Machine Learning Algorithms for Regression and Global Optimization of Risk Measures

Léonard Torossian

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *17 décembre 2019* **par :**

Léonard Torossian

Méthodes d'apprentissage statistique pour la régression et l'optimisation globale de mesures de risque

---

JURY

Luc Pronzato	CNRS Nice-Sophia Antipolis	Rapporteur
Odalric-Ambrym Maillard	INRIA Lille - Nord Europe	Rapporteur
Robert Faivre	INRA Toulouse	Directeur de thèse
Aurélien Garivier	ENS Lyon	Directeur de thèse
Victor Picheny	Prowler.io Cambridge	Encadrant
Sébastien Da Veiga	Safran Tech Saclay	Examinateur
Béatrice Laurent-Bonneau	INSA Toulouse	Examinateur
Céline Helbert	Ecole Centrale Lyon	Examinateur

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :
INRA-MIAT (UR875), Institut de Mathématiques de Toulouse (UMR 5219)

Directeur(s) de Thèse :
Aurélien Garivier, Robert Faivre et *Victor Picheny (Encadrant)*

Rapporteurs :
Odalric-Ambrym Maillard et *Luc Pronzato*

À mes grands-parents:

Simone, France, René et Maurice

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Robert Faivre, Aurélien Garivier et Victor Picheny pour m'avoir guidé et soutenu tout au long de ces trois années. Je me sens chanceux et fier d'avoir appris à conduire un projet scientifique sous votre direction. Robert, ce fut un vrai plaisir de discuter de la philosophie des méthodes statistiques et de découvrir ce sujet de thèse passionnant avec toi. Aurélien, je garderai un excellent souvenir de mes visites à Lyon, travailler ensemble face aux tableaux à craies était vraiment génial. Ton investissement quasi surhumain dans le monde académique restera une source de motivation pour moi. Victor, je te suis profondément reconnaissant pour ces trois années. Tes idées originales, ton suivi quasi quotidien et ta bonne humeur m'ont permis de ne jamais (ou très peu) perdre le fil de ce travail. Je n'oublierai pas ton pragmatisme lorsqu'il s'agit de traiter une question scientifique.

Je souhaite également remercier Luc Pronzato et Odalric-Ambrym Maillard pour avoir rapporté cette thèse. Merci à Béatrice Laurent-Bonneau, Céline Helbert et Sébastien Da Veiga d'avoir accepté de faire partie de mon jury. Sébastien je te suis également reconnaissant pour les nombreux échanges que l'on a pu avoir pendant ces trois années. Enfin, merci à tout le jury d'avoir bravé les incertitudes de la grève et d'être arrivé (entier ?).

Pendant ces trois ans j'ai eu la chance de baigner dans un environnement scientifique riche et stimulant. Pour cela, côté MIAT, je souhaite remercier : Ronan pour ton enthousiasme et pour ta disponibilité ; Pierre pour les discussions passionnantes autour de l'agronomie ; Nathalie P. pour m'avoir initié en douceur aux méthodes variationnelles ; Marie-Jo et Matthias pour votre bienveillance ; Damien pour tes tutos informatiques, grâce à toi j'impressionne les utilisateurs Windows avec Xkill ; Sylvain pour ton habileté à mener l'unité d'une main de fer dans un gant de velours ; Régis pour ton investissement sans faille dans l'amical du flag MIAT ; Patrick C., Hélène et Simon pour votre gentillesse, c'était toujours une joie de discuter avec vous ; Thomas pour ton lifehack qui me permettra de boire de l'eau fraîche si un jour je me trouve dans la Vallée de la Mort ; Frédéric pour les discussions passionnantes en allant à Digitag et puis tous les autres David ; Nathalie V., Christine, Annick, Patrick.T, Roger, Stéphane, Michaël pour faire de ce labo un lieu plaisant à vivre. Côté MIAT toujours, je n'oublie pas le trio Alain, Nathalie J., Fabienne ô combien compétent qui tous les jours nous rend la vie plus facile. Je vous remercie d'avoir été présents avec votre bonne humeur pendant mon séjour dans l'unité. Coté IMT j'aimerais remercier : Edouard P., je garderai un

souvenir impérissable de ton mini cours sur l'optimisation et de ta réactivité à mes questions ; Sébastien.G pour l'exemple de rigueur que tu me donnes depuis maintenant 5 ans; Francois B. et Gersende pour nos discussions toujours passionnantes. Et enfin côté Sorbonne Université, merci beaucoup Gérard pour ton soutien dans certains moments clefs de mon parcours scientifique, pour ta disponibilité et sourtout pour l'exemple de pédagogie et de rigueur dont nous profitons lors de chacune de tes présentations.

To the people of Prowler.io I want to address a very big thank you. Thank you Vincent.D, I hope we will play chess together soon. Thank you Alan for introducing me to the Chain GP and for showing me the English countryside. I will remember that bouldering on gritstone is not easy especially if it is rainy. Thank you Sergio for being a wonderful lunch partner. Thank you Vincent.A the future 7a's crusher, I hope your double sparse technique will be recognized by the whole world some days. Thank you Nicolas for the discussions on the black board, it was very fun. Thank you Artem, James, Ti, Mark, Stephanos, Jordi, Peter, Egor, Sergio V. and all the team for your expertise, kindness and availability.

Ma pensée va également à tous les doctorants et post-doctorants avec qui j'ai partagé un bout de chemin. Je souhaite remercier : Clément V. pour les pauses déjeuner à parler de SNK, les marches digestives et les tutos "pour devenir quelqu'un" ; David G. pour l'enthousiasme avec lequel tu partages avec moi les secrets de ton organisation à l'Allemande, j'ai malheureusement encore beaucoup de travail pour arriver à rivaliser avec toi ; Adrien pour les bons moments passés en conf' et à Toulouse, j'espère que tu garderas un bon souvenir du canal du midi ; thank you Denis for sharing your passion about nature and for helping me to find the city that fits me the best ; Pierre pour les discussions sur les compétitions de ML, un jour on gagnera; ; Jelena pour m'avoir laissé de la place sur le podium du concours JCF; Sara, Lise, Ivana, Fulya, Manon, Ludovic, Jérome, Eva, Anouar, Baptiste pour tous nos échanges.

Ces trois années ont également été passionnantes sportivement notamment grâce au groupe d'escalade de Paul Sabatier. Pour cela je souhaite remercier en premier lieu Max pour la dynamique qu'il a créé autour de l'escalade à la fac et pour la qualité de ses entraînements. Mes remerciements vont ensuite à l'ensemble de l'équipe : merci Marylou de nous partager ton expertise de la grimpe statique ; Fantine pour avoir toujours rigolé à mes blagues, pour ton éternelle bonne humeur et surtout ton implication dans la team trois mousquetons ; Lisa pour le côté quali que tu apportes partout avec toi ; Agnès pour ta discrétion quand tu me ridiculises sur les blocs ; merci Pierrick de nous montrer que manger 500g de fromage par jour et faire 8b+ n'est pas incompatible ; Boris pour le soutien de mes goûts musicaux et pour ton expertise sur les volumes ; Paul pour les méthodes et ta faculté à grimper avec classe tout en portant la moustache ; Valentin et Théo pour les bons moments passés en salle et en extérieur, j'espère que nous aurons l'occasion de croîter en terre bellifontaine d'ici peu ; merci à Hugo, Alexandre, Julien P., Étienne, Margaux et toutes les autres machines pour nous montrer la voie.

Je tiens a remercier deux personnes qui ont été très importantes pour moi pendant cette thèse. L'un est Gentleman-Rider[1], l'autre était mon voisin de bureau à l'IMT.

[1]Titre officiel indispensable pour peser un maximum sur les hippodromes

Ces trois années à partager la grimpe, la cuisine et vos appartements étaient vraiment plaisantes. Alexis et Julien merci à vous deux pour ces excellents moments !

Certaines amitiés se sont créées, d'autres étaient déjà présentes et se sont renforcées. Merci à Emeric et Maxence, tout a démarré au basket mais tout s'est accéléré en terminale. En écrivant ces lignes je me remémore tellement de bons moments qu'il me tarde de vivre les prochains. Merci à Clément pour ces longues années de rigolade, de culture physique et d'échanges très sérieux sur OSS117 et Dikkenek, j'espère que tu trouveras ce que tu cherches dans ton nouveau métier. Merci à Coriolan pour accepter d'être mon partenaire de baroude et un homme de l'ombre aux nombreuses compétences quand j'en ai besoin. J'attends le double Half Cab Roll de pied ferme ! Enfin merci Matthieu, je pense que nos échanges et discussions ont initié ma volonté de me lancer dans un doctorat. Discuter avec toi est vraiment enrichissant, j'espère pouvoir continuer à pousser nos réflexions longtemps. Je n'ai qu'une chose à dire : que ton chemin soit mon déclin.

Je ne vois pas le doctorat comme l'aboutissement de trois ans de travail mais comme le résultat d'un long processus. C'est pourquoi j'aimerais remercier quelques enseignants qui m'ont beaucoup apporté pendant mes années scolaires : merci à M. Lambert avec qui j'ai découvert les joies des mathématiques au travers du calcul d'aires et périmètres de polygones ; merci à Mme Duvelle pour m'avoir introduit à la rigueur des mathématiques ; merci à M. Chollet pour m'avoir enseigné les mathématiques avec une passion mêlée à un soupçon de cynisme pendant deux années ; merci à Dr. Franses pour m'avoir conté l'algèbre linéaire avec virtuosité ; merci à M. Normand qui m'a fait apprécier l'histoire autant que les échecs.

Parce que je ne suis pas le plus organisé, j'ai quelques remerciements en vrac : merci à Antho le rigolo de partager avec moi sa passion pour les automobiles de qualité #voyagevoyage ; merci à Emeline et Axel pour tous ces bons moments passées sur les stades d'athlé et dans la coloc #teamplaisir ; merci à Pauline et Adèle pour votre générosité et pour soutenir ma passion de la chocolatine ; merci Nadine pour ta sincérité et ta passion ; merci Fabio pour ton amour de la gastronomie ; merci Aïcha pour ton pragmatisme et ta détermination ; merci Sébastien, Nicolas, François, Benjamin, Kévin, Pierre, Alexandre, Quentin pour toutes ces aventures vécues depuis les classes prépa.

Enfin, j'aimerais remercier ma famille pour tout ce que j'ai reçu et tout ce que nous avons pu partager jusqu'à aujourd'hui : merci à mon Père, je te remercie de m'avoir toujours laissé la liberté de m'investir dans mes projets ; merci à ma mère ; merci à ma petite soeur pour sa sensibilité et sa gentillesse ; merci à ma grande soeur de partager avec moi ses connaissances médicale toujours avec une grande énergie ; merci à Paul Erner aka Bobby Fischer pour sa passion des échecs, je suis convaincu que nos parties de l'été 2007 ont eu un impact sur mon investissement dans les sciences. Et surtout, je tiens à remercier mes grands-parents qui m'ont toujours accordé leur plus grande attention et avec qui j'ai toujours pu échanger dans la plus grande simplicité.

Côté financement je remercie la région Occitanie/Pyrénées-Méditerranée et le département MIA de l'INRA pour avoir financé cette thèse ainsi que l'institut de convergence #digitag pour avoir financé mon voyage au Japon pour la conférence ACML.


INRA
SCIENCE & IMPACT

département
mia

La Région
Occitanie
Pyrénées - Méditerranée

Résumé

Cette thèse s'inscrit dans le contexte général de l'estimation et de l'optimisation de fonctions de type boîte noire dont la sortie est une variable aléatoire. Motivé par la nécessité de quantifier l'occurrence d'événements extrêmes dans des disciplines comme la médecine, l'agriculture ou la finance, dans cette thèse des indicateurs sur certaines propriétés de la distribution en sortie, comme la variance ou la taille des queues de distribution, sont étudiés. De nombreux indicateurs, aussi connus sous le nom de mesure de risque, ont été proposés dans la littérature ces dernières années. Dans cette thèse nous concentrons notre intérêt sur les quantiles, CVaR et expectiles. Dans un premier temps, nous comparons les approches K-plus proches voisins, forêts aléatoires, régression dans les RKHS, régression par réseaux de neurones et régression par processus gaussiens pour l'estimation d'un quantile conditionnel d'une fonction boite noire. Puis, nous proposons l'utilisation d'un modèle de régression basé sur le couplage de deux processus gaussiens estimés par une méthode variationnelle. Nous montrons que ce modèle, initialement développé pour la régression quantile, est facilement adaptable à la régression d'autres mesures de risque. Nous l'illustrons avec l'expectile. Dans un second temps, nous analysons le problème relatif à l'optimisation d'une mesure de risque. Nous proposons une approche générique inspirée de la littérature $\mathcal{X}$-armed bandits, permettant de fournir un algorithme d'optimisation, ainsi qu'une borne supérieure sur le regret, adaptable au choix de la mesure de risque. L'applicabilité de cette approche est illustrée par l'optimisation d'un quantile ou d'une CVaR. Enfin, nous proposons des algorithmes d'optimisation utilisant des processus gaussiens associés aux stratégies UCB et Thompson sampling, notre objectif étant l'optimisation d'un quantile ou d'un expectile.

Mots-clés : Fonction boîte noire stochastique, mesure de risque, métamodèle, optimisation bandit, optimisation bayésienne, regret simple, inférence variationnelle.

Abstract

This thesis presents methods for estimation and optimization of stochastic black box functions. Motivated by the necessity to take risk-averse decisions in medecine, agriculture or finance, in this study we focus our interest on indicators able to quantify some characteristics of the output distribution such as the variance or the size of the tails. These indicators also known as measure of risk have received a lot of attention during the last decades. Based on the existing literature on risk measures, we chose to focus this work on quantiles, CVaR and expectiles. First, we will compare the following approaches to perform quantile regression on stochastic black box functions: the K-nearest neighbors, the random forests, the RKHS regression, the neural network regression and the Gaussian process regression. Then a new regression model is proposed in this study that is based on chained Gaussian processes inferred by variational techniques. Though our approach has been initially designed to do quantile regression, we showed that it can be easily applied to expectile regression. Then, this study will focus on optimisation of risk measures. We propose a generic approach inspired from the $\mathcal{X}$-armed bandit which enables the creation of an optimiser and an upper bound on the simple regret that can be adapted to any risk measure. The importance and relevance of this approach is illustrated by the optimization of quantiles and CVaR. Finally, some optimisation algorithms for the conditional quantile and expectile are developed based on Gaussian processes combined with UCB and Thompson sampling strategies.

Keywords: Stochastic black box function, measure of risk, metamodel, bandit optimisation, Bayesian optimisation, simple regret, variational inference.

Contents

Chapter 1

Introduction générale

Contents

1.1 Contexte

Comprendre le monde réel et ses phénomènes observables est, depuis plusieurs siècles, un des objectifs majeurs de notre civilisation. En écrivant "Le livre de l'Univers est écrit en langue mathématique" (L'essayeur, 1623), Galilée initie la physique mathématique qui sera au centre de découvertes majeures dans les siècles qui suivent. Mais c'est Newton qui met cette idée en application le premier. Dans Principes Mathématiques de la Philosophie Naturelle (1687) il établit un lien formel entre équations mathématiques et réalité empirique. Un des exemples le plus célèbre est l'utilisation des résultats de Newton pour établir l'équation de la chute d'un corps dans le vide. A partir de son principe fondamental de la dynamique

$$m \times \vec{a} = \sum_{i \in I} \vec{f}_i,$$

où m représente la masse de l'objet étudié, a l'accélération et f_i les forces qui s'appliquent au système, il est possible d'établir un bilan des forces sur un objet en chute libre pour décrire son évolution en fonction du temps. Le modèle obtenu est

$$\begin{cases} a_z = -g \\ v_z(t) = -gt + v_0 \\ z(t) = -\frac{1}{2}gt^2 + v_0 t + z_0 \end{cases} \tag{1.1}$$

avec a_z l'accélération, v_z la vitesse, g l'accélération de la pesanteur et v_0, z_0 la vitesse et l'altitude initiale.

Pour établir les principes et lois sur le mouvement, Newton a défini des hypothèses sur les systèmes étudiés, puis par l'intermédiaire d'un raisonnement logique il a abouti à une théorie. Cette façon de procéder a largement été utilisée jusqu'à ce jour et a contribué à la création d'une vaste littérature scientifique allant de l'électromagnétisme Jones [2013], à la physique quantique Feynman [1950] en passant par la mécanique des fluides Temam and Chorin [1978].

Cependant bien qu'il soit relativement facile d'interpréter le modèle (1.1), ou par exemple la célèbre équation d'Einstein $E = mc^2$, il existe d'autres modèles ou formulations de l'évolution de systèmes physiques plus difficiles à interpréter par un cerveau humain normalement constitué. C'est le cas de l'équation de Navier-Stokes. Cette analogie au principe fondamental de la dynamique pour un flux se formule comme l'équation aux dérivées partielles suivantes :

$$\rho\left(\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v}\right) = -\nabla \vec{p} + \mu \nabla^2 \vec{v},$$

avec ∇ l'opérateur gradient, $\vec{v}$ le champ de vitesse, p la pression, ρ la masse volumique du fluide et μ sa viscosité. En raison du terme $\vec{v} \cdot \nabla \vec{v}$, l'équation est non linéaire et donc difficile à analyser mathématiquement, si bien que démontrer l'existence et l'unicité d'une solution régulière à conditions initiales fixées reste un des problèmes du millénaire.

De plus, ce terme non linéaire (rendant compte des turbulences), rend très difficile la prévision, même approximative, de l'évolution d'un flux uniquement par une lecture très attentive de l'équation. Pour surmonter ces difficultés et réussir à prédire l'évolution d'un flux, il est possible de simuler cette équation à l'aide d'un ordinateur (voir Qian et al. [1992] pour plus de détails sur la simulation des équations de Navier-Stokes). La simulation numérique des équations de Navier-Stokes est utilisée en météorologie mais aussi pour simuler des écoulements d'air autour d'une structure par exemple. Enfin, il faut garder à l'esprit que la simulation numérique ne s'arrête pas à Navier-Stokes, d'autres équations complexes sont simulées numériquement dans différents domaines. C'est le cas par exemple d'équations stochastiques en biologie Meng et al. [2004] ou d'équations de la théorie cinétique en physique Birdsall and Langdon [2018].

Dans certains cas, pour atteindre un haut niveau de précision, les simulateurs doivent procéder à un très grand nombre d'opérations arithmétiques. Cette masse d'opérations créée des simulations très gourmandes en terme de temps de calcul ou en terme de consommation énergétique. Obtenir le résultat d'une simulation peut donc être coûteux mais obtenir des résultats précis reste absolument nécessaire dans certains domaines. Par exemple dans l'exploitation pétrolière des compagnies investissent énormément sur la base de simulations. Le but pouvant être la localisation des meilleures zones de forage et la sélection de la meilleure politique d'exploitation d'un réservoir pour maximiser la quantité de fluide extraite tout en minimisant les coups. Or, pour comprendre les liens entre les paramètres du modèle et les résultats, il est souvent nécessaire de simuler un grand nombre d'expériences. Dans ce cas l'approche par simulations peut rapidement être limitée car réaliser un grand nombre de simulations avec un haut niveau de précision implique de disposer d'une puissance de calcul extrêmement élevée. Explorer un modèle dans le but de sélectionner des paramètres pour optimiser un critère d'intérêt ou bien pour comprendre les liens entrées du modèle/résultats devient une tache trop coûteuse. **Des stratégies doivent êtres mises en place pour diminuer les coups tout en conservant un haut niveau de précision.**

Parallèlement au développement de modèles mathématiques et de simulateurs, ces dernières décennies ont vu se renforcer la méthode qui vise à extraire des connaissances à partir des observations réalisées directement dans la vie réelle. Cette stratégie apporte un nouveau paradigme. On ne tente plus de poser des hypothèses sur un phénomène puis d'utiliser des principes mathématiques pour en extraire des lois. Sous ce formalisme il est supposé que les lois physiques sont exprimées par les données et qu'en les collectant il est possible d'extraire des lois sans formulation d'aucunes (ou avec un jeu très limité) hypothèses mathématiques sur le système observé. C'est le formalisme sur lequel les algorithmes d'*intelligence artificielle* de Google, Facebook, Netflix et autre acteurs du monde numérique se basent. Les domaines d'application sont nombreux, nous pouvons citer notamment la reconnaissance d'image, la recommandation de contenu ou même la création d'algorithmes pour le pilotage automatique de voitures.

Pour illustrer le fonctionnement de ce paradigme, revenons à l'exemple de la chute d'un corps dans le vide. Imaginons un expérimentateur placé dans le vide, réalisant des mesures sur la vitesse et l'altitude d'un objet en chute libre, en fonction du temps.

Modulo des potentielles erreurs de mesure, sur Terre les observations vont suivre le modèle (1.1). A partir de ces observations, l'expérimentateur pourra estimer une loi puis faire une prédiction là où il n'avait pas fait d'observations. Un exemple illustrant ce que pourrait être des données collectées et des lois estimées est représenté Figure 1.1.

