distributional regression (Gneiting and Katzfuss, 2014) suggest to estimate the full conditional distribution of Y given X, called the predictive distribution.

In the last decades, weather forecasting has been a major motivation for the development of probabilistic forecasts. Ensemble forecasts are based on a given number of deterministic models whose parameters vary slightly in order to take into account observation errors and incomplete physical representation of the atmosphere. This leads to an ensemble of different forecasts that overall also assess the uncertainty of the forecast. Ensemble forecasts suffer from bias and underdispersion (Hamill and Colucci, 1997) and need to be statistically postprocessed in order to be improved. Different postprocessing methods have been proposed, such as Ensemble Model Output Statistics (Gneiting et al., 2005), Quantile Regression Forests (Taillardat et al., 2019) or Neural Networks (Schulz and Lerch, 2021) among others. Distributional regression is now widely used beyond meteorology and recent methodological works include deep distribution regression by Li et al. (2021), distributional random forest by Ćevid et al. (2022) or isotonic distributional regression by Henzi et al. (2021b).

The purpose of the present paper is to provide an extension to the framework of distributional

regression of the celebrated Stone's theorem (Stone, 1977) that states the consistency of the local weight algorithm for the estimation of the regression function. The strength of Stone's theorem is that it is fully non-parametric and model-free, with very mild assumptions that covers many important cases such as kernel algorithms and nearest neighbor methods, see e.g. Györfi et al. (2002) for more details. We prove that Stone's theorem has a natural and elegant extension to distributional regression with error measured by the Wasserstein distance of order $p \geq 1$. Our result covers not only the case of a one-dimensional output $Y \in \mathbb{R}$ where the Wasserstein distance has a simple explicit form, but also the case of a multivariate output $Y \in \mathbb{R}^d$. The use of the Wasserstein distance is motivated by recent works revealing that it is a useful and powerful tool in statistics, see e.g. the review by Panaretos and Zemel (2020). Besides this main result, we characterize, in the case $d = 1$ and $p = 1$, the optimal minimax rate of convergence on suitable classes of distributions. We also discuss implications of our results to estimate various statistics of possible interest such as the expected shortfall or the probability weighted moment.
The structure of the paper is the following. In Section 2, we present the required background on Stone's theorem and Wasserstein spaces. Section 3 gathers our main results, including the extension of Stone's theorem to distributional regression (Theorem 4.5), the characterization of optimal minimax rates of convergence (Theorem 4.10) and some applications (Proposition 4.12 and the subsequent examples). All the technical proofs are postponed to Section 4.

2 Background

2.1 Stone's theorem

In a regression framework, we observe a sample (X_i, Y_i), $1 \leq i \leq n$, of independent copies of $(X, Y) \in \mathbb{R}^k \times \mathbb{R}^d$ with distribution P. Based on this sample and assuming Y integrable, the goal is to estimate the regression function

$$r(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}^k.$$

Local average estimators take the form

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \tag{4.1}$$

with $W_{n1}(x), \ldots, W_{nn}(x)$ the *local weights* at x. The local weights are assumed to be measurable functions of x and $X_1, \ldots, X_n$ but not to depend on $Y_1, \ldots, Y_n$, that is

$$W_{ni}(x) = W_{ni}(x; X_1, \ldots, X_n), \quad 1 \leq i \leq n. \tag{4.2}$$

For the convenience of notation, the dependency on $X_1, \ldots, X_n$ is implicit. In this paper, we focus only on the case of *probability weights* satisfying

$$W_{ni}(x) \geq 0, \ 1 \leq i \leq n, \quad \text{and} \quad \sum_{i=1}^{n} W_{ni}(x) = 1. \tag{4.3}$$

Stone's Theorem states the universal consistency of the regression estimate in L^p-norm.

Theorem 4.1 (Stone (1977))**.** *Assume the probability weights* (4.3) *satisfy the following three conditions :*

i) there is $C > 0$ *such that* $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) g(X_i)\right] \leq C\mathbb{E}[g(X)]$ *for all* $n \geq 1$ *and measurable* $g : \mathbb{R}^k \to [0, +\infty)$ *such that* $\mathbb{E}[g(X)] < \infty$ *;*

ii) for all $\varepsilon > 0$, $\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > \varepsilon\}} \to 0$ *in probability as* $n \to +\infty$ *;*

iii) $\max_{1 \leq i \leq n} W_{ni}(X) \to 0$ *in probability as* $n \to +\infty$.

Then, for all $p \geq 1$ *and* $(X, Y) \sim P$ *such that* $\mathbb{E}[\|Y\|^p] < \infty$,

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{4.4}$$

Conversely, if Equation (4.4) *holds, then the probability weights must satisfy conditions* $i) - iii)$.

Remark 4.2. Stone's theorem is usually stated in dimension $d = 1$. Since the convergence of random vectors $\hat{r}_n(X) \to r(X)$ in L^p is equivalent to convergence in L^p of all the components, the extension to the dimension $d \geq 2$ is straightforward. Furthermore, more general weights than probability weights can be considered : condition (4.3) can be dropped and replaced by the weaker assumptions that

$$|W_{ni}(X)| \leq M \quad \text{a.s. for some } M > 0.$$

and

$$\sum_{i=1}^{n} W_{ni}(X) \to 1 \text{ in probability.}$$

Such general weights will not be considered in the present paper and we therefore stick to probability weights. The reader can refer to Biau and Devroye (2015) for a complete proof of Stone's theorem together with a discussion.

Example 4.3. The following two examples of kernel weights and nearest neighbor weights are the most important ones in the literature and we refer to Györfi et al. (2002) Chapter 5 and 6, respectively, for more details.

- The kernel weights are defined by

$$W_{ni}(x) = \frac{K\left(\frac{x - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}, \quad 1 \leq i \leq n \tag{4.5}$$

  if the denominator is nonzero, and $1/n$ otherwise. Here the bandwidth $h_n > 0$ depends only on the sample size n and the function $K : \mathbb{R}^k \to [0, +\infty)$ is called a kernel. In this case, the estimator (4.1) corresponds to the Nadaraya-Watson estimator of the regression function (Nadaraya, 1964; Watson, 1964). We say that K is a boxed kernel if there are constants $R_2 \geq R_1 > 0$ and $M_2 \geq M_1 > 0$ such that

$$M_1 \mathbb{1}_{\{\|x\| \leq R_1\}} \leq K(x) \leq M_2 \mathbb{1}_{\{\|x\| \leq R_2\}}, \quad x \in \mathbb{R}^k.$$

  Theorem 5.1 in Györfi et al. (2002) states that, for a boxed kernel, the kernel weights (4.5) satisfy conditions $i) - iii)$ of Theorem 4.1 if and only if $h_n \to 0$ and $nh_n^k \to +\infty$ as $n \to +\infty$.

- The nearest neighbor (NN) weights are defined by

$$W_{ni}(x) = \begin{cases} \frac{1}{\kappa_n} & \text{if } X_i \text{ belongs to the } \kappa_n\text{-NN of } x \\ 0 & \text{otherwise} \end{cases}, \tag{4.6}$$

  where the number of neighbors $\kappa_n \in \{1, \ldots, n\}$ depends only on the sample size. Recall that the κ_n-NN of x within the sample $(X_i)_{1 \leq i \leq n}$ are obtained by sorting the distances $\|X_i - x\|$ in increasing order and keeping the κ_n points with the smallest distances – as discussed in Györfi et al. (2002) Chapter 6, several rules can be used to break ties such as lexicographic or random tie breaking. Theorem 6.1 in the same reference states that the nearest neighbor weights (4.6) satisfy conditions $i) - iii)$ of Theorem 4.1 if and only if $\kappa_n \to +\infty$ and $\kappa_n / n \to 0$ as $n \to +\infty$.

Example 4.4. Interestingly, some variants of the celebrated Breiman's Random Forest (Breiman, 2001) produce probability weights satisfying the assumptions of Stone's theorem. In Breiman's Random Forest, the splits involve both the covariates and the response variable so that the associated weights $W_{ni}(x) = W_{ni}(x; (X_l, Y_l)_{1\leq l\leq n})$ are not in the form (4.2). Scornet (2016) considers two simplified versions of infinite random forests where the associated weights $W_{ni}(x)$ do not depend on the response values and satisfy the so called X-property, that is they are in the form (4.2). For totally non adaptive forests, the trees are grown thanks to a binary splitting rule that does not use the training sample and is totally random ; the author shows that the probability weights associated to the infinite forest satisfy the assumptions of Stone's theorem under the condition that the number of leaves grows to infinity at a rate smaller than n and the leaf volume tends to zero in probability (see Theorem 4.1 and its proof). For q-quantile forests, the binary splitting rules involves only the covariates and the author shows that the weights associated to the infinite forest satisfy the assumptions of Stone's theorem provided the subsampling number a_n satifies $a_n \to +\infty$ and $a_n/n \to 0$ (see Theorem 5.1 and its proof).

2.2 Wasserstein spaces

We recall the definition and some elementary facts on Wasserstein spaces on $\mathbb{R}^d$. More details and further results on optimal transport and Wasserstein spaces can be found in the monograph by Villani (2009), Chapter 6.
For $p \geq 1$, the Wasserstein space $\mathcal{P}^p(\mathbb{R}^d)$ is defined as the set of Borel probability measures on $\mathbb{R}^d$ having a finite moment of order p, i.e. such that

$$M_p(\mu) = \Big(\int_{\mathbb{R}^d} \|y\|^p \, \mu(\mathrm{d}y) \Big)^{1/p} < \infty. \tag{4.7}$$

It is endowed with the distance defined, for $Q_1, Q_2 \in \mathcal{P}^p(\mathbb{R}^d)$, by

$$W_p(Q_1, Q_2) = \inf_{\pi \in \Pi(Q_1, Q_2)} \left(\int \|y_1 - y_2\|^p \, \pi(\mathrm{d}y_1 \mathrm{d}y_2) \right)^{1/p}, \tag{4.8}$$

where $\Pi(Q_1, Q_2)$ denotes the set of measures on $\mathbb{R}^d \times \mathbb{R}^d$ with margins Q_1 and Q_2. A couple (Z_1, Z_2) of random variables with distributions Q_1 and Q_2 respectively is called a *coupling*. The Wasserstein distance is thus the minimal distance $\|Z_1 - Z_2\|_{\mathrm{L}^p} = \mathbb{E}[\|Z_1 - Z_2\|^p]^{1/p}$ over all possible couplings. Existence of optimal couplings is ensured since $\mathbb{R}^d$ is a complete and separable metric space so that the infimum is indeed a minimum.
Wasserstein distances are generally difficult to compute, but the case $d = 1$ is the exception. A simple optimal coupling is provided by the probability inverse transform : for $i = 1, 2$, let $Q_i \in \mathcal{P}^p(\mathbb{R})$, F_i denotes its cumulative distribution function and F_i^{-1} its generalized inverse (quantile function). Then, starting from a uniform random variable $U \sim \mathrm{Unif}(0, 1)$, an optimal coupling is given by $(Z_1, Z_2) = (F_1^{-1}(U), F_2^{-1}(U))$. Therefore, the Wasserstein distance is explicitly given by

$$W_p(Q_1, Q_2) = \left(\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p \mathrm{d}u \right)^{1/p}. \tag{4.9}$$

When $p = 1$, a simple change of variable yields

$$W_1(Q_1, Q_2) = \int_{-\infty}^{+\infty} |F_1(u) - F_2(u)| \mathrm{d}u. \tag{4.10}$$

3 Main results

3.1 Stone's theorem for distributional regression

We now present the main result of the paper which is a natural extension of Stone's theorem to the framework of distributional regression. Given a distribution $(X, Y) \sim P$ on $\mathbb{R}^k \times \mathbb{R}^d$, we

denote by F the marginal distribution of Y and by F_x its conditional distribution given $X = x$. This conditional distribution can be estimated on a sample $(X_i, Y_i)_{1\leq i\leq n}$ of independent copies of (X, Y) by the weighted empirical distribution

$$\hat{F}_{n,x} = \sum_{i=1}^{n} W_{ni}(x)\delta_{Y_i} \tag{4.11}$$

where δ_y denotes the Dirac mass at point $y \in \mathbb{R}^d$. For probability weights satisfying (4.3), $\hat{F}_{n,x}$ is a probability measure and can be viewed as a random element in the complete and separable space $\mathcal{P}^p(\mathbb{R}^d)$. We recall that the weights $W_{ni}(x) = W_{ni}(x; X_1, \ldots, X_n)$ implicitly depend on $X_1, \ldots, X_n$ but not on $Y_1, \ldots, Y_n$.

Theorem 4.5. *Assume the probability weights satisfy conditions i) − iii) from Theorem 4.1. Then, for all $p \geq 1$ and (X, Y) such that $\mathbb{E}[\|Y\|^p] < \infty$,*

$$\mathbb{E}\big[W_p^p(\hat{F}_{n,X}, F_X)\big] \longrightarrow 0 \quad \text{as } n \to +\infty. \tag{4.12}$$

Conversely, if Equation (4.12) holds, then the probability weights must satisfy conditions i) − iii).

It is worth noticing that

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \leq \mathbb{E}\big[W_p^p(\hat{F}_{n,X}, F_X)\big]$$

so that Theorem 4.5 implies Theorem 4.1 in a straightforward way. The proof of Theorem 4.5 is postponed to Section 4. It first considers the case $d = 1$ where the Wasserstein distance is explicitly given by formula (4.9). Then, the results is extended to higher dimension $d \geq 2$ thanks to the notion of max-sliced Wasserstein distance (Bayraktar and Guo, 2021a) which allows to reduce the convergence of measures on $\mathbb{R}^d$ to the convergence of their one dimensional projections (a precise statement is given in Theorem 4.17 below).

3.2 Rates of convergence

We next consider rates of convergence in the minimax sense. Note that similar questions and results have been established in Pic et al. (2022), where the second order Cramér's distance was considered, i.e.

$$\|\hat{F}_{n,X} - F_X\|_{L_2}^2 = \int_{\mathbb{R}} |\hat{F}_{n,X}(y) - F_X(y)|^2 \, \mathrm{d}y.$$

We focus here on the Wasserstein distance $W_p(\hat{F}_{n,X}, F_X)$ and consider only the case $d = 1$ and $p = 1$ which allows the explicit expression (4.10). The other cases seem harder to analyze and are beyond the scope of the present paper. Our first result considers the error in Wasserstein distance when $X = x$ is fixed.

Proposition 4.6. *Assume $d = 1$ and $(X, Y) \sim P$ such that $\mathbb{E}[|Y|] < \infty$. Then,*

$$\mathbb{E}\big[W_1(\hat{F}_{n,x}, F_x)\big] \leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x) W_1(F_{X_i}, F_x)\Big] + M(x)\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2},$$

where $M(x) = \int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\mathrm{d}z$.

The first term corresponds to an approximation error due to the fact that we use a biased sample to estimate F_x. The more regular the model is, the smaller the approximation error is. The second term is an estimation error due to the fact that we use an empirical mean to estimate F_x. This estimator error is smaller if the distribution error has a lower dispersion (as measured by $M(x)$) or if $\sum_{i=1}^n W_{ni}^2(x)$ is small. Note that in the case of nearest neighbor weights, $1/\sum_{i=1}^n W_{ni}^2(x)$ is

exactly equal to κ so that this quantity is often referred to as the *effective sample size* and the estimation error is proportional to the square root of the expected reciprocal effective sample size.
In view of Proposition 4.6, we introduce the following classes of functions.

Definition 4.7. *Let $\mathcal{D}(H, L, M)$ be the class of distributions $(X, Y) \sim P$ on $\mathbb{R}^k \times \mathbb{R}$ satisfying :*

a) $X \in [0,1]^k$ a.s. and $\mathbb{E}|Y| < \infty$,
b) for all $x, x' \in [0,1]^k$, $W_1(F_x, F_{x'}) \leq L\|x - x'\|^H$,
c) for all $x \in [0,1]^k$, $\int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\,\mathrm{d}z \leq M$.

The definition of the class together with Proposition 4.6 entails that the expected error is uniformly bounded on the class $\mathcal{D}(H, L, M)$ by

$$\begin{aligned}&\mathbb{E}\Big[W_1(\hat{F}_{n,X}, F_X)\Big] \\ &\leq L\mathbb{E}\Big[\sum_{i=1}^n W_{ni}(X)\|X_i - X\|^H\Big] + M\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(X)\Big]^{1/2}. \end{aligned} \tag{4.13}$$

As a consequence, Proposition 4.6 allows to derive explicit bounds uniformly on $\mathcal{D}(H, L, M)$ for the kernel and nearest neighbor methods from Example 4.3. For the sake of simplicity, we consider the uniform kernel only.

Corollary 4.8. *Let $\hat{F}_{n,X}$ be given by the kernel method with uniform kernel $K(x) = \mathbb{1}_{\{\|x\| \leq 1\}}$ and weights given by Equation (4.5). If $P \in \mathcal{D}(H, L, M)$, then*

$$\mathbb{E}\big[W_1(\hat{F}_{n,X}, F_X)\big] \leq Lh_n^H + M\sqrt{(2 + 1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}$$

with $c_k = k^{k/2}$.

Corollary 4.9. *Let $\hat{F}_{n,X}$ be given by the nearest neighbor method with weights given by Equation (4.6) and assume $P \in \mathcal{D}(H, L, M)$. Then,*

$$\mathbb{E}\big[W_1(\hat{F}_{n,X}, F_X)\big] \leq \begin{cases} L8^{H/2}(\kappa_n/n)^{H/2} + M\kappa_n^{-1/2} & \text{if } k = 1, \\ L\widetilde{c}_k^{H/2}(\kappa_n/n)^{H/k} + M\kappa_n^{-1/2} & \text{if } k \geq 2, \end{cases}$$

where $\widetilde{c}_k$ depends only on the dimension k and is defined in Biau and Devroye (2015, Theorem 2.4).

One can see that consistency holds — i.e. the expected error tends to 0 as $n \to +\infty$ — as soon as $h_n \to 0$ and $nh_n^k \to +\infty$ for the kernel method and $\kappa_n/n \to 0$ and $\kappa_n \to +\infty$ for the nearest neighbor method.

The next theorem provides the optimal minimax rate of convergence on the class $\mathcal{D}(H, L, M)$. We say that two sequences of positive numbers (a_n) and (b_n) have the same rate of convergence, noted $a_n \asymp b_n$, if the ratios a_n/b_n and b_n/a_n remain bounded as $n \to +\infty$.

Theorem 4.10. *The optimal minimax rate of convergence on the class $\mathcal{D}(H, L, M)$ is given by*

$$\inf_{\hat{F}_n} \sup_{P \in \mathcal{D}(H,L,M)} \mathbb{E}[W_1(\hat{F}_{n,X}, F_X)] \asymp n^{-H/(2H+k)}.$$

Theorem 4.10 is the counterpart of Pic et al. (2022, Theorem 1) where the minimax rate of convergence for the second order Cramér's distance has been considered. The strategy of proof is similar : i) we prove a lower bound by considering a suitable class of binary distributions where the error in Wasserstein distance corresponds to an absolute error in point regression for which the minimax lower rate of convergence is known ; ii) we check that the upper bound for the kernel

and/or nearest neighbor algorithm has the same rate of convergence as the lower bound, which proves that the optimal minimax rate of convergence has been identified. In particular, our proof shows that the kernel method defined in Equation (4.5) reaches the minimax rate of convergence in any dimension $k \geq 1$ with the choice of bandwidth $h_n \asymp n^{-1/(2H+k)}$; the nearest neighbor method defined in Equation (4.6) reaches the minimax rate of convergence in any dimension $k \geq 2$ with the number of neighbors $\kappa_n \asymp n^{H/(H+k/2)}$.

Remark 4.11. Our estimate of the minimax rate of convergence holds only for $d = p = 1$ and we briefly discuss what can be expected in other cases.
When $p = 1$ and $d \geq 2$, one may hope to use the strong equivalence between the max-sliced Wasserstein distance and the Wasserstein distance (Bayraktar and Guo, 2021a, Theorem 2.3.ii). This requires to estimate the expectation of a supremum over the sphere and this line of research is left for further work.
When $p > 1$, even in dimension $d = 1$, it seems difficult to obtain bounds for the Wasserstein distance of order p without very strong assumptions. Bobkov and Ledoux (2019) consider the rate of convergence of the empirical distribution $\hat{F}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for an i.i.d. sample $Y_1, \dots, Y_n$ with distribution F on $\mathbb{R}$. A first consistency result (Theorem 2.14) states that $\mathbb{E}[W_p^p(\hat{F}_n, F)] \to 0$ as soon as F has a finite moment of order $p \geq 1$. Regarding rates of convergence, they show (Corollary 3.9) that for $p = 1$ the standard rate of convergence holds, i.e. $\mathbb{E}[W_1(\hat{F}_n, F)] = O(1/\sqrt{n})$, if and only if

$$J_1(F) = \int_{\mathbb{R}} \sqrt{F(z)(1 - F(z))}\mathrm{d}z < \infty.$$

On the other hand, rate of convergences for higher order $p > 1$ require the condition

$$J_p(F) = \int_{\mathbb{R}} \frac{[F(z)(1 - F(z))]^{p/2}}{f(z)^{p-1}}\mathrm{d}z < \infty,$$

where f is the density of the absolutely continuous component of F. They show (Corollary 5.5) that the standard rate holds, i.e. $\mathbb{E}[W_p^p(\hat{F}_n, F)] = O(n^{-p/2})$, if and only if $J_p(F) < \infty$. However, this condition is very strong : it does not hold for the Gaussian distribution or for distributions with disconnected support.

3.3 Applications

We briefly illustrate Theorem 4.5 with some applications and examples. In statistics, we commonly face the following generic situation : we are interested in a summary statistic S with real values, e.g. quantiles or tail expectation, and we want to assess the effect of X on Y through S, that is we want to assess $S_{Y|X=x}$. Assuming that S is well-defined for distributions on $\mathbb{R}^d$ with a finite moment of order $p \geq 1$, it can be seen as a map $S : \mathcal{P}^p(\mathbb{R}^d) \to \mathbb{R}$ and then $S_{Y|X=x} = S(F_x)$ with F_x the conditional distribution of Y given $X = x$. A natural plug-in estimate of $S_{Y|X=x}$ is

$$\hat{S}_{n,x} = S(\hat{F}_{n,x}) \quad \text{with } \hat{F}_{n,x} \text{ defined by (4.11)}.$$

In this generic situation, our extension of Stone's theorem directly implies the following proposition. Recall that $M_p(\mu)$ is defined in Equation (4.7).

Proposition 4.12. *Assume $\mathbb{E}[\|Y\|^p] < \infty$ and $\mathbb{P}(F_X \in \mathcal{C}) = 1$ where $\mathcal{C} \subset \mathcal{P}^p(\mathbb{R}^d)$ denotes the continuity set of the statistic $S : \mathcal{P}^p(\mathbb{R}^d) \to \mathbb{R}$. Then weak consistency holds, i.e.*

$$\hat{S}_{n,X} \longrightarrow S_{Y|X} \quad \textit{in probability as } n \to +\infty.$$

If furthermore the statistic S admits a bound of the form

$$|S(\mu)| \leq a M_p^q(\mu) + b, \quad \textit{with } a, b \geq 0 \textit{ and } 0 < q \leq p, \tag{4.14}$$

then consistency holds in $\mathrm{L}^{p/q}$*, i.e.*

$$\mathbb{E}\big[|\hat{S}_{n,X} - S_{Y|X}|^{p/q}\big] \longrightarrow 0 \quad \textit{as } n \to +\infty$$

Example 4.13. (quantile). For a distribution G on $\mathbb{R}$, we define the associated quantile function

$$G^{-1}(\alpha) = \inf\{z \in \mathbb{R} : G(z) \geq \alpha\}, \quad 0 < \alpha < 1.$$

It is well-known that the weak convergence $G_n \xrightarrow{d} G$ implies the quantile convergence $G_n^{-1}(\alpha) \to G^{-1}(\alpha)$ at each continuity point α of G^{-1}. Equivalently, considering $\mathcal{P}(\mathbb{R})$ endowed with the weak convergence topology, the α-quantile statistic $S_\alpha(G) = G^{-1}(\alpha)$ is continuous at G as soon as G^{-1} is continuous at α.
In view of this, we let $\mathcal{C} = \{G \in \mathcal{P}(\mathbb{R})\colon G^{-1}$ continuous on $(0,1)\}$ and assume that the conditional distribution satisfies $\mathbb{P}(F_X \in \mathcal{C}) = 1$. Then weak convergence holds for the conditional quantiles, i.e.