Sans entrer dans des détails de fond sur la difficulté à faire la différence entre corrélations et causalités avec cette approche, cette deuxième méthode soulève deux problèmes. Le premier est le temps et l'argent nécessaire pour collecter les données, ce qui tend à limiter le nombre d'expériences et donc la précision de la loi extrapolée. Le second problème est la potentielle difficulté à estimer la loi sous-jacente. Sur la Figure 1.1 il est relativement facile d'estimer une loi pour la vitesse (courbe rouge, graphique de gauche), un simple modèle linéaire peut être défini sans l'aide de l'outil informatique. Cependant, quand la fonction possède davantage de non linéarités, comme c'est le cas pour la loi régissant l'évolution de l'altitude en fonction du temps (courbe verte, graphique de droite), ou si la dimension des valeurs d'entrées est plus grande que deux [1], l'extrapolation de la loi est nettement plus difficile. **Dans ce contexte une méthode doit être mise en place pour estimer une loi à partir des observations et cela de manière autonome et avec le moins d'observations possibles.**

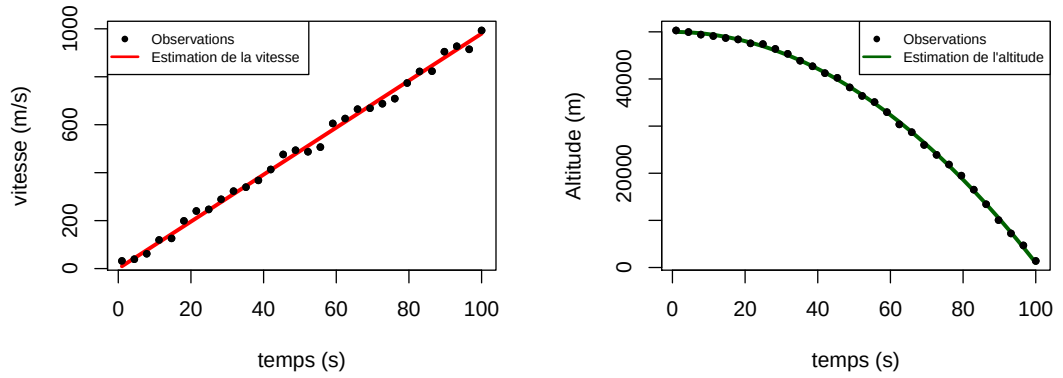


Figure 1.1: Estimation d'une loi pour la vitesse et l'altitude d'un objet lâché dans le vide sans vitesse initiale à une altitude de 50km. A gauche la droite rouge d'équation $y = 9.8 \times t$ donne une estimation de la vitesse en fonction du temps (suivant l'axe $-0z$). A droite la courbe verte d'équation $y = -4.9t^2 + 50000$ donne une estimation de l'altitude en fonction du temps.

Le cadre de cette thèse s'articule autour des deux problématiques soulevées. Dans la suite les deux formalismes introduits sont confondus et leur étude est formalisée comme l'analyse d'un système dit *boîte noire* mais l'étude de cette boîte noire dépend

[1] A partir de cette valeur les données ne peuvent plus êtres représentées dans leur intégralité sur un graphe en 3D, il est donc plus difficile pour l'Homme de les visualiser.

des système considérés. Dans le cas où les systèmes sont totalement non contrôlables, l'objectif de cette thèse est l'étude d'outils statistiques permettant la prédiction dans le but d'anticiper et de s'adapter au mieux aux événements futurs. Cette approche fait sens par exemple pour la prévision du trafic routier. Dans les cas où le système peut être influencé sensiblement par l'action humaine, alors les travaux de cette thèse visent à développer des outils permettant la sélection de stratégies optimales. Les applications sont nombreuses, nous pouvons citer la recherche clinique où la composition des médicaments est modifiable dans le but de les rendre le plus efficace possible, l'agriculture où le choix des variétés cultivées est personnalisable dans le but d'optimiser un rendement, on peut également citer la finance où l'optimisation des stratégies d'investissement occupe une place centrale dans la maximisation des profits.

1.2 Système boîte noire aléatoire

1.2.1 Formalisation

Si nous observons le résultat d'une simulation faite par un ordinateur dont nous ignorons le modèle sous-jacent, ou bien si nous observons directement l'évolution d'un système dans la vie réelle sur lequel aucune théorie n'existe, alors nous dirons que le système étudié est une boîte noire. Aucune information autre que celles que nous observons ne peut nous être apportée sur le fonctionnement du système étudié. Dans cette thèse nous séparons deux types de boîtes noires.

Le premier type correspond aux boîtes noires déterministes, que nous décrivons par une fonction inconnue Ψ définie comme :

$$\Psi : \mathcal{X} \to \mathbb{R},$$

avec $\mathcal{X}$ un espace compact inclus dans $\mathbb{R}^D$. Par exemple, la chute d'un corps dans le vide peut être décrite par une fonction de ce type. Si différents objets sont lâchés à une altitude z_0 et vitesse initiale v_0, alors au temps t_1 les vitesses et les altitudes mesurées seront identiques pour tous les objets, il n'y a pas d'aléa ici.

Le second type correspond aux boîtes noires stochastiques. Dans ce cas le système dépend d'un aléa. La boîte noire possède une entrée rendant compte de ce caractère aléatoire. Nous décrivons les boîtes noires stochastiques par une fonction inconnue Ψ définie comme :

$$\Psi : \mathcal{X} \times \Omega \to \mathbb{R},$$

avec $\mathcal{X}$ un espace compact inclus dans $\mathbb{R}^D$ et Ω l'espace représentant l'aléa du système. La Figure 1.2 schématise la définition d'une telle fonction. Contrairement au cas déterministe, pour des entrées du modèle fixées dans l'espace $\mathcal{X}$, si nous évaluons plusieurs expériences alors plusieurs résultats différents vont êtres observés. En effet, à x fixé, la sortie d'une boîte noire stochastique est une variable aléatoire notée Y_x de loi (inconnue) $\mathbb{P}_x$.

Pour mieux comprendre, considérons le système illustré Figure 1.3. Ce système boîte noire a comme sortie le rendement d'une variété de tournesol en fonction d'un

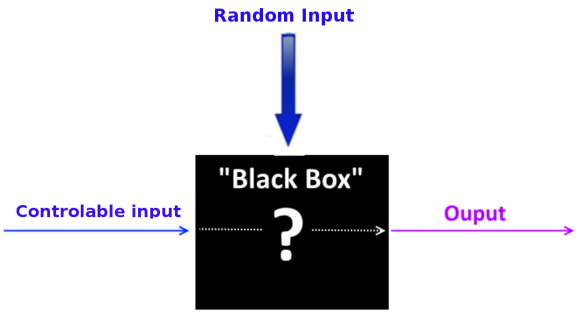


Figure 1.2: Illustration d'une boîte noire stochastique.

trait phénotypique potentiel (la longueur de tige maximale) défini par un paramètre $x \in \mathcal{X} = [0, 1]$, et du climat sur un an, *i.e* Ω est l'ensemble des climats possibles sur une année. A trait phénotypique x fixé, d'une année sur l'autre le rendement ne sera pas le même du fait du caractère aléatoire du climat et cela se traduit par des observations différentes pour le même point x.

Dans la suite de cette thèse nous considérons uniquement des problèmes boîte noire stochastique. Notre approche sera la suivante, nous considérons les entrées dans l'espace $\mathcal{X}$ comme contrôlables. Par exemple $\mathcal{X}$ peut être l'espace des choix d'une variété dans une exploitation agricole, le choix d'une stratégie financière, le choix du design d'une pièce en industrie ou encore une politique d'exploitation d'une forêt ou d'un verger. Les entrées prises dans Ω sont non contrôlables et potentiellement de très grande taille, elles représentent le caractère aléatoire (stochastique) du modèle. Ces entrées peuvent êtres une météo subie, des aléas dans la chaîne de montage d'un produit industriel, les stratégies d'investissement de tous les autres acteurs de la place boursière ou encore des caractéristiques génétiques d'un individu.

1.2.2 Sortie d'une boîte noire stochastique et mesures de perte

Dès lors que la sortie est une distribution de probabilité, nous pouvons être intéressé par différentes quantités pour identifier ses caractéristiques. Mais avant toutes choses nous définissons deux termes centraux dans la suite de cette thèse. Nous disons que la boîte noire est homoscédastique si la variance de Y_x ne dépend pas de x. A l'inverse nous disons que la boîte noire est hétéroscédastique si la variance de Y_x dépend de x. Maintenant définissons la fonction g

$$g(x) = \rho(\mathbb{P}_x) \tag{1.2}$$

avec ρ une fonctionnelle définie sur les mesures de probabilité et à valeurs réelles. Différentes fonctions g peuvent êtres définies. Une approche classique consiste à fixer g comme étant la moyenne conditionnelle. Dans ce cas l'estimation de g a largement été

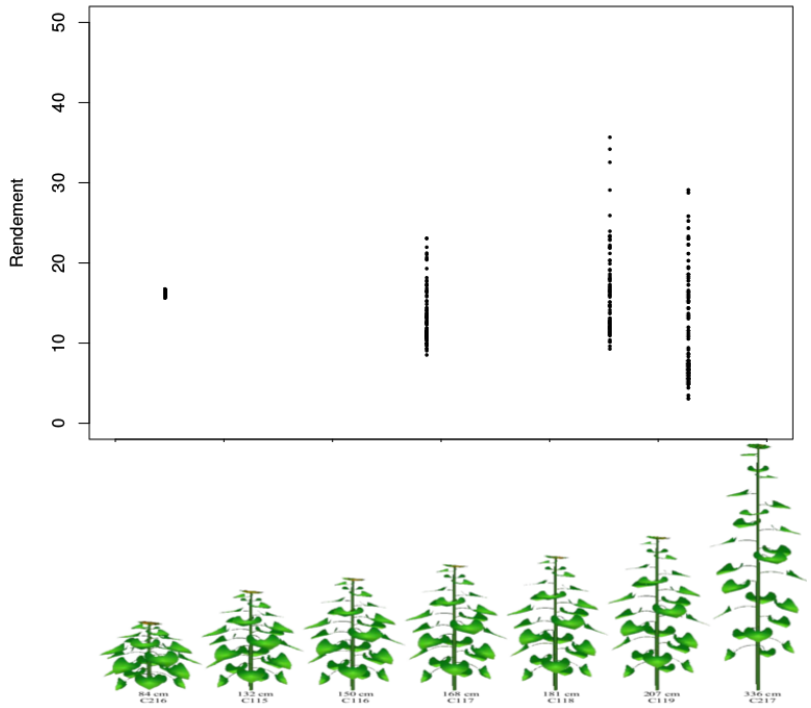


Figure 1.3: Illustration d'une boîte noire aléatoire avec pour chaque point testé le résultat de 100 évaluations correspondant à 100 climats différents. Ici la boîte noire représente le rendement d'une variété de tournesol définie par sa longueur de tige potentielle. A longueur de tige fixée, on observe différents rendements du fait d'un climat (aléatoire) différent subit par la variété.

développée dans la littérature (voir Györfi et al. [2006], Härdle [1990] par exemple) mais l'inconvénient majeur de ce choix est que la moyenne conditionnelle n'informe pas sur les queues de distribution. Cela peux être désavantageux si l'objectif est de prendre une décision basée sur une aversion aux cas extrêmes dans un contexte hétéroscédastique. Cet inconvénient est visible Figure 1.4, graphique de gauche. La courbe bleue représente la moyenne conditionnelle du problème boîte noire sous-jacent et ne rend pas compte de la modification de la variance en fonction de l'entrée. Bien que la moyenne soit quasiment la même pour $x = -0.6$ et $x = 2$, en $x = 2$ des événements extrêmes peuvent survenir, ce qui peut impliquer des cas très défavorables qui auraient pu êtres évités en sélectionnant $x = -0.6$. De ce fait la moyenne conditionnelle n'est pas un bon indicateur pour prendre des décisions "robustes" dans un contexte hétéroscédastique, on dit que c'est une mesure neutre par rapport au risque. A noter que dans un cas homoscédastique, la question de prendre une décision qui protège des cas extrêmes ne se pose pas.

D'autres fonctionnelles de la loi peuvent être utilisées pour prendre en compte les valeurs extrêmes résultant de queues de distribution potentiellement larges. Le graphique de droite de la Figure 1.4 représente en rouge les quantiles conditionnels d'ordre 0.1 et 0.9. En utilisant ces indicateurs on remarque facilement que la variance dépend des entrées et que la variance de la distribution $\mathbb{P}_2$ est beaucoup plus grande que celle de la distribution $\mathbb{P}_{-0.6}$. Il est donc possible de prendre des décisions basées sur une aversion pour les cas extrêmes avec des choix de g autres que la moyenne conditionnelle.

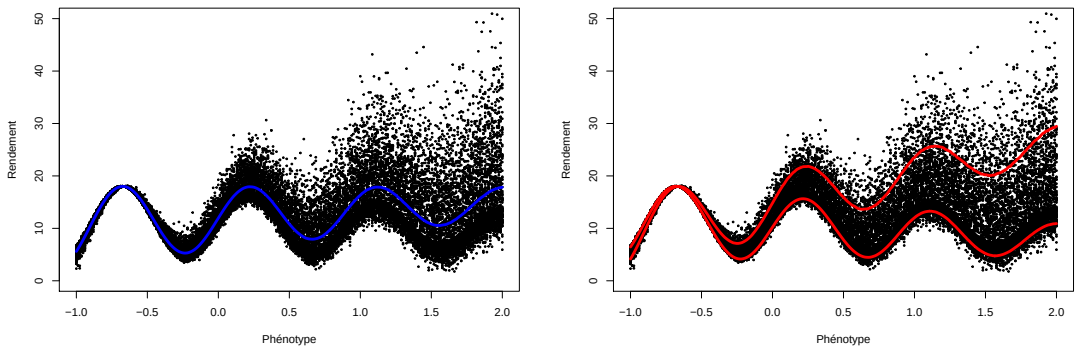


Figure 1.4: Estimation par méthode de Monte Carlo de la moyenne en bleue à gauche et des quantiles d'ordre 0.1 et 0.9 en rouge à droite d'un problème boîte noire stochastique.

Dans la suite nous présentons plus en détails le quantile (aussi appelé Value at Risk), le superquantile (aussi appelé conditional value at risk), l'expected shortfall et l'expectile, qui sont des indicateurs utilisés dans la littérature pour prendre des décisions en fonction des queues de distribution.

1.3 Mesures de risque et positionnement

Il est bien connu que la maîtrise et l'optimisation des risques est une préoccupation centrale du monde de la finance. Se protéger d'une forte perte est au fondamental pour ses acteurs s'ils veulent définir des stratégies profitables sur le long terme. Pour cela la notion de mesure de risque définie comme présenté en (2.16) a été introduite. Cependant la notion de risque est loin de se limiter au monde de la finance. Concernant la production d'énergie nucléaire, un réacteur ne sera pas conservé si la probabilité qu'il explose est trop élevée, un vaccin ne sera pas commercialisé si les risques de pathologies liés aux effets secondaires sont trop grands, un design de turbine de réacteur d'avion ne sera pas sélectionné si la probabilité que la turbine casse en vol est trop importante. Ainsi, mesurer la probabilité d'occurrence d'un événement aléatoire extrême est un problème central dans beaucoup de domaines. Cette mesure permet de prendre des décisions qui protègent contre des événements fortement mauvais.

Basé sur la définition (2.16), il existe une infinité de mesures de risque g. Certaines

mesures de risque sont utilisées car elle permettent une très bonne compréhension de ce qu'elles mesurent, c'est le cas du quantile par exemple. D'autres sont utilisées car elles sont dites cohérente au sens de l'article Rockafellar [2007]. C'est à dire qu'elles vérifient les propriétés :

- $g(C) = C$ pour toute constante C,
- $g(Y + Y') \leqslant g(Y) + g(Y')$ on parle de sous additivité,
- $g(Y) \leqslant g(Y')$ si $Y \leqslant Y'$,
- Si pour tout suite de variables aléatoires $(Y_h)_{h\in\mathbb{R}}$, $g(Y_h) \leqslant 0$ et $\lim_{h\to 0} ||Y_h - Y|| = 0$, alors $g(Y) \leqslant 0$
- $g(\lambda Y) = \lambda g(Y)$ pour $\lambda > 0$.

A noter en particulier qu'à partir de ces propriétés on peut en déduire la propriété de convexité $g\big((1-\lambda)Y + \lambda Y'\big) \leqslant (1-\lambda)g(Y) + \lambda g(Y')$. Avoir une mesure de risque vérifiant ces propriétés est en effet très utile quand l'objectif est l'optimisation du risque associé à un portefeuille (voir Rockafellar et al. [2000], Krokhmal et al. [2002] pour des exemples). Parmi les mesures de risque cohérentes il existe la conditional value at risk (CVaR), l'expected shortfall (ES) et l'expectile. Dans la suite nous définissons ces mesures de risque et discutons de leur interprétabilité et d'une façon de les utiliser pour prendre des décisions protégeant d'événements extrêmes indésirables.

1.3.1 Quantile et Value at risk

Le quantile et la value at risk conditionnel d'ordre τ sont des mesures de risque définies comme

$$q_\tau(x) = \inf\{q \in \mathbb{R},\ F_x(q) \geqslant \tau\},$$

avec F_x la fonction de répartition associée à la distribution $\mathbb{P}_x$, ou d'une manière équivalente (si F_x est strictement croissante) comme

$$q_\tau(x) = \arg\min_{q\in\mathbb{R}} \mathbb{E}(l_\tau(Y_x - q)),$$

avec l_τ la fonction pinball définie comme

$$l_\tau(\xi) = (\tau - \mathbb{1}_{(\xi<0)})\xi, \quad \xi \in \mathbb{R}. \tag{1.3}$$

La pinball est présentée Figure 1.6, graphique de gauche. Le quantile d'ordre τ peut être vu comme un seuil qui est dépassé seulement avec une probabilité $1-\tau$ [2]. Les exemples

[2] Ou d'une manière équivalente, ce seuil ne sera pas dépassé avec probabilité τ.

d'un tel seuil dans la vie réelle sont nombreux. Ce seuil peut être la hauteur d'une digue garantissant avec probabilité τ que les flots ne la dépasseront pas, dans ce cas τ est pris proche de 1 pour se protéger le plus possible des inondations. Le quantile d'ordre τ peut également représenter le pire rendement possible d'une exploitation agricole dans les $1-\tau$ cas les plus favorables. Dans cette situation il est raisonnable de considérer des valeurs petites de τ pour se protéger d'une année extrêmement mauvaise. En finance c'est sensiblement identique. Considérons Y une variable aléatoire représentant les pertes (valeurs négatives) et profits (valeurs positives). Le quantile d'ordre τ représente la perte minimale dans les τ cas les pires ou la perte maximale dans les $1-\tau$ cas les meilleurs. Une illustration des quantiles d'ordre 0.1 et 0.9 d'une loi log-normale est proposée Figure 1.5.

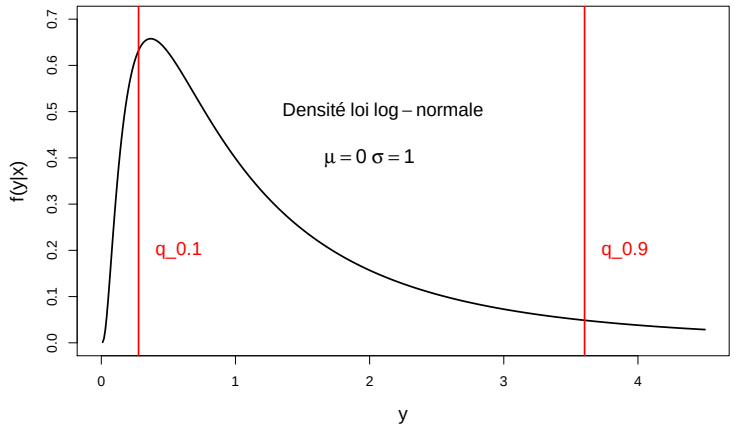


Figure 1.5: Les quantiles d'ordre 0.1 et 0.9 de la loi log-normale de paramètres $(0, 1)$ sont égaux à l'intersection des droites rouges avec l'axe des abscisses.

Outre la facilité d'interprétation de cette mesure de risque, le quantile d'ordre τ à quelques propriétés qui le rendent très utile en pratique. Il ne prend pas en compte les *outliers* et les quantiles caractérisent la distribution sous-jacente. Cependant le quantile n'est pas sous additif ce qui fait que cette mesure de risque n'est pas cohérente. C'est un mauvais point pour l'utilisation de cette mesure de risque dans le cadre de l'optimisation d'un portefeuille.

1.3.2 Conditional value at risk et Expected shortfall

La CVAR est définie comme

$$\mathrm{CVaR}_\tau(x) = \inf_{c\in\mathbb{R}} \{c + \frac{1}{1-\tau}\mathbb{E}[(Y_x - c)^+]\}, \text{ avec } (y-c)^+ = \max(y-c, 0),$$

ce qui est équivalent (voir Ben-Tal and Teboulle [2007] pour plus de détails) pour une distribution continue à

$$\text{CVaR}_\tau(x) = \mathbb{E}[Y_x | Y_x \geqslant q_\tau(x)].$$

De nombreuses propriétés sur la CVaR sont discutés dans Pflug [2000]. La CVaR_τ peut être vue comme la moyenne de la hauteur des crues dans les τ cas les pires. Dans les cas où Y représente des pertes et profits alors la définition de la CVaR_τ ne permet pas de quantifier les pertes moyennes dans les pires cas. Pour cela on peut utiliser l'expected shortfall définie comme

$$E_\tau(x) = \inf_{c \in \mathbb{R}} \{ \frac{1}{\tau} \mathbb{E}[(Y_x - c)^-] - c \}, \text{ avec } (y-c)^- = \min(y - c, 0),$$

qui dans le cas continu donne

$$E_\tau(x) = \mathbb{E}[Y_x | Y_x \leqslant q_\tau(x)].$$

Ainsi dans le cas d'une exploitation agricole, E_τ représente la moyenne des pertes subies dans les τ cas les pires. Contrairement au quantile, ces mesures de risque apportent une indication sur l'amplitude de la distribution au-delà du quantile. De plus cet indicateur a toutes les propriétés qu'une mesure de risque doit avoir pour être cohérente. Cela lui donne un net avantage dans le cas de l'optimisation de portefeuilles. Cependant ce type de mesure de risque va être sensible aux outliers et ne caractérise pas la distribution. De plus connaître la CVaR_τ ne donne aucune indication sur ce qui se passe dans les τ autres cas possibles (à gauche du quantile d'ordre τ) et cette mesure est "non élicitable" Barone Adesi [2016], ce qui veut dire qu'elle ne peut s'écrire comme le minimiseur d'une fonction *score*. Cette propriété rend les méthodes d'estimation classique (voir Section 1.5.2 pour plus de détails) non utilisables directement. En revanche des travaux récents ont montré que le couple quantile/CVaR était lui "élicitable" (voir Acerbi and Szekely [2014]), ce qui ouvre la porte à de nombreuses possibilités d'estimation (non développés dans cette thèse). Ainsi la CVaR possède de très bons atouts mais comme le quantile cette mesure de risque possède des points négatifs.

1.3.3 Expectile

Les expectiles sont une variante des quantiles en le sens qu'ils sont définis comme

$$e_\tau = \arg\min_{e \in \mathbb{R}} \mathbb{E}\big(l_\tau^e(Y_x - e)\big), \tag{1.4}$$

avec l_τ^e une fonction pinball modifiée

$$l_\tau^e(\xi) = \mathbb{1}_{(\xi<0)}(1-\tau)\xi^2 + \mathbb{1}_{(\xi>0)}\tau\xi^2, \quad \xi \in \mathbb{R}. \tag{1.5}$$

A noter que pour $\tau = 0.5$, l'expectile est égal à la moyenne. La fonction de perte (1.5) est présentée Figure 1.6, graphique de droite. L'expectile est une mesure de risque cohérente Ziegel [2016] et les expectiles caractérisent la distribution Abdous and Remillard [1995].

En effet, contrairement à la CVaR les expectiles prennent en compte les queues de distribution à droite et à gauche. Pour s'en convaincre considérons la condition d'optimalité de premier ordre sur (1.4) :

$$\tau \int_e^{+\infty} |Y - e| dF_Y = (1 - \tau) \int_{-\infty}^{e} |Y - e| dF_Y. \tag{1.6}$$

Ce qui implique que l'amplitude moyenne de Y pour des valeurs inférieures à e est égale à l'amplitude moyenne de Y pour des valeurs supérieures à e multiplié par le facteur $\tau/(1-\tau)$. De plus, dans Kuan et al. [2009], partant de (1.6) les auteurs établissent l'égalité

$$\frac{\int_{-\infty}^{e} |Y - e| dF_Y}{\int_{-\infty}^{+\infty} |Y - e| dF_Y} = \tau.$$

Celle-ci implique que la dispersion moyenne à gauche de e_τ représente une proportion τ de la dispersion totale autour de e_τ.

Bien que les expectiles regroupent de nombreux avantages, ils souffrent d'une faible interprétabilité. Pour réduire le problème dans Jones [1994], Yao and Tong [1996], les auteurs montrent que les expectiles peuvent êtres vus comme des quantiles d'ordre donnés. Plus précisément une bijection $q_\tau = e_{h(\tau)}$ est explicitement fournie avec

$$h(\tau) = \frac{-\tau q_\tau + G(\tau)}{-e_{0.5} + 2G(q_\tau) + (1 - 2\tau) q_\tau},$$

et $G(q) = \int_{-\infty}^{q} y dF$. Mais cette équivalence reste difficile à transmettre pour des personnes non initiées et la dépendance en $\mathbb{P}_x$ rend cette interprétation relativement difficile à exploiter dans un contexte de régression.

1.3.4 Autres mesures de risque

Moyenne-variance

Supposons qu'il existe des traitements A et B pour soigner une pathologie. Le traitement A a un taux de réussite moyen $\mu_A = 0.4$ avec une variance $\sigma_A = 0.1$. Le traitement B a un taux de réussite moyen $\mu_B = 0.5$ et une variance $\sigma_B = 0.25$. En moyenne le traitement B est meilleur mais parfois il sera bien moins. Le choix entre le traitement A et B peut se faire à l'aide du critère *moyenne-variance* introduit par Markowitz [1952] défini comme

$$\text{mv}(Y) = \sigma_Y - \rho \mu_Y, \text{ avec } \rho > 0.$$

Ici le compromis entre la performance et le risque se fait à l'aide du paramètre ρ. Dans le cas où les distributions sont gaussiennes cette mesure de risque fait sens même si elle n'est pas cohérente (si $\rho \in]0, 1[$, $\text{mv}(C) = \rho C \neq C$). Il convient de noter que dans d'autres cas, notamment pour des distributions asymétriques, cette mesure est discutable.

Mesure de risque entropique

Une autre mesure de risque possédant un paramètre permettant à l'utilisateur de moduler son rapport au risque est la mesure de risque *log-exponentielle* ou *entropique* définie comme

$$\kappa_{\lambda,Y} = \frac{1}{\lambda} \log \mathbb{E} \exp(\lambda Y), \ \lambda \neq 0. \tag{1.7}$$

Dans Rockafellar [2007] il est montré que cette mesure de risque pour $\lambda < 0$ est cohérente en un sens faible car elle ne satisfait pas le cinquième point énoncé plus haut. Cependant elle possède de bonnes propriétés théoriques. En effet dans Maillard [2013] l'auteur montre, sous réserve que $G(t) = \log \mathbb{E}[\exp(tY)]$ soit bien définie au voisinage de 0, que pour tout $\delta \in (0, 1)$:

$$\mathbb{P}\left[Y \geqslant \inf\left\{\frac{1}{\lambda} \log \mathbb{E} \exp(\lambda Y) + \frac{\log(1/\delta)}{\lambda} : \ \lambda > 0\right\}\right] \leqslant \delta,$$

$$\mathbb{P}\left[Y \leqslant \sup\left\{ -\frac{1}{\lambda} \log \mathbb{E} \exp(-\lambda Y) - \frac{\log(1/\delta)}{\lambda} : \ \lambda > 0\right\}\right] \leqslant \delta.$$

La première équation contrôle la probabilité qu'une réalisation de Y soit grande alors que la seconde contrôle la probabilité qu'une réalisation de Y soit petite. Étudier 1.7 fait donc sens dans un contexte où la prise de décisions en fonction des queues de distribution est souhaitée.

1.3.5 Positionnement rapport au risque

On peut distinguer trois positionnements par rapport au risque. Le premier est le point de vue neutre par rapport au risque. Si ce positionnement est choisi, la réalisation de valeurs extrêmes n'est pas un critère d'intérêt. L'utilisation de la moyenne est parfaitement sensé pour prendre des décisions.

Le second point de vue est l'aversion au risque. Dans ce cas on veut se protéger d'un événement défavorable. Cet événement peut se manifester de deux manières. La première est l'observation d'un gain, une quantité qui, plus elle est grande, plus l'environnement est profitable. C'est le cas d'un rendement agricole, du rendement d'un portefeuille financier ou du temps de survie d'un produit industriel. Dans ce cadre nous souhaitons que la valeur observée soit la plus rarement possible petite. Cela équivaut à maximiser un quantile d'ordre inférieur à 0.5 ou maximiser l'expected shortfall associée. Ou bien cela revient à vouloir que la balance entre la dispersion à gauche d'un expectile soit petite par rapport à la dispersion totale, ce qui revient à maximiser un expectile d'ordre $\tau < 0.5$. L'autre possibilité est l'observation d'une variable aléatoire Y rendant compte d'événements défavorables. Par exemple la perte subit par un portefeuille, la hauteur d'une crue, le nombre d'occurrences ou l'intensité des effets secondaires d'un médicament. Dans ce cas on souhaite que les observations soient le moins souvent grandes. Se protéger d'un risque implique donc minimiser un quantile d'ordre supérieur

à 0.5 ou minimiser la CVaR associée ou bien minimiser un expectile d'ordre supérieur à 0.5.

Enfin il y a le point de vu d'appétence au risque. Ce point de vue est en quelque sorte une position élitiste. Par exemple on peut souhaiter que la performance d'un petit groupe soit la plus élevée possible sans se préoccuper des résultats obtenus par le reste de la population. Cela peut être le cas lorsqu'on veut maximiser les résultats d'un petit groupe d'athlètes lors d'une compétition sans se soucier des sportifs moins performants. Dans ce contexte il s'agit de maximiser un quantile haut ou maximiser une CVaR ou un expectile d'ordre élevée. Un autre positionnement serait d'observer des pertes et de prendre une décision pour que la perte d'un petit groupe soit faible sans se soucier du reste de la population. Dans ce cas on minimiserait un quantile bas ou l'ES associée ou bien on minimiserait un expectile bas.

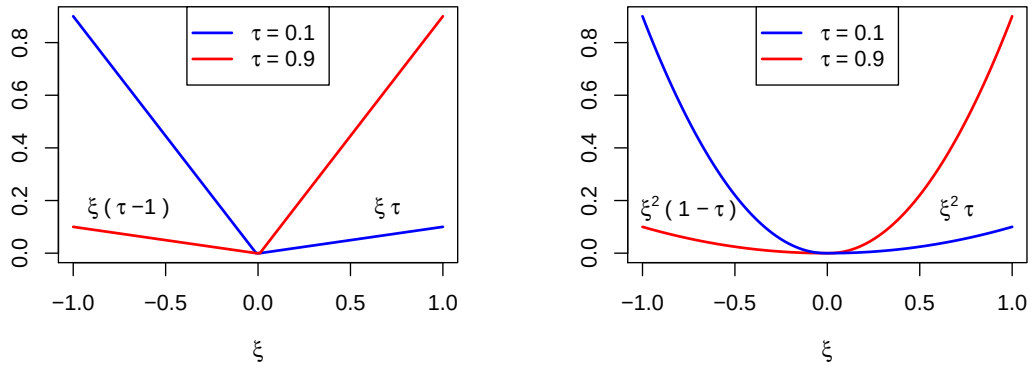


Figure 1.6: A gauche l_τ (fonction de perte associée au quantile d'ordre τ) à droite l^e_τ (fonction de perte associée a l'expectile d'ordre τ).

1.4 Problématiques et contributions apportées par cette thèse

Dans cette thèse nous proposons l'études de méthodes statistiques permettant d'estimer et d'optimiser des mesures de risque d'un modèle boîte noire sous la contrainte que la taille de l'échantillon est limitée. Pour l'estimation nous travaillons avec des méta-modèles (aussi appelé émulateurs statistiques) et pour l'optimisation nous utilisons des techniques inspirées de la littérature dite *bandit* ou bien des techniques d'optimisation à base de métamodèles gaussiens. Les contributions de cette thèse sont détaillées dans cette section.

1.4.1 Métamodèles

Nous avons vu plus haut que connaître une mesure de risque g d'un code boîte noire aléatoire est utile dans de nombreux domaines. Dans le cas où l'évaluation du modèle boîte noire est coûteuse, il est classique d'utiliser des outils statistiques appelés émulateurs statistiques ou métamodèles pour estimer une telle fonction. Il est possible de créer des métamodèles pour un large éventail de mesures de risque différentes. En dehors de l'estimation de la moyenne conditionnelle, qui, rappelons-le, est un point de vue neutre par rapport au risque, dans la littérature c'est certainement la création de métamodèles de quantile qui a été le plus développée. Un métamodèle étant une approximation basée sur un nombre fini de points, il est légitime de se poser les questions suivante :

- Y-a-t-il un type de métamodèle meilleur que tous les autres ?
- Commment évolue la précision d'estimation en fonction de la taille de l'échantillon d'apprentissage ?
- Quel est l'impact de la dimension de l'espace d'entrée $\mathcal{X}$ sur l'estimation ?
- Quel est l'impact du rapport signal sur bruit sur la qualité d'estimation ?
- Sachant que le théorème central limite pour l'estimation de quantiles s'écrit (dans un cadre asymptotique) :

$$\frac{\sqrt{n}(\widehat{q_\tau} - q_\tau)}{\sqrt{\tau(1-\tau)}} \to \mathcal{N}\Big(0, \frac{1}{f^2(q_\tau)}\Big),$$

  avec $\widehat{q_\tau}$ la $\lceil n\tau \rceil$-ième statistique d'ordre d'un échantillon $Y_1, \cdots, Y_n$ de variables aléatoires i.i.d de densité f. Quel impact a la valeur de la densité (en le quantile ou dans un voisinage centré en le quantile) sur l'estimation pour un échantillon de petite taille ?
- Comment les métamodèles réagissent quand la forme de la distribution varie fortement en espace ou dans des cas très hétérocédastiques ?

Or ces questions sont très peu prises en compte dans la conception et l'évaluation des métamodèles dans la littérature existante. Dans cette thèse nous proposons une synthèse des métamodèles créés pour estimer un quantile conditionnel d'ordre fixé. Puis nous testons ces méthodes sur différents cas tests afin de répondre aux questions soulevées plus haut et dans le but d'extraire des comportement propres à chacune des méthodes. Une synthèse de nos conclusions est présentée Figure 2.21.

Suite à cette première analyse, en plus des conclusions présentées Figure 2.21, nous avons retenu deux informations. La première est que l'utilisation des processus gaussiens associés aux méthodes variationnelles (pour la procédure d'inférence) était parmi les méthodes produisant les meilleurs résultats sur notre ensemble de cas tests. Le second point est que toutes les méthodes étudiées avaient des difficultés pour estimer un quantile

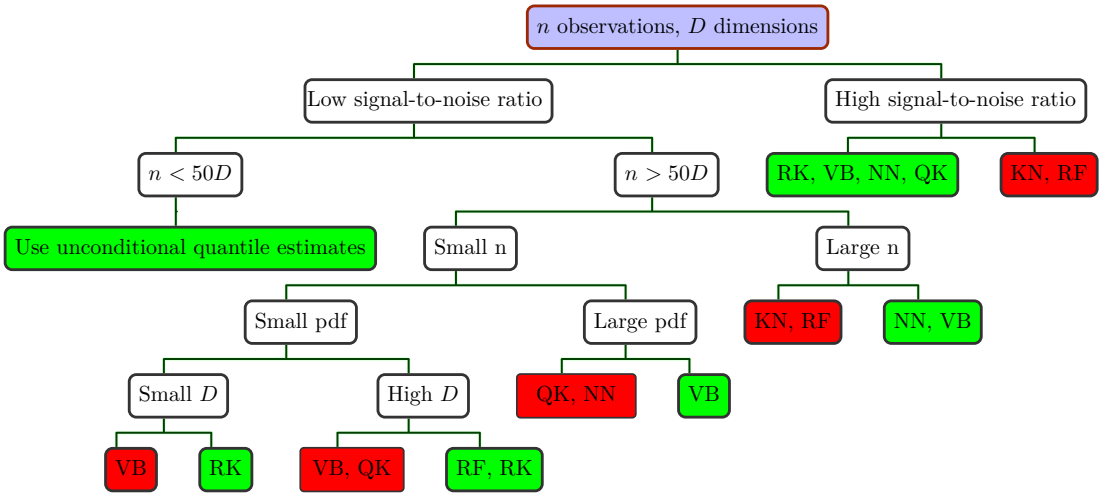


Figure 1.7: Recommendation des méthodes de régression quantile en fonction des caractéristiques des problèmes (vert : méthodes recommandées, rouge : méthodes à éviter). KN: K-plus proches voisins, RF: Forêts aléatoires, NN: Réseaux de neurones, RK: Régression dans un RKHS, QK: Quantile kriging, VB: méthode bayesienne variationnelle.

dans un cadre fortement hétéroscédastique. Une illustration de cette pathologie est présentée Figure 1.8, graphique de gauche.

Ainsi une autre contribution de cette thèse consiste à mettre une place un métamodèle de quantile basé sur les processus gaussiens qui soit suffisamment flexible pour estimer les quantiles conditionnels d'un système boîte noire fortement hétéroscédastique. La capacité de notre modèle à estimer des quantiles dans le cas très fortement hétéroscédastique est représenté Figure 1.8, graphique de droite.

Pour réaliser l'inférence dans un cadre fortement hétéroscédastique nous utilisons une méthode variationnelle de type boîte noire. Cette technique d'inférence ne dépend pas des propriétés de la vraisemblance utilisée. De ce fait la procédure suivie est suffisamment robuste pour permettre l'estimation de mesures de risque autre que le quantile. Pour illustrer ce propos nous définissons une vraisemblance spécifique à l'estimation de l'expectile et mettons en place notre stratégie pour créer un métamodèle gaussien d'expectile conditionnel.

A noter que dans les Chapitres 2 et 4, nous orientons notre approche vers l'estimation de modèles boîte noire représentant des simulateurs numériques. Or nous sommes convaincu que les outils développés peuvent s'appliquer à des systèmes observés directement in vivo.

1.4.2 Optimisation

Supposons qu'il existe au moins un point $x^* \in \mathcal{X}$ tel que

$$g(x^*) = \max_{x \in \mathcal{X}} g(x).$$

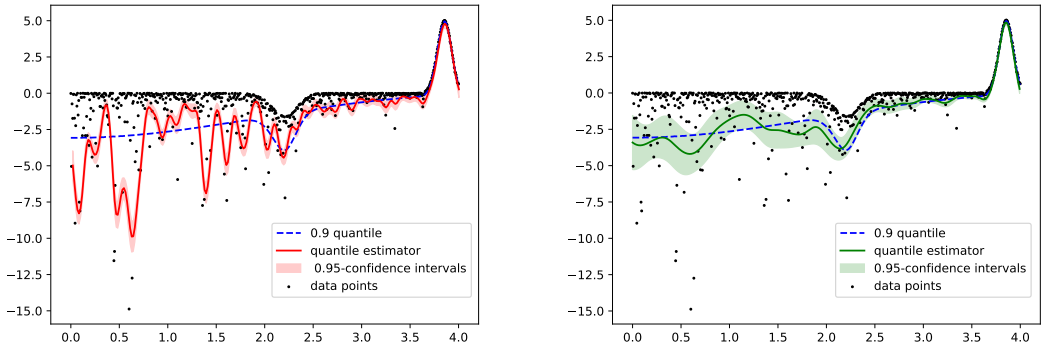


Figure 1.8: A gauche le métamodèle issu de Abeywardana and Ramos [2015], incapable d'estimer correctement le quantile dans la zone de gauche. A droite le métamodèle de quantile développé dans cette thèse qui se trouve être plus apte à traiter les cas fortement hétéroscédastique.

En optimisation, un objectif classique est de trouver l'argument $x \in \mathcal{X}$ qui maximise g. C'est à dire trouver x^* tel que

$$x^* = \arg\max_{x \in \mathcal{X}} g(x), \tag{1.8}$$

ou bien trouver un point $\widehat{x}^*$ qui retourne une valeur de g la plus proche possible de $g(x^*)$. Pour réaliser cet objectif, l'optimisation classique utilise des informations sur la fonction cible telles que, le gradient ou sa hessienne, pour conduire la recherche d'un optimum. L'optimisation d'un problème boîte noire se démarque du cadre classique par le fait qu'il est uniquement possible d'évaluer le modèle point par point et qu'aucune autre information n'est disponible. Il faut donc établir des stratégies alternatives pour guider l'optimisation. De plus, dans le cadre de l'optimisation d'une mesure de risque g d'une boîte noire stochastique, la quantité g n'est pas directement observée. L'optimiseur aura seulement à disposition des réalisations d'une variable aléatoire de loi inconnue. Enfin chaque évaluation de la fonction est supposée coûteuse donc leur nombre est limité.

Notre objectif est donc de définir des stratégies efficaces pour conduire l'optimisation vers un point optimal uniquement à l'aide d'un nombre **limité** d'évaluations **ponctuelles** et possiblement fortement **bruitées**.

Naturellement, pour mener à bien l'optimisation, il faudra utiliser des outils statistiques, pour, dans un premier temps, estimer g avant de retourner un point potentiellement optimal. Tout au long de cette thèse nous supposons qu'à tout temps nous avons la possibilité d'évaluer n'importe quel point $x \in \mathcal{X}$. Cette hypothèse est complètement justifiée lorsque nous travaillons avec des simulateurs numériques car les entrées sont totalement contrôlables et elle permet d'utiliser des méthodes statistiques séquentielles qui apportent un bon formalisme pour la résolution de notre problème. En effet, plutôt

que d'utiliser l'ensemble du budget d'un seul coup pour estimer la quantité d'intérêt et retourner un point supposé optimal, l'utilisation d'outils issus de la statistique séquentielle va permettre de mettre à jour nos connaissances de la fonction cible pas à pas, pour ainsi concentrer la recherche d'information proche des points à fort potentiel au prochain pas de temps. Sous le formalisme séquentiel, le budget total est divisé en $k > 0$ parts et à chaque pas de temps une part du budget est utilisée pour raffiner la recherche d'optimum. Aussi, à chaque itération le but de la stratégie est de trouver le bon compromis entre allouer du budget pour permettre une bonne estimation de la quantité bruitée là où la fonction est identifiée comme potentiellement optimale et utiliser du budget pour permettre une bonne exploration de l'espace pour s'assurer de ne ne pas avoir manqué un potentiel maximum global. Cet équilibre est connu sous le nom de compromis exploration exploitation. Il trouve des applications dans de nombreux domaines comme l'AB testing Kaufmann et al. [2014], ou les essais clinique Garivier et al. [2017].

Il existe une large gamme d'algorithmes d'optimisation utilisant des statistiques séquentielles. Dans cette thèse nous en considérons deux : l'optimisation basée sur des stratégies bandits et l'optimisation à base de métamodèles gaussiens qui est une branche de l'optimisation bayésienne. Avec ces deux approches le compris exploration exploitation est géré par une fonction d'acquisition f_a. L'Algorithme 1 schématise leur fonctionnement.

Algorithm 1: Optimisation basée sur les statistiques séquentielles

Input: Echantillon initial $\mathcal{D}_I$; fonction d'acquisition f_a; critère de sélection du point final c_f;

for $t = 1$ **to** T **do**

  Chercher l'argument qui maximise $f_a(x, \mathcal{D}_n)$;

  $x(t) \leftarrow \arg\max_{x \in \mathcal{X}} f_a(x, \mathcal{D}_n)$;

  Mettre à jour l'échantillon;

  $\mathcal{D}_{I(t+1)} = \mathcal{D}_{I(t)} \cup (x(t), y_{x(t)})$;

end

Output: Retourner un point $\widehat{x}^*$ maximisant c_f.