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in probability.}$$

Note that no integrability condition is needed here because we can apply Proposition 4.12 on the transformed data $(X_i, \widetilde{Y}_i)_{1\leq i\leq n}$, where $\widetilde{Y}_i = \tan^{-1}(Y_i)$ is bounded so that convergence in Wasserstein distance is equivalent to weak convergence. If furthermore Y is p-integrable, then the bound

$$\begin{aligned} |S_\alpha(G)|^p &\leq \frac{1}{\alpha}\int_0^\alpha |G^{-1}(u)|^p \mathrm{d}u + \frac{1}{1-\alpha}\int_\alpha^1 |G^{-1}(u)|^p \mathrm{d}u \\ &\leq \Big(\frac{1}{\alpha} + \frac{1}{1-\alpha}\Big) M_p^p(G) \end{aligned}$$

implies the strengthened convergence

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in } \mathrm{L}^p.$$

Example 4.14. (tail expectation) The tail expectation above level $\alpha \in (0,1)$ is the risk measure defined for $G \in \mathcal{P}^1(\mathbb{R})$ by

$$S_\alpha(G) = \frac{1}{1-\alpha}\int_\alpha^1 G^{-1}(u)\,\mathrm{d}u.$$

The name comes from the equivalent definition

$$S_\alpha(G) = \mathbb{E}[Y \mid Y > G^{-1}(\alpha)], \quad Y \sim G,$$

which holds when G^{-1} is continuous at α. One can see that

$$\begin{aligned} |S_\alpha(G_1) - S_\alpha(G_2)| &\leq \frac{1}{1-\alpha}\int_\alpha^1 |G_1^{-1}(u) - G_1^{-1}(u)|\,\mathrm{d}u \\ &\leq \frac{1}{1-\alpha}\int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|\,\mathrm{d}u \\ &= \frac{1}{1-\alpha} W_1(G_1, G_2). \end{aligned}$$

so that S_α is Lipschitz continuous with respect to the Wasserstein distance W_1. As a consequence, the conditional tail expectation $S_\alpha(F_x)$ can be estimated in a consistent way by the plug-in estimator $S_\alpha(\hat{F}_{n,x})$ since

$$\mathbb{E}[|S_\alpha(\hat{F}_{n,X}) - S_\alpha(F_X)|] \leq \frac{1}{1-\alpha}\mathbb{E}[W_1(\hat{F}_{n,X}, F_X)] \longrightarrow 0.$$

Example 4.15. (probability weighted moments) A similar result holds for the probability weighted moment of order $p, q > 0$ defined by

$$S_{p,q}(G) = \int_0^1 G^{-1}(u) u^p (1-u)^q \, \mathrm{d}u, \quad G \in \mathcal{P}^1(\mathbb{R}).$$

(Greenwood et al. (1979)). The name comes from the equivalent definition

$$S(G) = \mathbb{E}[Y G(Y)^p (1 - G(Y))^q], \quad Y \sim G,$$

which holds when G^{-1} is continuous on $(0, 1)$. One can again check that the statistic $S_{p,q}$ is Lipschitz continuous with respect to the Wasserstein distance W_1 since

$$\begin{aligned} |S_{p,q}(G_1) - S_{p,q}(G_2)| &\leq \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)| u^p (1-u)^q \, \mathrm{d}u \\ &\leq \max_{0 \leq u \leq 1} u^p (1-u)^q \times \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)| \, \mathrm{d}u \\ &= \Big(\frac{p}{p+q}\Big)^p \Big(\frac{q}{p+q}\Big)^q W_1(G_1, G_2). \end{aligned}$$

Example 4.16. (covariance) We conclude with a simple example in dimension $d = 2$ where the statistic of interest is the covariance between the two components of $Y = (Y_1, Y_2)$ given $X = x$. Here, we consider

$$S(G) = \int_{\mathbb{R}^2} y_1 y_2 \, \mathrm{d}G - \int_{\mathbb{R}^2} y_1 \, \mathrm{d}G \int_{\mathbb{R}^2} y_2 \, \mathrm{d}G, \quad G \in \mathcal{P}^2(\mathbb{R}^2).$$

Considering square integrable random vectors $Y = (Y_1, Y_2)$ and $Z = (Z_1, Z_2)$ with distribution G and H respectively, we compute

$$\begin{aligned} &|S(G) - S(H)| \\ &= \big|\mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(Z_1, Z_2)\big| \\ &= \big|\mathrm{Cov}(Y_1, Y_2 - Z_2) - \mathrm{Cov}(Z_1 - Y_1, Z_2)\big| \\ &\leq \mathrm{Var}(Y_1)^{1/2} \mathrm{Var}(Y_2 - Z_2)^{1/2} + \mathrm{Var}(Z_2)^{1/2} \mathrm{Var}(Z_1 - Y_1)^{1/2} \end{aligned}$$

were the last line is a consequence of the Cauchy-Schwartz inequality. We have the upper bounds

$$\mathrm{Var}(Y_1)^{1/2} \leq M_2(G), \quad \mathrm{Var}(Z_2)^{1/2} \leq M_2(H)$$

and, choosing an optimal coupling (Y, Z) between G and H,

$$\mathrm{Var}(Z_1 - Y_1)^{1/2} \leq \|Y - Z\|_{L^2} = W_2(G, H), \quad \mathrm{Var}(Y_2 - Z_2)^{1/2} \leq W_2(G, H).$$

Altogether, we obtain,

$$|S(G) - S(H)| \leq \big(M_2(G) + M_2(H)\big) W_2(G, H).$$

This proves that S is locally Lipschitz and hence continuous with respect to the distance W_2. Taking $H = \delta_0$, we obtain

$$|S(G)| \leq M_2(G)^2$$

and the bound (4.14) holds with $q = 2$. Thus Proposition 4.12 implies that the plug-in estimator

$$S(\hat{F}_{n,x}) = \sum_{i=1}^n W_{ni}(x) Y_{1i} Y_{2i} - \sum_{i=1}^n W_{ni}(x) Y_{1i} \sum_{i=1}^n W_{ni}(x) Y_{2i}$$

is consistent in absolute mean for the conditional covariance

$$S(F_x) = \mathbb{E}(Y_1 Y_2 \mid X = x) - \mathbb{E}(Y_1 \mid X = x) \mathbb{E}(Y_2 \mid X = x),$$

i.e. $\mathbb{E}[|S(\hat{F}_{n,X}) - S(F_X)|] \longrightarrow 0$ as $n \to +\infty$.

4 Proofs

4.1 Proof of Theorem 4.5

Proof of Theorem 4.5 - case $d = 1$. We first consider the case when Y is uniformly bounded and takes its values in $[-M, M]$ for some $M > 0$. Then, it holds

$$F_x(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \geq M \end{cases} \quad \text{and} \quad \hat{F}_{n,x}(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \geq M \end{cases}.$$

and the generalized inverse functions (quantile functions) are bounded in absolute value by M. As a consequence,

$$\begin{aligned} \mathbb{E}\left[W_p^p(\hat{F}_{n,X}, F_X)\right] &= \mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|^p \mathrm{d}z\right] \\ &\leq (2M)^{p-1}\mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|\mathrm{d}u\right] \\ &= (2M)^{p-1}\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z. \end{aligned} \tag{4.15}$$

In this lines, we have used Equations (4.9) and (4.10) together with Fubini's theorem.
Consider the regression model $(X, \mathbb{1}_{\{Y \leq z\}}) \in \mathbb{R}^d \times \mathbb{R}$ where $z \in [-M, M]$ is fixed. The corresponding regression function is

$$x \mapsto \mathbb{E}[\mathbb{1}_{\{Y \leq z\}} | X = x] = F_x(z)$$

and the local weight estimator associated with the sample $(X_i, \mathbb{1}_{\{Y_i \leq z\}})$, $1 \leq i \leq n$ is

$$x \mapsto \sum_{i=1}^{n} W_{ni}(x)\mathbb{1}_{\{Y_i \leq z\}} = \hat{F}_{n,x}(z).$$

An application of Stone's theorem with $p = 1$ yields

$$\mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \longrightarrow 0, \quad \text{as } n \to +\infty,$$

whence we deduce, by the dominated convergence theorem,

$$\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z \longrightarrow 0.$$

The upper bound (4.15) finally implies

$$\mathbb{E}\left[W_p^p(\hat{F}_{n,X}, F_X)\right] \longrightarrow 0.$$

We next consider the general case when Y is not necessarily bounded. For $M > 0$, we define the truncation Y^M of Y by

$$Y^M = \begin{cases} -M & \text{if } Y < -M \\ Y & \text{if } -M \leq Y < M \\ M & \text{if } Y \geq M \end{cases}.$$

We define similarly $Y_1^M, \ldots, Y_n^M$ the truncations of $Y_1, \ldots, Y_n$ respectively. The conditional distribution associated with Y^M is

$$F_x^M(z) = \mathbb{P}(Y^M \leq z | X = x) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq Y < M \\ 1 & \text{if } z \geq M \end{cases}.$$

The local weight estimation built on the truncated sample is

$$\hat{F}^M_{n,x}(z) = \sum_{i=1}^{n} W_{ni}(x) \mathbb{1}_{\{Y_i^M \leq z\}}.$$

By the triangle inequality,

$$W_p(\hat{F}_{n,x}, F_x) \leq W_p(\hat{F}_{n,x}, \hat{F}^M_{n,x}) + W_p(\hat{F}^M_{n,x}, F^M_x) + W_p(F^M_x, F_x),$$

whence we deduce

$$\begin{aligned} &\mathbb{E}[W_p^p(\hat{F}_{n,x}, F_x)] \\ &\leq 3^{p-1} \left(\mathbb{E}[W_p^p(\hat{F}_{n,X}, \hat{F}^M_{n,X})] + \mathbb{E}[W_p(\hat{F}^M_{n,X}, F^M_X)] + \mathbb{E}[W_p^p(F^M_X, F_X)] \right). \end{aligned}$$

By the preceding result in the bounded case, for any fixed M, the second term converge to 0 as $n \to +\infty$. We next focus on the first and third term.
For fixed $X = x$, there is a natural coupling between the distribution $\hat{F}_{n,x}$ and $\hat{F}^M_{n,x}$ given by (Z_1, Z_2) such that

$$(Z_1, Z_2) = (Y_i, Y_i^M) \quad \text{with probability } W_{ni}(x).$$

Clearly $Z_1 \sim \hat{F}_{n,x}$ and $Z_2 \sim \hat{F}^M_{n,x}$ and this coupling provides the upper bound

$$W_p^p(\hat{F}_{n,x}, \hat{F}^M_{n,x}) \leq \|Z_1 - Z_2\|^p_{\mathrm{L}^p} = \sum_{i=1}^{n} W_{ni}(x) |Y_i - Y_i^M|^p. \tag{4.16}$$

Let us introduce the function $g_M(x)$ defined by

$$g_M(x) = \mathbb{E}\left[|Y - Y^M|^p \mid X = x \right].$$

Using the fact that, conditionally on $X_1, \ldots, X_n$, the random variables $Y_1, \ldots, Y_n$ are independent with distribution $F_{X_1}, \ldots, F_{X_n}$, we deduce

$$\mathbb{E}\left[W_p^p(\hat{F}_{n,x}, \hat{F}^M_{n,x}) \right] \leq \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(x) g_M(X_i) \right].$$

The condition $i)$ on the weights in Stone's Theorem then implies

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) g_M(X_i) \right] \leq C \mathbb{E}[g_M(X)].$$

Because $|Y - Y^M|^p$ converges almost surely to 0 as $M \to +\infty$ and is bounded by $2^p |Y|^p$ which is integrable, Lebesgue's convergence theorem implies

$$\mathbb{E}[g_M(X)] = \mathbb{E}\left[|Y - Y^M|^p \right] \longrightarrow 0 \quad \text{as } M \to +\infty.$$

We deduce that the first term satisfies

$$\mathbb{E}\left[W_p^p(\hat{F}_{n,X}, \hat{F}^M_{n,X}) \right] \leq C \mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

where the convergence is uniform in n.
We now consider the third term. Since Y^M is obtained from Y by truncation, the distribution functions and quantile functions of Y and Y^M are related by

$$F^M_x(z) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq z < M \\ 1 & \text{if } z \geq M \end{cases}$$

and

$$(F_x^M)^{-1}(u) = \begin{cases} -M & \text{if } F_x^{-1}(u) < -M \\ (F_x)^{-1}(u) & \text{if } -M \leq F_x^{-1}(u) < M \\ M & \text{if } F_x^{-1}(u) \geq M \end{cases}.$$

As a consequence

$$\begin{aligned} W_p^p(F_x^M, F_x) &= \int_0^1 |(F_x^M)^{-1}(u) - F_x^{-1}(u)|^p \mathrm{d}u \\ &= \mathbb{E}\left[|Y^M - Y|^p \mid X = x\right] = g_M(x). \end{aligned}$$

We deduce

$$\mathbb{E}\left[W_p^p(F_X^M, F_X)\right] = \mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

where the convergence is uniform in n.
We finally combine the three terms. The sum can be made smaller than any $\varepsilon > 0$ by first choosing M large enough so that the first and third terms are smaller than $\varepsilon/3$ and then choosing n large enough so that the second term is smaller than $\varepsilon/3$. This proves Equation (4.12) and concludes the proof. □

In order to extend the proof from $d = 1$ to $d \geq 2$, we need the notion of *sliced Wasserstein distance*, see Bayraktar and Guo (2021a) for instance. Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\| = 1\}$ be the unit sphere in $\mathbb{R}^d$ and, for $u \in \mathbb{R}^d$, let $u_* : \mathbb{R}^d \to \mathbb{R}$ be the linear form defined by $u_*(x) = u \cdot x$. The projection in direction u of a measure μ on $\mathbb{R}^d$ is defined as the pushforward $\mu \circ u_*^{-1}$ which is a measure on $\mathbb{R}$. The inequality $|u \cdot x| \leq \|x\|$ implies that $\mu \circ u_*^{-1} \in \mathcal{P}^p(\mathbb{R})$ for all $\mu \in \mathcal{P}^p(\mathbb{R}^d)$ and $u \in \mathbb{S}^{d-1}$. The sliced and max-sliced Wasserstein distances between $\mu, \nu \in \mathcal{P}^p(\mathbb{R}^d)$ are then defined respectively by

$$SW_p(\mu, \nu) = \left(\int_{\mathbb{S}^{d-1}} W_p^p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1})\, \sigma(\mathrm{d}u)\right)^{1/p},$$

where σ denotes the uniform measure on $\mathbb{S}^{d-1}$ and

$$\overline{SW}_p(\mu, \nu) = \max_{u \in \mathbb{S}^{d-1}} W_p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1}).$$

In plain words, the sliced and max-sliced Wasserstein distance are respectively the average and the maximum over all the 1-dimensional Wasserstein distances between the projections of μ and ν. The following result is crucial in our proof.

Theorem 4.17 (Bayraktar and Guo (2021a))**.** *For all $p \geq 1$, SW_p and $\overline{SW}_p$ are distances on $\mathcal{P}^p(\mathbb{R}^d)$ which are equivalent to W_p, i.e. for all sequence $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}^p(\mathbb{R}^d)$*

$$SW_p(\mu_n, \mu) \to 0 \iff \overline{SW}_p(\mu_n, \mu) \to 0 \iff W_p(\mu_n, \mu) \to 0.$$

Proof of Theorem 4.5 - case $d \geq 2$. For the sake of clarity, we divide the proof into three steps :

1) we prove that the result holds in max-sliced Wasserstein distance, i.e. $\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)] \to 0$;
2) we deduce that $W_p(\hat{F}_{n,X}, F_X) \to 0$ in probability ;
3) we show that the sequence $W_p^p(\hat{F}_{n,X}, F_X)$ is uniformly integrable.

Points 2) and 3) together imply $\mathbb{E}[W_p^p(\hat{F}_{n,X}, F_X)] \to 0$ as required.

Step 1). For all $u \in \mathbb{S}^{d-1}$, the projection $\hat{F}_{n,X} \circ u_*^{-1}$ is the weighted empirical distribution

$$\hat{F}_{n,X} \circ u_*^{-1} = \sum_{i=1}^n W_{ni}(X) \delta_{Y_i \cdot u}.$$

An application of Theorem 4.5 to the 1-dimensional sample $(Y_i \cdot u)_{i\geq 1}$ yields

$$\mathbb{E}[W_p^p(\hat{F}_{n,X} \circ u_*^{-1}, F_X \circ u_*^{-1})] \longrightarrow 0. \tag{4.17}$$

Note indeed that $\mathbb{E}[|Y|^p] < \infty$ implies $\mathbb{E}[|Y \cdot u|^p] < \infty$ and that the conditional laws of $Y \cdot u$ are the pushforward of those of Y, i.e. $\mathcal{L}(Y \cdot u \mid X) = F_X \circ u_*^{-1}$.
We next consider the max-sliced Wasserstein distance. Regularity in the direction $u \in \mathbb{S}^{d-1}$ will be useful and we recall that the Wasserstein distance between projections depends on the direction in a Lipschitz way. More precisely, according to Bayraktar and Guo (2021a, Proposition 2.2),

$$|W_p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1}) - W_p(\mu \circ v_*^{-1}, \nu \circ v_*^{-1})| \leq (M_p(\mu) + M_p(\nu))\|u - v\|,$$

for all $\mu, \nu \in \mathcal{P}^p(\mathbb{R}^d)$ and $u, v \in \mathbb{S}^{d-1}$ (recall Equation (4.7) for the definition of $M_p(\mu)$, $M_p(\nu)$). The sphere $\mathbb{S}^{d-1}$ being compact, for all $\varepsilon > 0$, one can find $K \geq 1$ and $u_1, \dots, u_K \in \mathbb{S}^{d-1}$ such that the balls $B(u_i, \varepsilon)$ with centers u_i and radius ε cover the sphere. Then, due to the Lipschitz property, the max-sliced Wasserstein distance is controlled by

$$\begin{aligned}
&\overline{SW}_p(\hat{F}_{n,X}, F_X) \\
&= \max_{u \in \mathbb{S}^{d-1}} W_p^p(\hat{F}_{n,X} \circ u_*^{-1}, F_X \circ u^{-1}) \\
&\leq \max_{1 \leq k \leq K} W_p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1}) + \varepsilon(M_p(\hat{F}_{n,X}) + M_p(F_X)).
\end{aligned}$$

Elevating to the p-th power and taking the expectation, we deduce

$$\begin{aligned}
&\mathbb{E}\big[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)\big] \\
&\leq 3^{p-1}\mathbb{E}\big[\max_{1 \leq k \leq K} W_p^p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1})\big] + 3^{p-1}\varepsilon^p(\mathbb{E}\big[M_p^p(\hat{F}_{n,X})\big] + \mathbb{E}\big[M_p^p(F_X)\big]).
\end{aligned}$$

The first term converges to 0 thanks to Eq. (4.17), i.e.

$$\mathbb{E}[\max_{1 \leq i \leq K} W_p^p(\hat{F}_{n,X} \circ u_{i*}^{-1}, F_X \circ u_{i*}^{-1})] \longrightarrow 0.$$

The second term is controlled by a constant times ε^p since

$$\mathbb{E}[M_p^p(\hat{F}_{n,X})] = \mathbb{E}\big[\sum_{i=1}^n W_{ni}(X)\|Y_i\|^p\big] \leq C\mathbb{E}[\|Y\|^p]$$

(by property i) of the weights) and

$$\mathbb{E}[M_p^p(F_X)] = \mathbb{E}\big[\mathbb{E}[\|Y\|^p \mid X]\big] = \mathbb{E}[\|Y\|^p]$$

(by the tower property of conditional expectation). Letting $\varepsilon \to 0$, the second term can be made arbitrarily small. We deduce $\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)] \to 0$.

Step 2). As a consequence of step 1), $\overline{SW}_p(\hat{F}_{n,X}, F_X) \to 0$ in probability, or equivalently $\hat{F}_{n,X} \to F_X$ in probability in the metric space $(\mathcal{P}^p(\mathbb{R}^d), \overline{SW}_p)$. Theorem 4.17 implies that the identity mapping is continuous from $(\mathcal{P}^p(\mathbb{R}^d), \overline{SW}_p)$ into $(\mathcal{P}^p(\mathbb{R}^d), W_p)$. The continuous mapping theorem implies that $\hat{F}_{n,X} \to F_X$ in probability in the metric space $(\mathcal{P}^p(\mathbb{R}^d), W_p)$. Equivalently, $W_p(\hat{F}_{n,X}, F_X) \to 0$ in probability.