Toutefois, dans la littérature bandits et optimisation à base de métamodèles, les travaux spécifiques à l'optimisation d'une mesure de risque dans le cas boîte noire sont peu nombreux. Dans un contexte où l'espace d'entrée est discret des travaux ont été développés avec une approche de type bandit mais dans le cas où $\mathcal{X}$ est un espace continu, alors il n'y a pas de méthodes existantes. L'optimisation à base de métamodèles propose peu de choses également. Certain algorithmes existent comme Browne et al. [2016] mais le nombre d'observations à disposition du métamodèle dans cet article est bien trop grand par rapport à ce que nous pouvons nous permettre dans notre approche.

Dans cette thèse nous apportons trois contributions dans le domaine de l'optimisation d'une mesure de risque d'un code boîte noire stochastique. Premièrement, basé sur les modèles bayesiens de quantile et d'expectiles développés dans cette thèse (et mentionnés plus haut), nous adaptons deux stratégies d'optimisation bayésienne pour l'optimisation

d'un quantile ou d'un expectile conditionnel d'ordre fixé. Puis nous adaptons l'algorithme StoOO (voir Munos [2014]) pour l'optimisation d'un quantile conditionnel ou d'une CVaR conditionnel. Ensuite, toujours dans le cadre optimisation bandit, nous proposons un formalisme générique pour l'obtention d'une borne supérieure sur le regret pour l'optimisation de n'importe quelle fonction g, sous réserve de la connaissance d'inégalités de déviations sur cette quantité.

1.5 Métamodèles

Un métamodèle est littéralement un modèle du modèle. C'est une version simplifiée du modèle, qui, généralement, traite le système sous un angle différent du formalisme initialement adopté par le modèle. En effet, alors que le modèle tente de rendre compte d'un phénomène de la réalité empirique en utilisant des hypothèses et théories liées à ce phénomène, le métamodèle tente, lui, de rendre compte du comportement du modèle avec des hypothèses établies directement sur ce dernier. Des hypothèses classiques sur le modèle peuvent êtres :

- Une hypothèse de régularité en fonction des valeurs prises dans l'espace $\mathcal{X}$. Par exemple une hypothèse peut être que g est deux fois dérivable et possède des dérivées secondes continues ($C^2(\mathcal{X})$) ou bien une hypothèse de linéarité ;
- Une hypothèse sur la distribution conditionnelle de la sortie du modèle, par exemple $\mathbb{P}_x$ peut être supposée gaussienne pour tout $x \in \mathcal{X}$, de variance dépendante ou non de $x \in \mathcal{X}$.

En plus des ces hypothèses, pour créer le métamodèle nous disposons d'un échantillon $\mathcal{D}_n = \big((x_1, y_1), \cdots, (x_n, y_n)\big)$, avec pour tout $1 \leqslant i \leqslant n$, $x_i \in \mathcal{X}$ et $y_i \in \mathbb{R}$. Cette échantillon correspond à n évaluations ponctuelles du système boîte noire. Dans le cadre de cette thèse une évaluation du modèle est supposée relativement coûteuse donc n est supposé relativement petit, nous ne sommes pas dans un contexte *Big Data*.

Une intuition sur le formalisme introduit est que, pour avoir un métamodèle de qualité, il est nécessaire d'utiliser le bon jeu d'hypothèses en fonction du modèle sous-jacent et des données. En effet plus les hypothèses formulées sont fortes, plus le métamodèle sera rigide et pourra introduire un biais dans l'estimation. A l'inverse des hypothèses trop faibles permettrons trop de flexibilité et conduira à un métamodèle qui sur-interprétera les données.

Il existe une vaste gamme de métamodèles issue du formalisme introduit plus haut. Dans cette thèse nous séparons les metamodèles en trois grandes catégories. La première est basée sur les statistiques empiriques, elle regroupe la méthode des plus proches voisins et les méthodes basées sur des arbres de décisions (arbre de régression, bagging, forêts aléatoires). La seconde combine le point de vue fonctionnel avec les M-estimateurs, elle regroupe la régression linéaire, la régression par des splines, la régression dans les espaces de Hilbert à noyaux auto-reproduisants (RKHS), et la régression par réseaux de neurones. La troisième est le point de vue bayesien de la seconde, elle regroupe entre

autre la régression linéaire bayésienne et la régression par processus gaussiens. A noter que la régression par processus gaussiens peut également être vue comme une forme de régression dans les RKHS (voir Rasmussen and Williams [2006]). Cependant dans la troisième catégorie un *a priori* est défini sur les paramètres du métamodèle ce qui permet d'obtenir une distribution *a posteriori* sur ces mêmes paramètres et donc de fournir une interprétation probabiliste du métamodèle.

Remarquons que cette classification n'est pas stricte, des liens peuvent êtres faits entre ces classes, notamment entre les forêts aléatoires et les méthodes à noyaux comme la régression dans un RKHS (voir Scornet [2016]). Cependant, dans la suite de ce travail nous utilisons cette classification pour simplifier la description des méthodes et leur analyse.

Dans ce qui suit, pour chaque classe de métamodèle nous rappelons comment conduire l'estimation dans le cas classique où g est la moyenne conditionnelle, puis nous élargissons la procédure à l'estimation d'autres mesures de risque d'un système boîte noire, *i.e* le quantile, l'expectile, la CVaR et l'expected shorfall. A noter que cette section utilise la classification utilisée dans le Chapitre 2 mais développe les idées sous un angle moins formel tout en proposant un cadre qui n'est pas spécifique à l'estimation des quantiles. L'objectif étant de mettre en perspective la métamodélisation de différentes mesures de risque sans introduire de détails trop techniques.

1.5.1 Métamodèles basés sur les statistiques empiriques

Supposons que nous disposons d'un échantillon $\mathcal{Y}_n(x) = (y_1, \cdots, y_n)$ contenant des variables indépendantes et identiquement distribuées (i.i.d) suivant la loi $\mathbb{P}_x$. Sous cette hypothèses, à partir de $\mathcal{Y}_n(x)$ il est possible de créer différents estimateurs associés à différentes mesures de risque pour $\mathbb{P}_x$. Pour le cas de la moyenne conditionnelle, on peut définir un estimateur comme :

$$\widehat{m}(x) = \underset{m\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2, \text{ qui est égal à } \widehat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} y_i^2. \tag{1.9}$$

Un estimateur de l'expectile conditionnel d'ordre τ est

$$\widehat{e}_\tau(x) = \underset{e\in\mathbb{R}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{y_i\leqslant e} - \tau)(y_i - e)^2 + \mathbb{1}_{y_i>e}\tau(y_i - e)^2. \tag{1.10}$$

Un estimateur du quantile conditionnel d'ordre τ est

$$\widehat{q}_\tau(x) = \inf\{q \in \mathbb{R},\ \widehat{F}_x(q) \geqslant \tau\}, \text{ avec } \widehat{F}_x(q) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{y_i\geqslant q}. \tag{1.11}$$

Un estimateur de la CVaR conditionnelle d'ordre τ est

$$\widehat{\text{CVaR}}_\tau(x) = \min_{c\in\mathbb{R}}\left\{c + \frac{1}{n(1-\tau)}\sum_{i=1}^{n}(y_i - c)^+\right\}. \tag{1.12}$$

Enfin un estimateur de l'expected shortfall conditionnelle d'ordre τ est

$$\widehat{E}_\tau(x) = \min_{c\in\mathbb{R}} \Big\{ \frac{1}{n\tau} \sum_{i=1}^{n} (y_i - c)^- - c \Big\}. \tag{1.13}$$

Cependant dans le cadre de la régression deux points sont à noter :

- On dispose d'un échantillon $\mathcal{D}_n$ dont les composantes sont supposées indépendantes mais en général (en absence de répétitions) non distribuées suivant la même loi.
- On souhaite pouvoir prédire une valeur en un point x qui n'est pas forcément associée à une entrée représentée dans $\mathcal{D}_n$.

Pour gérer ces deux points et prédire une valeur en un point $x \in \mathcal{X}$, l'idée est de pondérer chaque point de l'échantillon d'apprentissage en fonction de sa localisation par rapport à x. Intuitivement un point x_i très éloigné de x doit très peu contribuer à l'estimation et inversement pour un point proche. Dans cette classe de métamodèles nous faisons une distinction entre deux types d'approches se basant sur ce principe : l'approche de type K-plus proches voisins et les approches à base d'arbres de partitionnement.

La méthode de type K-plus proches voisins donne un poids $1/K$ aux K points de l'échantillon les plus proches de x et un poids nul pour les points plus éloignés. Utiliser cette technique implique définir une distance. En pratique le choix de la distance est l'élément central de cette méthode car c'est elle qui va permettre l'extraction d'informations cohérentes à partir de $\mathcal{D}_n$. Parmi les distances classiques il y a la distance euclidienne (pondérée ou non) et la distance de Mahalanobis. Bien que ces distances apportent des résultats satisfaisants. En pratique trouver la distance optimale reste la difficulté majeur lorsque l'on utilise cette méthode.

Les méthodes basées sur les arbres de partitionnement permettent d'échapper (en parti) au problème de définition de distance. L'idée est d'utiliser une règle de classification séquentielle pour créer des classes dans l'échantillon $\mathcal{D}_n$. Ces classes sont définies sur l'espace $\mathcal{X}$ et en forment un partitionnement sensé être adapté au problème. Une illustration du processus de classification séquentielle et d'un partitionnement est présenté Figure 1.9. Pour prédire une valeur en x, les méthodes basées sur des arbres vont allouer des poids aux points de $\mathcal{D}_n$ en fonction de leur position dans l'arbre. Ainsi chaque point de l'échantillon partageant une classe avec x se voit attribuer un poids égal à l'inverse de la population de la classe. Si bien que la prédiction sera constante dans chaque cellule du partitionnement. L'estimation de g avec un seul arbre est connue comme une approche souffrant d'une trop grande dépendance en $\mathcal{D}_n$. Pour diminuer cette dépendance il existe différentes méthodes qui utilisent des arbres, notamment le Bagging et les forêts aléatoires. L'idée directrice de ces méthodes et est qu'il est possible d'introduire de l'aléa dans la création d'un arbre et donc dans la création d'un partitionnement. Comme chaque partitionnement permet de faire une prédiction, il est possible de faire ce qu'on appelle *une méta-prédiction* en agrégeant les résultats des différents arbres. Les méthodes utilisant la méta-prédiction sont connues pour améliorer les résultats en réduisant la variance de prédiction.

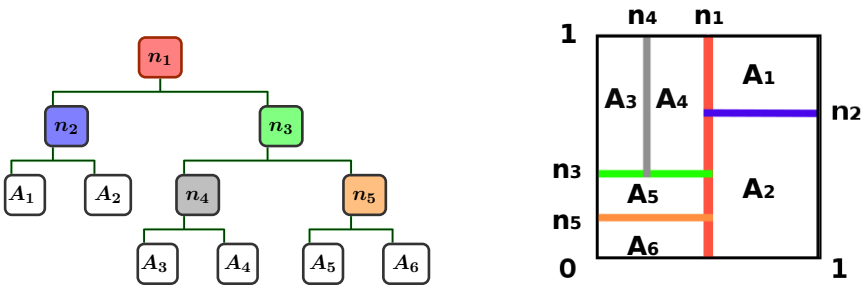


Figure 1.9: A gauche un arbre de partitionnement. Les noeuds n_i $(1 \leqslant i \leqslant 5)$ représentent une étape de la classification séquentielle, les feuilles A_i $(1 \leqslant i \leqslant 6)$ représentent les classes obtenues. A droite: $\mathcal{X} = [0,1]^2$ partitionné par l'arbre de gauche. La prédiction est constante dans chaque classe A_i.

En pratique une des difficultés majeurs de cette méthode est la définition d'une règle de classification adaptée à la quantité que nous souhaitons estimer. Il existe différentes règles de classification (voir Ishwaran [2015] pour divers exemples) mais aucune ne semble adaptée à l'estimation de mesures de risque autres que la moyenne.

1.5.2 Méthodes basées sur l'analyse fonctionnelle et les M-estimateurs

Avec ce type de méthode, l'hypothèse principale est que la fonction g visée vie dans un espace fonctionnel $\mathcal{H}$. Cet espace $\mathcal{H}$ peut être un espace de Hilbert de dimension finie, par exemple l'ensemble des fonctions linéaires. Dans ce cas g s'écrira

$$g_\alpha(x) = \alpha_0 + x \cdot \alpha_1 \quad \text{avec} \quad (\alpha_0, \alpha_1) \in \mathbb{R}^{d+1}.$$

La fonction g peut également être supposée vivre dans l'espace des fonctions polynomiales de degré au plus $m \in \mathbb{N}^*$, c'est à dire

$$g_\alpha(x) = \alpha_0 + \alpha_1 \cdot x + \alpha_2 \cdot x^2 + \cdots + \alpha_m x^m \text{avec} \quad \alpha_0 \in \mathbb{R} \text{ et } \forall 1 \leqslant i \leqslant m, \ \alpha_i \in \mathbb{R}^d.$$

L'espace $\mathcal{H}$ peut également être un espace de Hilbert de dimension infinie, par exemple un RKHS. Dans ce cas g s'écrira

$$g_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) \text{ avec } \forall 1 \leqslant i \leqslant n, \ \alpha_i \in \mathbb{R}^d,$$

et $k(.,.)$ le noyau associé à $\mathcal{H}$.

L'avantage d'utiliser des espaces de Hilbert est qu'ils sont équipés d'un produit scalaire, ce qui facilite l'estimation des paramètres du modèle. En effet avec les formalismes introduits, il suffit d'estimer un nombre fini de paramètres α_i (ensembles des coefficients de g dans une base associée à $\mathcal{H}$ ou à un sous espace de dimension fini engendré par les données) pour estimer la fonction ciblée.

Une autre possibilité pour créer un métamodèle de g flexible tout en estimant un nombre fini de paramètres est l'utilisation des réseaux de neurones. Sous l'hypothèse que g peut s'écrire comme la sortie d'un réseau de neurones *feedforward* (voir Bishop [1995] pour plus d'informations sur les réseaux de neurones), dans le cas d'un réseau à 3 couches g s'écrira

$$g_\alpha(x) = g_3\left(\sum_{j=1}^{J_2} g_2(\sum_{i=1}^{J_1} g_1(\left\langle \alpha_i^{(h_1)}, x \right\rangle + b_i^{(1)})\alpha_j^{(h_2)} + b_j^{(2)})\alpha^{(h_3)} + b^{(3)}\right), \tag{1.14}$$

avec $\alpha_i^{(h_1)} = (\alpha_{i1}^{(h_1)}, ..., \alpha_{id}^{(h_1)}) \in \mathbb{R}^d$, $\alpha^{(h_3)}$ et $\alpha_j^{(h_2)} \in \mathbb{R}$, $b^{(3)} \in \mathbb{R}$ et $b_i^{(1)}$, $b_j^{(2)} \in \mathbb{R}$ des poids à optimiser et g_i une fonction possiblement non linéaire appelée fonction transfert. Les entiers J_1 et J_2 représentent le nombre de neurones dans la première et la seconde couche.

Que g soit supposée vivre dans un espace de Hilbert ou soit supposée s'écrire comme la sortie d'un réseau de neurones, l'estimation du métamodèle s'écrit comme un problème d'optimisation. En effet la moyenne conditionnelle peut être définie comme :

$$m(x) = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}\Big((Y_x - \mu)^2\Big). \tag{1.15}$$

L'expectile conditionnel peut être défini comme

$$e_\tau(x) = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}\Big(l_e^\tau(Y_x - e)\Big). \tag{1.16}$$

Le quantile conditionnel peut être défini comme

$$q_\tau(x) = \arg\min_{\mu \in \mathbb{R}} \mathbb{E}\Big(l_q^\tau(Y_x - \mu)\Big). \tag{1.17}$$

et enfin la CVaR et l'expected shortfall conditionnelles comme

$$\mathrm{CVaR}_\tau(x) = \min_{c \in \mathbb{R}} \mathbb{E}\Big((Y_x - c)^2 | Y_x \geqslant q_\tau(x)\Big),$$

et

$$E(x) = \min_{c \in \mathbb{R}} \mathbb{E}\Big((Y_x - c)^2 | Y_x \leqslant q_\tau(x)\Big).$$

Dans tous ces cas il est possible de remplacer l'espérance par son estimateur empirique et d'estimer les paramètres du métamodèle comme l'argument minimisant ce risque. Cependant, les modèles cités étant très flexibles, optimiser directement le risque empirique peut conduire à un métamodèle interpolant les données. Par exemple (sous réserve d'utiliser une bonne fonction transfert) un réseau de neurones avec plus de paramètres qu'il n'y a de données peut interpoler les données ce qui produira un risque empirique nul. Or rappelons que dans le cas boîte noire stochastique, g est observée avec du bruit. De ce fait un métamodèle interpolant les observations ne fournira pas une bonne estimation de g. En effet le métamodèle sera uniquement bon sur l'échantillon

$\mathcal{D}_n$ (sur lequel le risque empirique à été calculé) et très mauvais sur un échantillon indépendant, c'est ce qu'on appelle le sur-apprentissage. Pour éviter cela il est possible d'ajouter une pénalité sur la flexibilité du modèle. Dans le cas où $\mathcal{H}$ est un espace de Hilbert, cette pénalité est une norme sur $\mathcal{H}$, pour les réseaux de neurones on choisit ce qui s'en rapproche le plus, à savoir une norme euclidienne définie sur l'espace des poids du réseau. Dans ce cas le risque utilisé sera un risque empirique régularisé de la forme

$$\mathcal{R}_{r,e}[s] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i - s(x_i)\big) + \lambda \|g\|^{\beta}, \tag{1.18}$$

avec l la fonction de perte qui convient et $\beta > 0$.

Il existe deux choix standards pour β qui sont $\beta = 1$ et $\beta = 2$.

- L'utilisation de $\beta = 2$ est classique pour pénaliser des modèles linéaires (voir Hastie et al. [2009]), des réseaux de neurones (voir Bishop [1995]) ou la pénalisation des méthodes à noyaux (voir Steinwart and Christmann [2008]). Cette pénalisation fournie des solutions avec des coefficients d'amplitude similaire, ce qui la rend utile pour estimer les paramètres d'un modèle lorsque ces derniers ont tous une influence comparable. De plus cette pénalisation permet l'obtention de solution analytiques dans de nombreux cas en raison du caractère différentiable de la norme L^2.

- L'utilisation de $\beta = 1$ est privilégiée quand seulement un sous ensemble des coefficients sont réellement influents car la norme L^1 fournie des solutions *sparses* (voir Bach et al. [2012] pour plus de détails). En effet elle contraint un sous ensemble des paramètres à valoir zéro. Cette norme est utilisée pour pénaliser les régressions linéaires, auquel cas on parlera de méthode du Lasso (voir Tibshirani [1996]), mais aussi pour pénaliser les réseaux de neurones (voir Ye and Sun [2018]) ou pour pénaliser les méthodes à noyaux Lopez-Martinez [2017].

Enfin d'autre pénalisations existent. Une introduction à différentes pénalisations ainsi qu'une discussion sur leur impact peut être trouvée dans Bach et al. [2012].

A noter que le (hyper)paramètre $\lambda \in \mathbb{R}^+$ est une quantité qui doit être fixée par l'utilisateur. Ce paramètre règle le compromis biais variance. Plus précisément prendre λ petit autorisera une grande flexibilité du modèle et donc introduit peu de biais. A l'inverse prendre λ grand forcera le modèle à être très régulier quitte à avoir un risque empirique très élevé. La sélection de λ se fait généralement par validation croisée et sera discutée dans le Chapitre 2.

Une fois la question du sur-apprentissage en partie réglée (λ fixé), les paramètres d'un estimateur de la moyenne conditionnelle peuvent s'obtenir comme

$$\boldsymbol{\alpha} = \arg\min \frac{1}{n}\sum_{i=1}^{n} \Big(y_i - s_\alpha(x_i)\Big)^2 + \lambda \|m\|^{\beta}.$$

Pour l'expectile conditionnel ils peuvent s'obtenir comme

$$\boldsymbol{\alpha} = \arg\min \frac{1}{n}\sum_{i=1}^{n} l^e_\tau(y_i - s_\alpha(x_i)) + \lambda \|e\|^{\beta}.$$

Pour le quantile conditionnel ils peuvent s'écrire

$$\boldsymbol{\alpha} = \arg\min \frac{1}{n} \sum_{i=1}^{n} l_\tau(y_i - s_\alpha(x_i)) + \lambda \|q\|^\beta .$$

Pour la CVaR et l'expected shortfall conditionnelles c'est différent. Comme ces quantités ne sont pas "élicitables", leurs coefficients ne peuvent s'obtenir comme l'argument minimisant un risque. En revanche une propriété intéressante de la CVaR fait qu'elle peut être considérée comme le minimum d'un risque. Ainsi pour un jeu de données i.i.d la CVaR peut s'obtenir sous la forme 1.12. Cependant le minimum est une quantité scalaire qui ne permet pas d'estimer une fonction sur un compact autre que par une valeur constante.

Bien que les méthodes présentés dans cette section ne sont pas applicables directement, dans la littérature certaines pistes ont été développées. Dans Rockafellar et al. [2014] une méthode basée sur les *quadrangle* propose un formalisme pour faire de la régression CVaR. Dans Acerbi and Szekely [2014] les auteurs montrent qu'une estimation conjointe du quantile et de la CVaR sont possibles par des méthodes classiques et proposent une fonction de perte pour cette tache.

1.5.3 De l'estimation par maximum de vraisemblance à l'inférence Bayesiennes

Maximum de vraisemblance

Commençons par définir un modèle statistique sur les observations de la forme :

$$Y_x = g(x) + \varepsilon(x), \tag{1.19}$$

avec g la fonction ciblée et ε une variable aléatoire rendant compte de la distribution des observations autour de g. Sous ce formalisme les paramètres du modèle peuvent êtres estimés par maximum de vraisemblance, avec la vraisemblance définie comme

$$p(\mathcal{Y}_n|\mathcal{X}_n, g_\alpha) = \prod_{i=1}^{n} p\big(\varepsilon|x_i, \alpha\big) = \prod_{i=1}^{n} p\big(y_i - g_\alpha|x_i, \alpha\big), \tag{1.20}$$

si les observations sont i.i.d. Plus précisément les paramètres sont définis comme

$$\boldsymbol{\alpha} = \arg\max p(\mathcal{Y}_n|\mathcal{X}_n, g_\alpha).$$

C'est à dire que les paramètres sélectionnés sont ceux qui sous l'hypothèse de modèle (1.19), donnent la plus grande probabilité d'observer $\mathcal{D}_n$.

Ici le terme ε est central pour au moins deux raisons :

- Il rend compte des caractéristiques de l'erreur d'observation. Pour caricaturer, il pose un a priori sur les queues ou sur la forme de la distribution de l'erreur. Par exemple il indique au modèle si des valeurs extrêmes peuvent êtres observées. Plus

précisément, le terme ε peut contraindre le modèle à prendre en compte les valeurs extrêmes dans l'estimation ou non. Un exemple simple pour comprendre ce point est le suivant : supposons que l'on observe dans un premier temps une variable aléatoire Y avec un bruit gaussien ε_1 de moyenne a et variance b puis avec un bruit Cauchy ε_2 de paramètre a', b'. La Figure 1.10 montre les deux distributions. Bien que ces deux distributions soient très semblables sur l'intervalle $[-1.5, 1.5]$, en dehors de celui-ci, les queues de distributions sont très différentes. Observer un point $y = -2$ avec la loi gaussienne est supposé très rare car la densité est quasi nulle dans ce voisinage. Ainsi dans la procédure d'inférence le modèle va être modifié pour que cette observation soit plus plausible sous l'hypothèse gaussienne et donc modifier la moyenne de ε_1. En effet en translatant un peu la moyenne alors l'observation devient plus vraisemblable. En revanche à moyenne fixée, il faudrait considérablement augmenter la variance si l'on souhaite sensiblement augmenter la probabilité de cette observation. Dans le cas de la distribution suivant une loi de Cauchy, la probabilité d'observer $y = 2$ reste élevée. En effet la réalisation $y = -2$ est seulement 10 fois moins probable que d'observer un point en 0. Avec ce prior, dans la procédure d'estimation la moyenne ne vas pas être énormément modifiée. Concrètement cela implique que le modèle avec l'hypothèse ε_1 risque de sur-apprendre (si des valeurs extrêmes sont présentes dans $\mathcal{D}_n$) car pour maximiser la vraisemblance il faudra largement modifier le paramètre de moyenne. Inversement dans le cas de la distribution Cauchy, des valeurs extrêmes sont supposées largement probables donc le modèle ne va pas sur-apprendre dans ce cas. Le point négatif à l'utilisation d'une distribution de Cauchy est que si aucune valeur extrême n'est présente dans l'échantillon alors utiliser cet hypothèse donnera un modèle qui aura tendance à sous-apprendre. Une description plus formelle de ce phénomène est proposée dans la section 1.5.3.

- Si l'objectif est l'estimation d'une mesure de risque bien identifiée, alors la loi de la variable aléatoire ε doit être sélectionnée en connaissance de cause. En effet les paramètres du modèle sont estimés par maximum de vraisemblance. Or maximiser la vraisemblance (1.20) peut être vu comme la maximisation d'un risque empirique. Ainsi pour estimer une mesure de risque g fixée, il peut être commode de sélection une distribution ε qui produira une vraisemblance liée à un risque empirique connu.

Maintenant discutons plus en détails du lien entre l'hypothèse sur ε et la mesure de risque g estimée. Dans le cas classique où g est la moyenne conditionnelle nous proposons deux possibilités pour ε. La première est l'hypothèse classique d'une distribution gaussienne. En effet la vraisemblance associée à une distribution gaussienne s'écrit

$$\begin{aligned} p(\mathcal{Y}_n|\mathcal{X}_n, g) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{\big(y_i - g(x_i)\big)^2}{2\sigma^2} \right) \\ &= \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(- \frac{\sum_{i=1}^{n} \big(y_i - g(x_i)\big)^2}{2\sigma^2} \right). \end{aligned} \tag{1.21}$$

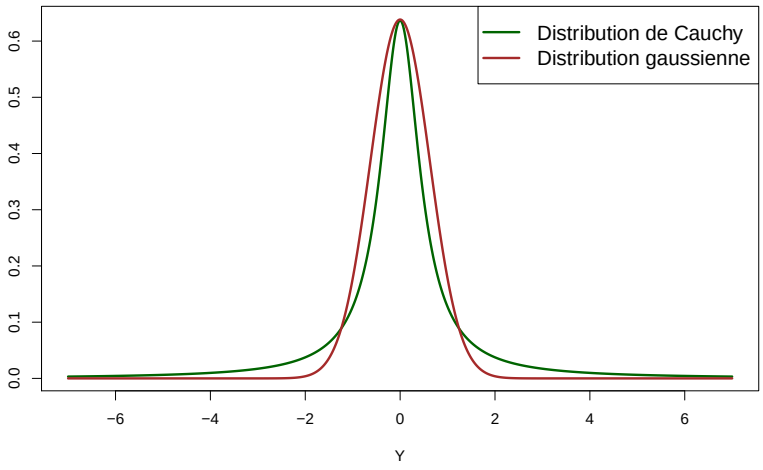


Figure 1.10: Illustration de la différence des queues de distribution entre une loi de Cauchy et une loi normale.

Maximiser (1.21) par rapport à g est donc équivalent à minimiser le risque empirique (1.15) et donc g sera un estimateur de la moyenne conditionnelle. C'est l'approche majoritairement utilisée dans la littérature. Une étude détaillée sur l'utilisation de cette vraisemblance est disponibles dans Rasmussen and Williams [2006].

Pour l'estimation de l'expectile il est possible d'utiliser la loi normale asymétrique dont la vraisemblance associée est donnée par

$$\begin{aligned} p(\mathcal{Y}_n|\mathcal{X}_n, g) =& \prod_{i=1}^{n} \frac{\sqrt{2\tau(1-\tau)}}{\sigma\sqrt{\pi}(\sqrt{\tau}+\sqrt{1-\tau})} \exp\Big(-\frac{l_e^\tau(y_i - g(x_i))}{2\sigma^2}\Big) \\ =& \Big(\frac{\sqrt{2\tau(1-\tau)}}{\sigma\sqrt{\pi}(\sqrt{\tau}+\sqrt{1-\tau})}\Big)^n \exp\Big(-\frac{\sum_{i=1}^n l_e^\tau(y_i - g(x_i))}{2\sigma^2}\Big). \end{aligned} \tag{1.22}$$

Maximiser la vraisemblance associée à une distribution gaussienne asymétrique est donc équivalent à maximiser le risque empirique (1.16). Une illustration de la loi gaussienne asymétrique est disponible Figure 1.11.

Pour l'estimation d'un quantile conditionnel, un raisonnement analogue à ce qui précède conduit à la possibilité d'utiliser la distribution Laplace asymétrique. En effet la densité d'une variable aléatoire suivant une loi Laplace asympétrique est donnée par

$$f(x) = \frac{\tau(1-\tau)}{\sigma} \exp\Big(-\frac{\tau x \mathbb{1}_{x\geqslant 0} + (\tau-1)x\mathbb{1}_{x<0}}{\sigma}\Big).$$

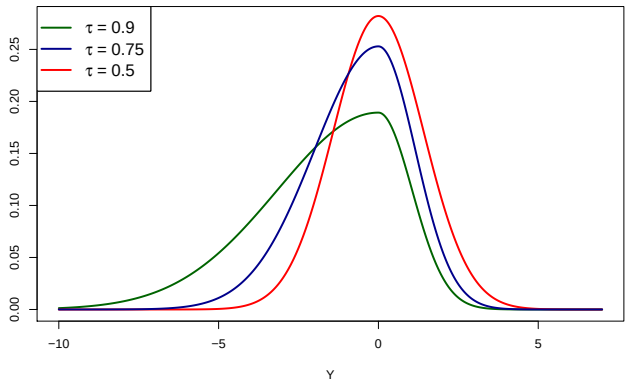


Figure 1.11: Lois gaussiennes asymétriques pour différentes valeurs de τ.

Utiliser la vraisemblance associée

$$p(\mathcal{Y}_n|\mathcal{X}_n, g) = \prod_{i=1}^{n} \frac{\tau(1-\tau)}{\sigma} \exp\left(-\frac{l_q^\tau(y_i - q(x_i))}{\sigma}\right) \tag{1.23}$$

est équivalent à minimiser le risque empirique (1.17), ce qui produira un estimateur du quantile. L'utilisation de cette vraisemblance est proposée dans Yu and Moyeed [2001], Kozumi and Kobayashi [2011], Abeywardana and Ramos [2015].

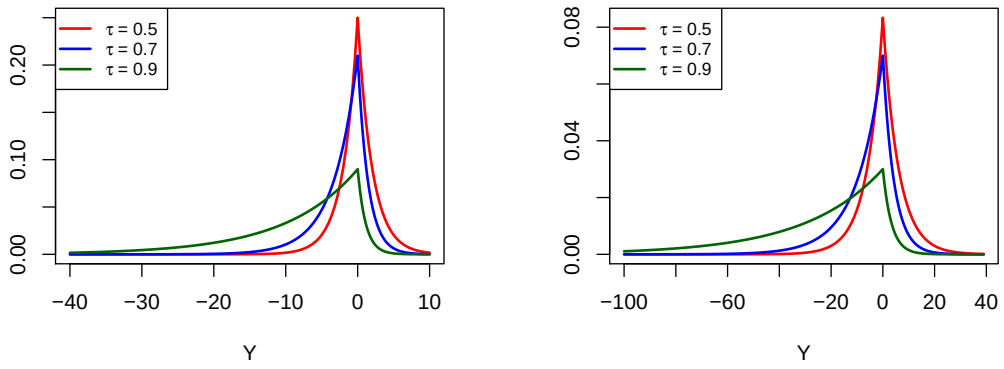


Figure 1.12: Distributions Laplace asymétriques avec différentes valeurs de τ et de σ. A droite $\sigma = 1$, à gauche $\sigma = 3$. On voit l'impact fort du paramètre σ sur l'étalement de la loi.

Pour estimer la CVaR et l'expected shortfall avec ce formalisme la question reste ouverte.

Inférence Bayesienne

Dans certains cas de l'information sur les paramètres $\boldsymbol{\alpha}$ des métamodèles introduits Section 1.5.2 est disponible ou bien une simple estimation scalaire des paramètres est insuffisante. Le premier cas de figure peut être classique en biologie et écologie (voir McCARTHY and Masters [2005], Isci et al. [2013]) alors que le second l'est pour les procédures d'optimisation bayésienne par exemple (voir Jones et al. [1998], Shahriari et al. [2015]). Dans ce contexte on utilisera un *a priori* sur ces paramètres inconnus, puis en utilisant la formule de Baye:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)},$$

il est possible de fournir une distribution de probabilité sur $\boldsymbol{\alpha}$ sachant les données. En effet on peut écrire

$$\mathbb{P}(\boldsymbol{\alpha}|\mathcal{D}_n) \propto \mathbb{P}(\mathcal{D}_n|\boldsymbol{\alpha})\mathbb{P}(\boldsymbol{\alpha}),$$

où $\mathbb{P}(\boldsymbol{\alpha})$ est un *a priori* sur $\boldsymbol{\alpha}$. Comme dit plus haut un a priori peut être sélectionné en fonction de connaissances sur les paramètres à estimer mais il peut aussi être sélectionné dans le but de fournir des garanties sur la convergence de la procédure d'estimation ou bien pour simplifier la procédure d'inférence en utilisant des quantités conjuguées (voir Ghaderinezhad and Ley [2019], Diaconis and Ylvisaker [1979] pour plus de détails sur les deux dernières motivations). Avec cette approche nous disposons d'une distribution a posteriori sur $\boldsymbol{\alpha}$ et non une estimation scalaire comme ce qui est le cas avec les méthodes fréquentistes introduites précédemment. Cependant ce gain d'information est souvent coûteux d'un point de vue calculatoire car la majorité des méthodes utilise des techniques de type Markov Chain Monte Carlo qui doivent simuler des marches aléatoires pour un grand nombre de pas de temps (voir Andrieu et al. [2003] pour une introduction à ces méthodes en apprentissage statistique).

Robustesse

Supposons

$$Y = m + \varepsilon,$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ avec σ connue et l'a priori $\mu \sim \mathcal{N}(m_0, \sigma_0^2)$. Il est possible de montrer que si nous disposons d'une unique observation $Y = y$ alors la distribution a posteriori sur m est gaussienne de moyenne

$$\mu_1 = \sigma_1^2(m_0[\sigma_0^{-2} + \sigma^{-2}] + [y - m_0]\sigma^{-2}),$$

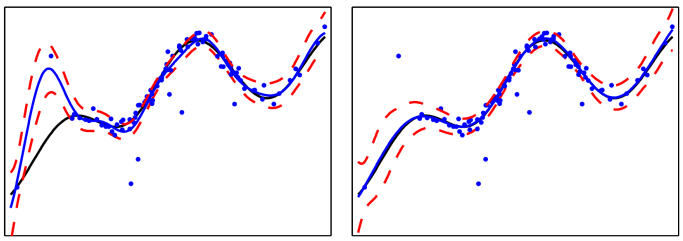
avec $\sigma_1^2 = \dfrac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}$. En d'autres termes la moyenne de la distribution a posteriori dévie de m_0 proportionnellement à $y - m_0$ et cela linéairement d'un facteur $\dfrac{\sigma_0^2}{\sigma^2 + \sigma_0^2}$. Cette pro-

priété peut être source de sur-apprentissage si une valeur extrême est présente dans le jeu d'apprentissage. Avec les méthodes de la deuxième catégorie (fonctionnelles d'inspiration fréquentiste) le sur-apprentissage est géré en parti par le paramètre régularisant λ qui est choisi à l'aide des données par validation croisée. Avec le formalisme bayésien une possibilité pour gérer la présence de valeurs extrêmes est l'utilisation d'autres distributions sur ε, par exemple la loi de Cauchy (aussi appelée Student-t). La vraisemblance associée à une loi de Cauchy de paramètre $\gamma > 0$ s'écrit :

$$p(\mathcal{Y}_n|\mathcal{X}_n, g) = \prod_{i=1}^{n} \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\big(y_i - g(x_i)\big)^2 + \gamma^2}\right). \tag{1.24}$$

Les propriétés liées à cette vraisemblance ont été étudiées dans Dawid [1973] où l'auteur montrent entre autre que si ε suit une loi de Cauchy et qu'un a priori gaussien est posé sur g alors conditionnellement à une observation à l'infinie la moyenne de la distribution a posteriori tend vers la moyenne de l'a priori. Une étude plus générale est détaillée dans O'Hagan [1979]. L'utilisation de cette vraisemblance est proposée dans Jylänki et al. [2011] et une illustration de l'utilité de la méthode est proposée Figure 1.13. Cependant contrairement aux cas où on utilise des distributions gaussiennes, gaussiennes asymétrique ou Laplace, ici g ne peut pas être interprétée comme une mesure de risque classique.

Dans la suite de cette thèse nous ne traitons pas les aspects relatifs à la *robustesse*. Toutefois des cas de sur apprentissage avec des méthodes bayésiennes sont relevés. S'intéresser à cet aspect *robustesse* semble donc tout à fait pertinent lorsqu'il s'agit d'estimer des mesures de risque par des métamodèles d'inspiration bayésienne.



(a) Gaussian observation model. (b) Student-t observation model.

Figure 1.13: Illustration issue de Vanhatalo et al. [2009]. En noir la fonction à estimer, en bleu l'estimateur. A droite estimation faite avec une distribution Cauchy sur ε, à gauche une distribution gaussienne.

1.6 Optimisation bandit

L'optimisation de type bandit a initialement été introduite sous le formalisme du problème du bandit stochastique (voir Bubeck et al. [2012] pour une étude détaillée). Sa

formulation initiale est la suivante : un joueur est face à K machines à sous [3] sur lesquelles il n'a aucune connaissance. Partant avec une somme initiale T, en jouant les machines une par une, son objectif est de trouver la machine qui lui permettra de cumuler le plus de gains. Cet objectif se traduit mathématiquement en un problème d'optimisation du regret cumulé défini par

$$R_T = T\mu^* - \mathbb{E}\Big(\sum_{t=1}^{T} Y_{K(t),t}\Big),$$

avec μ^* le maximum des moyennes sur l'ensemble des K machines et $Y_{K(t)}$ la variable aléatoire de loi inconnue représentant les gains retournés par la machine sélectionnée au temps t avec $1 \leqslant K(t) \leqslant K$. Pour maximiser R_T, une stratégie basique consiste à jouer dans un premier temps toutes les machines un nombre fixé de fois, pour se faire une idée de leur moyenne. Puis, dans une seconde phase, il faut trouver une stratégie permettant majoritairement de collecter des gains avec peu de risques. Pour cela il faut jouer les machines identifiées comme bonnes, c'est l'exploitation. Mais il faut également jouer suffisamment toutes les machines pour trouver les meilleures, c'est l'exploration. Une bonne stratégie permet de gérer au mieux ce compromis exploration exploitation.

Sur le problème du bandit stochastique, un autre objectif peut être l'identification de la meilleure machine avec un budget fixé T, sans se préoccuper des pertes ou gains enregistrés pendant la séquence de jeu. Ce point de vue transforme le problème initial en problème purement exploratoire. Dans ce cas l'objectif sera l'optimisation du regret simple r_T défini comme

$$r_T = \mu^* - \mu_T,$$

avec μ_T la moyenne de la machine sélectionnée comme étant la meilleure après avoir utilisé l'ensemble du budget. Ici le compromis exploration exploitation est toujours présent mais davantage dissimulé. Dans un cadre stochastique, l'exploitation permettra de fournir une estimation plus précise de la moyenne et donc de retourner une machine identifiée comme optimale avec de plus grandes garanties probabilistes. Tout l'enjeu est d'exploiter des machines à fort potentiel pour avoir une estimation précise de leur moyenne tout en explorant et en échantillonnant juste ce qu'il faut les machines sous-optimales pour s'assurer de leur sous-optimalité.

A noter qu'il existe un lien entre le regret simple et le regret cumulé qui borne le regret simple par le regret cumulé :

$$\mathbb{E}(r_T) \leqslant \frac{\mathbb{E}(R_T)}{T}.$$

Cette propriété est largement utilisée pour donner des garanties en terme de regret simple à partir du regret cumulé mais nous ne l'utiliserons pas dans cette thèse.

Comme présenté plus haut, l'objectif initial de ce genre d'approche a été d'optimiser la moyenne sur un espace discret et fini. Les algorithmes efficaces initialement définis pour cette tache sont : Thompson sampling Thompson [1935, 1933], UCB et ε-greedy Auer et al. [2002a]. Depuis ce formalisme a été élargi suivant cinq directions :

[3] "bandit slot machines", en anglais.

- la recherche d'algorithmes asymptotiquement optimaux pour l'optimisation du regret cumulé, par exemple l'algorithme KL-UCB Garivier and Cappé [2011], ou minimax optimaux, comme l'algorithme Minimax Optimal Strategy in Stochastic case (MOSS) Audibert and Bubeck [2009] ;
- l'optimisation de différentes mesures de risque comme le quantile Szorenyi et al. [2015], David and Shimkin [2016], la CVaR Kolla et al. [2019], la mesure de risque entropique Maillard [2013] ou encore le critère moyenne-variance Sani et al. [2012] dans un espace discret ;
- l'optimisation dans un contexte où l'espace de recherche est discret mais largement plus grand que le budget T et qu'aucune structure existe sur l'espace de recherche Wang et al. [2009], Carpentier and Valko [2015]. Dans ce cas l'objectif est l'optimisation de la moyenne ;
- le cadre *contextuel* où la distribution des gains dépend d'un paramètre décrivant l'état du système Li et al. [2010] ;
- le cadre *adversarial* où la distribution des gains n'est pas supposée aléatoire mais régie par un adversaire Bubeck et al. [2012]. Sous ce formalisme le point de vue *worst-case* (voir Auer et al. [2002b]) se rapproche des considérations *risk-averse* mentionnées dans cette thèses car l'intérêt porte sur le regret cumulé dans le pire cas possible à stratégie fixée ;
- l'optimisation du regret cumulé et regret simple (g étant la moyenne conditionnelle) sur des espaces continus avec une hypothèse sur la régularité (connue ou non) sur la fonction g. Ce formalisme porte le nom de $\mathcal{X}$-armed bandit et regroupe notament les algorithmes HOO Bubeck et al. [2011], StoOO Munos [2014], HCT Azar et al. [2014].

Dans le Chapitre 3 nous proposons un formalisme qui lie les points 2 et 6. Nous cherchons à optimiser une mesure de risque autre que la moyenne dans le cas où l'espace $\mathcal{X}$ est continu borné et que la fonction g présente une régularité en espace. Nous proposons un algorithme capable de réaliser cette tache et montrons sont efficacité quand l'objectif est l'optimisation d'un quantile conditionnel d'ordre fixé ou d'une CVaR conditionnelle d'ordre fixée. Nous détaillons l'étude d'une borne supérieure sur le regret simple pour ces deux cas. De plus nous proposons une écriture générique du regret simple. Cette écriture permet d'établir une borne supérieure sur le regret simple pour différentes mesures de risque, pourvu qu'on soit capable de borner l'erreur d'estimation sur g en fonction du nombre d'observations. Dans la suite de cette section nous proposons d'introduire les outils et idées utilisés dans le Chapitre 3.