Step 3). By the triangle inequality,

$$W_p(\hat{F}_{n,X}, F_X) \leq W_p(\hat{F}_{n,X}, \delta_0) + W_p(\delta_0, F_X)$$

with δ_0 the Dirac mass at 0. Furthermore, for any $\mu \in W_p(\mathbb{R}^d)$,

$$W_p(\mu, \delta_0) = \left(\int_{\mathbb{R}^d} \|x\|^p \, \mu(\mathrm{d}x)\right)^{1/p} = M_p(\mu).$$

We deduce

$$W_p^p(\hat{F}_{n,X}, F_X) \leq 2^{p-1} M_p^p(\hat{F}_{n,X}) + 2^{p-1} M_p^p(F_X).$$

In order to prove the uniform integrability of the left hand side, it is enough to prove that

$$M_p^p(F_X) \text{ is integrable and } M_p^p(\hat{F}_{n,X}),\ n \geq 1, \text{ is uniformly integrable.} \tag{4.18}$$

We have

$$M_p^p(F_X) = \mathbb{E}[\|Y\|^p \mid X]$$

which is integrable because $\mathbb{E}[\|Y\|^p] < \infty$. Furthermore,

$$M_p^p(\hat{F}_{n,X}) = \sum_{i=1}^n W_{ni}(X)\|Y_i\|^p$$

and Stone's Theorem ensures that

$$\sum_{i=1}^n W_{ni}(X)\|Y_i\|^p \longrightarrow \mathbb{E}[\|Y\|^p \mid X] \quad \text{in } L^1.$$

Since the sequence $M_p^p(\hat{F}_{n,X})$ converges in L^1, it is uniformly integrable and the claim follows. □

4.2 Proof of Proposition 4.6, Corollaries 4.8-4.9 and Theorem 4.10

Proof of Proposition 4.6. The proof of the upper bound relies on a coupling argument. Without loss of generality, we can assume that the Y_i's are generated from uniform random variables U_i's by the inversion method – i.e. we assume that U_i, $1 \leq i \leq n$, are independent identically distributed random variables with uniform distribution on $(0,1)$ that are furthermore independent from the covariates X_i, $1 \leq i \leq n$ and we set $Y_i = F_{X_i}^{-1}(U_i)$. Then the sample (X_i, Y_i) is i.i.d. with distribution P. In order to compare $\hat{F}_{n,x}$ and F_x, we introduce the random variables $\widetilde{Y}_i = F_x^{-1}(U_i)$ and we define

$$\widetilde{F}_{n,x}(z) = \sum_{i=1}^n W_{ni}(x)\mathbb{1}_{\{\widetilde{Y}_i \leq z\}}.$$

By the triangle inequality,

$$W_1(\hat{F}_{n,x}, F_x) \leq W_1(\hat{F}_{n,x}, \widetilde{F}_{n,x}) + W_1(\widetilde{F}_{n,x}, F_x).$$

In the right hand side, the first term is interpreted as an *approximation error* comparing the weighted sample $(Y_i, W_{ni}(x))$ to $(\widetilde{Y}_i, W_{ni}(x))$ where the $\widetilde{Y}_i$ have the target distribution F_x. The second term is an *estimation error* where we use the weighted sample $(\widetilde{Y}_i, W_{ni}(x))$ with the correct distribution to estimate F_x.
We first consider the approximation error. A similar argument as for the proof of Equation (4.16) implies

$$W_1(\hat{F}_{n,x}, \widetilde{F}_{n,x}) \leq \sum_{i=1}^n W_{ni}(x)|Y_i - \widetilde{Y}_i|.$$

Introducing the uniform random variables U_i's, we get

$$\begin{aligned}
\mathbb{E}[W_1(\hat{F}_{n,x}, \widetilde{F}_{n,x})] &\leq \mathbb{E}\Big[\sum_{i=1}^n W_{ni}(x)|F_{X_i}^{-1}(U_i) - F_x^{-1}(U_i)|\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^n W_{ni}(x)\int_0^1 |F_{X_i}^{-1}(u) - F_x^{-1}(u)|\,\mathrm{d}u\Big] \quad \text{by independence} \\
&= \mathbb{E}\Big[\sum_{i=1}^n W_{ni}(x) W_1(F_{X_i}, F_x)\Big],
\end{aligned}$$

where the equality relies on Equation (4.9). Note that this control of the approximation error is very general and could be extended to the Wasserstein distance of order $p > 1$.
We next consider the estimation error and our approach works for $p = 1$ only. By Equation (4.10),

$$\mathbb{E}[W_1(\widetilde{F}_{n,x}, F_x)] = \mathbb{E}\Big[\int_{\mathbb{R}} \Big|\sum_{i=1}^n W_{ni}(x)\big(\mathbb{1}_{\{\widetilde{Y}_i \le z\}} - F_x(z)\big)\Big| \mathrm{d}z\Big].$$

Applying Fubini's theorem and using the upper bound

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{i=1}^n W_{ni}(x)\big(\mathbb{1}_{\{\widetilde{Y}_i \le z\}} - F_x(z)\big)\Big|\Big] \\
&\le \mathbb{E}\Big[\Big|\sum_{i=1}^n W_{ni}(x)\big(\mathbb{1}_{\{\widetilde{Y}_i \le z\}} - F_x(z)\big)\Big|^2\Big]^{1/2} \\
&= \mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(x)\Big]^{1/2} \sqrt{F_x(z)(1 - F_x(z))},
\end{aligned}$$

we deduce

$$\mathbb{E}[W_1(\widetilde{F}_{n,x}, F_x)] \le \mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(x)\Big]^{1/2} \int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))} \mathrm{d}z.$$

Collecting the two terms yields Proposition 4.6. □

Proof of Corollary 4.8. For the kernel algorithm with uniform kernel and weights (4.5), we denote by

$$N_n(X) = \sum_{i=1}^n \mathbb{1}_{\{X_i \in B(X, h_n)\}}$$

the number of points in the ball $B(X, h_n)$ with center X and radius h_n. If $N_n \ge 1$, only the points in $B(X, h_n)$ have a nonzero weight which is equal to $1/N_n$. If $N_n = 0$, then by convention all the weights are equal to $1/n$. Thus we deduce

$$\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(X)\Big] = \mathbb{E}\Big[\frac{1}{N_n(X)} \mathbb{1}_{\{N_n(X) \ge 1\}}\Big] + \frac{1}{n}\mathbb{P}(N_n(X) = 0)$$

and

$$\mathbb{E}\Big[\sum_{i=1}^n W_{ni}(X)\|X_i - X\|^H\Big] \le h_n^H \mathbb{P}(N_n(X) \ge 1) + k^{H/2}\mathbb{P}(N_n(X) = 0)$$

because the distance to X for the points with non zero weight can be bounded from above by h_n if $N_n(X) \ge 1$ and by $\sqrt{k}$ otherwise (note that $\sqrt{k}$ is the diameter of $[0,1]^k$).
Next, we use the fact that, conditionally on $X = x$, $N_n(x)$ has a binomial distribution with parameters n and $p_n(x) = \mathbb{P}(X_1 \in B(x, h_n))$. This implies

$$\mathbb{E}\Big[\frac{1}{N_n(X)} \mathbb{1}_{\{N_n(X) \ge 1\}}\Big] \le \mathbb{E}\Big[\frac{2}{np_n(X)}\Big] \le \frac{2c_k}{nh_n^k}$$

where the first inequality follows from Györfi et al. (2002, Lemma 4.1) and the second one from Györfi et al. (2002, Equation 5.1) where the constant $c_k = k^{k/2}$ can be taken. Similarly,

$$\begin{aligned}
\mathbb{P}(N_n(X) = 0) = \mathbb{E}[(1 - p_n(X))^n] &\le \mathbb{E}[e^{-np_n(X)}] \\
&\le \big(\max_{u>0} ue^{-u}\big) \times \mathbb{E}\Big[\frac{1}{np_n(X)}\Big] \\
&\le \frac{c_k}{nh_n^k}.
\end{aligned}$$

In view of these different estimates, Equation (4.13) entails

$$\begin{aligned}\mathbb{E}\big[W_1(\hat{F}_{n,X}, F_X)\big] &\leq L\Big(h_n^H + k^{H/2}\frac{c_k}{nh_n^k}\Big) + M\left(\frac{(2+1/n)c_k}{nh_n^k}\right)^{1/2} \\ &\leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.\end{aligned}$$

□

Proof of Corollary 4.9. For the nearest neighbor weights (4.6), there are exactly κ_n non-vanishing weights with value $1/\kappa_n$ whence

$$\sum_{i=1}^{n} W_{ni}^2(X) = \frac{1}{\kappa_n}.$$

Furthermore, the κ_n nearest neighbors of X satisfy

$$\|X_{i:n}(X) - X\| \leq \|X_{\kappa_n:n}(X) - X\|, \quad i = 1, \ldots, \kappa_n.$$

In view of this, Equation (4.13) entails

$$\begin{aligned}\mathbb{E}\big[W_1(\hat{F}_{n,X}, F_X)\big] &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^H\big] + M\kappa_n^{-1/2} \\ &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^2\big]^{H/2} + M\kappa_n^{-1/2}\end{aligned}$$

where the last line relies on Jensen's inequality. We conclude thanks to Biau and Devroye (2015, Theorem 2.4) stating that

$$\mathbb{E}\left[\|X_{\kappa_n:n}(X) - X\|^2\right] \leq \begin{cases} 8(\kappa_n/n) & \text{if } k = 1, \\ \widetilde{c}_k(\kappa_n/n)^{2/k} & \text{if } k \geq 2. \end{cases}$$

□

Proof of Theorem 4.10 (lower bound). The proof of a lower bound for the minimax risk in Wasserstein distance is adapted from the proof of Proposition 3 in Pic et al. (2022, Appendix C) and we give only the main lines.
Consider the subclass of $\mathcal{D}(H, L, M)$ where Y is a binary variable with possible values 0 and B. Note that condition c) of Definition 4.7 is automatically satisfied if $B \leq 4M$. The conditional distribution of Y given $X = x$ is characterized by

$$p(x) = \mathbb{P}(Y = B \mid X = x)$$

and the Wasserstein distance by

$$W_1(F_x, F_{x'}) = B|p(x) - p(x')|,$$

so that property b) of Definition 4.7 is equivalent to

$$B|p(x) - p(x')| \leq L\|x - x'\|^H. \tag{4.19}$$

Similarly as in Pic et al. (2022, Lemma 1), one can show that a general prediction with values in $\mathbb{R}$ can always be improved (in terms of Wasserstein error) into a binary prediction with values in $\{0, B\}$. Indeed, for a given prediction $\hat{F}_{n,x}$, the binary prediction

$$\widetilde{F}_{n,x} = (1 - \widetilde{p}_n(x))\delta_0 + \widetilde{p}_n(x)\delta_B$$

with

$$\widetilde{p}_n(x) = \frac{1}{B}\int_0^B \big(1 - \hat{F}_{n,x}(z)\big)\mathrm{d}z$$

always satisfies

$$\mathbb{E}[W_1(\widetilde{F}_{n,X}, F_X)] \leq \mathbb{E}[W_1(\hat{F}_{n,X}, F_X)].$$

This simple remark implies that, when considering the minimax risk on the restriction of the class $\mathcal{D}(H, L, M)$ to binary distributions, we can focus on binary predictions. But for binary predictions,

$$\mathbb{E}[W_1(\widetilde{F}_{n,X}, F_X)] = B|\widetilde{p}_n(X) - p(X)|,$$

showing that the minimax rate of convergence for distributional regression in Wasserstein distance is equal to the minimax rate of convergence for estimating the regression function $\mathbb{E}[Y|X = x] = Bp(x)$ in absolute error under the regularity assumption (4.19) . According to Stone (1980), a lower bound for the minimax risk in L^1-norm is $n^{-H/(2H+k)}$ (in the first paper, we consider the Bernoulli regression model referred to as Model 1 Example 5 and the L^q distance with $q = 1$). □

Proof of Theorem 4.10 (upper bound). For the kernel method, Corollary 4.8 states that the expected Wasserstein error is upper bounded by

$$Lh_n^H + M\sqrt{(2 + 1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.$$

Minimizing the sum of the first two terms in the right-hand side with respect to h_n leads to $h_n \propto n^{1/(2H+1)}$ and implies that right-hand side is of order $n^{-H/(2H+k)}$ (the last term is negligible). This matches the minimax lower rate of convergence previously stated previously and proves that the optimal minimax risk is of order $n^{-H/(2H+k)}$.
For the nearest neighbor method, minimizing the upper bound for the expected Wasserstein error from Corollary 4.9 leads to

$$\kappa_n \propto \begin{cases} n^{H/(H+1)} & \text{if } k = 1 \\ n^{H/(H+k/2)} & \text{if } k \geq 2 \end{cases},$$

with a corresponding risk of order

$$\begin{cases} n^{-H/(2H+2)} & \text{if } k = 1 \\ n^{-H/(2H+k)} & \text{if } k \geq 2 \end{cases},$$

whence the nearest neighbor method reaches the optimal rate when $k \geq 2$. □

4.3 Proof of Proposition 4.12

Proof of Proposition 4.12. The first point follows from the fact that composition by a continuous application respects convergence in probability. Indeed, as the estimator $\hat{F}_{n,X}$ converges to F_X in probability for the Wasserstein distance W_p, $S(\hat{F}_{n,X})$ converges to $S(F_X)$ in probability.
In order to prove the consistency in $\mathrm{L}^{p/q}$, it is enough to prove furthermore the uniform integrability of $|S(\hat{F}_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$. With the convexity inequality of power functions as $p/q \geq 1$, Equation (4.14) entails

$$\begin{aligned} |S(\hat{F}_{n,X}) - S(F_X)|^{p/q} &\leq 2^{p/q-1}\left(|S(\hat{F}_{n,X})|^{p/q} + |S(F_X)|^{p/q}\right) \\ &\leq 2^{p/q-1}\left((aM_p^q(\hat{F}_{n,X}) + b)^{p/q} + (aM_p^q(F_X) + b)^{p/q}\right) \\ &\leq 2^{2(p/q-1)}\left(a^{p/q}M_p^p(\hat{F}_{n,X}) + a^{p/q}M_p^p(F_X) + 2b^{p/q}\right). \end{aligned}$$

This upper bound together with Equation (4.18) implies the uniform integrability of $|S(\hat{F}_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$, which concludes the proof. □

Chapitre 5

Characterization of translation invariant MMD on $\mathbb{R}^d$ and connections with Wasserstein distances

Ce dernier chapitre est constitué du préprint Modeste and Dombry (2022), soumis dans *Journal of Machine Learning Research*. Nous construisons des distances équivalentes aux distances de Wasserstein. Cette construction est faite à partie des RKHS, un outils utilisé en Machine Learning.

Abstract : Kernel mean embeddings and maximum mean discrepancies (MMD) associated with positive definite kernels are important tools in machine learning that allow to compare probability measures and sample distributions. Two kernels are said to be equivalent if their associated MMDs are equal. We characterize the equivalence of kernels in terms of their variogram and deduce that MMDs are in one to one correspondance with conditionally negative definite functions. As a consequence, we provide a full characterization of translation invariant MMDs on $\mathbb{R}^d$ that are parametrized by a spectral measure and a semi-definite symmetric matrix. Furthermore, we investigate the connections between translation invariant MMDs and Wasserstein distances on $\mathbb{R}^d$. We show in particular that convergence with respect to the MMD associated with the Energy Kernel of order $\alpha \in (0, 1)$ implies convergence with respect to the Wasserstein distance of order $\beta < \alpha$. We also provide examples of kernels metrizing the Wasserstein space of order $\alpha \geq 1$.

1 Introduction

Background. Many problems in statistics and machine learning require comparing several probability measures and/or sample distributions : goodness-of-fit testing compares a sample distribution to a reference distribution (Chwialkowski et al., 2016) ; two-sample testing compares two sample distributions (Gretton et al., 2012) ; independence testing compares a joint distribution to a product distribution (Gretton et al., 2005) ; generative model fitting compares the distributions of real and fake data (Dziugaite et al., 2015; Sutherland et al., 2017). The different methods proposed in these references all rely on the important notion of Minimum Mean Discrepancy (MMD).

MMDs are semi-metrics between probability measures and their definition relies on the theory of Reproducing Kernel Hilbert Spaces (RKHS) and Kernel Mean Embeddings (KME). Given a symmetric positive definite kernel k and its associated RKHS $\mathcal{H}_k$, the KME is a map $\mu \mapsto K(\mu)$ that assigns a function $K(\mu) \in \mathcal{H}_k$ to each signed measure μ in a suitable subspace $\mathcal{M}_k$ (defined in Equation (5.4) below). The corresponding MMD between two measures μ and ν is defined as the RKHS distance between their embeddings, i.e. $d_k(\mu, \nu) := \|K(\mu) - K(\nu)\|_{\mathcal{H}_k}$. When the KME is injective, in which case the kernel is called characteristic, the MMD defines a proper

distance that can be used to compare probability measures and/or sample distributions. Due to their theoretical tractability and computational efficiency, KMEs and MMDs are widely used in many areas of machine learning. We refer to Smola et al. (2007) for an overview on distribution Hilbert space embeddings and their applications in machine learning.

Related works. In the last decade, an important line of research has focused on theoretical properties of KMEs and MMDs. Sriperumbudur et al. (2010) and Sriperumbudur et al. (2011) consider conditions ensuring that a kernel is characteristic, meaning that the associated kernel mean embedding is injective. In the particular case of invariant kernels on $\mathbb{R}^d$, the question can be adressed thanks to Fourier analysis and the kernel is shown to be characteristic if and only if the spectral measure has a full support on $\mathbb{R}^d \setminus \{0\}$ (Sriperumbudur et al., 2010, Theorem 9). Already considered in the latter references, the question of whether MMD can metrize weak convergence of distributions has been fully addressed by Simon-Gabriel and Schölkopf (2018) and Simon-Gabriel et al. (2021). The main result is that, for a continuous kernel with RKHS included in the space of continuous functions vanishing at infinity, the MMD metrizes weak convergence if and only if the kernel is characteristic.
Although weak convergence is an important concept and a minimal requirement, this notion of convergence is very weak, as its name suggests. A stronger notion of convergence, which has turned out to be very useful and successful in machine learning, is the convergence in Wasserstein space. The Wasserstein distance is related to optimal transport (Villani, 2009) and has recently been considered in several learning algorithms (Frogner et al., 2015). One of the main question addressed in the present paper is whether a MMD can metrize the Wasserstein space. We show that the answer is positive and that the use of unbounded kernels is needed. In a slightly different perspective, Auricchio et al. (2020) and Vayer and Gribonval (2021) establish non-asymptotic inequalities relating MMD and Wasserstein distances.

Main contributions. Our main findings are the following :

- The notion of equivalent kernels is introduced (Definition 5.6) and characterized via the variogram (Proposition 5.8), showing that MMDs are in one-to-one correspondence with negative semi-definite functions.
- The class of translation invariant MMD on $\mathbb{R}^d$ is characterized by a spectral measure and a symmetric positive semi-definite matrix (Corollary 5.14). Extending the results of Sriperumbudur et al. (2010), we provide an explicit formula for the MMD in terms of Fourier transform (Proposition 5.18) and provide a necessary and sufficient condition for the kernel to be characteristic over probability measures (Proposition 5.21).
- Strong connections between Energy kernels and Wasserstein distances are established (Theorem 5.25). More precisely, for $\alpha \in (0, 1)$, we denote by d_α the MMD associated with the energy kernel of order α and by W_α the Wasserstein distance of order α ; we prove that convergence of probability measures with respect to W_α implies convergence with respect to d_β for all $0 < \beta \leq \alpha$ and, conversely, that convergence with respect to d_α implies convergence with respect to W_β for all $0 < \beta < \alpha$.
- We exhibit new families of kernels that metrize the Wasserstein spaces of order $\alpha \geq 1$ (Theorem 5.30).
- We provide non-asymptotic inequalities between W_α and d_α for tight subsets of probability measures (Proposition 5.33).

Potential applications. Although our focus here is mostly on theoretical properties, we believe that the present work advocates for further and possibly more applied research to connect MMD- and Wasserstein-based learning. Due to its implicit definition as the minimum of the transport cost, the computation of Wasserstein distances remains challenging, even if efficient algorithms have been designed and surrogate distances have been considered to reduce the computational burden (Kolouri et al., 2019; Bayraktar and Guo, 2021b). Interestingly, in the framework of

Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), both MMD and Wasserstein distances have been studied (Li et al., 2015; Arjovsky et al., 2017). Based on the relationships between Wasserstein distances and Energy Kernel MMDs discussed in this paper, it would for instance be interesting to compare Wasserstein-GAN and MMD-GAN based on the Energy Kernel.

Structure of the paper. Section 2 gathers our main results on translation invariant MMD. We first introduce some background on reproducing kernel Hilbert spaces, kernel mean embeddings and maximum mean discrepancies in Section 2.1. The notion of equivalent kernels and its characterization via variograms are the purpose of Section 2.2. Translation invariant MMDs and their properties are studied in Section 2.3. Section 3 focuses on the connections between MMDs and Wasserstein distances. Some background on Wasserstein spaces is presented in Section 3.1 and some preliminary results in Section 3.2. The relationships between MMDs associated with Energy Kernel of order $\alpha < 1$ and Wasserstein distances of order $\alpha < 1$ are investigated in Section 3.3. New families of kernels metrizing the Wasserstein spaces of order $\alpha \geq 1$ are studied in Section 3.4. Finally, some nonasymptotic inequalities relating MMDs and Wasserstein distances are established in Section 3.5. All the proofs are postponed to Section 4.

Notation. In Sections 2.1 and 2.2, $(\mathcal{X}, \mathcal{B})$ denotes a measurable space and $\mathcal{M}$ (resp. $\mathcal{P}$) the sets of signed measures (resp. probability measures) on $(\mathcal{X}, \mathcal{B})$. The total variation measure of a signed measure $\mu \in \mathcal{M}$ is denoted by $|\mu|$. In the rest of the paper, we take $\mathcal{X} = \mathbb{R}^d$ endowed with its Borel sigma-field and $\mathcal{M}$ (resp. $\mathcal{P}$) denotes the space of Borel signed measures (resp. probability measures) on $\mathbb{R}^d$. We equip $\mathbb{R}^d$ with its canonical Euclidean structure and we write $\|x\|$ and $x \cdot y$ respectively for the norm of x and the inner product between x and y. For $\alpha > 0$, we define

$$\mathcal{M}^\alpha = \left\{ \mu \in \mathcal{M} \colon \int_{\mathbb{R}^d} \|x\|^\alpha \, |\mu|(\mathrm{d}x) < \infty \right\} \quad \text{and} \quad \mathcal{P}^\alpha = \mathcal{M}^\alpha \cap \mathcal{P} \tag{5.1}$$

the set of signed measures (resp. probability measures) with finite α-moment.

2 Kernel Mean Embeddings and Maximum Mean Discrepancy

2.1 Hilbert space embedding of measures

We present some basic elements of the theory of Reproducing Kernel Hilbert Space (RKHS), Kernel Mean Embeddings (KME) and Maximum Mean Discrepancy (MMD). For more details, the reader could refer to Berlinet and Thomas-Agnan (2004), Smola et al. (2007) or Steinwart and Christmann (2008, Section 4).

Reproducing Kernel Hilbert Spaces (RKHS). Let $\mathcal{X}$ be an arbitrary space and $\mathcal{F}(\mathcal{X}, \mathbb{R})$ denote the space of real valued function on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if it is symmetric and positive definite. The latter conditions means that

$$\sum_{1 \leq i,j \leq n} a_i a_j k(x_i, x_j) \geq 0, \quad \text{for all } n \geq 1,\ x_1, \ldots, x_n \in \mathcal{X},\ a_1, \ldots, a_n \in \mathbb{R}.$$

Definition 5.1. *A Hilbert space $\mathcal{H} \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ is called an RKHS if, for all $x \in \mathcal{X}$, the evaluation map $f \mapsto f(x)$ is continuous.*

By the Riesz representation theorem, there exists, for all $x \in \mathcal{X}$, a unique representer $K(x) \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H},\ f(x) = \langle f, K(x) \rangle.$$

Then, the function $k(x, y) = \langle K(x), K(y) \rangle$ is a kernel and is called the *reproducing kernel* of $\mathcal{H}$ because of the following *reproducing property* : for all $x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$ and

$$\forall f \in \mathcal{H},\ f(x) = \langle f, k(x, \cdot) \rangle. \tag{5.2}$$

In particular, we have $K(x) = k(x, \cdot)$. The reproducing kernel characterizes the RKHS and conversely, according to Aronszajn's theorem, any kernel defines a unique RKHS.

Theorem 5.2 (Aronszajn's theorem)**.** *For any kernel k on $\mathcal{X} \times \mathcal{X}$, there exists an unique RKHS, noted $\mathcal{H}_k$, with reproducing kernel k.*

Kernel Mean Embedding (KME). We assume that $(\mathcal{X}, \mathcal{B})$ is a measurable space and the kernel k is measurable on $\mathcal{X} \times \mathcal{X}$. The space of signed finite measures (resp. probability measures) μ on $(\mathcal{X}, \mathcal{B})$ is denoted by $\mathcal{M}$ (resp. $\mathcal{P}$) and the total variation measure of μ by $|\mu|$. The reproducing kernel property (5.2) readily implies that for any finite discrete measure $\mu = \sum_{i=1}^n a_i \delta_{x_i}$, the function $K(\mu) = \sum_{i=1}^n a_i K(x_i) \in \mathcal{H}_k$ satisfies

$$\forall f \in \mathcal{H}_k, \ \langle f, K(\mu) \rangle = \int_{\mathcal{X}} f(x) \, \mu(\mathrm{d}x). \tag{5.3}$$

The KME extends this property to the class of measures

$$\mathcal{M}_k = \Big\{ \mu \in \mathcal{M} \colon \int_{\mathcal{X}} \sqrt{k(x,x)} \, |\mu|(\mathrm{d}x) < +\infty \Big\}. \tag{5.4}$$

The following proposition defines the KME on $\mathcal{M}_k$. See, e.g., Steinwart and Christmann (2008, Theorem 4.26) for the proof – note that, the kernel k being measurable, all functions $f \in \mathcal{H}_k$ are measurable (Steinwart and Christmann, 2008, Lemma 4.24).

Proposition 5.3. *For all $\mu \in \mathcal{M}_k$, $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$ and there exists an unique $K(\mu) \in \mathcal{H}_k$ satisfying Equation (5.3).*

The map $K : \mathcal{M}_k \to \mathcal{H}_k$ is the KME associated with k; the vector $K(\mu)$ represents the measure μ in the same way as the vector $K(x)$ represents the point x (identified with the Dirac measure δ_x). One of the main argument in the proof of Proposition 5.3 is the continuity of the linear form $f \in \mathcal{H}_k \mapsto \int f \, \mathrm{d}\mu$ for all $\mu \in \mathcal{M}_k$. It follows from the inequality

$$|f(x)| = |\langle f, k(x, .) \rangle_{\mathcal{H}_k}| \leq \|f\|_{\mathcal{H}_k} \|k(x, \cdot)\|_{\mathcal{H}_k} = \|f\|_{\mathcal{H}_k} \sqrt{k(x,x)}$$

which entails

$$\Big| \int_{\mathcal{X}} f \, \mathrm{d}\mu \Big| \leq \|f\|_{\mathcal{H}_k} \int_{\mathcal{X}} \sqrt{k(x,x)} \, |\mu|(\mathrm{d}x).$$

Remark 5.4. One can find in the literature a different construction of the KME on an extended space of measures in terms of the Pettis integral (Diestel and Uhl, 1977, Section 2.3). Using the Closed Graph Theorem for Banach spaces, Steinwart and Ziegel (2021, Section 2) proves the continuity of the linear form $f \in \mathcal{H}_k \mapsto \int_{\mathcal{X}} f \mathrm{d}\mu$ as soon as $\mathcal{H}_k \subset \mathcal{L}^1(\mu)$. The KME is then defined on the subspace $\mathcal{M}'_k = \{\mu \in \mathcal{M} \colon \mathcal{H}_k \subset \mathcal{L}^1(\mu)\}$. This subspace always contains $\mathcal{M}_k$ and the two constructions of the KME coincide there.

Maximum Mean Discrepancy (MMD). To compare two measures in $\mathcal{M}_k$, we compare their images in $\mathcal{H}_k$ under the KME : the MMD is defined by

$$d_k(\mu, \nu) = \|K(\mu) - K(\nu)\|_{\mathcal{H}_k}, \quad \mu, \nu \in \mathcal{M}_k.$$

The reproducing kernel property (5.3) - applied twice - implies

$$\begin{aligned} d_k^2(\mu, \nu) &= \langle K(\mu - \nu), K(\mu - \nu) \rangle_{\mathcal{H}_k} \\ &= \int_{\mathcal{X} \times \mathcal{X}} k(x, y) \, (\mu - \nu) \otimes (\mu - \nu)(\mathrm{d}x \mathrm{d}y). \end{aligned} \tag{5.5}$$

For sample distributions $\mu_n = n^{-1}\sum_{k=1}^n \delta_{x_k}$ and $\nu_m = m^{-1}\sum_{l=1}^m \delta_{y_l}$, the MMD reduces to

$$d_k^2(\mu_n, \nu_m) = n^{-2} \sum_{1\le k,l\le n} k(x_k, x_l) + m^{-2} \sum_{1\le k,l\le m} k(y_k, y_l) - 2n^{-1}m^{-1} \sum_{1\le k\le n} \sum_{1\le l\le m} k(x_k, y_l)$$

and is easily computed (for sample of reasonable size). Furthermore, using the dual representation of the Hilbert norm in $\mathcal{H}_k$, the MMD can also be expressed as

$$d_k(\mu, \nu) = \sup_{\|f\|_{\mathcal{H}_k}\le 1} \left| \int_{\mathcal{X}} f \,\mathrm{d}\mu - \int_{\mathcal{X}} f \,\mathrm{d}\nu \right|.$$

This form corresponds to an Integral Probability Metric (Müller, 1997) with test functions belonging to the unit ball of the RKHS.

Example 5.5. When $\mathcal{X} = \mathbb{R}^d$, the Gaussian kernel is the most popular one in machine learning and is defined by

$$k(x, y) = \exp(-\|x - y\|_2^2/2), \quad x, y \in \mathbb{R}^d.$$

This kernel being bounded, we have $\mathcal{M}_k = \mathcal{M}$ and, using Fourier theory, the MMD can be rewritten as

$$d_k^2(\mu, \nu) = \int_{\mathbb{R}^d} |\hat{\mu}(t) - \hat{\nu}(t)|^2 \, \varphi(t) \,\mathrm{d}t,$$

where φ denotes the multivariate standard Gaussian density on $\mathbb{R}^d$ and $\hat{\mu}$ (resp. $\hat{\nu}$) the characteristic function of μ (resp. ν). By Theorem 7 of Simon-Gabriel et al. (2021), the MMD metrizes weak convergence on $\mathcal{P}$.

2.2 Equivalent kernels and variograms

Given different measurable kernels on $\mathcal{X} \times \mathcal{X}$, one can wonder in which case the associated MMDs are equal. This gives rise to the following definition.

Definition 5.6. *The measurable kernels k_1 and k_2 on $\mathcal{X} \times \mathcal{X}$ are said to be equivalent if*

$$\mathcal{M}_{k_1} = \mathcal{M}_{k_2} \quad \text{and} \quad d_{k_1}(\mu, \nu) = d_{k_2}(\mu, \nu) \qquad \text{for all } \mu, \nu \in \mathcal{M}_{k_1} \cap \mathcal{P}.$$

Let us stress that, in this definition, the equality of MMDs is required for probability measures only.

Our characterization of equivalent kernel relies on the notion of variogram that we now define.

Definition 5.7. *We call the variogram associated with a kernel k the function*

$$\rho(x, y) = \frac{1}{2}k(x, x) + \frac{1}{2}k(y, y) - k(x, y), \quad x, y \in \mathcal{X}.$$

Clearly, the variogram ρ is a symmetric function on $\mathcal{X} \times \mathcal{X}$ and vanishes on the diagonal, i.e.

$$\rho(x, x) = 0 \quad \text{for all } x \in \mathcal{X}.$$

Furthermore, according to Berg et al. (1984, Lemma 2.1 p.74), the variogram is a conditionally negative definite function on $\mathcal{X} \times \mathcal{X}$, meaning that

$$\sum_{1\le i,j\le n} a_i a_j \rho(x_i, x_j) \le 0$$

for all $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$. See Chapter 3 in Berg et al. (1984) for more details on the strong relationships between positive definite and negative definite functions.

In Sejdinovic et al. (2013), the authors define the equivalence between two kernels as the fact of defining the same variogram. In fact, these two definitions are equivalent (see Proposition 20 and Theorem 22). We give another proof of this fact based on the construction of a normalized kernel (Proposition 5.9).

Proposition 5.8. *Two measurable kernels are equivalent if and only if they have the same variogram.*

In order to have a form of uniqueness, we consider the notion of a normalized kernel. Fix an arbitrary origin $o \in \mathcal{X}$. A kernel k is said to be normalized (with origin o) if

$$k(x, o) = k(o, x) = 0 \quad \text{for all } x \in \mathcal{X}.$$

Proposition 5.9. *For any kernel k on $\mathcal{X} \times \mathcal{X}$, there exists a unique kernel k_0 which is equivalent to k and normalized (with origin o). It is given by*

$$k_0(x, y) = k(x, y) - k(x, o) - k(o, y) + k(o, o).$$

Denoting by ρ the common variogram of k and k_0, one can easily check that k_0 can be written as

$$k_0(x, y) = \rho(x, o) + \rho(o, y) - \rho(x, y).$$

Remark 5.10. The term *variogram* comes from the theory of stochastic processes and geostatistics (Cressie, 1993). Let $(B(x))_{x\in\mathcal{X}}$ be a square integrable stochastic process on $\mathcal{X}$. The covariance function is a symmetric and positive definite function on $\mathcal{X} \times \mathcal{X}$, that is

$$k(x, y) = \mathrm{Cov}(B(x), B(y))$$

is a kernel. The associated variogram

$$\begin{aligned}\rho(x, y) &= \frac{1}{2}k(x, x) + \frac{1}{2}k(y, y) - k(x, y)\\ &= \frac{1}{2}\mathrm{Var}(B(y) - B(x))\end{aligned}$$

corresponds to half the variance of the increment $B(y)-B(x)$. Given an origin $o \in \mathcal{X}$, the process $(B(x) - B(o))_{x\in\mathcal{X}}$ of increments at the origin has covariance

$$\begin{aligned}k_0(x, y) &= \mathrm{Cov}(B(x) - B(o), B(y) - B(o))\\ &= k(x, y) - k(x, o) - k(o, y) + k(o, o),\end{aligned}$$

which is the unique normalized kernel with variogram ρ. We focus next on the class of Gaussian processes. If the process B is centered and Gaussian, then its distribution is fully characterized by its covariance function. It follows that, given an origin o and a variogram ρ, there exists a (unique in distribution) centered Gaussian process $B = (B(x))_{x\in\mathcal{X}}$ such that

$$\mathrm{Var}(B(y) - B(x)) = 2\rho(x, y) \quad \text{and} \quad B(o) = 0 \text{ a.s.}$$

The process B is called the Gaussian process with variogram ρ and origin o.

2.3 Translation invariant MMD on $\mathbb{R}^d$

In the rest of the paper, we consider $\mathcal{X} = \mathbb{R}^d$ endowed with its Borel sigma-field. We study now translation invariant MMDs as in the following definition. For $h \in \mathbb{R}^d$, we note $\tau_h : \mathbb{R}^d \to \mathbb{R}^d$ the translation defined by $\tau_h(x) = x + h$.

Definition 5.11. *The MMD associated with a kernel k on $\mathbb{R}^d \times \mathbb{R}^d$ is said to be translation invariant if, for all $h \in \mathbb{R}^d$, $\mu \in \mathcal{M}_k$ implies $\mu \circ \tau_h^{-1} \in \mathcal{M}_k$ and*

$$d_k(\mu \circ \tau_h^{-1}, \nu \circ \tau_h^{-1}) = d_k(\mu, \nu) \quad \textit{for all } \mu, \nu \in \mathcal{M}_k. \tag{5.6}$$

Clearly, if the kernel k is translation invariant, i.e. satisfies

$$k(x+h, y+h) = k(x, y), \quad \text{for all } x, y, h \in \mathbb{R}^d,$$

then the associated MMD is invariant. Such kernels are of the form $k(x, y) = \psi(x - y)$ with ψ a positive definite function and were studied by Sriperumbudur et al. (2010, Section 3.2). Note that a translation invariant kernel is always bounded since

$$|k(x, y)| \leq \sqrt{k(x, x)}\sqrt{k(y, y)} = \psi(0).$$

Interestingly, the class of translation invariant MMDs is much larger and is fully characterized in the next theorem. A function $\gamma : \mathbb{R}^d \to \mathbb{R}$ is said to be conditionally negative definite if

$$\sum_{i=1}^{n} a_i a_j \gamma(x_i - x_j) \leq 0$$

for all $x_1, \ldots, x_n \in \mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$.

Theorem 5.12. *The MMD associated with the kernel k is translation invariant if and only if there exists a negative definite function $\gamma : \mathbb{R}^d \to [0, \infty)$ such that the variogram ρ associated with k satisfies $\rho(x, y) = \gamma(y - x)$.*
Conversely, for all negative definite function $\gamma : \mathbb{R}^d \to [0, \infty)$ such that $\gamma(0) = 0$, the MMD associated with the normalized kernel $k_0(x, y) = \gamma(x) + \gamma(y) - \gamma(y - x)$ is translation invariant and its variogram is $\rho(x, y) = \gamma(y - x)$.

In other words, Theorem 5.12 establishes a one-to-one correspondence between translation invariant MMDs and negative definite functions.

Remark 5.13. As a continuation of Remark 5.10 relating kernels, variograms and stochastic processes, one can relate translation invariant MMDs with stationary increment processes. A process $(B(x))_{x \in \mathbb{R}^d}$ is said to have stationary increments if for any $x_0, \ldots, x_n$ and $h \in \mathbb{R}^d$, we have

$$(B(x_i) - B(x_0))_{1 \leq i \leq n} \overset{d}{=} (B(x_i + h) - B(x_0 + h))_{1 \leq i \leq n},$$

where $\overset{d}{=}$ stands for equality in distribution. We can reformulate Theorem 5.12 as follows : let k be a kernel on $\mathbb{R}^d \times \mathbb{R}^d$ and ρ the associated variogram ; then the MMD associated with k is translation invariant if and only if the Gaussian process B with origin 0 and variogram ρ has stationary increments.

Using the previous remark and the characterization of stationary increment Gaussian processes by Yaglom and Silverman (1962) (see Formula (3.59) in Section 3.18), we can characterize all normalized kernels associated with a translation invariant MMD.

Corollary 5.14. *Let k be a normalized (with origin 0) and continuous kernel on $\mathbb{R}^d \times \mathbb{R}^d$. If the MMD associated to k is translation invariant, then there exists a symmetric Borel measure Λ, i.e.*

$$\forall A \in \mathcal{B}(\mathbb{R}^d), \ \Lambda(A) = \Lambda(-A), \tag{5.7}$$

on $\mathbb{R}^d \setminus \{0\}$ satisfying

$$\int_{\mathbb{R}^d} \left(\|\xi\|^2 \wedge 1\right) \Lambda(\mathrm{d}\xi) < \infty \tag{5.8}$$

and a $d \times d$ symmetric positive semi-definite matrix Σ such that

$$k(x, y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix \cdot \xi}\right)\left(1 - \mathrm{e}^{-iy \cdot \xi}\right) \Lambda(\mathrm{d}\xi) + x^T \Sigma y. \tag{5.9}$$

Conversely, for any such Λ and Σ, the kernel k defined by (5.9) is continuous on $\mathbb{R}^d \times \mathbb{R}^d$, normalized, and the associated MMD is translation invariant.

Note that the integrability condition (5.8) ensures that the integral in Equation (5.9) is well-defined because

$$\left|\left(1-\mathrm{e}^{ix\cdot\xi}\right)\left(1-\mathrm{e}^{-iy\cdot\xi}\right)\right| \leq (\|x\|\|y\|\|\xi\|^2) \wedge 4.$$

The symmetry condition (5.7) implies that the kernel is real-valued and given by

$$k(x,y) = \int_{\mathbb{R}^d} \big(1-\cos(x\cdot\xi)-\cos(y\cdot\xi)+\cos((x-y)\cdot\xi)\big)\Lambda(\mathrm{d}\xi) + x^T\Sigma y. \tag{5.10}$$

Clearly, in the case when $\Sigma = 0$ and Λ is finite, the kernel k is equivalent to

$$\widetilde{k}(x,y) = \int_{\mathbb{R}^d} \mathrm{e}^{i(x-y)\cdot\xi}\,\Lambda(\mathrm{d}\xi).$$

This class of bounded translation invariant kernels is studied in Sriperumbudur et al. (2010, Section 3.2). Most of the available literature on KME and MMD focuses on bounded kernels; the following lemma characterizes the boundedness of k in terms of Λ and Σ.

Lemma 5.15. *Let k be the kernel defined by (5.9). The following statements are equivalent :*

i) k is bounded on $\mathbb{R}^d \times \mathbb{R}^d$;

ii) Λ is a finite measure and $\Sigma = 0$.

We next provide examples of translation invariant MMDs associated with unbounded kernels, the so-called Energy Kernels, that will be the focus of Section 3.3.

Example 5.16. Brownian motion is a classical stationary increment process. In dimension 1, its covariance function $k(x,y) = \min(x,y)$ for $x, y \geq 0$ can be rewritten as

$$k(x,y) = \frac{1}{2}(|x| + |y| - |x-y|).$$

The Energy Kernels can be seen as an extension of this formula. Let $H \in (0,1)$ and define, for $x, y \in \mathbb{R}^d$,

$$k_H(x,y) = \|x\|^{2H} + \|y\|^{2H} - \|x-y\|^{2H}. \tag{5.11}$$

This kernel corresponds to the covariance of so-called Fractional Brownian Motion, see Herbin and Merzbach (2007) or Cohen and Istas (2013, Section 3). It is a well-studied family of kernels in statistics and is connected with the α-distance correlation for independence tests (Székely and Rizzo, 2009, Section 4). Lemma 1 in Szekely (2003) gives us the spectral representation of these kernels, for $H \in (0,1)$, $x \in \mathbb{R}^d$,

$$\|x\|_2^{2H} = \frac{1}{C(d,2H)} \int_{\mathbb{R}^d} \frac{1-\cos(\xi\cdot x)}{\|\xi\|^{d+2H}}\,\mathrm{d}\xi,$$

where $C(d,2H)$ is a constant depending only on d and H. Then by a direct computation with Equation (5.10),

$$k_H(x,y) = \frac{1}{C(d,2H)_{\mathbb{R}^d}} \int_{\mathbb{R}^d} \frac{\left(1-\mathrm{e}^{ix\cdot\xi}\right)\left(1-\mathrm{e}^{-iy\cdot\xi}\right)}{\|\xi\|^{d+2H}}\,\mathrm{d}\xi. \tag{5.12}$$

This shows that the Energy Kernel corresponds to the spectral measure

$$\Lambda(\mathrm{d}\xi) = \frac{1}{C(d,2H)}\|\xi\|^{-d-2H}\,\mathrm{d}\xi.$$

We next discuss the domain of definition $\mathcal{M}_k$ of the KME associated with k and the form of the corresponding MMD d_k. Since the kernel k decomposes into two terms

$$k_\Lambda(x,y) = \int_{\mathbb{R}^d} \left(1-\mathrm{e}^{ix\cdot\xi}\right)\left(1-\mathrm{e}^{-iy\cdot\xi}\right)\Lambda(\mathrm{d}\xi) \tag{5.13}$$

$$k_\Sigma(x,y) = x^T\Sigma y, \tag{5.14}$$

the following lemma will be useful. It is still present in the literature (Steinwart and Ziegel, 2021, Lemma 3.3).

Lemma 5.17. *Let k_1, k_2 be two kernels and $k = k_1 + k_2$. Then $\mathcal{M}_k = \mathcal{M}_{k_1} \cap \mathcal{M}_{k_2}$ and*

$$d_k^2(\mu,\nu) = d_{k_1}^2(\mu,\nu) + d_{k_2}^2(\mu,\nu), \quad \text{for all } \mu,\nu \in \mathcal{M}_k.$$

Lemma 5.17 suggests that one can study k_Λ and k_Σ separately. For the sake of readability, we use the short notation $\mathcal{M}_\Lambda$ and d_Λ (resp. $\mathcal{M}_\Sigma$ and d_Σ) instead of $\mathcal{M}_{k_\Lambda}$ and d_{k_Λ} (resp. $\mathcal{M}_{k_\Sigma}$ and d_{k_Σ}). Recall the definition (5.1) of the set $\mathcal{M}^\alpha$ of finite signed measures with a finite absolute moment of order $\alpha > 0$.

Proposition 5.18. *Let k_Λ and k_Σ be the kernels defined by Equations (5.13) and (5.14) respectively.*

- *If $\alpha > 0$ is such that $\int_{\mathbb{R}^d}(\|\xi\|^\alpha \wedge 1)\,\Lambda(\mathrm{d}\xi) < \infty$, then $\mathcal{M}^{\alpha/2} \subset \mathcal{M}_\Lambda$; in particular, Equation (5.8) implies that we always have $\mathcal{M}_1 \subset \mathcal{M}_\Lambda$. For $\mu,\nu \in \mathcal{M}_\Lambda$,*

$$d_\Lambda^2(\mu,\nu) = \int_{\mathbb{R}^d} \left|\hat\mu(\xi) - \hat\nu(\xi) - \mu(\mathbb{R}^d) + \nu(\mathbb{R}^d)\right|^2 \Lambda(\mathrm{d}\xi),$$

  where $\hat\mu(\xi) = \int_{\mathbb{R}^d} e^{i\xi\cdot x}\,\mu(\mathrm{d}x)$ (resp. $\hat\nu$) denotes the characteristic function of μ (resp. ν).
- *The space $\mathcal{M}_\Sigma$ is characterized by*

$$\mathcal{M}_\Sigma = \Big\{\mu \in \mathcal{M}\colon \int_{\mathbb{R}^d} |e_j \cdot x|\,|\mu|(\mathrm{d}x) < \infty \quad \text{for all } 1 \le j \le r\Big\},$$

  where r denotes the rank of Σ and $(e_1,\dots,e_r)$ an orthonormal system of eigenvectors associated with the positive eigenvalues $\lambda_1 \le \dots \le \lambda_r > 0$. For $\mu,\nu \in \mathcal{M}_\Sigma$,

$$d_\Sigma^2(\mu,\nu) = \sum_{j=1}^r \lambda_j \Big| \int_{\mathbb{R}^d} (e_j \cdot x)\,\mu(\mathrm{d}x) - \int_{\mathbb{R}^d} (e_j \cdot x)\,\nu(\mathrm{d}x)\Big|^2.$$

Remark 5.19. The following special cases are important :

1. If Λ is finite, then $\mathcal{M}_\Lambda = \mathcal{M}$; this corresponds exactly to the case when the kernel k_Λ is bounded and this case has been studied in Sriperumbudur et al. (2010, Section 3.2).
2. For $\mu,\nu \in \mathcal{M}_\Lambda$ with the same total mass, in particular for probability measures,

$$d_\Lambda^2(\mu,\nu) = \int_{\mathbb{R}^d} \left|\hat\mu(\xi) - \hat\nu(\xi)\right|^2 \Lambda(\mathrm{d}\xi) = \|\hat\mu - \hat\nu\|_{L^2(\Lambda)}^2.$$

   The MMD is equal to the norm in $L^2(\Lambda)$ distance between characteristic functions and the spectral measure Λ puts more or less weight to the different frequencies in the spectral domain.
3. If Σ is strictly positive definite, then $\mathcal{M}_\Sigma = \mathcal{M}_1$ and, for $\mu,\nu \in \mathcal{M}_1$,

$$d_\Sigma^2(\mu,\nu) = \|e(\mu) - e(\nu)\|_\Sigma^2$$

   where $e(\mu) = \int_{\mathbb{R}^d} x\,\mu(\mathrm{d}x)$ is the expectation of μ and $\|x\|_\Sigma^2 = x^T\Sigma x$ the squared norm associated with Σ. This quadratic kernel has been considered in Sriperumbudur et al. (2010, Example 2).

Example 5.20. As a continuation of Example 5.16, consider the Energy Kernel with index $H \in (0,1)$ defined in Equation (5.11). We have $\Sigma = 0$ and $\Lambda(\mathrm{d}\xi) = C(d,2H)^{-1}\|\xi\|^{-d-2H}\,\mathrm{d}\xi$. The equality $k(x,x) = 2\|x\|^H$ implies $\mathcal{M}_\Lambda = \mathcal{M}_H$. For probability measures $\mu,\nu \in \mathcal{M}_H \cap \mathcal{P}$,

$$d_H^2(\mu,\nu) = \frac{1}{C(d,2H)} \int_{\mathbb{R}^d} \frac{|\hat\mu(\xi) - \hat\nu(\xi)|^2}{\|\xi\|^{d+2H}}\,\mathrm{d}\xi. \tag{5.15}$$

We finally focus on conditions ensuring that the kernel k is characteristic over probability measures, meaning that d_k defines a proper distance (and not only a semi-metric) on $\mathcal{M}_k \cap \mathcal{P}$. This happens exactly when the KME $K : \mathcal{M}_k \cap \mathcal{P} \to \mathcal{H}_k$ is injective. The following Theorem generalizes Theorem 9 in Sriperumbudur et al. (2010) which considers bounded kernels only. Proposition 5.28 states a similar result and we will prove only this latter one.

Proposition 5.21. *The MMD d_k is a distance on $\mathcal{M}_k \cap \mathcal{P}$ if and only if* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$.

Note that the kernel k is not characteristic on $\mathcal{M}_k$ – i.e. the KME is not injective on $\mathcal{M}_k$ – because $d_k^2(\mu, \mu + \alpha\delta_0) = 0$ for all $\mu \in \mathcal{M}_k$ and $\alpha \in \mathbb{R}$.

3 Metrizing the Wasserstein space with MMD

The MMD associated with a characteristic kernel defines a distance on the space of probability measures. Understanding the notion of convergence, or equivalently the topology, associated with this distance is an important question which has been investigated in particular by Sriperumbudur et al. (2010) and Simon-Gabriel and Schölkopf (2018). Most of the results in this line of research show the equivalence between weak convergence and convergence in MMD for bounded kernels. In this section, we investigate whether convergence in Wasserstein spaces can be metrized by an MMD.