1.6.1 Partitionnement hiérarchique de l'espace $\mathcal{X}$

Comme la plupart des algorithmes $\mathcal{X}$-armed bandits, l'algorithme Stochastic Risk Optimistic Optimisation (StoROO) développé dans cette thèse utilise un partitionnement

hiérarchique de $\mathcal{X}$. Formellement un partitionnement hiérarchique infini $\mathcal{P} = \{\mathcal{P}_{h,j}\}_{h,j}$ de $\mathcal{X}$ peut se définir comme

$$\mathcal{P}_{0,1} = \mathcal{X}, \quad \mathcal{P}_{h,j} = \bigcup_{i=0}^{K-1} \mathcal{P}_{h+1,Kj-i},$$

avec K le nombre de sous-régions obtenues après avoir explosé une cellule et $\mathcal{P}_{h,j}$ la j-ème cellule à profondeur h. L'une des utilités d'un tel partitionnement est de structurer la recherche d'optima. Basé sur un partitionnement hiérarchique, l'algorithme va explorer l'espace en intensifiant les divisions dans les zones à fort potentiel. Cette intensification de la division permettra d'avoir une quantité plus dense de points dans les zones contenant potentiellement un optimum global. Une fois un partitionnement hiérarchique

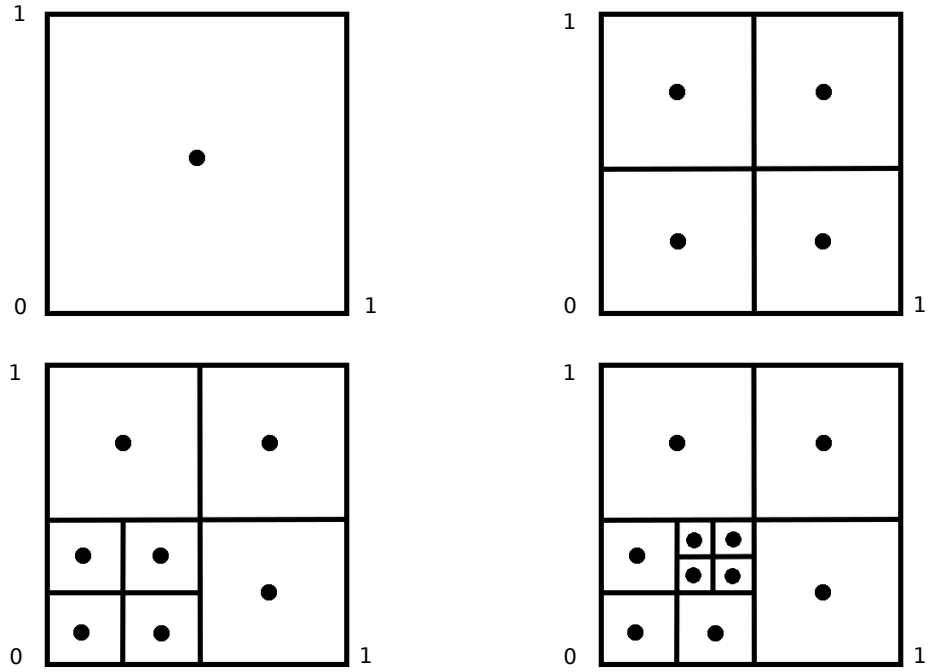


Figure 1.14: Exemple d'un partitionnement hiérarchique avec $d = 2$, en bas à droite l'arbre de partitionnement $\mathcal{T}_3$ obtenu après trois explosions de cellules $\mathcal{T}_3 = \{\mathcal{P}_{1,1},\ \mathcal{P}_{1,2},\ \mathcal{P}_{1,4},\ \mathcal{P}_{2,9},\ \mathcal{P}_{2,10},\ \mathcal{P}_{2,11},\ \mathcal{P}_{2,12}\}$.

sélectionné, il faut définir une stratégie qui indiquera dans quelle cellule échantillonner et quand exploser la cellule. Dans les deux sections suivantes nous introduisons les outils pour définir une telle stratégie.

1.6.2 Stratégie UCB

Le choix de la cellule à échantillonner se fait par l'optimisation d'une fonction d'acquisition f_a. Ici nous utilisons une fonction d'acquisition de type *Upper Confidence Bound* (UCB),

qui doit donner un majorant de la fonction cible avec grande probabilité. Ainsi, à chaque pas de temps, le point ayant la plus grande UCB sera échantillonné. L'idée derrière cette stratégie est *l'optimisme* devant l'incertain. Nous échantillons un point, qui, dans une configuration qui sera la meilleure pour lui, fait qu'il sera l'argument maximisant g. La Figure 4.10 montre un exemple de création et d'utilisation d'UCB dans le cas où g est une fonction non bruitée. A noter qu'il existe différentes manières d'échantillonner la cellule sélectionnée. Dans Bubeck et al. [2011] l'échantillonnage se fait aléatoirement dans la cellule alors que dans Munos [2014] et dans cette thèse le point échantillonné correspond au centre de la cellule.

Avant de considérer directement le cas où g est observée avec un bruit, nous considérons le cas deterministe pour donner une intuition sur la façon de créer f_a. Premièrement, il sera montré dans le Chapitre 3 que pour obtenir un algorithme convergent vers x^*, il suffit de définir une fonction f_a majorant g uniquement dans la cellule contenant x^*. Pour fournir une telle fonction d'acquisition, partant d'une observation $g(x)$, une majoration du maximum de g dans la cellule contenant x^* peut être obtenue en ajoutant à $g(x)$ une quantité majorant la croissance potentielle de g entre x et le bord de la cellule. Or une majoration de l'accroissement de g dans la cellule contenant x^*

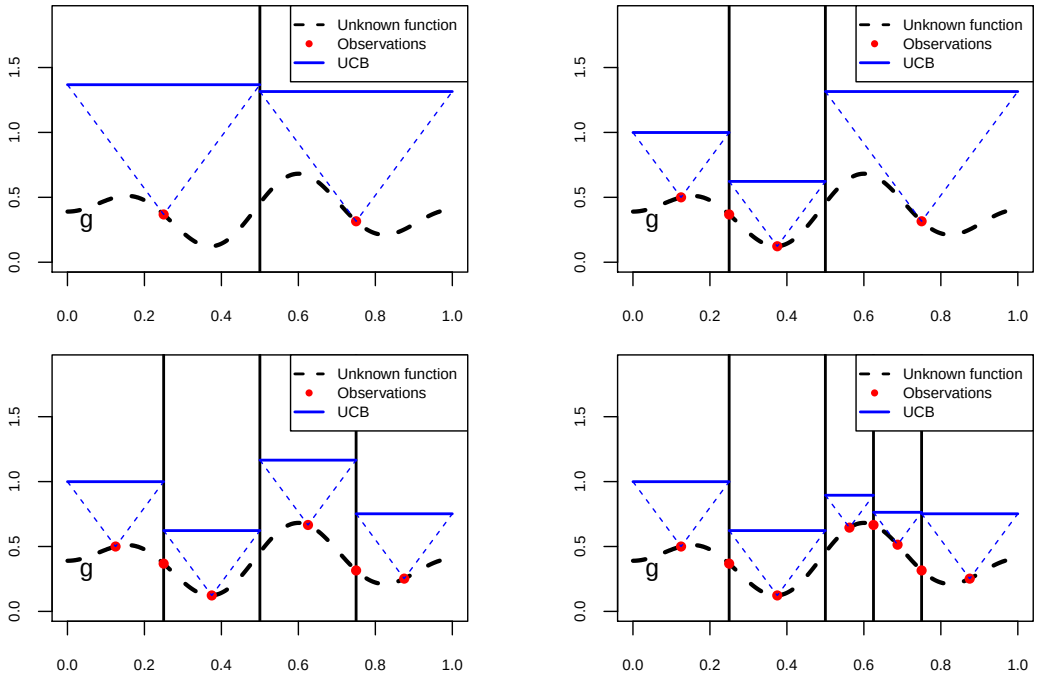


Figure 1.15: Illlustration d'une UCB dans le cas où l'observation de g est non bruitée. A gauche initialisation, à droite la cellule de gauche a été sélectionnée car son UCB était la plus grande. Sélectionner la cellule de gauche à conduit à son explosion en deux nouvelles cellules, chacune contenant une nouvelle observations. En bas à gauche sont représentées les cellules après 3 explosions, en bas à droites cellules après 4 explosions.

peut être obtenue pour les fonctions vérifiant une condition hölderienne par rapport à un argument maximisant la fonction, *i.e*

$$\forall x \in \mathcal{X}, \quad g(x) \geqslant g(x^*) - \beta ||x - x^*||^\gamma \text{ with } \gamma, \ \beta \ > 0. \tag{1.25}$$

Ainsi sous réserve que cette hypothèse soit valide et que g soit observée sans bruit, nous sommes capable de fournir une fonction d'acquisition qui, au moins dans la cellule contenant x^*, est une UCB pour g.

Dans le cas où g est observée avec du bruit (le point de vue de cette thèse) il nous faut utiliser un outil statistique permettant de fournir des intervalles de confiance à partir des observations. Pour construire f_a, nous remplaçons g par une borne supérieure de confiance U à laquelle nous ajouterons un majorant de la croissance maximale de g dans la cellule. Cette idée est représentée Figure 1.16.

La création d'une borne supérieure de confiance $U(x)$ sur $g(x)$ demande d'échantillonner plusieurs fois la cellule. Ainsi contrairement au cas déterministe, une fois une cellule sélectionnée, il est possible de soit l'échantillonner davantage pour raffiner les bornes de confiance sur g, soit d'exploser la cellule pour réduire le biais introduit sur f_a (biais lié au diamètre de la cellule et à l'hypothèse (1.25)). Cette décision peut être prise grâce à une règle très intuitive : tant que l'erreur de biais est plus petite que le diamètre de l'intervalle de confiance autour de g (i.e l'erreur d'estimation locale sur g), alors on échantillonne la cellule, mais dès que les proportions s'inversent, alors on explose la cellule. Cette règle est représentée Figure 1.16.

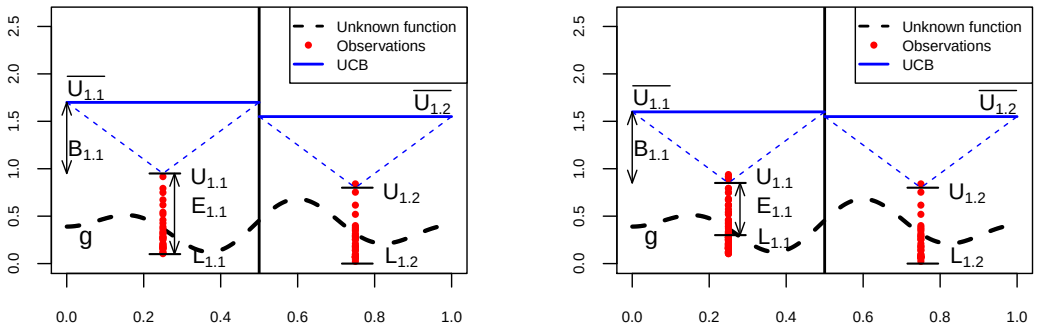


Figure 1.16: A gauche les deux erreurs de biais $B_{1.1}$ et $B_{1.2}$ sont plus petites que les erreurs d'estimation locales $E_{1.1}$ et $E_{1,2}$, donc la cellule sélectionnée doit être échantillonnée davantage pour réduire l'erreur d'estimation avant d'être explosée. A droite la cellule de gauche a été sélectionnée car elle possède l'UCB la plus grande et son erreur de biais est plus grande que son erreur d'estimation, la cellule va être explosée. Le résultat est présenté Figure 1.17

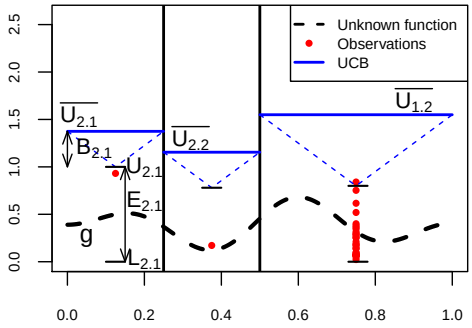


Figure 1.17: Figure 1.16 après explosion.

1.6.3 Création d'intervalles de confiance

Dans cette thèse nous considérons qu'un appel au code est relativement coûteux ce qui implique que les échantillons sont supposés êtres de taille limitée. Ainsi pour créer des intervalles de confiance pour g nous utiliserons des outils statistiques propres au régime non asymptotique. Pour construire des intervalles de confiance sur des données i.i.d [4] en régime non asymptotique, il est possible d'utiliser des inégalités dites de déviation. La plus connue est sûrement l'inégalité de Hoeffding qui borne la déviation de la moyenne de n réalisations de variables aléatoires indépendantes $Y_1, \cdots, Y_n$ bornées par l'intervalle $[0, 1]$. En définissant la moyenne empirique des réalisations

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

tout $\varepsilon > 0$ l'inégalité de Hoeffding donne

$$\mathbb{P}(|\bar{Y}_n - \mathbb{E}(\bar{Y}_n)| \geqslant \varepsilon) \leqslant 2\exp(-2n\varepsilon^2).$$

L'inégalité de Hoeffding donne explicitement des intervalles de confiance autour de la moyenne d'un échantillon. Dans le cadre où g n'est pas la moyenne mais une autre mesure de risque, alors des intervalles de confiance ne s'obtiennent pas de manière immédiate, il faut adapter cette inégalité.

Schématisons un cheminement possible pour obtenir des intervalles de confiance sur le quantile. Dans cette introduction nous notons $\boldsymbol{q}_x(\tau)$ le quantile d'ordre τ de la loi $\mathbb{P}_x$. Supposons que l'on dispose d'une suite de variables aléatoires $\mathcal{Y}_n(x) = (Y_1, \cdots, Y_n)$ i.i.d, l'astuce consiste à considérer la variable aléatoire

$$Z = \mathbb{1}_{Y \leqslant \boldsymbol{q}(\tau)}, \quad \text{et sa moyenne empirique} \quad \widehat{F}^n\big(\boldsymbol{q}(\tau)\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \leqslant \boldsymbol{q}(\tau)},$$

[4] Ou non i.i.d mais nous ne considérerons que le cas i.i.d dans ce travail.

qui n'est autre que la fonction de répartition empirique. Bien évidemment ici nous ne connaissons pas $\boldsymbol{q}(\tau)$, donc calculer la fonction de répartition empirique n'est pas possible. Heureusement ce calcul n'est pas nécessaire, nous avons seulement besoin de son inverse pour créer un intervalle de confiance autour de $\boldsymbol{q}(\tau)$. En effet en définissant $\widehat{F}^{n\ -}$ comme l'inverse généralisée de la fonction de répartition empirique $\widehat{F}^n$ et en combinant les équivalences suivantes

$$\begin{aligned} &\forall\ \varepsilon > 0 \text{ such that } \ \tau + \varepsilon < 1, \quad \widehat{F}^n(\boldsymbol{q}(\tau)) \geqslant \tau + \varepsilon \quad \Leftrightarrow \quad \boldsymbol{q}(\tau) \geqslant \widehat{F}^{n\ -}(\tau + \varepsilon), \\ &\forall\ \varepsilon > 0 \text{ such that } \ \tau + \varepsilon > 0, \quad \widehat{F}^n(\boldsymbol{q}(\tau)) < \tau - \varepsilon \quad \Leftrightarrow \quad \boldsymbol{q}(\tau) \leqslant \widehat{F}^{n\ -}(\tau - \varepsilon), \end{aligned} \tag{1.26}$$

avec l'inégalité de Hoeffding on peut montrer que

$$\mathbb{P}\Big(\widehat{F}^{n\ -}(\tau - \varepsilon) \leqslant \boldsymbol{q}(\tau) \leqslant \widehat{F}^{n\ -}(\tau + \varepsilon)\Big) \geqslant 1 - 2\exp(-2n\varepsilon^2). \tag{1.27}$$

La Figure 1.18 illustre la première équivalence. L'inégalité (1.27) est obtenue lorsque

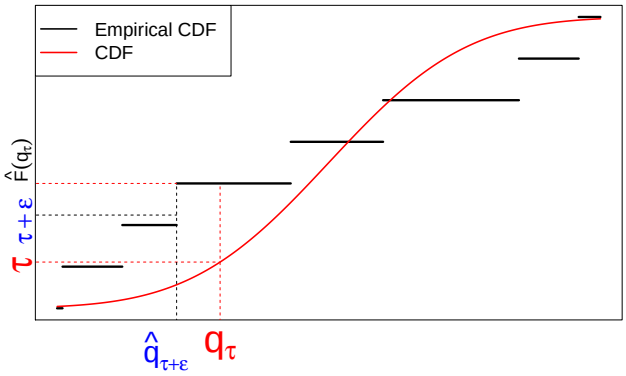


Figure 1.18: Illustration de l'équivalence (1.26)

nous considérons un échantillon i.i.d de taille fixée. Cependant comme présenté plus haut, le nombre de fois qu'une cellule va être échantillonnée dépend de l'UCB qui elle va dépendre des valeurs des observations. De ce fait nous perdons le caractère indépendant des observations. De plus au temps $t \geqslant 0$, le nombre de fois qu'une cellule a été échantillonnée et le nombre de cellules explosées sont des quantités aléatoires. Prendre en compte ces deux facteurs est primordial et sera fait de deux manières dans le Chapitre 3. La manière la plus simple étant de faire une *double borne d'union*. Dans ce cas on obtient, avec l'inégalité de Hoeffding, les bornes supérieures et inférieures de confiance suivantes pour la cellule (h, j) au temps t pour tout $\eta > 0$:

$$U^{\eta}_{h,j}(t) = \begin{cases} \min\left\{q,\ \widehat{F}^t_{h,j}(q) \geqslant \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)} < 1 \\ +\infty & \text{sinon,} \end{cases}$$

$$L^{\eta}_{h,j}(t) = \begin{cases} \max\left\{q,\ \widehat{F}^{t}_{h,j}(q) \geqslant \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)} > 0 \\ -\infty & \text{sinon,} \end{cases}$$

avec $\widehat{F}^{t}_{h,j}$ l'estimateur de la fonction de répartition conditionnelle du point au centre de la cellule (h, j) au temps t, $N_{h,j}(t)$ le nombre de fois que la cellule (h, j) a été échantillonnée et

$$\varepsilon^{\eta,T}_{N_{h,j}(t)} = \sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}}.$$

A noter que l'inégalité de Hoeffding est obtenue en bornant la variance de la loi de Y par $1/4$, ce qui est sous optimal dans de nombreux cas. En particulier dans notre étude comme nous nous ramenons à des variables aléatoires Z suivant une loi de Bernoulli de paramètre τ, la variance est égale à $\tau(1-\tau)$ et est seulement égale à $1/4$ quand $\tau = 0.5$. L'inégalité de Hoeffding est donc très souvent sous optimale pour nous et utiliser d'autres inégalités plus adaptées peut améliorer les bornes de confiance obtenues. Dans le Chapitre 3 nous utiliserons l'inégalité de Bernstein et l'inégalité de Chernoff. En supposant que le moment d'ordre deux des variables aléatoires Y_i est borné, l'inégalité de Bernstein s'énonce comme suit :

$$\mathbb{P}(|\bar{Y}_n - \mathbb{E}(\bar{Y}_n)| \geqslant \varepsilon) \leqslant 2\exp\left(\frac{n\varepsilon^2}{2\mathbb{E}(\bar{Y}_n^2) + 2\varepsilon/3}\right).$$

L'inégalité de Chernoff s'obtient en suivant le même schéma de preuve que celui utilisé pour l'inégalité de Hoeffding mais en utilisant une mesure de dissimilarité entre les lois (au lieu de borner la variance). Lorsque les variables considérées sont des Bernoulli de paramètre τ, alors il est possible d'utiliser la divergence de Kullbach-Leibler noté $\mathrm{kl}(.,.)$, pour mesurer cette dissimilarité. Dans ce cas l'inégalité de Chernoff est donnée par :

$$\mathbb{P}(\bar{Y}_n \geqslant \varepsilon) \leqslant \exp\big(-n\,\mathrm{kl}(\varepsilon, \tau)\big).$$

Dans le Chapitre 3 nous dérivons des inégalités de déviation pour le quantile puis nous les utilisons pour créer des UCB et LCB, quantités indispensables pour définir une fonction d'acquisition pour l'algorithme StoROO. Cependant dans la littérature des inégalités de déviation existent pour d'autres mesures de risque. C'est le cas notamment pour la CVaR (voir Brown [2007], Thomas and Learned-Miller [2019]). Nous montrons dans le Chapite 3 qu'une adaptation de ces inégalités permet de créer une version de StoROO directement opérationnelle pour l'optimisation de la CVaR d'une boîte noire aléatoire.

Enfin toutes les inégalités de déviation ne se valent pas quand nous les utilisons dans la routine de StoROO. Des expérience numériques réalisées dans le Chapitre 3 montrent qu'utiliser des inégalités plus fines, *i.e* Chernoff dans le cas du quantile et celle établie dans Thomas and Learned-Miller [2019] pour la CVaR, permettent d'accélérer l'optimisation.

1.6.4 Borne supérieure sur le regret simple

Un avantage non négligeable des algorithmes d'optimisation de type bandit est la possibilité d'obtenir des bornes supérieures sur le regret. Dans la littérature $\mathcal{X}$-armed une quantité centrale et générique à toutes les approche est la *near-optimality dimension* introduite par Bubeck et al. [2011] et Munos [2014].

Definition 1.6.1. Soit ℓ une semi métrique sur $\mathcal{X}$. La ν-near-optimality dimension est la plus petite constante $d \geqslant 0$ tel qu'il existe $C > 0$ tel que pour tout $\varepsilon > 0$, le nombre maximum de ℓ-boules de rayon $\nu\varepsilon$ et centré en $\mathcal{X}_\varepsilon = \{x \in \mathcal{X}, g(x) \geqslant g^* - \varepsilon\}$ est plus petit que $C\varepsilon^{-d}$.