3.1 Background on Wasserstein spaces

We first provide the necessary background on Wasserstein spaces. For the purpose of this paper, the underlying space will always be $\mathbb{R}^d$ and we therefore restrict our presentation to this case. More general results as well as proofs can be found in (Villani, 2003, Section 7).
Recall from Equation (5.1) the notation $\mathcal{M}^\alpha$ (resp. $\mathcal{P}^\alpha$) for the set of signed measures (resp. probability measures) with a finite absolute moment of order $\alpha > 0$. Given two probability measures μ, ν on $\mathbb{R}^d$, we denote by $\Gamma(\mu, \nu)$ the set of couplings between μ and ν, that is the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that

$$\gamma(B \times \mathbb{R}) = \mu(B) \quad \text{and} \quad \gamma(\mathbb{R} \times B) = \nu(B),$$

for all Borel set $B \subset \mathbb{R}^d$. The Wasserstein distance of order α is defined, for $\alpha \geq 1$, by

$$W_\alpha(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \|x - y\|^\alpha \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/\alpha}, \quad \mu, \nu \in \mathcal{P}^\alpha.$$

For $\alpha \in (0, 1)$, it is defined by

$$W_\alpha(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d} \|x - y\|^\alpha \, \gamma(\mathrm{d}x, \mathrm{d}y).$$

For all $\alpha > 0$, the Wasserstein space $(\mathcal{P}^\alpha, W_\alpha)$ is a complete and separable metric space. The case $\alpha < 1$ is somewhat less usual and we stress that the Wasserstein distance W_α is then equal to the Wasserstein distance of order 1 on the metric space $(\mathbb{R}^d, \rho_\alpha)$ with the alternative distance $\rho_\alpha(x, y) = \|x - y\|^\alpha$. To see that this application defines a distance, it is sufficient to note that $a \mapsto (a + b)^\alpha - a^\alpha - b^\alpha$ for $b \geq 0$ is non increasing on $(0, +\infty)$.
An important result in the theory of Wasserstein spaces is the Kantorovitch-Rubinstein duality which states that

$$W_1(\mu, \nu) = \sup \left\{ \int_{\mathbb{R}^d} \varphi \, \mathrm{d}(\mu - \nu) \; : \; \varphi : \mathbb{R}^d \to \mathbb{R} \;\; 1\text{-Lipschitz} \right\}.$$

In the case $\alpha > 1$, a more involved duality theory, called Kantorovitch duality, holds but it will not be needed here. In the case $\alpha < 1$, we have

$$W_\alpha(\mu,\nu) = \sup\Big\{ \int_{\mathbb{R}^d} \varphi \,\mathrm{d}(\mu-\nu) \;:\; \varphi : \mathbb{R}^d \to \mathbb{R} \;\; (\alpha,1)\text{-Hölder}\Big\}, \tag{5.16}$$

where a function φ is said to be $(\alpha,1)$-Hölder if $|\varphi(x)-\varphi(y)| \le \|x-y\|^\alpha$ for all $x, y \in \mathbb{R}^d$. Note that the set of $(\alpha,1)$-Hölder functions is equal to the set of 1-Lipschitz functions on $\mathbb{R}^d$ equipped with the distance ρ_α, so that the duality in the case $\alpha < 1$ is a straightforward consequence from the Kantorovitch-Rubinstein duality.
We finally discuss the notion of convergence in Wasserstein spaces. Let $\alpha > 0$ and $(\mu_n)_{n\ge1}, \mu \in \mathcal{P}^\alpha$. According to (Villani, 2003, Theorem 7.12), the following statements are equivalent :

i) $W_\alpha(\mu_n,\mu) \to 0$;

ii) the sequence $(\mu_n)_{n\ge1}$ converges weakly to μ and

$$\int_{\mathbb{R}^d} \|x\|^\alpha \,\mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \|x\|^\alpha \,\mu(\mathrm{d}x);$$

iii) for each continuous function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $|\varphi(x)| = O_{x\to\infty}(\|x\|^\alpha)$, we have

$$\int_{\mathbb{R}^d} \varphi(x)\,\mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} \varphi(x)\,\mu(\mathrm{d}x).$$

Note that the convergence in $\mathcal{P}_\alpha$ is stronger for larger values of α. More precisely, $\beta < \alpha$ implies $\mathcal{P}^\alpha \subset \mathcal{P}^\beta$, and for all $(\mu_n)_{n\ge1}, \mu \in \mathcal{P}^\alpha$,

$$W_\alpha(\mu_n,\mu) \to 0 \quad \text{implies} \quad W_\beta(\mu_n,\mu) \to 0. \tag{5.17}$$

Equivalently, the injection $(\mathcal{P}^\alpha, W_\alpha) \to (\mathcal{P}^\beta, W_\beta)$ is continuous.

3.2 Some negative answers

Our main question is whether an MMD can metrize the Wasserstein distance according to the following definition.

Definition 5.22. *Let k be a kernel on $\mathbb{R}^d$ and $\alpha > 0$. We say that the MMD d_k associated with the kernel k metrizes the Wasserstein space of order α if $\mathcal{P} \cap \mathcal{M}_k = \mathcal{P}_\alpha$ and, for all $(\mu_n)_{n\ge1}, \mu \in \mathcal{P}_\alpha$,*

$$d_k(\mu_n,\mu) \to 0 \quad \textit{if and only if} \quad W_\alpha(\mu_n,\mu) \to 0.$$

The following proposition is elementary but it emphasizes the need for unbounded kernels.

Proposition 5.23. *Assume the kernel k metrizes the Wasserstein space of order $\alpha > 0$. Then k is unbounded on $\mathbb{R}^d \times \mathbb{R}^d$.*

Another negative result focuses on translation invariant MMDs associated with kernels of the form (5.9). According to Proposition 5.18, such kernels satisfy $\mathcal{P}^1 \subset \mathcal{M}_k$ so that it is natural to ask whether d_k can metrize the Wasserstein space of order 1.

Proposition 5.24. *There exists no kernel k of the form (5.9) such that d_k metrizes the Wasserstein space of order 1.*

More generally, as a straightforward adaptation of the proof of Proposition 5.24 shows, there exists no translation invariant MMD metrizing the Wasserstein space of order $\alpha \ge 1$.

3.3 Energy kernels and Wasserstein spaces of order $\alpha < 1$

We focus in this section on the special class of Energy Kernels, see Example 5.16. We recall that, for $\alpha \in (0,1)$, the Energy Kernel is defined by

$$k_\alpha(x,y) = \|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x-y\|^{2\alpha}, \quad x,y \in \mathbb{R}^d,$$

and that the associated MMD is defined on $\mathcal{M}^\alpha$ and translation invariant. For clarity of notation, we denote by $d_\alpha = d_{k_\alpha}$ the MMD associated with k_α. The following theorem links Energy Kernels and Wasserstein distances.

Theorem 5.25. *Let $\alpha \in (0,1)$ and $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$.*

i) $W_\alpha(\mu_n,\mu) \to 0$ implies $d_\alpha(\mu_n,\mu) \to 0$.

ii) $d_\alpha(\mu_n,\mu) \to 0$ implies $W_\beta(\mu_n,\mu) \to 0$ for all $\beta < \alpha$.

The theorem reveals the close relationship between the Wasserstein distance W_α and the MMD d_α. The first point states that W_α is stronger than d_α, while the second point states that d_α is stronger than W_β for all $\beta < \alpha$. Since W_α can be seen as the limit of W_β as $\beta \uparrow \alpha$, this suggests that d_α and W_α are *almost equivalent*. However, we conjecture that the two distances are not equivalent on $\mathcal{P}^\alpha$.

Conjecture 1. Let $\alpha \in (0,1)$. There exist $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$ such that

$$d_\alpha(\mu_n,\mu) \to 0 \quad \text{and} \quad W_\alpha(\mu_n,\mu) \nrightarrow 0.$$

Remark 5.26. It is easy to show that, for $\beta < \alpha < 1$, there exist $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}^\alpha$ such that

$$W_\beta(\mu_n,\mu) \to 0 \quad \text{and} \quad d_\alpha(\mu_n,\mu) \nrightarrow 0, \tag{5.18}$$

or, similarly,

$$d_\beta(\mu_n,\mu) \to 0 \quad \text{and} \quad W_\alpha(\mu_n,\mu) \nrightarrow 0. \tag{5.19}$$

We construct simple examples by considering

$$\mu_n = (1-p_n)\delta_0 + p_n\delta_{nx} \quad \text{and} \quad \mu = \delta_0,$$

where $x \in \mathbb{R}^d \setminus \{0\}$ and $p_n \in (0,1)$ is suitably chosen. We easily compute

$$d_\alpha(\mu_n,\mu) = \sqrt{2}p_n n^\alpha \|x\|^\alpha \quad \text{and} \quad W_\alpha(\mu_n,\mu) = p_n n^\alpha \|x\|^\alpha.$$

Similar equations hold for d_β and W_β. Taking $p_n = 1/n^\alpha$, we obtain an example for Equation (5.18). Taking $p_n = 1/(n^\beta \log n)$, we obtain an example for Equation (5.19).

3.4 MMD metrizing the Wasserstein space for $\alpha \geq 1$

In view of the negative result from Proposition 5.24, we wish to exhibit an MMD that metrizes the Wasserstein space of order 1, or more generally, of order $\alpha \geq 1$. The issue evidenced in the proof of Proposition 5.24 is that the matrix part d_Σ controls the expectation and not the absolute moment, suggesting the following modification of Equation (5.9).
Consider the symmetric positive definite kernel

$$k(x,y) = \int_{\mathbb{R}^d} \left(1 - \mathrm{e}^{ix\cdot\xi}\right)\left(1 - \mathrm{e}^{-iy\cdot\xi}\right) \Lambda(\mathrm{d}\xi) + |x|^{\alpha T}\Sigma|y|^\alpha, \tag{5.20}$$

where Λ is a symmetric measure on $\mathbb{R}^d\setminus\{0\}$ satisfying condition (5.8), Σ is a $d\times d$ symmetric positive semi-definite matrix, $\alpha \geq 1$ and $|x|^\alpha = (|x_1|^\alpha, \ldots, |x_1|^\alpha)$ denotes the componentwise absolute α-power. Note that the introduction of this absolute power breaks the translation invariance of the associated MMD.
We first consider the domain of definition.

Lemma 5.27. *Let k be the kernel defined by Equation (5.20).*

1. *$\mathcal{M}_k$ contains the set of measures $\mathcal{M}^\alpha$ that have a finite moment of order α.*
2. *If $\ker \Sigma \cap (\mathbb{R}^+)^d = \{0\}$ then $\mathcal{M}_k = \mathcal{M}^\alpha$.*

Lemma 5.17 and similar arguments as in the proof of Proposition 5.18 show that, for $\mu, \nu \in \mathcal{M}_k$,

$$d_k^2(\mu,\nu) = \int_{\mathbb{R}^d} |\hat{\mu}(\xi) - \hat{\nu}(\xi) - \mu(\mathbb{R}^d) + \nu(\mathbb{R}^d)|^2 \, \Lambda(\mathrm{d}\xi) + \|m_\alpha(\mu) - m_\alpha(\nu)\|_\Sigma^2, \tag{5.21}$$

where $m_\alpha(\mu) = \int_{\mathbb{R}^d} |x|^\alpha \, \mu(\mathrm{d}x) \in \mathbb{R}^d$ is the absolute α-moment of μ. Similarly as in Proposition 5.21, one can easily characterize characteristic kernels in this class.

Proposition 5.28. *Let k be the kernel defined by Equation (5.20). Then the MMD d_k is a distance on $\mathcal{M}_k \cap \mathcal{P}$ if and only if* $\mathrm{supp}(\Lambda) = \mathbb{R}^d$.

Remark 5.29. The condition on the support is not sufficient even for metrizing weak convergence. Indeed, consider

$$\Lambda = \sum_{n=1}^{+\infty} \frac{1}{n^2} \delta_{x_n} \text{ and } \Sigma = 0,$$

where $(x_n)_{n\geq 1}$ is an enumeration of the countable set

$$\left\{ \pm \frac{2a+1}{2^b} \pi \mid a, b \in \mathbb{N} \right\}.$$

This set is dense in $\mathbb{R}$ but for $\mu_n := \delta_{2^n}$, one notes that for all $j \in \mathbb{N}$, $\widehat{\mu_n}(x_j) = 1$ for n large enough. Then $\xi \mapsto 1 - \hat{\mu}_n(\xi)$ converges $\Lambda - ae$ to 0. By the Dominated Convergence Theorem $d_\Lambda(\mu_n, \delta_0) \to 0$, but the sequence $(\mu_n)_{n\geq 1}$ does not converge weakly to δ_0. Note that this kernel verifies all the assumptions of Theorem 7 of Simon-Gabriel et al. (2021), except $\mathcal{H}_k \subset \mathcal{C}_0$ where $\mathcal{C}_0$ is the subspace of functions that vanish at infinity.

The following theorem is the main result of this section. It provides an example of an MMD that metrizes the Wasserstein space of order $\alpha \geq 1$.

Theorem 5.30. *Let k be the kernel defined by Equation (5.20). Then the MMD d_k metrizes the Wasserstein space of order $\alpha \geq 1$ if and only if* $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ *and* $\ker \Sigma \cap (\mathbb{R}^+)^d = \{0\}$.

Example 5.31. The Gaussian Kernel, Example 5.5, is generalizable in this way by considering

$$k(x,y) = \exp(-\|x-y\|_2^2/2) + |x| \cdot |y|, \quad x, y \in \mathbb{R}^d.$$

The previous theorem states that this kernel metrizes the convergence in Wasserstein distance W_1.

3.5 Non asymptotic inequalities for the control of Wasserstein distances

The problem of the Wasserstein distance is its computational cost. That is why a strong equivalence with another distance could be more useful than a topological equivalence. The translation invariant MMD is an L^2-distance of Fourier Transform. The link between the Wasserstein distance and this L^2-distance, for a measure Λ, has already been established in Auricchio et al. (2020) for discrete measures on a regular grid of $[0,1]^d$. The current equivalences, present in the literature, do not allow us to conclude in our case, as we do not want to consider a specific subset of $\mathbb{R}^d$. We pay the cost of the lack of assumption on the support of the measures by the uniform integrability assumption. Moreover, we do not prove a strong equivalence, i.e. an upper bound of a distance by another but only a partial upper bound. This type of inequality has already been introduced and obtained for the MMD in the Section 4 of Vayer and Gribonval (2021). The

authors treat the case where the kernel k is bounded and especially the case where the kernel k is translation invariant. Our results concern only the Energy Kernels, Equation (5.11). The first proposition concerns the Fortet-Mourier distance d_{FM}, i.e. a distance which metrizes the weak convergence, defined by

$$d_{FM}(\mu,\nu) = \sup\Big\{ \int_{\mathbb{R}^d} \varphi \,\mathrm{d}(\mu-\nu) \;:\; \varphi : \mathbb{R}^d \to \mathbb{R} \;\text{ 1-Lipschitz and } \|f\|_\infty \leq 1 \Big\}.$$

We recall the dual formulation of W_1

$$W_1(\mu,\nu) = \sup\Big\{ \int_{\mathbb{R}^d} \varphi \,\mathrm{d}(\mu-\nu) \;:\; \varphi : \mathbb{R}^d \to \mathbb{R} \;\text{ 1-Lipschitz} \Big\}.$$

Proposition 5.32. *Let $\alpha \in (0,1)$ and $\mathcal{T} \subset \mathcal{P}^\alpha(\mathbb{R}^d)$ be a tight subset, i.e.*

$$\forall \varepsilon > 0,\ \exists K \subset \mathbb{R}^d \text{ compact},\ \forall \mu \in \mathcal{T},\ \mu(K^c) \leq \varepsilon.$$

Then, for all $\varepsilon > 0$, there exists $C > 0$ such that

$$\forall \mu,\nu \in \mathcal{T},\ d_{FM}(\mu,\nu) \leq C d_\alpha(\mu,\nu) + \varepsilon.$$

With a stronger assumption, we can get a similar result for the Wasserstein distance W_1.

Proposition 5.33. *Let $\alpha \in (0,1)$ and $\mathcal{T} \subset \mathcal{P}^1(\mathbb{R}^d)$ be a uniformly integrable subset, i.e.*

$$\forall \varepsilon > 0,\ \exists K \subset \mathbb{R}^d \text{ compact},\ \forall \mu \in \mathcal{T},\ \int_{K^c} \|x\| \ \mu(\mathrm{d}x) \leq \varepsilon.$$

Then, for all $\varepsilon > 0$, there exists $C > 0$ such that

$$\forall \mu,\nu \in \mathcal{T},\ W_1(\mu,\nu) \leq C d_\alpha(\mu,\nu) + \varepsilon.$$

Remark 5.34. For both propositions, the constant C depends on ε and on the set $\mathcal{T}$. If we assume that the absolute moments of order $\beta > 1$ are bounded by a constant M, we can give an explicit form of the constant C only in terms of ε. Indeed, Markov's and Hölder's inequality allow to quantify the tighness and the uniform integrability of the set $\mathcal{T}$, i.e. one has an explicit expression of the compact K in function of ε.

4 Proofs

4.1 Proofs related to Section 2

Proof of Proposition 5.8. For a kernel k, Equation (5.5) implies that

$$d_k^2(\delta_x,\delta_y) = k(x,x) + k(y,y) - 2k(x,y) = 2\rho(x,y),$$

for all $x, y \in \mathcal{X}$. It follows that if k_1 and k_2 are equivalent kernels, then they have the same variogram.
Conversely, we prove that kernels with the same variograms are equivalent. Let k be a kernel with variogram ρ. We fix an origin $o \in \mathcal{X}$ and consider the kernel

$$k_0(x,y) = k(x,y) - k(x,o) - k(o,y) + k(o,o) = \rho(x,o) + \rho(o,y) - \rho(x,y)$$

which has the same variogram ρ. The application k_0 is indeed a kernel by the Lemma 2.1 of Berg et al. (1984) cause $-k$ is negative definite. We show that $\mathcal{M}_k = \mathcal{M}_{k_0}$ and $d_k = d_{k_0}$ on $\mathcal{M}_k \cap \mathcal{P}$. Since k_0 depends only on the variogram ρ, this implies that two kernels with the same variogram are equivalent.

The inequality

$$k(x,x) - 2|k(x,o)| + k(o,o) \leq k_0(x,x) \leq k(x,x) + 2|k(x,o)| + k(o,o)$$

together with the Cauchy Schwarz inequality entail

$$\left(\sqrt{k(x,x)} - \sqrt{k(o,o)}\right)^2 \leq k_0(x,x) \leq \left(\sqrt{k(x,x)} + \sqrt{k(o,o)}\right)^2.$$

It follows that $\int_{\mathcal{X}} \sqrt{k(x,x)}\,|\mu|(\mathrm{d}x) < \infty$ if and only if $\int_{\mathcal{X}} \sqrt{k_0(x,x)}\,|\mu|(\mathrm{d}x) < \infty$ so that $\mathcal{M}_{k_0} = \mathcal{M}_k$. Let μ and ν be probability measures in $\mathcal{M}_k$. By Equation (5.5),

$$\begin{aligned} d^2_{k_0}(\mu,\nu) &= \int_{\mathcal{X}\times\mathcal{X}} \Big(k(x,y) - k(x,o) - k(o,y) + k(o,o)\Big)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y) \\ &= \int_{\mathcal{X}\times\mathcal{X}} k(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y) \\ &= d^2_k(\mu,\nu). \end{aligned}$$

The second equality uses that $\mu-\nu$ has total mass 0 (since μ and ν are probability measures) so that only $k(x,y)$ yields a non null integral. Interestingly, a similar computation shows that the MMD can be directly written in terms of the variogram : for $\mu,\nu \in \mathcal{M}_k$ with the same mass,

$$\begin{aligned} d^2_k(\mu,\nu) &= d^2_{k_0}(\mu,\nu) \\ &= \int_{\mathcal{X}\times\mathcal{X}} \Big(\rho(x,o) + \rho(o,y) - \rho(x,y)\Big)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y) \\ &= -\int_{\mathcal{X}\times\mathcal{X}} \rho(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x,\mathrm{d}y). \end{aligned}$$

□

Proof of Proposition 5.9. Let k be a kernel and $o \in \mathcal{X}$ an arbitrary origin. The kernel k_0 is naturally normalized with origin o. It is easy to check that k_0 has the same variogram than k, then these two kernels are equivalent by Proposition 5.8. Conversely, let K_0 be another kernel equivalent to k normalized, then K_0 and k_0 have the same variogram then for $x \in \mathcal{X}$,

$$k_0(x,x) = 2\rho(o,x) = K_0(x,x).$$

So for any $x,y \in \mathcal{X}$, the equality of the variogram implies $K_0(x,y) = k_0(x,y)$, then this kernel is unique. □

Proof of Theorem 5.12. Assume the MMD associated with k is translation invariant. For $h \in \mathbb{R}^d$, define the translated kernel $k_h(x,y) = k(x+h, y+h)$. Clearly, we have

$$d_k(\mu \circ \tau_h^{-1}, \nu \circ \tau_h^{-1}) = d_{k_h}(\mu,\nu)$$

and Equation (5.6) implies that the kernel k and k_h are equivalent (in the sense of Definition 5.6). Proposition 5.8 implies that k_h and k have the same variogram, which implies

$$\rho(x,y) = \rho(x+h, y+h), \quad \text{for all } x,y \in \mathbb{R}^d.$$

Since h is arbitrary, we can take $h = y-x$ and define the function $\gamma(h) = \rho(0,h)$ so as to obtain $\rho(x,y) = \rho(0,y-x) = \gamma(y-x)$. The function γ is negative definite because ρ is negative definite. Furthermore, $\gamma(0) = \rho(0,0) = 0$.

Conversely, given a negative definite function $\gamma : \mathbb{R}^d \to [0, \infty)$ such that $\gamma(0) = 0$, the function $\rho(x, y) = \gamma(y - x)$ is negative definite on $\mathbb{R}^d \times \mathbb{R}^d$ and

$$k_0(x, y) = \rho(x, 0) + \rho(0, y) - \rho(x, y) - \rho(0, 0)$$

is positive definite, see Berg et al. (1984, Lemma 2.1 p.74). One can easily check that $k_0(x, y) = \gamma(x) + \gamma(y) - \gamma(y - x)$. Furthermore, the translated kernel

$$k_h(x, y) = k_0(x + h, y + h) = \gamma(x + h) + \gamma(y + h) - \gamma(y - x)$$

has variogram

$$\rho_h(x, y) = \frac{1}{2}k_h(x, x) + \frac{1}{2}k_h(y, y) - k_h(x, y) = \gamma(y - x).$$

The kernels k_h and k have the same variogram and are thus equivalent, which proves that the MMD is translation invariant. □

Proof of Lemma 5.15. If Λ is finite then k_Λ is bounded. Now, assume that Λ is not finite. Let $R > 0$, we denote by B_R the ball with center 0 and radius R in $\mathbb{R}^d$ and by λ_R its volume for the Lebesgue measure λ. By Fubini-Tonelli Theorem

$$\frac{1}{\lambda_R}\int_{B_R} k_\Lambda(x, x)\,\lambda(\mathrm{d}x) = \frac{1}{\lambda_R}\int_{\mathbb{R}^d}\int_{B_R} |1 - \mathrm{e}^{ix\cdot\xi}|^2\,\lambda(\mathrm{d}x)\Lambda(\mathrm{d}\xi).$$

We consider

$$f_R(\xi) = \frac{1}{\lambda_R}\int_{B_R} |1 - \mathrm{e}^{ix\cdot\xi}|^2\,\lambda(\mathrm{d}x) = \frac{1}{\lambda_R}\int_{B_R} \big(2 - 2\cos(x \cdot \xi)\big)\,\lambda(\mathrm{d}x).$$

By Fatou's Lemma, as $R \to +\infty$,

$$\liminf \frac{1}{\lambda_R}\int_{B_R} k_\Lambda(x, x)\,\lambda(\mathrm{d}x) = \liminf \int_{\mathbb{R}^d} f_R(\xi)\,\Lambda(\mathrm{d}\xi) \geq \int_{\mathbb{R}^d} \liminf f_R(\xi)\,\Lambda(\mathrm{d}\xi).$$

If $\xi \neq 0$, Riemann-Lebesgue Lemma entails, as $R \to +\infty$,

$$\lim f_R(\xi) = \lim \frac{1}{\lambda_R}\int_{B_R} \big(2 - 2\cos(x \cdot \xi)\big)\,\lambda(\mathrm{d}x) = 2,$$

whence we deduce

$$\liminf \frac{1}{\lambda_R}\int_{B_R} k_\Lambda(x, x)\,\lambda(\mathrm{d}x) \geq 2\Lambda(\mathbb{R}^d) = +\infty.$$

This shows that k_Λ is not bounded. We have proven that k_Λ is bounded if and only if Λ is bounded. The condition on $k = k_\Lambda + k_\Sigma$ follows easily. □

Proof of Lemma 5.17. The proof of $\mathcal{M}_k = \mathcal{M}_{k_1} \cap \mathcal{M}_{k_2}$ relies on the inequality

$$\max\Big(\sqrt{k_1(x, x)}, \sqrt{k_2(x, x)}\Big) \leq \sqrt{k_1(x, x) + k_2(x, x)} \leq \sqrt{k_1(x, x)} + \sqrt{k_2(x, x)},$$

which implies that $\sqrt{k_1(x, x) + k_2(x, x)}$ is $|\mu|(\mathrm{d}x)$-integrable if and only if both $\sqrt{k_1(x, x)}$ and $\sqrt{k_2(x, x)}$ are. Then, for $\mu, \nu \in \mathcal{M}_k$, we can compute $d_k^2(\mu, \nu)$ according to Equation (5.5) with k replaced by k_1 and k_2 ; since $\mu, \nu \in \mathcal{M}_{k_1} \cap \mathcal{M}_{k_2}$, the integral can be split into two integrals, one for k_1 and one for k_2, and we obtain $d_k^2(\mu, \nu) = d_{k_1}^2(\mu, \nu) + d_{k_2}^2(\mu, \nu)$. □

The following Lemma gives an upper bound on the growth of the kernel k_Λ and will be useful in the proof of Proposition 5.18.

Lemma 5.35. *Let k_Λ be a kernel of the form (5.13) and assume that, for some $0 < \alpha \leq 2$, we have $\int_{\mathbb{R}^d}(\|\xi\|^\alpha \wedge 1)\,\Lambda(\mathrm{d}\xi) < +\infty$. Then $k_\Lambda(x,x) = o(\|x\|^\alpha)$, as $\|x\| \to +\infty$, and $\mathcal{M}^{\alpha/2} \subset \mathcal{M}_\Lambda$.*

Proof of Lemma 5.35. Assume $\int_{\mathbb{R}^d}(\|\xi\|^\alpha \wedge 1)\Lambda(\mathrm{d}\xi) < \infty$ with $0 < \alpha \leq 2$. We show that for all $\varepsilon > 0$, there exists $C > 0$ such that

$$|k_\Lambda(x,x)| \leq C + \varepsilon\|x\|^\alpha, \quad x \in \mathbb{R}^d. \tag{5.22}$$

Since ε can be chosen arbitrary small, this shows $k_\Lambda(x,x) = o(\|x\|^\alpha)$ as $\|x\| \to +\infty$.
We compute

$$k_\Lambda(x,x) = \int_{\mathbb{R}^d} \left|1 - \mathrm{e}^{ix\cdot\xi}\right|^2 \Lambda(\mathrm{d}\xi) \leq 4\int_{\mathbb{R}^d} \left((\|x\|\|\xi\|)^2 \wedge 1\right)\Lambda(\mathrm{d}\xi)$$

and divide the integral into two parts, depending whether $\|\xi\|$ is larger or smaller than some $\eta > 0$ that will be fixed later. The inequality $u^2 \wedge 1 \leq 1$ implies

$$\int_{\{\|\xi\|\geq\eta\}} \left((\|x\|\|\xi\|)^2 \wedge 1\right)\Lambda(\mathrm{d}\xi) \leq \Lambda(\|\xi\| \geq \eta).$$

For $0 < \alpha \leq 2$, the inequality $u^2 \wedge 1 \leq |u|^\alpha$ implies

$$\int_{\{\|\xi\|<\eta\}} \left((\|x\|\|\xi\|)^2 \wedge 1\right)\Lambda(\mathrm{d}\xi) \leq \int_{\{\|\xi\|<\eta\}} (\|x\|\|\xi\|)^\alpha\,\Lambda(\mathrm{d}\xi) \leq \|x\|^\alpha \int_{\{\|\xi\|<\eta\}} \|\xi\|^\alpha\,\Lambda(\mathrm{d}\xi).$$

Since $\int_{\mathbb{R}^d}(\|\xi\|^\alpha \wedge 1)\Lambda(\mathrm{d}\xi) < \infty$, for any fixed $\varepsilon > 0$, one can find $\eta > 0$ small enough such that $\int_{\{\|\xi\|<\eta\}} \|\xi\|^\alpha\,\Lambda(\mathrm{d}\xi) < \varepsilon/4$. Setting $C = 4\Lambda(\|\xi\| \geq \eta)$, the upper bounds for the two terms above entail Equation (5.22).
As a direct consequence of Equation (5.22), any measure $\mu \in \mathcal{M}$ satisfying $\int_{\mathbb{R}^d} \|x\|^\alpha\,|\mu|(\mathrm{d}x) < \infty$ satisfies also $\int_{\mathbb{R}^d} \sqrt{k_\Lambda(x,x)}\,|\mu|(\mathrm{d}x) < \infty$. In other words, $\mathcal{M}^\alpha \subset \mathcal{M}_\Lambda$ and this concludes the proof of the Lemma. □

Proof of Proposition 5.18. • The inclusion $\mathcal{M}^{\alpha/2} \subset \mathcal{M}_\Lambda$ is proven in Lemma 5.35. Assumption (5.8) implies that $\mathcal{M}^1 \subset \mathcal{M}_\Lambda$. The computation of the MMD in terms of characteristic functions follows the lines of Sriperumbudur et al. (2010, Corollary 4 and its proof). For $\mu, \nu \in \mathcal{M}_\Lambda$,

$$\begin{aligned}
d_\Lambda^2(\mu,\nu) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} k_\Lambda(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d}\int_{\mathbb{R}^d} \left(1-\mathrm{e}^{ix\cdot\xi}\right)\left(1-\mathrm{e}^{-iy\cdot\xi}\right)\Lambda(\mathrm{d}\xi)(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\
&= \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left(1-\mathrm{e}^{ix\cdot\xi}\right)(\mu-\nu)(\mathrm{d}x)\int_{\mathbb{R}^d}\left(1-\mathrm{e}^{-iy\cdot\xi}\right)(\mu-\nu)(\mathrm{d}y)\right]\Lambda(\mathrm{d}\xi) \\
&= \int_{\mathbb{R}^d}\left(\mu(\mathbb{R}^d)-\nu(\mathbb{R}^d)-\hat{\mu}(\xi)+\hat{\nu}(\xi)\right)\overline{\left(\mu(\mathbb{R}^d)-\nu(\mathbb{R}^d)-\hat{\mu}(\xi)+\hat{\nu}(\xi)\right)}\,\Lambda(\mathrm{d}\xi) \\
&= \int_{\mathbb{R}^d}\left|\mu(\mathbb{R}^d)-\nu(\mathbb{R}^d)-\hat{\mu}(\xi)+\hat{\nu}(\xi)\right|^2\Lambda(\mathrm{d}\xi).
\end{aligned}$$

In these lines, we have used successively Equations (5.5) and (5.13), Fubini's theorem and the definition of the characteristic function.
• The Spectral Theorem for the symmetric positive semidefinite matrix Σ implies

$$k_\Sigma(x,y) = x^T\Sigma y = \sum_{j=1}^r \lambda_j x^T e_j e_j^T y, \quad x,y \in \mathbb{R}^d,$$

where $\lambda_1 \geq \ldots \geq \lambda_r > 0$ are the positive eigenvalues of Σ associated with the orthonormal eigenvectors $(e_1, \ldots, e_r)$. Together with the elementary inequality $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$, for $a, b \geq 0$, we deduce

$$\sqrt{\lambda_l}\,|e_l^T x| \leq \sqrt{k_\Sigma(x,x)} \leq \sum_{j=1}^{r} \sqrt{\lambda_r}\,|e_j^T x|\,, \quad l = 1, \ldots, r.$$

We deduce that $\int_{\mathbb{R}^d} \sqrt{k_\Sigma(x,x)}\,|\mu|(\mathrm{d}x)$ is finite if and only if $\int_{\mathbb{R}^d} |e_j^T x|\,|\mu|(\mathrm{d}x)$ is finite for all $j = 1, \ldots, r$. This proves the characterization of M_Σ. On the other hand, a direct computation gives, for $\mu, \nu \in \mathcal{M}_\Sigma$,

$$\begin{aligned} d_\Sigma^2(\mu,\nu) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} k_\Lambda(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\ &= \sum_{j=1}^{r} \lambda_j \int_{\mathbb{R}^d \times \mathbb{R}^d} \left(x^T e_j e_j^T y\right)(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\ &= \sum_{j=1}^{r} \lambda_j \left| \int_{\mathbb{R}^d} (e_j^T x)\,\mu(\mathrm{d}x) - \int_{\mathbb{R}^d} (e_j^T x)\,\mu(\mathrm{d}x)\right|^2. \end{aligned}$$

□

4.2 Proofs related to Section 3

Proofs of Subection 3.2

Proof of Proposition 5.23. The proof is done by contraposition. Assume that the kernel k is bounded and let $\alpha > 0$. We prove that d_k does not metrize the Wasserstein space of order α. The assumption that k is bounded implies $\mathcal{M}_k = \mathcal{M}$. For $x \in \mathbb{R}^d \setminus \{0\}$ and $n \geq 1$, we consider the probability measures

$$\mu_n = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_{n^{1/\alpha}x} \quad \text{and} \quad \mu = \delta_0.$$

Then, since k is bounded,

$$d_k^2(\mu_n,\mu) = \frac{1}{n^2}\Big(k(0,0) + k(n^{1/\alpha}x, nn^{1/\alpha}x) - 2k(n^{1/\alpha}x,0)\Big) \to 0.$$

On the other hand,

$$W_\alpha(\mu_n,\mu) = \int_{\mathbb{R}^d} \|y\|^\alpha\,\mu_n(\mathrm{d}y) = \|x\| \nrightarrow 0.$$

This shows that d_k does not metrize the Wasserstein space of order α. □

Proof of Proposition 5.24. For $x \in \mathbb{R}^d \setminus \{0\}$ and $n \geq 2$, we consider the probability measures

$$\mu_n = \frac{n-2}{n}\delta_0 + \frac{1}{n}\delta_{-nx} + \frac{1}{n}\delta_{nx} \quad \text{and} \quad \mu = \delta_0.$$

On the one hand, the measures μ_n and μ are symmetric and thus have expectation 0. It follows that $e(\mu) = e(\mu_n) = 0$ and $d_\Sigma(\mu_n, \delta_0) = 0$ according to Proposition 5.18. Furthermore, we compute

$$d_\Lambda^2(\mu_n,\mu) = \frac{1}{n^2}\big(k_\Lambda(nx,nx) + k_\Lambda(-nx,-nx) + 2k_\Lambda(nx,-nx)\big)$$

and, according to Lemma 5.35, $|k_\Lambda(nx,nx)| = o(n^2)$, $|k_\Lambda(-nx,-nx)| = o(n^2)$ and

$$|k_\Lambda(-nx,nx)| \leq \sqrt{k_\Lambda(nx,nx)}\sqrt{k_\Lambda(-nx,-nx)} = o(n^2).$$

We deduce $d_k(\mu_n,\mu) = d_\Lambda(\mu_n,\mu) \to 0$. On the other hand,

$$W_1(\mu_n,\mu) = \int_{\mathbb{R}^d} \|y\|\,\mu_n(\mathrm{d}y) = \|x\| \nrightarrow 0.$$

This proves that no kernel of the form (5.9) can metrize the Wasserstein space of order 1. □

Proof of Theorem 5.25

For $\alpha \in (0,1)$, we recall that the Energy Kernel is defined by

$$k_\alpha(x,y) = \|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x-y\|^{2\alpha}$$

and we denote by $\mathcal{H}_\alpha = \mathcal{H}_{k\alpha}$ and $d_\alpha = d_{k_\alpha}$ the associated RKHS and the MMD. We recall that $\mathcal{M}_{k_\alpha} = \mathcal{M}_\alpha$. The kernel mean embedding is denoted by $K_\alpha : \mathcal{M}_\alpha \to \mathcal{H}_\alpha$ and is defined by

$$K_\alpha(\mu)(x) = \int_{\mathbb{R}^d} k_\alpha(x,y)\,\mu(\mathrm{d}y), \quad x \in \mathbb{R}^d.$$

For the sake of clarity, we divide the proof of Theorem 5.25 into two parts. The next two lemma will be useful for the first part.

Lemma 5.36. *For all $\mu \in \mathcal{M}_\alpha$, the kernel mean embedding $K_\alpha(\mu)$ is α-Hölder continuous with constant $c_\alpha(\mu) = 2\int_{\mathbb{R}^d} \|y\|^\alpha\,|\mu|(\mathrm{d}y)$, i.e.*

$$|K_\alpha(\mu)(x) - K_\alpha(\mu)(x')| \le c_\alpha(\mu)\|x-x'\|^\alpha, \quad x, x' \in \mathbb{R}^d.$$

Proof of Lemma 5.36. We have, for $x, x' \in \mathbb{R}^d$,

$$\begin{aligned}|K_\alpha(\mu)(x) - K_\alpha(\mu)(x')| &= \Big|\int_{\mathbb{R}^d} k_\alpha(x,y)\,\mu(\mathrm{d}y) - \int_{\mathbb{R}^d} k_\alpha(x',y)\,\mu(\mathrm{d}y)\Big| \\ &\le \int_{\mathbb{R}^d} \big|k_\alpha(x,y) - k_\alpha(x',y)\big|\,|\mu|(\mathrm{d}y).\end{aligned}$$

Using the reproducing kernel property and Cauchy-Schwartz inequality, the integrand satisfies

$$\begin{aligned}|k_\alpha(x,y) - k_\alpha(x',y)| &= |\langle K_\alpha(x), K_\alpha(y)\rangle - \langle K_\alpha(x'), K_\alpha(y)\rangle| \\ &= |\langle K_\alpha(x) - K_\alpha(x'), K_\alpha(y)| \\ &\le \|K_\alpha(x) - K_\alpha(x')\|\|K_\alpha(y)\| \\ &= \sqrt{k_\alpha(x,x) + k_\alpha(x',x') - 2k_\alpha(x,x')}\sqrt{k_\alpha(y,y)} \\ &= 2\|x-x'\|^\alpha\|y\|^\alpha.\end{aligned}$$

Integrating with respect to $|\mu|(\mathrm{d}y)$, we deduce

$$|K_\alpha(\mu)(x) - K_\alpha(\mu)(x')| \le 2\|x-x'\|^\alpha \int_{\mathbb{R}^d} \|y\|^\alpha\,|\mu|(\mathrm{d}y),$$

whence the function $K_\alpha(\mu)$ is Hölder-continuous with exponent α. □

Lemma 5.37. *For all $\mu, \nu \in \mathcal{P}_\alpha$, we have*

$$d_\alpha^2(\mu,\nu) \le \big(c_\alpha(\mu) + c_\alpha(\nu)\big)W_\alpha(\mu,\nu).$$

Proof of Lemma 5.37. We recall that, for $\alpha \in (0,1)$, the Kantorovitch-Rubinstein duality implies that

$$W_\alpha(\mu,\nu) = \sup\Big|\int_{\mathbb{R}^d} \varphi(x)\,(\mu-\nu)(\mathrm{d}x)\Big| \tag{5.23}$$

with the supremum taken over the set of Hölder-continuous function with exponent α and constant 1.
Starting from Equation (5.5) and integrating with respect to y, we get

$$\begin{aligned}d_\alpha^2(\mu,\nu) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} k_\alpha(x,y)\,(\mu-\nu)\otimes(\mu-\nu)(\mathrm{d}x\mathrm{d}y) \\ &= \int_{\mathbb{R}^d} K_\alpha(\mu-\nu)(x)\,(\mu-\nu)(\mathrm{d}x).\end{aligned}$$

According to Lemma 5.36, the function $K_\alpha(\mu-\nu)$ is Hölder continuous with exponent α and constant $c_\alpha(\mu-\nu)$. Then, Equation (5.23) implies

$$\begin{aligned} d_\alpha^2(\mu,\nu) &= \int_{\mathbb{R}^d} K_\alpha(\mu-\nu)(x)\,(\mu-\nu)(\mathrm{d}x) \\ &\leq c_\alpha(\mu-\nu) W_\alpha(\mu,\nu). \end{aligned}$$

We conclude by using the fact that

$$\begin{aligned} c_\alpha(\mu-\nu) &= 2\int_{\mathbb{R}^d} \|y\|^\alpha |\mu-\nu|(\mathrm{d}y) \\ &\leq 2\int_{\mathbb{R}^d} \|y\|^\alpha \mu(\mathrm{d}y) + 2\int_{\mathbb{R}^d} \|y\|^\alpha \nu(\mathrm{d}y) \\ &= c_\alpha(\mu) + c_\alpha(\nu). \end{aligned}$$

□

Proof of Theorem 5.25 (first point). Let $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_\alpha$ be such that $W_\alpha(\mu_n,\mu) \to 0$. By Lemma 5.37,

$$d_\alpha^2(\mu_n,\mu) \leq \big(c_\alpha(\mu_n) + c_\alpha(\mu)\big) W_\alpha(\mu_n,\mu).$$

It is enough to prove that the sequence $(c_\alpha(\mu_n))_{n\geq 1}$ remains bounded in order to conclude $d_\alpha(\mu_n,\mu) \to 0$. This is indeed the case since the convergence $\mu_n \to \mu$ in Wasserstein space of order α implies the convergence of absolute moments

$$\int_{\mathbb{R}^d} \|x\|^\alpha \,\mu_n(\mathrm{d}x) \longrightarrow \int_{\mathbb{R}^d} \|x\|^\alpha \,\mu(\mathrm{d}x),$$

which yields $c_\alpha(\mu_n) \to c_\alpha(\mu)$. Being convergent, the sequence $(c_\alpha(\mu_n))_{n\geq 1}$ is bounded. □

We next consider the proof of the second point in Theorem 5.25. The following lemma is the key of the proof.

Lemma 5.38. *For $r > 0$, we define the measure $\mu_r(\mathrm{d}s) = (1+\|s\|)^{-d-r}\mathrm{d}s$. Then, for $r > \alpha$, $\mu_r \in \mathcal{M}_\alpha$. Furthermore, for $\alpha < r < 1 \wedge 2\alpha$, the kernel mean embedding satisfies*

$$K_\alpha(\mu_r)(x) \sim d(\alpha,r)\|x\|^{2\alpha-r}, \quad \text{as } \|x\| \to +\infty,$$

with $d(\alpha,r) > 0$.

Proof of Lemma 5.38. As $r > \alpha$, the function $\sqrt{k_\alpha(x,x)} = \sqrt{2}\|x\|^\alpha$ is μ_r-integrable and hence $\mu_r \in \mathcal{M}_\alpha$. The KME $K_\alpha(\mu_r) \in \mathcal{H}_\alpha$ is defined by

$$\begin{aligned} K(\mu_r)(x) &= \int_{\mathbb{R}^d} k_\alpha(x,y)\,\mu_r(\mathrm{d}y) \\ &= \int_{\mathbb{R}^d} \Big(\|x\|^{2\alpha} + \|y\|^{2\alpha} - \|x-y\|^{2\alpha}\Big)\big(1+\|y\|\big)^{-(d+r)}\,\mathrm{d}y. \end{aligned}$$

The change of variable $z = y/\|x\|$ yields

$$K(\mu_r)(x) = \|x\|^{2\alpha+d} \int_{\mathbb{R}^d} \Big(1 + \|z\|^{2\alpha} - \|x/\|x\| - z\|^{2\alpha}\Big)\big(1+\|x\|\|z\|\big)^{-(d+r)}\,\mathrm{d}z.$$

By the rotationnal invariance of the Euclidean norm and the Lebesgue measure, the integral does not change if we replace the unit vector $x/\|x\|$ by $e_1 = (1,0,\ldots,0)$. This yields

$$K(\mu_r)(x) = \|x\|^{2\alpha+d} \int_{\mathbb{R}^d} \Big(1 + \|z\|^{2\alpha} - \|e_1 - z\|^{2\alpha}\Big)\big(1+\|x\|\|z\|\big)^{-(d+r)}\,\mathrm{d}z.$$

Note that $K_\alpha(\mu_r)(x)$ is rotation invariant and depends only on $\|x\|$. We next consider the asymptotic as $\|x\| \to +\infty$. In order to ease the analysis, we use the following form

$$K(\mu_r)(x) = \|x\|^{2\alpha-r} \int_{\mathbb{R}^d} \left(\frac{\|x\|\|z\|}{1+\|x\|\|z\|}\right)^{d+r} \frac{1+\|z\|^{2\alpha}-\|e_1-z\|^{2\alpha}}{\|z\|^{d+r}}\,\mathrm{d}z.$$

Using this expression, the proof of the Lemma is reduced to the proof of the convergence

$$\int_{\mathbb{R}^d} \left(\frac{u\|z\|}{1+u\|z\|}\right)^{d+r} \frac{1+\|z\|^{2\alpha}-\|e_1-z\|^{2\alpha}}{\|z\|^{d+r}}\,\mathrm{d}z \to d(\alpha,r) > 0, \quad \text{as } u \to +\infty. \tag{5.24}$$

We observe that, for all $z \in \mathbb{R}^d \setminus \{0\}$,

$$\left(\frac{u\|z\|}{1+u\|z\|}\right)^{d+r} \longrightarrow 1, \quad \text{as } u \to \infty,$$

suggesting the convergence with limit

$$d(\alpha,r) = \int_{\mathbb{R}^d} \frac{1+\|z\|^{2\alpha}-\|e_1-z\|^{2\alpha}}{\|z\|^{d+r}}\,\mathrm{d}z.$$

This is justified by Lebesgue dominated convergence Theorem, since

$$\left(\frac{u\|z\|}{1+u\|z\|}\right)^{d+r} \le 1$$

and

$$g(z) = \frac{1+\|z\|^{2\alpha}-\|e_1-z\|^{2\alpha}}{\|z\|^{d+r}} \text{ is integrable.}$$

This last claim holds because :

- for $\|z\| > 1/2$, the upper bound

  $$|g(z)| = \|z\|^{-(d+r)}|k_\alpha(e_1,z)| \le \|z\|^{-(d+r)}\sqrt{k_\alpha(e_1,e_1)}\sqrt{k_\alpha(z,z)} = 2\|z\|^{\alpha-d-r},$$

  implies integrability on $\{z : \|z\| > 1/2\}$ since $r > \alpha$;
- for $\|z\| \le 1/2$, the function $z \mapsto 1-\|e_1-u\|^{2\alpha}$ is continuously differentiable on the compact ball $\{z : \|z\| \le 1/2\}$ and vanishes at 0 so that $\left|1-\|e_1-z\|^{2\alpha}\right| \le C\|z\|$ for some $C > 0$; we deduce

  $$|g(z)| \le \|z\|^{2\alpha-d-r} + C\|z\|^{1-d-r}$$

  which implies integrability on $\{z : \|z\| \le 1/2\}$ since $r < 1 \wedge 2\alpha$.

The convergence (5.24) is proved and it remains to show that the limit is positive. By rotation invariance,

$$d(\alpha,r) = \int_{\mathbb{R}^d} \frac{1+\|z\|^{2\alpha}-\|z-e_1\|^{2\alpha}}{\|z\|^{d+r}}\,\mathrm{d}z = \int_{\mathbb{R}^d} \frac{1+\|z\|^{2\alpha}-\|z+e_1\|^{2\alpha}}{\|z\|^{d+r}}\,\mathrm{d}z.$$

Then, taking the mean of the two expressions,we get

$$\begin{aligned}
d(\alpha,r) &= \int_{\mathbb{R}^d} \|z\|^{-(d+r)}\left(1+\|z\|^{2\alpha} - \frac{\|z-e_1\|^{2\alpha}+\|z+e_1\|^{2\alpha}}{2}\right)\mathrm{d}z \\
&\ge \int_{\mathbb{R}^d} \|z\|^{-(d+r)}\left(1+\|z\|^{2\alpha} - \left[\frac{\|z-e_1\|^2+\|z+e_1\|^2}{2}\right]^{\alpha}\right)\mathrm{d}z \\
&= \int_{\mathbb{R}^d} \|z\|^{-(d+r)}\left(1+\|z\|^{2\alpha} - \left(1+\|z\|^2\right)^{\alpha}\right)\mathrm{d}z \\
&> \int_{\mathbb{R}^d} \|z\|^{-(d+r)}\left(1+\|z\|^{2\alpha} - 1 - \|z\|^{2\alpha}\right)\mathrm{d}z \\
&= 0.
\end{aligned}$$

The first inequality uses the concavity of the function $u \mapsto u^\alpha$ on $(0,+\infty)$ and the second inequality uses $(1+u)^\alpha < 1+u^\alpha$ for $u > 0$. Both properties hold because $\alpha \in (0,1)$. □

In the proof of this second point, one will also need this technical lemma. It is a generalization of the classical characterization of the Wasserstein convergence (Theorem 7.12, Villani (2003)).

Lemma 5.39. *Let $f \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R})$ and $\beta \in (0,1)$ such that*

$$f(x) \sim C\|x\|^\beta, \quad \text{as } \|x\| \to +\infty,$$

with $C > 0$. Let $(\mu_n)_{n\geq 1}$ be a sequence of probability measures and $\mu \in \mathcal{P}_\beta$. If the sequence $(\mu_n)_{n\geq 1}$ converges weakly to μ and $\int_{\mathbb{R}^d} f(x)\ \mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^d} f(x)\ \mu(\mathrm{d}x)$ then $W_\beta(\mu_n, \mu) \to 0$.