Pour illustrer la dépendance de la near-optimality dimension en l'hypothèse hölderienne utilisée (1.25) dans la routine de StoROO, nous étudions un exemple légèrement modifié de ce qui est présenté dans Munos [2014]. Considérons la fonction $g(x) = 1 - \beta||x||_\infty^\gamma$, pour $\gamma,\ \beta > 0$. Il est immédiat que g vérifie la propriété hölderienne (1.25). Supposons que lors de l'utilisation de StoROO nous utilisons l'hypothèse de régularité $\ell(x, y) = \beta'||x - y||_\infty^{\gamma'}$ avec $\beta' > \beta$ et $\gamma' < \gamma$, de telle sorte que la régularité de g soit sous-estimée proche du maximum et donc que ℓ permette de majorer la croissance de g dans la cellule contenant x^*. L'optimum de g est en $x^* = 0$ et dans ce cas

$$\begin{aligned}\mathcal{X}_\varepsilon &= \{x \in \mathcal{X}, 1 - \beta||x||_\infty^\gamma \geqslant 1 - \varepsilon\} \\ &= \{x \in \mathcal{X}, ||x||_\infty \leqslant \frac{\varepsilon^{1/\gamma}}{\beta}\}.\end{aligned}$$

Donc $\mathcal{X}_\varepsilon$ est une L_∞-boule de rayon $\frac{\varepsilon^{1/\gamma}}{\beta}$. Pour connaître la near-optimality dimension associée il faut établir combien de ℓ-boules de rayon $\nu\varepsilon$ elle contient. En prenant $\nu = 1$ nous obtenons

$$\begin{aligned}B_\ell(\varepsilon) &= \{x \in \mathcal{X}, \beta'||x||_\infty^{\gamma'} \leqslant \varepsilon\} \\ &= \{x \in \mathcal{X}, ||x||_\infty \leqslant \frac{\varepsilon^{1/\gamma'}}{\beta'}\} \\ &= B_\infty\Big(\frac{\varepsilon^{1/\gamma'}}{\beta'}\Big).\end{aligned}$$

Ainsi il y a au moins $\frac{\beta'\varepsilon^{1/\gamma}}{\beta\varepsilon^{1/\gamma'}}$ L_∞-boules de diamètres ε dans $\mathcal{X}_\varepsilon$. Ce qui implique $d = D(1/\gamma' - 1/\gamma)$ et $C = (\beta'/\beta)^D$. Si la puissance γ' est égale à γ alors $d = 0$. La near-optimality dimension va donc dépendre de l'écart entre la vraie régularité de g proche du maximum et la régularité supposée.

A noter qu'il n'est pas toujours possible de trouver faire une hypothèse de régularité permettant d'obtenir $d = 0$. En effet certaines fonction g peuvent avoir des propriétés de régularité vérifiant (1.25) mais trop exotiques pour permettre $d = 0$. Une discussion sur ce sujet est disponible dans Grill et al. [2015].

Toutefois dans Munos [2014], Bubeck et al. [2011], une borne sur le regret simple pour l'optimisation de la moyenne conditionnelle est obtenue sous la forme

$$r_T = O\Big((\frac{\log T}{T})^{\frac{1}{d+2}}\Big).$$

Dans le Chapitre 3 nous démontrons qu'une borne supérieure sur le regret associée à l'optimisation de la CVaR ou du quantile suit la même vitesse de convergence modulo une constante. Ces constantes dépendent essentiellement de τ et de la valeur de la densité dans un voisinage du quantile visé.

1.7 Optimisation à base de métamodèles gaussiens

Nous avons vu que les méthodes $\mathcal{X}$-armed utilisaient un partitionnement hiérarchique fixé pour définir des fonctions d'acquisition. En pratique cela tend à limiter leur utilisation quand la dimension de $\mathcal{X}$ est sensiblement plus grande que 1. Pour traiter des problèmes en dimension plus grande (entre 5 et 20), il est nécessaire de trouver une autre méthode pour définir des fonctions d'acquisition capables de guider la recherche d'optimum global. Dans ce contexte, les métamodèles gaussiens ont largement été utilisés ces dix dernières années. Initialement utilisés pour l'optimisation d'un code boîte noire déterministe avec l'algorithme EGO (voir Močkus [1975], Jones et al. [1998]), ils ont rapidement été utilisés dans le cadre boîte noire stochastique, pour dans ce cas optimiser la moyenne conditionnelle.

1.7.1 Cadre standard

Les processus gaussiens sont populaires en optimisation boîte noire déterministe ou stochastique car combinés à une vraisemblance gaussienne ils fournissent une distribution a posteriori qui est analytiquement calculable. En effet, en supposant que g soit la réalisation d'un processus gaussien de moyenne m et de fonction de covariance k, *i.e*, $g(x) \sim \mathcal{GP}(m(x), k(x, x'))$ et que

$$g(x_i) = y_i + \varepsilon_i \text{ avec } \varepsilon_i \sim \mathcal{N}(0, \sigma_i),$$

conditionnellement à $\mathcal{D}_n$, la distribution en un point quelconque $x_* \in \mathcal{X}$ est gaussienne de moyenne et de variance connue. Plus précisément $g(x_*) \sim \mathcal{N}\big(m(x_*), \mathbb{V}(x_*)\big)$ avec

$$m(x_*) = K_{x_*,x}(K_{x,x} + \text{diag}(\boldsymbol{\sigma}^2))^{-1}\mathcal{Y}_n,$$

$$\mathbb{V}(x_*) = k(x_*, x_*) - K_{x_*,x}(K_{x,x} + \text{diag}(\boldsymbol{\sigma}^2))^{-1}K_{x,x_*},$$

où $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_n)$, $K_{x,x} \in \mathbb{R}^{n\times n}$ telle que pour tout $1 \leqslant j \leqslant n$ et $1 \leqslant i \leqslant n$, $K_{x,x}(i, j) = k(x_i, x_j)$ et $K_{x_*,x} \in \mathbb{R}^n$ tel que $K_{x_*,x}(i) = k(x_*, x_i)$.

Connaître la distribution a posteriori permet de fournir un estimateur de g mais également de l'incertitude locale. En effet, bien que la fonction visée soit déterministe, dans ce contexte elle est supposée être la réalisation d'un processus gaussien. Si bien

que dans les zones où la variance est élevée, l'information sur la fonction visée est faible en raison de la très grande variabilité des trajectoires réalisables. La Figure 1.19 illustre ce phénomène.

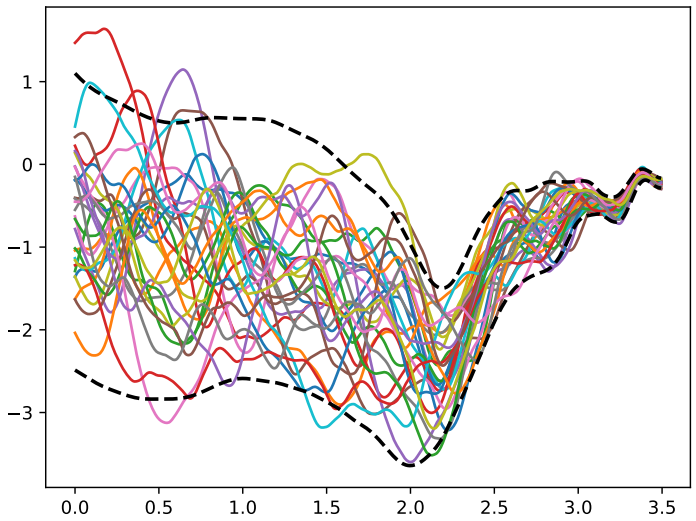


Figure 1.19: Traits pleins : réalisations d'un processus gaussiens dont la variance dépend de l'espace, traits pointillés : la l'écart type multiplié par un facteur 1.96. On remarque que l'incertitudes sur les trajectoires dépends de l'écart type. Pour $x \in [2.5, 3.5]$ les trajectoires sont très similaires, il n'y a pas beaucoup d'incertitudes. Pour $x \in [0, 1]$, les trajectoires sont très variables.

A partir de la connaissance de la distribution a posteriori, différentes fonctions d'acquisition peuvent être définies. La plus utilisée est l'expected improvement (voir Jones et al. [1998]) :

$$\mathrm{EI}(x) = \mathbb{E}\big((\widehat{Y}_x - \widehat{g}^*)^+\big),$$

avec $\widehat{Y}_x \sim \mathcal{N}(m(x), \mathbb{V}(x))$ et $\widehat{g}^*$ pouvant être différentes quantités. Dans Picheny et al. [2013] le choix de $\widehat{g}^*$ est discuté. Il peut être défini comme le maximum d'un quantile de la distribution a posteriori de g ou bien le maximum de m. Dans les deux cas, cette fonction permet la recherche du point x tel que l'espérance de $\widehat{Y}_x$ sachant $\widehat{Y}_x$ supérieur a une valeur considérée comme le maximum courant soit maximale. Sous l'hypothèse gaussienne du modèle, cette quantité peut se calculer analytiquement. La Figure 1.20 illustre l'algorithme EGO qui en découle, dans le cas de l'optimisation de la moyenne conditionnelle avec $\widehat{g}^*$ le maximum de m.

Il existe d'autres fonctions d'acquisition basées sur la distribution a posteriori, par exemple *knowledge gradient* Frazier et al. [2009], *upper confidence bound* Srinivas et al.

[2009], *entropy search* Hernández-Lobato et al. [2014]. Cependant la plupart d'entres elles (EI y compris) sont faites pour échantillonner le modèle séquentiellement en ajoutant à chaque pas de temps une seule observation. Or dans le cas où l'on cherche à optimiser une mesure de risque, il se peut que les observations de g soient très fortement bruitées. Dans ce cas nous conjecturons que mettre à jour le modèle uniquement avec une seule nouvelle observation peut conduire, au pire à des instabilités et au mieux à un très grand nombre de mises à jours inutiles et coûteuses. A noter que dans la littérature il existe des versions parallélisables des fonctions d'acquisition citées plus haut mais le nombre de points sélectionnables à chaque itération dépasse difficilement la dizaine. Ainsi deux stratégies s'offrent à nous :

- Soit nous conservons tous les formalismes développés dans le cadre de l'optimisation de la moyenne conditionnelle. Ce qui implique qu'il faut avoir des observations de g (qui seront supposée bruitées suivant un bruit gaussien centré de variance à estimer) à fournir au modèle. Pour cela une procédure intuitive est l'utilisation de répétitions pour extraire une estimation locale de g.
- Soit nous n'utilisons pas de répétitions dans le plan d'expérience et utilisons une méthode variationnelle pour estimer le modèle gaussien qui est la meilleure approximation de g. Dans ce cas nous souhaitons évaluer un nombre b de nouveaux points à chaque boucle de la procédure d'optimisation. Il nous faut donc utiliser une fonction d'acquisition qui soit parallélisable. Or dans le cas variationnel une expression du modèle est en générale non analytique (pour plus de détails voir la section 1.8 ou le Chapitre 4), ce qui impliquera que la plupart de fonctions d'acquisition existantes dans la littérature ne seront pas adaptables.

1.7.2 Fonctions d'acquisitions parallélisables pour l'optimisation à base de métamodèles non analytiques

Dans cette section nous détaillons la procédure d'optimisation dans le cas où nous utilisons un modèle dont les paramètres sont obtenus à l'aide d'une méthode d'inférence variationnelle. Une bonne stratégie d'échantillonnage consisterait à d'échantillonner g suivant la probabilité qu'un point x soit égal à x^*. Cette distribution étant généralement inconnue, l'objectif est de trouver une procédure d'échantillonnage se rapprochant le plus possible de la procédure qui serait décrite par l'échantillonnage suivant la loi du maximum. Parmi les fonctions d'acquisition listées plus haut, nous en identifions deux qui ne nécessitent pas une connaissance analytique du modèle et facilement parallélisables. Il s'agit de la fonction d'acquisition de l'algorithme GP-UCB et celle de l'algorithme Thompson sampling. De là nous dérivons deux algorithmes qui sont risk-parallele-ucb (RP-UCB) et risk-parallele-Thompson-sampling (RP-TS).

Dans l'article original, GP-UCB utilise la fonction d'acquisition

$$f_a(x,t) = \widehat{g}(x) + \beta_t \sqrt{\mathbb{V}(x)}$$

avec β_t un paramètre décroissant avec t et garantissant la convergence de l'algorithme vers un maximum global sous réserve que g soit un processus gaussien. Or dans notre

cas g n'a pas de raisons d'être un processus gaussien et donc démontrer un résultat de convergence sous ce formalisme semble hors de porté. Ici nous souhaitons uniquement définir une heuristique de recherche d'un maximum global. De ce fait une approche intuitive est de considérer $(\beta_1, \cdots, \beta_b)$ différentes valeurs de β à chaque pas de temps de telle sorte de garantir l'exploration avec les grandes valeurs de β et d'exploiter grâce aux valeurs de β petites. Ainsi l'algorithme RP-UCB sélectionne à chaque pas de temps b nouveaux points à échantillonner notés $(x_{n,1}, \cdots, x_{n,b})$ avec pour tout $1 \leqslant i \leqslant b$

$$x_{n,i} = \arg\max_{x \in \mathcal{X}} \widehat{g}(x) + \beta_i \sqrt{\mathbb{V}(x)}.$$

En ce qui concerne RP-TS l'idée est de simuler b trajectoires conditionnelles suivant le posterior sur g. Le compromis exploration exploitation est naturellement présent avec cette stratégie. En effet les trajectoires sont intrinsèquement stochastiques ce qui permet l'exploration. En revanche si nous tirons un très grand nombre de trajectoires alors nous pouvons estimer le posterior sur g. Ce qui implique que le maximum des trajectoires tend à être distribué suivant la loi du maximum de g. Ainsi l'exploitation se fait également naturellement. Pour tirer des trajectoires conditionnelles il y a deux approches. La première consiste à discrétiser l'espace sur une grille de taille M dont nous notons les points $\boldsymbol{x}$. Dans ce cas les valeurs d'une trajectoire en les points de la grille sont données par :

$$\mathrm{tr}(\boldsymbol{x}) = \widehat{g}(\boldsymbol{x}) + \Sigma^{1/2} \mathcal{N}^M,$$

avec $\mathcal{N}^M$ un vecteur contenant une réalisation d'une loi multivariée normale de dimension M centrée réduite et $\Sigma^{1/2}$ la matrice de Cholesky de la covariance conditionnelle de g évaluée en $\boldsymbol{x}$. Calculer $\Sigma^{1/2}$ coûte $O(M^3)$, ce qui représente un inconvénient majeur si nous souhaitons maximiser précisément une trajectoire en dimension quelconque. Une approche alternative consiste à revenir au point de vue paramétrique de la régression bayésienne, *i.e*

$$\mathrm{tr}(x) = \Phi(x)^T \Theta,$$

avec Θ un vecteur gaussien dont les paramètres sont à spécifier et Φ une fonction telle que $\langle \Phi(x), \Phi(x') \rangle = k(x, x')$. Or pour des noyaux classiques comme le noyau gaussien ou le noyau Matérn, une expression finie de Φ n'existe pas. Toutefois il est possible approcher Φ par une expression finie en utilisant les *random fourier features* Rahimi and Recht [2008]. Dans ce cas il devient possible d'obtenir une approximation des trajectoires. Dans le Chapitre 4 nous montrons qu'il est possible d'obtenir des approximations de trajectoires continues échantillonnée suivant le posterior sur g avec un coût de l'ordre de $O(mn^2)$, avec m le nombre de vecteurs de bases utilisés pour approcher Φ.

Ces deux stratégies seront comparées à une stratégie EI couplée à une modèle gaussien utilisant des répétitions dans le plan d'expérience pour estimer localement g dans le Chapitre 4 pour l'optimisation de quantiles conditionnels et d'expectiles conditionnels.

1.8 Methodes variationnelles

Sur diverses problèmes étudiés dans cette thèse les méthodes variationnelles ont montré de très bons résultats. C'est le cas dans le Chapitre 2 où un métamodèle de quantile basé sur les processus gaussiens estimés par un algorithme EM variationnel s'avère être un des meilleurs métamodèles de quantile conditionnel parmi les métamodèles testés. Dans le Chapitre 4 nous montrons que les méthodes variationnelles sont très flexibles et efficaces pour estimer des métamodèles utilisant diverses vraisemblances. Cela permet de créer un modèle gaussien pour l'estimation de diverses mesures de risque et d'en dériver des algorithmes d'optimisation bayesienne. Pour finir cette introduction à cette thèse, nous proposons dans cette section une présentation de l'approche variationnelle comme nous l'utilisons dans le Chapitre 2 et 4.

1.8.1 Cas général

Pour expliquer des observations, en statistique il n'est pas rare d'introduire des variables $z \in \mathbb{R}^N$ dites cachées ou latentes. Avec ce formalisme nous nous retrouvons avec la nécessité d'estimer la distribution a posteriori sur z sachant les observations. Ce posterior est donné par

$$p(z|\mathcal{Y}_n) = \frac{p(\mathcal{Y}_n, z)}{\int p(\mathcal{Y}_n, z)dz}.$$

Cependant cette probabilité est souvent non estimable car les variables sont de grande dimension, ce qui rend l'intégrale au dénominateur incalculable.

L'idée des méthodes variationnelles est d'approcher ce posterior par une distribution plus simples $\widetilde{p}(z;\lambda)$ avec λ des paramètres dits variationnels. Pour trouver cette distribution $\widetilde{p}$, partons de l'équation

$$\log p(\mathcal{Y}_n) = \log \int p(\mathcal{Y}_n, z)dz.$$

En multipliant le numérateur et le dénominateur dans l'intégrale par $\widetilde{p}$ on obtient :

$$\begin{aligned}\log p(\mathcal{Y}_n) &= \log \int \frac{p(\mathcal{Y}_n, z)\widetilde{p}(z;\lambda)}{\widetilde{p}(z;\lambda)}dz \\ &= \log \mathbb{E}_{\widetilde{p}(z;\lambda)}\left(\frac{p(\mathcal{Y}_n, z)}{\widetilde{p}(z;\lambda)}\right),\end{aligned}$$

puis par l'inégalité de Jensen on obtient :

$$\begin{aligned}\log p(\mathcal{Y}_n) &\geqslant \mathbb{E}_{\widetilde{p}(z;\lambda)}\left(\log \frac{p(\mathcal{Y}_n, z)}{\widetilde{p}(z;\lambda)}\right) \\ &= \mathcal{L}(\lambda).\end{aligned}$$

Ainsi la méthode variationnelle transforme le problème initial en un problème d'optimisation en les paramètres variationnels. La distribution recherchée étant celle maximisant l'*evidence*

lower bound (ELBO) $\mathcal{L}$ et donc proposant la plus grande borne inférieure sur la vraisemblance.

Cette approche est renforcée par une autre relation. En effet, il peut être établi que

$$\begin{aligned}\log p(\mathcal{Y}_n) - \mathcal{L} &= \log p(\mathcal{Y}_n) - \mathbb{E}_{\widetilde{p}(z;\lambda)}\Big(\log \frac{p(\mathcal{Y}_n, z)}{\widetilde{p}(z;\lambda)}\Big) \\ &= \mathbb{E}_{\widetilde{p}(z;\lambda)}\Big(\log p(\mathcal{Y}_n) - \log \frac{p(\mathcal{Y}_n, z)}{\widetilde{p}(z;\lambda)}\Big) \\ &= -\mathbb{E}_{\widetilde{p}(z;\lambda)}\Big(\log \frac{p(z|\mathcal{Y}_n)}{\widetilde{p}(z;\lambda)}\Big) \\ &= \mathrm{kl}(\widetilde{p}||p), \end{aligned} \tag{1.28}$$

avec kl la divergence de Kullbach-Leibler. Rappelons qu'en probabilités il est possible de mesurer la dissimilarité entre deux mesure de probabilité à l'aide de *divergence* $D(\widetilde{p}(z)||p(z))$ vérifiant deux propriétés

$$D(\widetilde{p}(z)||p(z)) \geqslant 0 \text{ et } D(\widetilde{p}(z)||p(z)) = 0 \Leftrightarrow \widetilde{p}(z) = p(z).$$

Ainsi l'équation (1.28) montre que minimiser $\mathcal{L}$ est équivalent à minimiser la divergence entre deux distributions, en l'occurrence entre la distribution ciblée et $\widetilde{p}$.

Sous cette forme maximiser $\mathcal{L}$ demande de calculer une intégrale sous la loi $\widetilde{p}$ ce qui apporte des contraintes. Des hypothèses doivent donc êtres formulées sur $\widetilde{p}$ pour rendre l'optimisation faisable. Cependant il est nécessaire de garder en tête qu'un modèle trop simple donnera une pauvre approximation de p. Un compromis classique est l'approximation champ moyen qui suppose que l'ensemble des variables latentes sont indépendante, *i.e*

$$\widetilde{p}(z;\lambda) = \prod_{i=1}^{N} \widetilde{p}_i(z_i;\lambda_i).$$

Une telle hypothèse permet de réécrire la borne $\mathcal{L}$ comme

$$\mathcal{L}(\lambda_i) = \int \widetilde{p}_i \mathbb{E}_{\widetilde{p}_{-z_i}}\Big(\log p(z_j, \mathcal{Y}_n|z_{-i})\Big)dz_i - \int \widetilde{p}_i(z_i;\lambda_i)\log\Big(\widetilde{p}_i(z_i;\lambda_i)\Big)dz_i + c_i.$$

Puis pour maximiser $\mathcal{L}$ il est classique d'utiliser des méthodes itératives optimisant $\mathcal{L}$ suivant chaque variable jusqu'à convergence. A noter que dans certains cas il peut être nécessaire d'estimer numériquement certaines quantités pour obtenir des estimations des dérivées. Dans la suite nous présentons comment se formalisme peut s'appliquer à l'estimation de processus gaussiens.

1.8.2 Méthode variationnelle pour les processus gaussiens

Pour estimer un quantile ou un expectile il est nécessaire d'utiliser des distributions sur ε qui dépendent d'un paramètre σ. Naturellement σ et g n'ont pas de raisons d'être

indépendants et g et σ on aucune raison d'être des processus gaussiens. Cependant en pratique supposer une propriété d'indépendance

$$p(g, \sigma) = p(g)p(\sigma),$$

combinée avec l'hypothèse $g \sim \mathcal{GP}$, permet d'obtenir des modèles flexibles dont les paramètres sont estimables numériquement.