Proof of Lemma 5.39. The purpose of this proof is to show a kind of Wasserstein tigntness as stated in point (ii) of (Theorem 7.12, Villani (2003)),

$$\lim_{R\to+\infty} \limsup_{n\to+\infty} \int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) = 0.$$

By this theorem, this condition will imply the Wasserstein convergence. Let $R > 1$ such that $\|x\|^\beta \leq 2Cf(x)$ for all $\|x\| \geq R-1$. Let $\chi_R : \mathbb{R}^d \to \mathbb{R}$ be the continuous function defined by

$$\chi_R(x) = \mathbb{1}_{\|x\|_2 \leq R-1} + (R - \|x\|_2)\mathbb{1}_{R-1<\|x\|_2<R}, \quad \text{for } x \in \mathbb{R}^d.$$

Let $n \geq 1$, noting that $1 - \chi_R(x) = 1$ for $\|x\| \geq R$ and $f(x) \geq 0$ for $\|x\| \geq R-1$,

$$\begin{aligned}\int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) &\leq 2C \int_{\mathbb{R}^d} (1-\chi_R(x))f(x)\ \mu_n(\mathrm{d}x) \\ &= 2C \int_{\mathbb{R}^d} f(x)\ \mu_n(\mathrm{d}x) - 2C \int_{\mathbb{R}^d} \chi_R(x)f(x)\ \mu_n(\mathrm{d}x).\end{aligned}$$

The function $\chi_R f$ is continuous and bounded then by the weak convergence

$$\limsup_{n\to+\infty} \int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) \leq 2C \int_{\mathbb{R}^d} f(x)\ \mu(\mathrm{d}x) - 2C \int_{\mathbb{R}^d} \chi_R(x)f(x)\ \mu(\mathrm{d}x).$$

As f is integrable, the Dominated Convergence Theorem gives

$$\lim_{R\to+\infty} \limsup_{n\to+\infty} \int_{\|x\|\geq R} \|x\|^\beta\ \mu_n(\mathrm{d}x) = 0.$$

This tightness condition implies $W_\beta(\mu_n, \mu) \to 0$. □

Proof of Theorem 5.25 (second point). Let $(\mu_n)_{n\geq 1}$ and μ be probability measures such that $d_\alpha(\mu_n, \mu) \to 0$. Then the sequence of KME $(K(\mu_n))_{n\geq 1}$ converges weakly (in Hilbert sense) to $K(\mu)$, i.e.

$$\forall f \in \mathcal{H}_\alpha,\ \langle f, K(\mu_n)\rangle = \int_{\mathbb{R}^d} f\ \mathrm{d}\mu_n \to \int_{\mathbb{R}^d} f\ \mathrm{d}\mu,$$

in particular, for any fonctions $K(\mu_r)$ of Lemma 5.38 with $\alpha < r < 1 \wedge 2\alpha$. For $\beta \in (2\alpha - 1 \vee 0, \alpha)$, let's consider $r := 2\alpha - \beta \in (\alpha, 1 \wedge 2\alpha)$. Again by the Lemma 5.38,

$$K(\mu_r)(x) \sim d(\alpha, r)\|x\|^\beta, \quad \text{as } \|x\| \to +\infty.$$

Hence there exists a constant $C > 0$ such that $\|x\|^\beta \leq C(K(\mu_r)(x) + 1)$ for all $x \in \mathbb{R}^d$. Then the sequence $(m_\beta(\mu_n))_{n\geq 1}$ of β-moment is bounded. The Markov Inequality ensures the tightness of the sequence $(\mu_n)_{n\geq 1}$. Let us recall the Equation (5.15) which gives the form of d_α^2

$$d_\alpha^2(\mu_n, \mu) = \frac{1}{C(d, 2\alpha)} \int_{\mathbb{R}^d} \frac{|\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2}{\|\xi\|^{d+2\alpha}}\ \mathrm{d}\xi.$$

As the convergence L^1 implies the converges almost everywhere to a sub-sequence and the characteristic function is continuous, the probability measure μ is the unique adherent point of the tight sequence $(\mu_n)_{n\geq 1}$, then by the Prokorhov Theorem, the sequence converges weakly to the measure μ.

The kernel k_α is continuous in its 2 variables, so it is separately continuous and locally bounded. Thus by the Corollary 3 of Simon-Gabriel and Schölkopf (2018), all functions $f \in \mathcal{H}_\alpha$ are continuous, including the function $K(\mu_r)$.

The assumptions of Lemma 5.39 are therefore satisfied, so $W_\beta(\mu_n, \mu) \to 0$. The continuity of the injection, recalled in formula (5.17), generalizes this convergence for any $\beta \in (0, \alpha)$. □

Proof of Subsection 3.4

Proof of Lemma 5.27. We first state a simple property that will be useful for the proof : there exists $M \geq 0$ such that

$$x^T \Sigma x \leq M\|x\|^2 \quad \text{for all } x \in (\mathbb{R}_+)^d, \tag{5.25}$$

and, if $\mathrm{Ker}(\Sigma) \cap (\mathbb{R}_+)^d) = \{0\}$, there exists also $m > 0$ such that

$$x^T \Sigma x \geq m\|x\|^2 \quad \text{for all } x \in (\mathbb{R}_+)^d. \tag{5.26}$$

To prove this, we consider $K = \{x \in (\mathbb{R}^+)^d \colon \|x\| = 1\}$ and we set

$$m = \min_{x\in K} x^T \Sigma x \quad \text{and} \quad M = \max_{x\in K} x^T \Sigma x.$$

The min and max are well defined because $x \mapsto x^T \Sigma x$ is continuous on K compact. Inequalities (5.25) and (5.26) are clearly satisfied for all $x \in K$, and, by a standard homogeneity argument, they also holds for all $x \in (\mathbb{R}_+)^d$. Finally, m and M are non negative because Σ is positive semi-definite and the conditions $\mathrm{Ker}(\Sigma) \cap (\mathbb{R}_+)^d = \emptyset$ implies that m and M are positive. We now prove Lemma 5.27. The kernel k defined by Equation (5.20) is the sum of two kernels

$$k(x, y) = k_\Lambda(x, y) + k_{\Sigma,\alpha}(x, y) \tag{5.27}$$

with k_Λ defined in Equation (5.13) and $k_{\Sigma,\alpha}(x, y) = |x|^{\alpha T} \Sigma |y|^\alpha$. Therefore Lemma 5.17 implies - with straightforward notation - $\mathcal{M}_k = \mathcal{M}_\Lambda \cap \mathcal{M}_{\Sigma,\alpha}$. According to Proposition 5.18, $\mathcal{M}_1 \subset \mathcal{M}_\Lambda$. According to Equation (5.25),

$$0 \leq k_{\Sigma,\alpha}(x, x) = |x|^{\alpha T} \Sigma |x|^\alpha \leq M\|x\|^{2\alpha},$$

which implies $\mathcal{M}_\alpha \subset \mathcal{M}_{\Sigma,\alpha}$. Then, for $\alpha \geq 1$, the inclusion $\mathcal{M}_\alpha \subset \mathcal{M}_1$ implies

$$\mathcal{M}_\alpha = \mathcal{M}_1 \cap \mathcal{M}_\alpha \subset \mathcal{M}_\Lambda \cap \mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_k.$$

When $\mathrm{Ker}(\Sigma) \cap (\mathbb{R}_+)^d = \{0\}$, Equations (5.25) and (5.26) together imply

$$m\|x\|^{2\alpha} \leq k_{\Sigma,\alpha}(x, x) \leq M\|x\|^{2\alpha},$$

and $\mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_\alpha$. Then, for $\alpha \geq 1$, the inclusions $\mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_\alpha \subset \mathcal{M}_1 \subset \mathcal{M}_\Lambda$ imply

$$\mathcal{M}_k = \mathcal{M}_\Lambda \cap \mathcal{M}_{\Sigma,\alpha} = \mathcal{M}_\alpha.$$

□

The key ingredient of the Proposition 5.28 is this following lemma. Our proof is largely inspired by the proof of Theorem 9 of Sriperumbudur et al. (2010).

Lemma 5.40. *Let $U \subset \mathbb{R}^d \setminus \{0\}$ be a symmetric open set, i.e. $U = -U$, and $\alpha \geq 1$. There exists a real-valued Schwartz function $\theta \neq 0$ which has a null Fourier transform outside U and satisfies*

$$\int_{\mathbb{R}^d} \theta(x)\, \mathrm{d}x = 0 \quad \textit{and} \quad \int_{\mathbb{R}^d} |x_i|^\alpha \theta(x)\, \mathrm{d}x = 0, \quad 1 \leq i \leq d.$$

Proof. For $w \in \mathbb{R}^d$ and $\varepsilon \in (0, +\infty)^d$, we define the function

$$f_{w,\varepsilon}(\xi) = \prod_{i=1}^{d} \mathrm{e}^{-\frac{\varepsilon_i^2}{\varepsilon_i^2 - (\xi_i - w_i)^2}} \mathbb{1}_{[-\varepsilon_i, \varepsilon_i]}(\xi_i - w_i), \quad \xi \in \mathbb{R}^d.$$

Clearly, $f_{w,\varepsilon}$ is a Schwartz function with support equal to the hypercube $[w-\varepsilon, w+\varepsilon]$. Because U is open and symmetric, there exist $w_1, \ldots, w_{d+1} \in U$ and $\varepsilon \in (0, +\infty)^d$ such that the symmetric sets $[w_j - \varepsilon, w_j + \varepsilon] \cup [-w_j - \varepsilon, -w_j + \varepsilon]$, $1 \leq j \leq d+1$, are all included in U and pairwise disjoint. Then the Schwartz functions

$$\widehat{\theta}_j = f_{w_j,\varepsilon} + f_{-w_j,\varepsilon}, \quad 1 \leq j \leq d+1,$$

are symmetric with disjoint support included in U. As the Fourier Transform is a bijection on the Schwartz class, there is a unique Schwartz function θ_j with Fourier transform $\widehat{\theta}_j$, $1 \leq j \leq d+1$. Note that the functions $\theta_1, \ldots, \theta_{d+1}$ are linearly independent because their Fourier transforms $\widehat{\theta}_1, \ldots, \widehat{\theta}_{d+1}$ have disjoint support and thus are linearly independent. Furthermore, θ_j is real-valued because $\widehat{\theta}_j$ is symmetric and its integral vanishes because the condition $0 \notin U$ implies

$$\int_{\mathbb{R}^d} \theta_i(x)\, \mathrm{d}x = \widehat{\theta}_i(0) = 0.$$

The $d+1$ vectors in dimension d

$$\left(\int_{\mathbb{R}^d} |x_i|^\alpha \theta_j(x)\, \mathrm{d}x \right)_{1 \leq i \leq d} \in \mathbb{R}^d, \quad 1 \leq j \leq d+1,$$

are not linearly independent so that there exist $u_1, \ldots, u_{d+1} \in \mathbb{R}$, non all zero, such that

$$\sum_{j=1}^{d+1} u_j \int_{\mathbb{R}^d} |x_i|^\alpha \theta_j(x)\, \mathrm{d}x = 0 \quad \text{for all } 1 \leq i \leq d.$$

Then the function $\theta = \sum_{j=1}^{d} u_j \theta_j$ satisfies the required properties. It is non null because the functions $\theta_1, \ldots, \theta_{d+1}$ are linearly independent. □

Proof of Proposition 5.28. Recall the decompostion $k = k_\Lambda + k_{\Sigma,\alpha}$ in Equation (5.27).
If $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, we prove that the kernel k_Λ is characteristic over probability measures and hence k is also characteristic. The proof is similar to the proof of Theorem 9 in Sriperumbudur et al. (2010) and we recall only the key arguments. By Proposition 5.18, as $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d) = 1$

$$d_\Lambda^2(\mu, \nu) = 0 \quad \text{if and only if} \quad \int_{\mathbb{R}^d} |\hat{\mu}(\xi) - \hat{\nu}(\xi)|^2 \, \Lambda(\mathrm{d}\xi) = 0.$$

Since Λ has a full support and the integrand is continuous, we must have

$$\hat{\mu}(\xi) - \hat{\nu}(\xi) = 0 \quad \text{for all } \xi \in \mathbb{R}^d.$$

We deduce $\mu = \nu$, showing that k_Λ is characteristic over probability measures.

Conversely, we now suppose that $\mathrm{supp}(\Lambda) \neq \mathbb{R}^d$ and show that $k = k_\Lambda + k_{\Sigma,\alpha}$ is not characteristic. Let $U \subset \mathbb{R}^d \setminus \{0\}$ be a symmetric open set such that $\Lambda(U) = 0$. By Lemma 5.40, there exists a Schwartz function $\theta \neq 0$ such that

$$\int_{\mathbb{R}^d} \theta(x)\, \mathrm{d}x = 0, \quad \int_{\mathbb{R}^d} |x_i|^\alpha \theta(x)\, \mathrm{d}x = 0, \quad 1 \leq i \leq d,$$

and $\widehat{\theta}(x) = 0$ for $x \notin U$. Let $n \geq 1$ and $C > 0$, such that the measure

$$\mu(\mathrm{d}x) = \frac{C}{1 + \|x\|^n}\, \mathrm{d}x$$

is a probability measure with a finite absolute moment of order p. As θ is continuous and with a fast decay at infinity, there exists $u > 0$, such that the function $C(1 + \|x\|)^{-n} + u\theta(x)$ remains positive on $\mathbb{R}^d$. Then the measure

$$\nu(\mathrm{d}x) = \left(\frac{C}{1 + \|x\|^n} + u\theta(x) \right) \mathrm{d}x$$

is probability measure (recall that θ has a vanishing integral on $\mathbb{R}^d$). By the properties of θ, the measures μ and ν have the same absolute moment of order p :

$$\int_{\mathbb{R}^d} |x_i|^\alpha\, \mu(\mathrm{d}x) = \int_{\mathbb{R}^d} |x_i|^\alpha\, \nu(\mathrm{d}x), \quad 1 \leq i \leq d,$$

so that $m_\alpha(\mu) = m_\alpha(\nu)$ and $d^2_{\Sigma,\alpha}(\mu, \nu) = 0$ (see Equation (5.21)). Furthermore, they have the same Fourier transforms outside U, and together with $\Lambda(U) = 0$, this entails

$$d^2_\Lambda(\mu, \nu) = \int_{\mathbb{R}^d} |\widehat{\mu}(\xi) - \widehat{\nu}(\xi)|^2\, \Lambda(\mathrm{d}\xi) = 0.$$

We conclude that $d^2_k(\mu, \nu) = d^2_\Lambda(\mu, \nu) + d^2_{\Sigma,\alpha}(\mu, \nu) = 0$, so that the MMD is not a distance on $\mathcal{M}_k \cap \mathcal{P}$ and k is not characteristic. □

The following Lemma is the sequential version of the Equations (5.25) and (5.26).

Lemma 5.41. *Let F be a non empty closed linear cone and $\Sigma \in \mathcal{S}_d(\mathbb{R})$ be a non negative matrix,*

$$\ker \Sigma \cap F = \{0\} \iff \left[\forall (x_n)_n \in F^{\mathbf{N}},\ (x_n^T \Sigma x_n \to 0 \implies x_n \to 0) \right]$$

Proof. $\boxed{\Leftarrow}$ This implication is proved by contraposition. If $\ker \Sigma \cap F \neq \{0\}$ then let $y \neq 0$ in this intersection. Let $(x_n)_n$ be te constant sequence equal to y. This sequence checks $x_n^T \Sigma x_n \to 0$ but $x_n \not\to 0$.
$\boxed{\Rightarrow}$ Let $(x_n)_n \in F^{\mathbf{N}}$ such that $x_n^T \Sigma x_n \to 0$ then by the Equation (5.26),

$$0 \leq m\|x_n\|_2^2 \leq x_n^T \Sigma x_n \to 0,$$

where $m > 0$ then $x_n \to 0$. □

Proof of Theorem 5.30. $\boxed{\Rightarrow}$ This implication is proved by contraposition. If $\mathrm{supp}(\Lambda) \neq \mathbb{R}^d$, then by the Proposition 5.28, the MMD d_k is not a distance So the MMD cannot metrize the Wasserstein space.

If $\ker \Sigma \cap (\mathbb{R}^+)^d \neq \{0\}$, let $x \in (\mathbb{R}^+)^d$ be a non null vector such that $(|x|^p)^T \Sigma |x|^p = 0$. Let's define the sequence of probabilty measures $\mu_n = \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_{nx}$. It is easy to see that $W_p(\mu_n, \delta_0) \not\to 0$ since the moment of order p does not coverge. But $d^2_k(\mu_n, \delta_0) = \frac{1}{n^2} k_\Lambda(nx, nx)$ cause $|x|^p \in \ker \Sigma$ and $|x| = x$. Then by Lemma 5.35,

$$d^2_k(\mu_n, \delta_0) = o_n(1),$$

then it vanishes. So the MMD does not metrize the Wasserstein space of order p.

$\boxed{\Leftarrow}$ First of all, by the Lemma 5.27, $\mathcal{M}_k \cap \mathcal{P} = \mathcal{P}_p$. Let $(\mu_n)_{n\geq 1}, \mu \in \mathcal{P}_p$, it must be shown that $W_p(\mu_n, \mu) \to 0$ if and only if $d_k(\mu_n, \mu) \to 0$.

- if $W_p(\mu_n, \mu) \to 0$, then $m_p(\mu_n) \to m_p(\mu)$. Then by the Equation (5.25),

$$d_\Sigma(\mu_n, \mu) = \|m_p(\mu_n) - m_p(\mu)\|_\Sigma \to 0.$$

Moreover, as $(\mu_n)_{n\geq 1}$ (resp. μ) have a first moment, their Fourier Transforms are $\|m_1(\mu_n)\|_2$ (resp. $\|m_1(\mu)\|_2$)-Lipschitz continuous and as the convergence for W_p implies the convergence of W_1,

$$\|m_1(\mu_n)\|_2 \to \|m_1(\mu)\|_2.$$

Then these Fourier Transforms are C-Lipschitz continuous with $C := \sup(\|m_1(\mu_n)\|_2)$. Then

$$|\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2 \leq 4(1 \wedge C^2\|\xi\|^2) \in \mathrm{L}^1(\Lambda). \tag{5.28}$$

As $(\mu_n)_{n\geq 1}$ converges weakly to μ, their Fourier transforms converge to $\hat{\mu}$. By (5.28) and Dominated Convergence Theorem,

$$d^2_\Lambda(\mu_n, \mu) = \int_{\mathbb{R}^d} |\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2 \ \Lambda(\mathrm{d}\xi) \to 0.$$

Then $d^2_k(\mu_n, \mu) = d^2_\Lambda(\mu_n, \mu) + d^2_\Sigma(\mu_n, \mu) \to 0$.

- if $d_k(\mu_n, \mu) \to 0$, then $d_\Sigma(\mu_n, \mu) = \|m_p(\mu_n) - m_p(\mu)\|_\Sigma \to 0$ so by the Lemma 5.41

$$m_p(\mu_n) \to m_p(\mu).$$

Then the sequence $(\mu_n)_{n\geq 1}$ is tight by the Markov Inequality. Moreover as

$$d^2_\Lambda(\mu_n, \mu) = \int_{\mathbb{R}^d} |\hat{\mu_n}(\xi) - \hat{\mu}(\xi)|^2 \ \Lambda(\mathrm{d}\xi) \to 0,$$

the measure μ is the unique adherent value of the sequence $(\mu_n)_{n\geq 1}$ then by the Prokhorov's theorem, $(\mu_n)_{n\geq 1}$ converges weakly to μ. However, the weak convergence and the convergence of the absolute moment of order p implies the W_p convergence, then $W_p(\mu_n, \mu) \to 0$.

□

Proof of Subsection 3.5

The proof of the Proposition 5.32 is based on this lemma. We denote by $*$ the convolution product.

Lemma 5.42. *For $\varphi \in \mathcal{C}^0(\mathbb{R}^d, \mathbb{R})$ with $\|\varphi\|_{\mathrm{Lip}}$ and let F be a probability on $\mathbb{R}^d$, we have*

$$\int_{\mathbb{R}^d} \varphi * h_\sigma \ \mathrm{d}F = (\sqrt{2\pi})^{-d} \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \hat{f}(t) h_1(\sigma t) \exp(-iy \cdot t) \ \mathrm{d}t\mathrm{d}y,$$

where $\hat{f}$ is the characteristic function of F and $h_\sigma(x) = (\sigma\sqrt{2\pi})^{-d} \exp(-\|x\|_2^2/2\sigma^2)$.

Proof. The proof of this lemma is present in Ouvrard (2004). This equality is not directly written. So we will quickly prove the equality. One has

$$\begin{aligned}\int_{\mathbb{R}^d} \varphi * h_\sigma \ \mathrm{d}F &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi(y) h_\sigma(t-y) \ \mathrm{d}y F(\mathrm{d}t) \\ &= \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} h_\sigma(t-y) \ F(\mathrm{d}t)\mathrm{d}y,\end{aligned}$$

by the Fubini Theorem. The lemma 12.5 of this reference states

$$\int_{\mathbb{R}^d} h_\sigma(t-y) \ F(\mathrm{d}t) = (\sqrt{2\pi})^{-d} \int_{\mathbb{R}^d} \hat{f}(t) h_1(\sigma t) \exp(-iy \cdot t) \ \mathrm{d}t.$$

And so using this last equality, we get the desired result. □

Proof of Proposition 5.32. Let $\varphi\colon \mathbb{R}^d \to [-1,1]$ be a 1-Lipschitz continuous function bounded by the constant 1. Let $\mu, \nu \in \mathcal{T}$, we define

$$\mu(\varphi) = \int_{\mathbb{R}^d} \varphi(x)\ \mu(\mathrm{d}x) \quad \text{and} \quad \nu(\varphi) = \int_{\mathbb{R}^d} \varphi(x)\ \nu(\mathrm{d}x).$$

Let $h_\sigma(x) = (\sigma\sqrt{2\pi})^{-d}\exp(-\|x\|_2^2/2\sigma^2)$ denote the multivariate Gaussian density function with standard deviation $\sigma > 0$. We use an approximation argument and consider, for a sequence $\sigma_n \to 0$, the approximations

$$\mu_n(\varphi) = \int_{\mathbb{R}^d} \varphi * h_{\sigma_n}\ \mathrm{d}\mu \quad \text{and} \quad \nu_n(\varphi) = \int_{\mathbb{R}^d} \varphi * h_{\sigma_n}\ \mathrm{d}\nu.$$

Note that the convolution is well-defined because φ is bounded and h_{σ_n} is integrable. Since the function φ is 1-Lipschitz continuous, we have

$$\|\varphi - \varphi * h_{\sigma_n}\|_\infty \leq \int_{\mathbb{R}^d} \|y\|_2 h_{\sigma_n}(y)\ \mathrm{d}y = \sigma_n \times m_d \to 0, \tag{5.29}$$

where m_d is the absolute moment of a d dimensional standard gaussian. Let $\varepsilon > 0$ and $N \in \mathbb{N}$ be such that $\|\varphi - \varphi * h_{\sigma_N}\|_\infty < \varepsilon$ then

$$\big|\mu(\varphi) - \mu_N(\varphi)\big| \leq \varepsilon \quad \text{and} \quad \big|\nu(\varphi) - \nu_N(\varphi)\big| \leq \varepsilon,$$

whence we deduce

$$\big|\mu(\varphi) - \nu(\varphi)\big| \leq \big|\mu_N(\varphi) - \nu_N(\varphi)\big| + 2\varepsilon. \tag{5.30}$$

Next, we introduce the characteristic function $\hat{\mu}$ (resp. $\hat{\nu}$) of μ (resp. ν). By Lemma 5.42

$$\mu_N(\varphi) = (\sqrt{2\pi})^{-d} \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \hat{\mu}(t) h_1(\sigma_N t) e^{-iy\cdot t}\ \mathrm{d}t\mathrm{d}y,$$

and the same equality holds for $\nu_N(\varphi)$ with $\hat{\mu}$ replaced by $\hat{\nu}$. Taking the difference, we get

$$\big|\mu_N(\varphi) - \nu_N(\varphi)\big| = \Big| \Big(\sqrt{2\pi}\Big)^{-d} \int_{\mathbb{R}^d} \varphi(y) \int_{\mathbb{R}^d} \big(\hat{\mu}(t) - \hat{\nu}(t)\big) h_1(\sigma_N t) e^{-iy\cdot t}\ \mathrm{d}t\mathrm{d}y \Big|.$$

Assuming that φ has compact support included in the ball $B(0,K)$ with center 0 and radius K, noted shortly $\mathrm{supp}(\varphi) \subset B(0,K)$, we deduce

$$\begin{aligned}
\big|\mu_N(\varphi) - \nu_N(\varphi)\big|^2 &\leq (2\pi)^{-d} \lambda_d(B(0,K))^2 \Big[\int_{\mathbb{R}^d} \big|\hat{\mu}(t) - \hat{\nu}(t)\big| h_1(\sigma_N t)\ \mathrm{d}t \Big]^2 \\
&\leq (2\pi)^{-d} \lambda_d(B(0,K))^2 \times \int_{\mathbb{R}^d} \|t\|^{d+2\alpha} h_1^2(\sigma_N t)\ \mathrm{d}t \times \int_{\mathbb{R}^d} \frac{|\hat{\mu}(t) - \hat{\nu}(t)|^2}{\|t\|^{d+2\alpha}}\ \mathrm{d}t \\
&= C^2 \times d_\alpha^2(\mu,\nu),
\end{aligned} \tag{5.31}$$

where C does not depend to μ and ν. In order to prove the result for the Fortet-Mourier distance, we have to remove the support constraint. The tightness of the subset $\mathcal{T}$ will be useful for this purpose. For any $\varepsilon > 0$, we can choose $K > 1$ such that $\mu(B(0,K)^c) < \varepsilon$ for all $\mu \in \mathcal{T}$. Let $\chi : \mathbb{R}^d \to \mathbb{R}$ be the 1-Lipschitz continuous function defined as in Lemma 5.39,

$$\chi(x) = \mathbb{1}_{\|x\|_2 \leq K-1} + (K - \|x\|_2)\mathbb{1}_{K-1<\|x\|_2<K}.$$

The decomposition $\varphi = \chi\varphi + (1 - \chi\varphi)$ implies

$$\begin{aligned}
|\mu(\varphi) - \nu(\varphi)| &\leq |\mu(\chi\varphi) - \nu(\chi\varphi)| + |\mu((1-\chi)\varphi) - \nu((1-\chi)\varphi)| \\
&\leq 2|\mu(\chi\varphi/2) - \nu(\chi\varphi/2)| + 2\varepsilon.
\end{aligned}$$

where $\chi\varphi/2$ is 1-Lipschitz continuous with values in $[-1,1]$ and support included in $B(0,K)$. Taking the supremum over the 1-Lipschitz continuous function $\varphi : \mathbb{R}^d \to [-1,1]$, we get

$$d_{FM}(\mu,\nu) \leq 2 \sup_{\mathrm{supp}(\varphi)\subset B(0,K)} |\mu(\varphi) - \nu(\varphi)| + 2\varepsilon. \tag{5.32}$$

By combining, the Equation (5.30) to (5.32),

$$d_{FM}(\mu,\nu) \leq 2C d_\alpha(\mu,\nu) + 4\varepsilon$$

□

Proof of the Proposition 5.33. Note that by the dual representation of the Wasserstein distance W_1, we can consider the supremum over 1-Lipschitz function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ and $\varphi(0) = 0$. Such functions satisfy $|\varphi(x)| \leq \|x\|$. Let $\varepsilon > 0$. By the definition of uniformly integrability, there is $K > 1$ such that

$$\int_{B(0,K)^c} \|x\| \ \mu(\mathrm{d}x) < \varepsilon \quad \text{for all } \mu \in \mathcal{T}.$$

Recall the function χ from the proof of Proposition 5.32 which is 1-Lipschitz continuous and such that

$$\mathrm{supp}(\chi) \subset B(0,K+1),\ 0 \leq \chi \leq 1 \text{ and } \chi \equiv 1 \text{ on B(0,K)}.$$

Then, in the decomposition $\varphi = \chi\varphi + (1-\chi)\varphi$, the function $\chi\varphi$ is uniformly bounded and $(K+2)-$Lipschitz with $\|\chi\varphi\|_\infty \leq K+1$. We deduce

$$\Big| \int_{\mathbb{R}^d} \chi\varphi \ \mathrm{d}\mu - \int_{\mathbb{R}^d} \chi\varphi \ \mathrm{d}\nu \Big| \leq (K+2) d_{FM}(\mu,\nu).$$

On the other hand, since $(1-\chi)\varphi$ vanishes on $B(0,K)$ and is bounded from above by the norm of x,

$$\Big| \int_{\mathbb{R}^d} (1-\chi)\varphi \ \mathrm{d}\mu - \int_{\mathbb{R}^d} (1-\chi)\varphi \ \mathrm{d}\nu \Big| \leq \varepsilon.$$

We deduce

$$W_1(\mu,\nu) \leq (K+2) d_{FM}(\mu,\nu) + \varepsilon.$$

Finally, Proposition 5.32 implies the desired inequality. □

Perspectives

Les perspectives de cette thèse concernent les trois derniers chapitres. Certaines de ces perspectives ont déjà été discutées dans les articles concernés. Nous les rappelons tout de même ici pour les réunir au même endroit.

Calibration : Nous avons élaboré dans cet article une famille de tests permettant de savoir si une prévision F prédisait un événement Y de manière parfaite avec l'information $\mathcal{F}$. C'est-à-dire que la prédiction probabiliste F est la loi conditionnelle de Y sachant $\mathcal{F}$. Pour construire nos tests, nous avons réécrit cette condition en terme de processus empirique. Malheureusement, en pratique cette manière de prédire est quasiment impossible. Ce critère n'est donc pas forcément pertinent pour définir ce qu'est une bonne prédiction. Plusieurs autres propriétés ont alors été introduites. Dans la Section 6 du Chapitre 3, nous avions aussi réécrit l'une des ces propriétés sous la forme de processus empirique. Ainsi, il est peut-être envisageable de prolonger la technique décrite dans l'article à cette nouvelle condition plus faible, et donc potentiellement plus réaliste dans les applications.

Théorème de Stone : En reprenant les notations du Chapitre 4, les poids $W_n(X; X_1, \dots, X_n)$ ne dépendaient que des variables explicatives $(X_i)_{1\leq i\leq n}$. Dans Elie-Dit-Cosaque and Maume-Deschamps (2022), les auteurs ont considéré des poids calculés à partir d'arbres de régression. Ainsi, les poids dépendaient aussi des observations $(Y_i)_{1\leq i\leq n}$. L'estimateur de la loi conditionnelle construite ainsi était uniformément consistant à x fixé. Un travail possible est de voir si notre consistance au sens de Wasserstein s'applique à ce nouveau cadre.

RKHS : Dans ce Chapitre, nous avons réussi à construire de nouvelles distances sur l'espace des probabilités de $\mathbb{R}^d$ et nous avons montré ces liens avec les distances de Wasserstein. Nous donnons ici plusieurs directions de perspectives.

- *transport optimal.* Il pourra être intéressant de comprendre un peu mieux la géométrie de ces distances en l'appliquant au domaine des EDP. Certaines preuves d'unicité en EDP se basent sur l'étude de la distance de Wasserstein de deux solutions. Pour l'instant, la convergence au sens de Wasserstein se fait à l'aide de la caractérisation par la convergence en loi et du moment absolue.
- *améliorer les résultats.* L'étude n'a été faite que sur des familles particulières de noyaux définis sur $\mathbb{R}^d$. L'amélioration peut se faire en trouvant de nouvelles familles sur $\mathbb{R}^d$ ou des espaces topologiques quelconques, ou bien en donnant des conditions nécessaires ou suffisantes, voire les deux en même temps.
- *appliquer ces distances à des problèmes de statistique.* La première application possible est le *Generative Adversarial Network* (GAN). Le but est de simuler une variable X d'une loi $\mathbb{P}$ en ayant observé un échantillon i.i.d. de cette même loi. Une autre application s'inspire de Hallin et al. (2021). Les auteurs relient les distances de Wasserstein à des tests d'adéquation. On peut espérer pouvoir obtenir des résultats équivalents avec nos distances. Si il est possible de le faire, les preuves seront sans doute plus simples car le MMD a l'avantage d'avoir une écriture explicite.

Annexe A

Appendice

1 Measurability of the score

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

Definition A.1. *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 A.2. *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 scoring rule.*

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 A.3. If $(\mathcal{Y}, d)$ is separable, the space of probability measure $\mathcal{M}_1(\mathcal{Y})$ 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 3.3. Moreover, a subset of a separable metric space remains separable.

In the Section 4, we introduced the family of kernel score. We will show the continuity of these scores for the Wasserstein distance introduced in Section 3.3. We need another writing of this distance, for $F, G \in \mathcal{P}_1(\mathbb{R}^d)$,

$$W_1(F, G) = \inf_{X \sim F, Y \sim G} \mathbb{E}[\|X - Y\|], \tag{A.1}$$

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

Proposition A.4. *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

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

As we do not do assumption on the dependence between X (resp. X') and Z (resp. Z'), we conclude with the formulation (A.1) of the Wasserstein distance that

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

Then

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

□

Proposition A.5. *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 (2.6). 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 $k \in \{1, \ldots, T\}$ and $C > 0$ be a Lipschitz constant of the kernel ρ, let's remember the definition of

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

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

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

The Proposition A.4 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 (A.1),

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

So the condition (2.6) is checked. □

2 Conditional distributions

In this Appendix, we describe briefly the theory of conditional distributions. The proof of the first general result can be found in Stroock (2010) section 9.2, and the other results are consequences of this result.

Theorem A.6. *Suppose that Ω is a Polish space and that $\mathcal{G}$ is the associated Borel σ-algebra. Let $\mathbb{P}$ be a probability measure on this σ-algebra. Then for every sub-σ-algebra $\mathcal{F}$ of $\mathcal{G}$, there is a kernel*

$$\mu^{\mathcal{F}} \colon \Omega \times \mathcal{F} \to [0, 1],$$

such that for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$,

$$\mathbb{P}(A \cap B) = \int_A \mu^{\mathcal{F}}(\omega, B) \, \mathbb{P}(\mathrm{d}\omega).$$

Moreover, the kernel $\mu^{\mathcal{F}} \colon \Omega \to \mathcal{M}_1(\Omega)$ is almost surely unique.

Thanks to this general result, the conditional distribution $\mathcal{L}(Y \mid \mathcal{F})$ can be defined and used for the definition of ideal forecast.

Corollary A.7. *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a Polish probability space, $\mathcal{F}$ be a sub-σ-algebra and let Y be a random variable with value in $(C, \mathcal{C})$. There exists an almost surely unique kernel*

$$\mu_Y^{\mathcal{F}} \colon \Omega \to \mathcal{M}_1(C),$$

such that for all $A \in \mathcal{F}$ and $B \in \mathcal{C}$,

$$\mathbb{P}(A \cap \{Y \in B\}) = \int_A \mu_Y^{\mathcal{F}}(\omega, B) \, \mathbb{P}(\mathrm{d}\omega).$$

The conditional distribution between two random variables $\mathcal{L}(Y \mid X)$ can also be derived in a similar way, as stated below.

Corollary A.8. *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space (not necessary Polish), X and Y be two random variables, respectively on $(E, \mathcal{E})$ a Polish space and on $(\mathcal{C}, C)$ a measurable space. There is a $\mathbb{P}_X$ almost surely unique kernel*

$$\mu_Y^X \colon E \to \mathcal{M}_1(C),$$

such that for all $A \in \mathcal{E}$ and $B \in \mathcal{C}$,

$$\mathbb{P}(X \in A, Y \in B) = \int_A \mu_Y^X(x, B) \, \mathbb{P}_X(\mathrm{d}x).$$

Let X, Y be random variables and $\mathcal{F}$ be a σ-algebra such that $\mathcal{F} = \sigma(X)$. There is a link between $\mu_Y^{\mathcal{F}}$ and μ_Y^X, which is explained in the following proposition.

Proposition A.9. *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a Polish probability space, $\mathcal{F}$ be a sub-σ-algebra and let X, Y be two random variables, respectively on $(E, \mathcal{E})$ a Polish space and on $(\mathcal{C}, C)$ a measurable space. If $\mathcal{F}$ is generated by X then*

$$\mu_Y^{\mathcal{F}}(\omega, \cdot) = \mu_Y^X(X(\omega), \cdot).$$

To conclude this appendix, let detail the expression of the conditional expectation, which follows from what precedes.

Proposition A.10. *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a Polish probability space, $\mathcal{F}$ be a sub-σ-algebra and let Y be a random variable with value in $(C, \mathcal{C})$. Let h be a measurable map between $(\Omega, \mathcal{F})$ and $(A, \mathcal{A})$ and f be a real measurable map defined on $(A \times C, \mathcal{A} \otimes \mathcal{C})$. If $f(h, Y) \in \ell^1$ then*

$$\mathbb{E}\left[f(h, Y) \mid \mathcal{F}\right] = \int_C f(h(\omega), y)\ \mu_Y^{\mathcal{F}}(\omega, \mathrm{d}y).$$

Bibliographie

Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837.

Anderson, J. L. (1996). A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *Journal of Climate*, 9(7) :1518–1530.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Auricchio, G., Codegoni, A., Gualandi, S., Toscani, G., and Veneroni, M. (2020). The equivalence of fourier-based and wasserstein metrics on imaging problems. *Rendiconti Lincei - Matematica e Applicazioni*.

Azaïs, R. and Ingels, F. (2020). The weight function in the subtree kernel is decisive. *Journal of Machine Learning Research*, 21(67) :1–36.

Bayraktar, E. and Guo, G. (2021a). Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none) :1 – 13.

Bayraktar, E. and Guo, G. (2021b). Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26 :1 – 13.

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. With a preface by Persi Diaconis.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

Bobkov, S. and Ledoux, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Mem. Amer. Math. Soc.*, 261(1259) :v+126.

Breiman, L. (2001). Random forests. *Machine Learning*, 45.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York, NY.

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. (2022). Uniform calibration tests for forecasting systems with small lead time. *Statistics and Computing*, 32(6).

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, A. (2007). Universal residuals : A multivariate transformation. *Statistics and Probability Letters*, 77(14) :1473 – 1478.

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

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

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA. PMLR.

Cohen, S. and Istas, J. (2013). *Fractional Fields and Applications*. Mathématiques et Applications. Springer. volume 76.

Costa, G., Manco, G., Ortale, R., and Tagarelli, A. (2004). A tree-based approach to clustering xml documents by structure. In Boulicaut, J.-F., Esposito, F., Giannotti, F., and Pedreschi, D., editors, *Knowledge Discovery in Databases : PKDD 2004*, pages 137–148, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cox, D. D. and Lee, J. S. (2008). Pointwise testing with functional data using the westfall-young randomization method. *Biometrika*, 95(3) :621–634.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons, Inc., New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication.

David, F. N. and Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35(1/2) :182–190.

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.

Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S., and Prieur, C. (2007). *Weak dependence : with examples and applications*, volume 190 of *Lectures Notes in Statistics*. Springer.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4) :863–883.

Diestel, J. and Uhl, Jr., J. J. (1977). *Vector measures*. American Mathematical Society, Providence, R.I. With a foreword by B. J. Pettis, Mathematical Surveys, No. 15.

Dombry, C., Modeste, T., and Pic, R. (2023). Stone's theorem for distributional regression in Wasserstein distance. working paper or preprint.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 258–267, Arlington, Virginia, USA. AUAI Press.

Elie-Dit-Cosaque, K. and Maume-Deschamps, V. (2022). Random forest estimation of conditional distribution functions and conditional quantiles. *Electronic Journal of Statistics*, 16(2) :6553 – 6583.

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

Epstein, E. S. (1969b). Stochastic dynamic prediction1. *Tellus*, 21(6) :739–759.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. (2015). Learning with a wasserstein loss. *Advances in Neural Information Processing Systems (NIPS)*, 28.

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

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, 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.

Gneiting, T. and Resin, J. (2021). Regression diagnostics meets forecast evaluation : Conditional calibration, reliability diagrams, and coefficient of determination.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments : Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5) :1049–1054.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25) :723–773.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70) :2075–2129.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer.

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

Hallin, M., Mordant, G., and Segers, J. (2021). Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15(1) :1328 – 1371.

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

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, 66.

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

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021b). Isotonic distributional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 83(5) :963–993.

Herbin, E. and Merzbach, E. (2007). *The Multiparameter Fractional Brownian Motion*, pages 93–101. Springer Berlin Heidelberg, Berlin, Heidelberg.

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.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. (2016). Probabilistic energy forecasting : Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3) :896–913.

Jolliffe, I. T. and Primo, C. (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6) :2133 – 2139.

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

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33(2) :270–281.

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

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized sliced wasserstein distances. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Le, S.-Y., Nussinov, R., and Maizel, J. V. (1989). Tree graphs of rna secondary structures and their comparisons. *Computers and Biomedical Research*, 22(5) :461–473.

Li, R., Reich, B. J., and Bondell, H. D. (2021). Deep distribution regression. *Computational Statistics & Data Analysis*, 159 :107203.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1718–1727. JMLR.org.

Martín-Delgado, M. A., Rodriguez-Laguna, J., and Sierra, G. (2002). Density-matrix renormalization-group study of excitons in dendrimers. *Phys. Rev. B*, 65 :155116.

Mitchell, J. and Wallis, K. (2011). Evaluating density forecasts : forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, 26(6) :1023–1040.

Modeste, T. and Dombry, C. (2022). Characterization of translation invariant MMD on R d and connections with Wasserstein distances. working paper or preprint.

Modeste, T., Dombry, C., and Fougères, A.-L. (2023). Modeling and scoring dynamic probabilistic forecasts. working paper or preprint.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14(1) :24 – 49.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2) :429–443.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1) :141–142.

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

Ouvrard, J.-Y. (2004). *Probabilité 2*. Cassini.

Panaretos, V. M. and Zemel, Y. (2020). *An invitation to statistics in Wasserstein space*. SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham.

Pic, R., Dombry, C., Naveau, P., and Taillardat, M. (2022). Distributional regression and its evaluation with the crps : Bounds and convergence of the minimax risk. *International Journal of Forecasting*.

Pohle, M.-O. (2020). The Murphy Decomposition and the Calibration-Resolution Principle : A New Perspective on Forecast Evaluation. Papers 2005.01835, arXiv.org.

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

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408.

Rosenblatt, M. (1961). Independence and dependence. In Neyman, J., editor, *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 2, pages 431–443. University of California Press, Berkeley, CA. (Berkeley, CA, 20–30 July 1960). Zbl :0105.11802. MR :133863.

Rubner, Y., Tomasi, C., and Guibas, L. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40 :99–121.

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. (2021). Machine learning methods for postprocessing ensemble forecasts of wind gusts : A systematic comparison. arXiv:2106.09512.

Scornet, E. (2016). On the asymptotics of random forests. *J. Multivariate Anal.*, 146 :72–83.

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).

Simon-Gabriel, C.-J., Barp, A., Schölkopf, B., and Mackey, L. (2021). Metrizing weak convergence with maximum mean discrepancies. (under review).

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings : Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44) :1–29.

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.

Smola, A. and Vishwanathan, S. (2002). Fast kernels for string and tree matching. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Sriperumbudur, B. (2016). On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3).

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 :2389–2410.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50) :1517–1561.

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.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.*, 5(4) :595–645. With discussion and a reply by the author.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6) :1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4) :1040–1053.

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

Stroock, D. (2010). *Probability Theory : An Analytic View (2nd ed.)*. Cambridge : Cambridge University Press.

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

Stute, W. (1986). On Almost Sure Convergence of Conditional Empirical Distribution Functions. *The Annals of Probability*, 14(3) :891 – 901.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*.

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

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4) :1236 – 1265.

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*, in press.

Taillardat, M., Fougères, A.-L., Naveau, P., and Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34.

Talagrand, O., Vautard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. *Workshop on Predictability, 20-22 October 1997*, pages 1–26.

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*.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes : With Applications to Statistics*. Springer Series in Statistics. Springer.

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.

Vayer, T. and Gribonval, R. (2021). Controlling Wasserstein distances by Kernel norms with application to Compressive Statistical Learning. working paper or preprint.

Villani, C. (2003). *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society.

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

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A (1961-2002)*, 26(4) :359–372.

Weigel, A. P. (2011). *Ensemble Forecasts*, chapter 8, pages 141–166. John Wiley & Sons, Ltd.

Yaglom, A. and Silverman, R. (1962). *An Introduction to the Theory of Stationary Random Functions*. Selected Russian publications in the mathematical sciences. Prentice-Hall.

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

Ćevid, D., Michel, L., Näf, J., Meinshausen, N., and Bühlmann, P. (2022). Distributional random forests : Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333) :1–79.

Évaluation et construction des prévisions probabilistes : score et calibration dans un cadre dynamique

Résumé : La prédiction est un domaine central en statistique. Pour prendre en compte toutes les incertitudes qui lui sont associées, la prédiction prend la forme d'une mesure de probabilité aléatoire, appelée prévision probabiliste. Plusieurs scénarios sont prédits, associés à chaque fois à une probabilité de réalisation. Évaluer la qualité d'une prévision probabiliste est une tâche délicate. Il faut comparer une prévision, mesure aléatoire, à une observation, qui sont deux objets différents. Parmi les différentes prévisions probabilistes, l'une d'entre elles se démarque. Il s'agit de la distribution conditionnelle. Elle correspond à la prévision parfaite sachant une certaine information.
Le premier outil étudié est le score. Il compare de manière directe et quantitative la mesure et la réalisation du phénomène que l'on souhaite prédire. Nous montrerons que le score permet de discriminer la prédiction parfaite dans un cadre séquentiel sans hypothèse de stationnarité. La seconde approche est plus qualitative. L'idée de la calibration est de définir plusieurs propriétés que doit vérifier une prévision de bonne qualité. Nous proposons de nouveaux tests basés sur les arbres de régression pour déterminer si une prévision est parfaite ou non.
Cette prévision particulière peut être approchée de manière classique par des moyennes pondérées des observations. Nous prolongeons le résultat du Théorème de Stone concernant l'estimation de l'espérance conditionnelle. Nous montrons que la distribution conditionnelle estimée par moyenne pondérée est consistante pour la distance de Wasserstein.
Cette distance sera la vedette de la dernière partie de cette thèse, où nous construirons de nouvelles distances qui lui seront topologiquement équivalentes. Cette construction sera basée sur la théorie des noyaux reproduisants en plongeant l'espace des mesures dans un espace de Hilbert.
Mots clés : Prédiction probabiliste, Score, Calibration, Théorème de Stone, Distance de Wasserstein, RKHS, MMD

Assessing and construction of probabilistic forecast : scoring rules and calibration in a dynamic framework

Abstract : Prediction is a central domain in statistics. To take into account all the uncertainties of a prediction, this latter takes the form of a random probability measure, called a probabilistic forecast. Several scenarios are predicted, each of them are associated with a probability of success. Assessing the quality of a probabilistic forecast is a delicate task. We have to compare a forecast, a random measure, to an observation, which are two different objects. Among the different probabilistic forecasts, the conditional distribution is the best one. It corresponds to the perfect forecast knowing a certain information.
The first tool studied is the score. It compares directly and quantitatively the measure and the realization of the phenomenon. We will prove that the score allows to discriminate the perfect prediction in a sequential framework without stationary condition. The second approach is more qualitative. The idea of calibration is to define several properties that sould be satisfied by a *good* prediction. Several tests based on regression trees (CART Algorithm) will be presented to determine if a forecast is ideal or not.
The ideal forecast can be classically estimated by weighted averages of observations. We extend the Stone's Theorem concerning the estimation of the conditional expectation. We show that the conditional distribution estimated using weighted averages is consistent for the Wasserstein distance.
The last part of this thesis is focused on this distance. New distances, topologically equivalent to the Wasserstein distance, will be constructed. This construction will be based on the theory of reproducing kernels by embedding the space of measures in a Hilbert space.
Keywords : Probabilistic Forecast, Scoring Rules, Calibration, Stone's Theorem, Wasserstein Distance, RKHS, MMD

UNIVERSITÉ DE LYON