Supposons la distribution de g en les x_i gaussienne multivariée $u_g \sim \mathcal{N}(\mu_g, S_g)$ avec $\mu_g \in \mathbb{R}^n$ et $S_g \in \mathbb{R}^{n\times n}$. Il y a donc deux ensembles de paramètres variationnels à estimer. Le premier ensemble est le vecteur des pseudo observations $\mu_g = (\mu_{g_1}, \cdots, \mu_{g_n})$ avec μ_{g_i} qui peut être interprété comme une estimation en les x_i de la valeur de la mesure de risque. Le second ensemble de paramètres correspond à l'intégralité des entrées de la matrice de covariance S_g.

La distribution a posteriori de g et σ en un nouveau point x_* peut s'écrire :

$$p(g(x_*), \sigma(x_*)|\mathcal{D}_n) = \int p(g(x_*), \sigma(x_*)|u_g, \sigma)p(u_g, \sigma|\mathcal{D}_n)du_g d\sigma. \tag{1.29}$$

Puis avec l'approximation champ moyen on obtient :

$$p(g(x_*)|\mathcal{D}_n) \approx \int p(g(x_*)|u_g)\widetilde{p}(u_g|\mathcal{D}_n)du_g. \tag{1.30}$$

Or une formule classique de conditionnement (voir Théorème 2 de Schön and Lindsten [2011]) donne

$$p(g(x_*)|u_g) = \mathcal{N}(K_{x_*,x}K_{x,x}^{-1}u_g, K_{x_*,x_*} - Q_g),$$

avec $Q_g = K_{x_*,x_j}K_{x,x}^{-1}K_{x_j,x_*}$. De plus le Corollaire 1 de Schön and Lindsten [2011] énonce que si

$$p(x_a) = \mathcal{N}(\mu_a, \Sigma_a)$$

et

$$p(x_b|x_a) = \mathcal{N}(Mx_a, \Sigma_{b|a}),$$

alors

$$p(x_b) = \mathcal{N}(M\mu_b, \Sigma_b), \quad \text{avec} \quad \Sigma_b = \Sigma_{b|a} + M\Sigma_a M^T.$$

Ici rappelons que u_g suit une distribution Gaussienne multivariée, il est donc possible d'obtenir :

$$p(g(x_*)|\mathcal{D}_n) \approx \mathcal{N}(K_{x_*,x}K_{x,x}^{-1}\mu_g, K_{x_*,x_*} + \widehat{Q})$$

où

$$\widehat{Q} = K_{x_*,x}K_{x,x}^{-1}(S_g - K_{x,x})K_{x,x}^{-1}K_{x,x_*}.$$

A partir de là, deux approches sont suivies dans cette thèse pour estimer les paramètres variationnels maximisant $\mathcal{L}$. La première est utilisée dans le Chapitre 2 et consiste à utiliser le plus possible des quantités conjuguées dans le but de proposer une expression

analytique de la majorité des dérivés de $\mathcal{L}$. L'autre méthode utilisée dans le Chapitre 4 est davantage boîte noire et nécessite une méthode d'estimation pour calculer certains gradients. Mais cette seconde méthode à l'avantage d'être une approche générique qui peut être adaptée très simplement à l'estimation de quantités autres que le quantile ou l'expectile, alors que ce qui est suivi dans le Chapitre 2 reste spécifique au quantile.

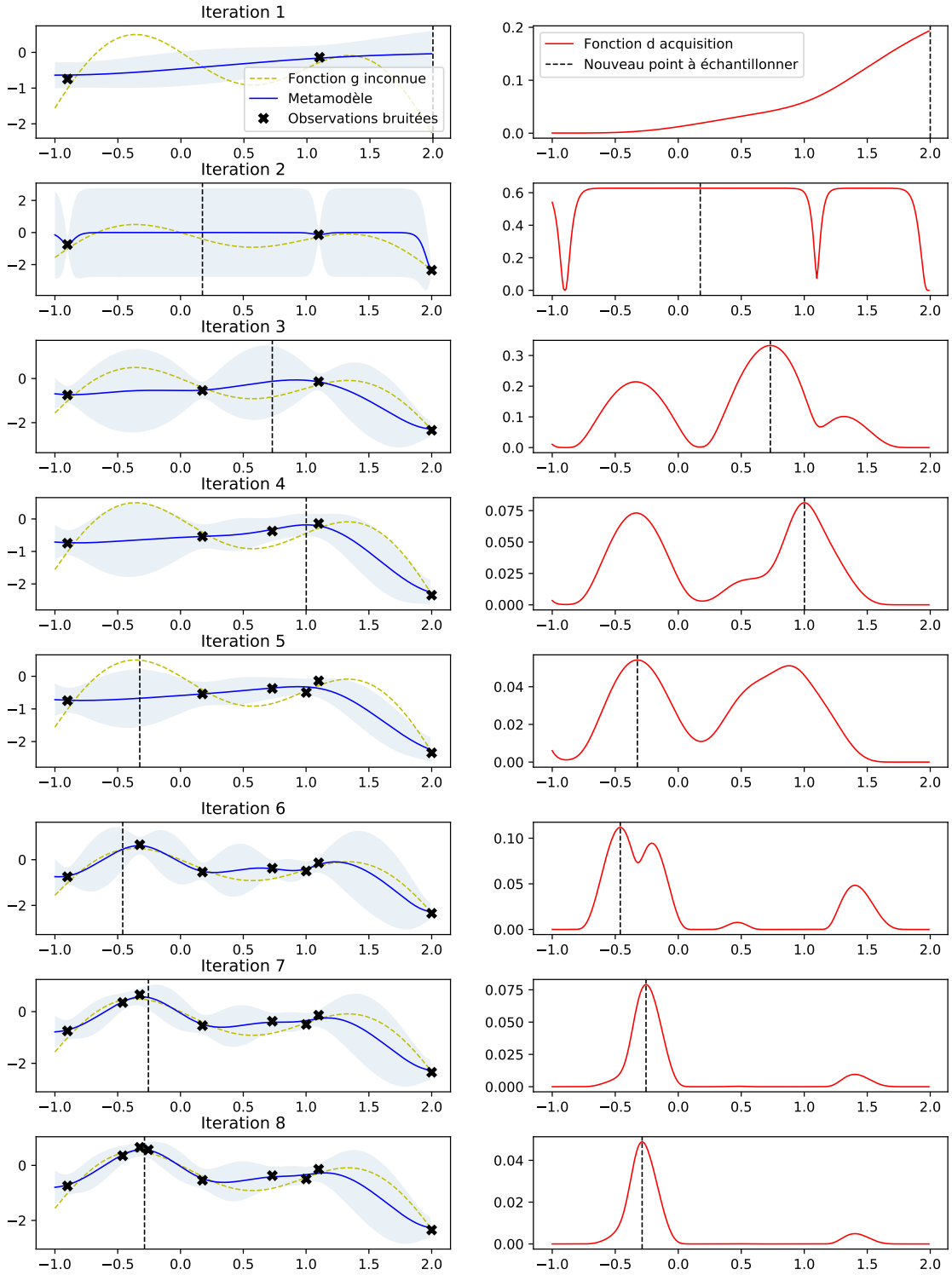


Figure 1.20: Illustration de l'algorithme EGO dans un cas faiblement bruité.

Chapter 2

A Review on Quantile Regression for Stochastic Computer Experiments

Contents

2.1 Résumé

Ce chapitre reprend l'article Torossian et al. [2019b] soumis pour publication dans la revue Reliability Engineering and Systeme Safety. Nous proposons une étude empirique des principales méthodes de régression quantile pour des codes de calcul stochastiques. Dans cette étude nous proposons l'analyse de six métamodèles que nous classifions en trois catégories : les méthodes basées sur les statistiques empiriques, les méthodes fonctionnelles et la dernière regroupant les méthodes d'inspiration bayésienne. Nous testons les métamodèles sur différents problèmes caractérisés par la taille de l'échantillon d'apprentissage, la dimension de $\mathcal{X}$, le ratio signal sur bruit et la valeur de la densité de la loi conditionnelle de la sortie en le quantile visé. Cette étude empirique présente certains contrastes nous permettant d'extraire des comportements propres à chaque méthode. Basé sur nos résultats, nous proposons des recommandations pour l'utilisation des différentes méthodes en fonction des caractéristiques du code de calcul considéré.

Ce travail a été réalisé en collaboration avec Victor Picheny, Robert Faivre et Aurélien Garivier.

2.2 Introduction

2.2.1 Stochastic experiment setting

Computer simulation models are now essential for performance evaluation, quality control and uncertainty quantification to assess decisions in complex systems. These computer simulators generally model systems depending on multiple input variables that can be divided into two categories: the controllable variables and the uncontrollable variables.

For example, in pharmacology, the optimal drug dosage depends on the drug formulation but also on the targeted individual (genetics, age, sex) and environmental interactions. The shelf life and performance of a manufacturing device depend on its design, but also on its environment and on some uncertainties during to the manufacturing process. The plant growth and yield depend on the genes of the plant and on the gardening techniques but also on the weather and potential diseases.

Evaluating the influence of the controllable and uncontrollable variables directly on the real-life problems can be costly and tedious. One solution is to encapsulate the systems into a computer simulation model which would reduce the cost and the time required for each test (see Herwig [2014] and reference therein for computer experiments applied to clinical trials, see Gijo and Scaria [2012] for a computer simulation model applied to industrial design and Van Maanen and Xu [2003], Casadebaig et al. [2011] for computer experiments applied to crop production).

In such computer simulation models, the links between the inputs and outputs may be too complex to be fully understood or to be formulated in a closed form. In this case, the system can be considered as a black box and formalized by an unknown function: $\Psi : \mathcal{X} \times \Omega \to \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^D$ denotes the compact space of controllable variables, and Ω denotes a probability space representing the uncontrollable variables. Note that some

black boxes have their outputs in a multidimensional space but this aspect is beyond the scope of the present paper.

Based on the standard constraints encountered in computer experiments, throughout this paper, we assume that the function is only accessible through pointwise evaluations $\Psi(x, \omega)$; no structural information is available regarding Ψ; in addition, we take into account the fact that evaluations may be expensive, which drastically limits the number of possible calls to Ψ.

If the space Ω is small enough (or highly structured), it may be possible to work directly on the space $\mathcal{X} \times \Omega$ (see Janusevskis and Le Riche [2013] for example). Often, however, Ω is too complex and working on the joint space is intractable. This is the case for intrinsic stochastic simulators (see Lei [2012], Ludkovski and Niemi [2010] for examples of biological systems, where the stochasticity is driven by stochastic equations such as the Fokker-Planck or the chemical Langevin equations), or for simulators associated with a very large space Ω (see Casadebaig et al. [2011] for the crop model SUNFLO that considers 5 weather indicators on 130 days, *i.e* Ω is a space of dimension 650).

In this paper we consider the case where Ω is too complex. We assume that Ω has any structure and its contribution is considered as random. In contrast to deterministic systems, for any fixed x, $\Psi(x, \cdot)$ is considered as a random variable of distribution $\mathbb{P}_x$; hence, such systems are often referred to as stochastic black boxes.

In order to understand the behavior of the system of interest or to take optimal decisions, information is needed about $\mathbb{P}_x$. An intuitive approach is to use a simple Monte-Carlo technique and evaluate $\Psi(x, \omega_1), \ldots, \Psi(x, \omega_n)$ to extract statistical moments, the empirical cumulative distribution function, etc. Unfortunately, such a stratified approach is not efficient when evaluating Ψ is expensive.

Instead, we focus on *surrogate models* (also referred to as *metamodels* or *statistical emulators*), which are appropriate approaches in a small data setting associated with a regularity hypothesis (with respect to $\mathcal{X}$) concerning the targeted statistics. Among the vast choice of surrogate models Storlie and Helton [2008], Villa-Vialaneix et al. [2012], the most popular ones include regression trees, Gaussian processes, support vector machines and neural networks. In the framework of stochastic black boxes, the standard approach consists in estimating the conditional expectation of Ψ. This case has been extensively treated in the literature and many applications, including Bayesian optimization Shahriari et al. [2016], have been developed. However, the conditional expectation is risk-neutral, whereas pharmacologists, manufacturers, asset managers, data scientists and agronomists may need to evaluate the worst case scenarios associated with their decisions.

Risk information can be introduced by using a surrogate expectation-variance model in which the distribution can be estimated by non-parametric kernels (see He [1997], Shim et al. [2009] for instance) or via heteroscedastic Gaussian processes Kersting et al. [2007], Lázaro-Gredilla and Titsias [2011]. However, such approaches usually imply that the shape of the distribution (e.g. normal, uniform, etc.) is the same for all $x \in \mathcal{X}$. Another possible approach would be to learn the whole distribution $\mathbb{P}_x$ with no strong structural hypotheses Moutoussamy et al. [2015], Hall et al. [2004], Efromovich [2010],

but this requires a large number of evaluations of Ψ. Here, we focus on the conditional quantile estimation of order τ, a flexible way to tackle cases in which the distribution of $\Psi(x, .)$ varies markedly in spread and shape with respect to $x \in \mathcal{X}$, and a classical risk-aware tool in decision theory Rostek [2010].

2.2.2 Paper Overview

Many metamodels originally designed to estimate conditional expectations have been adapted to estimate the conditional quantile. However, despite extensive literature on estimating the quantile in the presence of spatial structure, few studies have reported on the constraints associated with stochastic black boxes. The performance of a metamodel with high dimension input is treated in insufficient details, performance based on the number of points has rarely been tackled and, to our knowledge, dependence on specific aspects of the quantile functions has never been studied. The aim of the present paper is to review quantile regression methods under standard constraints related to the stochastic black box framework, so as to provide information on the performance of the selected methods, and to recommend which metamodel to use depending on the characteristics of the computer simulation model and the data.

A comprehensive review of quantile regression is of course beyond the scope of the present work. We limit our review to the approaches that are best suited for our framework, while ensuring the necessary diversity of metamodels. In particular, we have chosen six metamodels that are representative of three main categories: approaches based on statistical order (K-nearest neighbors [KN] regression Bhattacharya and Gangopadhyay [1990] and random forest [RF] regression Meinshausen [2006]), functional or frequentist approaches (neural networks [NN] regression Cannon [2011] and regression in reproducing kernel Hilbert space [RK] Takeuchi et al. [2006]), and Bayesian approaches based on Gaussian processes (Quantile Kriging [QK] Plumlee and Tuo [2014] and the variational Bayesian [VB] regression Abeywardana and Ramos [2015]). Each category has some specificities in terms of theoretical basis, implementation and complexity. We begin this presentation by describing the methods in full in sections 2.4, 2.5 and 2.6.

In order to identify the relevant areas of expertise of the different metamodels, an original benchmark system is designed based on four toy functions and an agronomical model Casadebaig et al. [2011]. The dimension of the problems ranges from 1 to 9 and the number of observations from 40 to 2000. Particular attention is paid to the performance of each metamodel according to the size of the learning set, the value of the probability density function at the targeted quantile $\widetilde{f}(., q_\tau)$ and the dimension of the problem. Sections 2.7 and 2.8 describe the benchmark system and detail its implementation, with particular focus on the tuning of the hyperparameters of each method. Full results and discussion are to be found in Sections 2.9 and 2.10, respectively.

2.3 Quantile emulators and design of experiments

We first provide the necessary definitions, objects and properties related to the quantile. The quantile of order $\tau \in (0,1)$ of a random variable Y can be defined either as the (generalized) inverse of a cumulative distribution function (CDF), or as the solution to an optimization problem:

$$q_\tau = \min\left\{q \in \mathbb{R} \ : F(q) \geqslant \tau\right\} = \arg\min_{q\in\mathbb{R}} \mathbb{E}\big[l_\tau(Y-q)\big], \tag{2.1}$$

$F(\cdot)$ is the CDF of Y and

$$l_\tau(\xi) = (\tau - \mathbb{1}_{(\xi<0)})\xi, \quad \xi \in \mathbb{R} \tag{2.2}$$

is the so-called pinball loss Koenker and Bassett Jr [1978] (Figure 2.1). In the following, we only consider situations in which F is continuous.

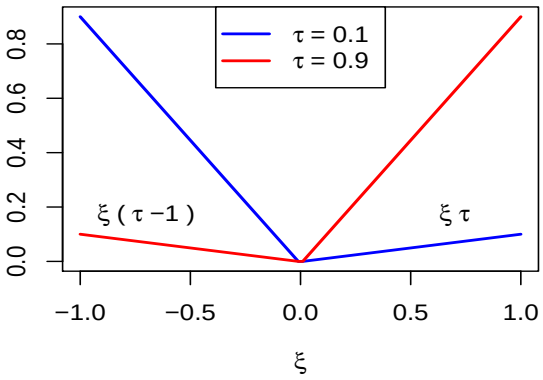


Figure 2.1: Pinball loss function with $\tau = 0.1$ and $\tau = 0.9$.

Given a finite observation set $\mathcal{Y}_n = (y_1, ..., y_n)$ composed of i.i.d samples of Y, the empirical estimator of q_τ can thus be introduced in two different ways:

$$\widehat{q}_\tau = \min\left\{y_i \in \mathcal{Y}_n : \widehat{F}(y_i) \geqslant \tau\right\} \tag{2.3}$$

or

$$\widehat{q}_\tau = \arg\min_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} l_\tau(y_i - q), \tag{2.4}$$

where $\widehat{F}$ denotes an estimator of the CDF function. In (2.3) $\widehat{q}_\tau$ coincides with an order statistic. For example, if $\widehat{F}$ is the empirical CDF function then $\widehat{q}_\tau = y_{([n\tau])}$, where $[n\tau]$ represents the smallest integer greater than or equal to $n\tau$ and $y_{(k)} = \mathcal{Y}_n(k)$ is the k-th smallest value in the sample $\{y_1, \ldots, y_n\}$. The estimators (2.4) and (2.3) may coincide, but are in general not equivalent.

Similarly to (2.1), the conditional quantile of order $\tau \in (0,1)$ can be defined in two equivalent ways:

$$q_\tau(x) = \min\left\{q : F(q|X=x) \geqslant \tau\right\} = \arg\min_{q\in\mathbb{R}} \mathbb{E}\big[l_\tau(Y_x - q)\big], \tag{2.5}$$

where Y_x is a random variable of distribution $\mathbb{P}_x$ and $F(.|X=x)$ is the CDF of Y_x.

In a quantile regression context, one only has access to a finite observation set $\mathcal{D}_n = \big\{(x_1,y_1),\ldots,(x_n,y_n)\big\} = (\mathcal{X}_n,\mathcal{Y}_n)$ with $\mathcal{X}_n$ a $n \times D$ matrix. Estimators for (2.5) are either based on the order statistic as in (2.3) (section 2.4), or on a minimizer of the pinball loss as in (2.4) (sections 2.5 and 2.6). Throughout this work, the observation set $\mathcal{D}_n$ is fixed (we do not consider a dynamic or sequential framework). Following the standard approach used in computer experiments, the training points x_i are chosen according to a space-filling design Cavazzuti [2013] over a hyperrectangle. In particular, we assume that there are no repeated experiments: $x_i \neq x_j, \forall i \neq j$; most of the methods chosen in this survey (KN, RF, RK, NN, VB) work under that setting.

However, as a baseline approach, one may decide to use a stratified experimental design $\mathcal{D}_{n',r}$ with r i.i.d samples for a given x_i, $i = 1,..,n'$, extract pointwise quantile estimates using (2.3) and fit a standard metamodel to these estimates. The immediate drawback is that for the same budget ($n' \times r = n$) such experimental designs cover much less of the design space than a design with no repetition. The QK method is based on this approach.

2.4 Methods based on order statistics

A simple way to compute a quantile estimate is to take an order statistic of an i.i.d. sample. A possible approach is to emulate such a sample by selecting all the data points in the neighborhood of the query point x, and then by taking the order statistic of this subsample as an estimator for the conditional quantile. One may simply choose a subsample of $\mathcal{D}_n$ based on a distance defined on $\mathcal{X}$: this is what the K-nearest neighbors approach does. It is common to use KN based on the Euclidean distance but of course any other distance can be used, such as Mahalanobis Verdier and Ferreira [2011] or weighted Euclidean distance Dudani [1976]. Alternatively, one may define a notion of neighborhood using some space partitioning of $\mathcal{X}$. That includes all the decision tree methods Breiman [2017], in particular regression trees, bagging or random forest Meinshausen [2006].

2.4.1 K-nearest neighbors

The K-nearest neighbors method was first proposed for the estimation of conditional expectations Stone [1975, 1977]. Its extension to the conditional quantile estimation can be found in Bhattacharya and Gangopadhyay [1990].

Quantile regression implementation

Define $\mathcal{X}_{\text{test}}$ as the set of query points. KN works as follows: for each $x_* \in \mathcal{X}_{\text{test}}$ define $\mathcal{X}^K(x_*)$ the subset of $\mathcal{X}_n$ containing the K points that are the closest to the query point x_*. Define $\mathcal{Y}_K^{x_*}$ the associated outputs, and define $\widehat{F}^K(y|X = x_*)$ as the associated empirical CDF. Following (2.3), the conditional quantile of order τ can be defined as the statistical order

$$\widehat{q}_\tau(x_*) = \mathcal{Y}_K^{x_*}([K\tau]). \tag{2.6}$$

Algorithm 2 details the implementation of the KN method.

Algorithm 2: K-nearest neighbors

Input: $\mathcal{D}_n$, τ, K, $\mathcal{X}_{\text{test}}$
for *each point in* $x_* \in \mathcal{X}_{\text{test}}$ **do**
    Compute all the distances between x_* and $\mathcal{X}_n$;
    Sort the computed distances ;
    Select the K-nearest points from x_* ;
    $\widehat{q}_\tau(x_*) = \mathcal{Y}_K^{x_*}([K\tau])$;
end

Computational complexity

For a naive implementation of such an estimator, one needs to compute $n \times N_{\text{new}}$ distances, where N_{new} is the number of query points, hence for a cost in $O(nN_{\text{new}}D)$. Moreover, sorting n distances in order to extract the K nearest points has a cost in $O(nN_{\text{new}} \log n)$. Combining the two operations implies a complexity of order

$$O(nN_{\text{new}}D) + O(nN_{\text{new}} \log n).$$

Note that some algorithms have been proposed in order to reduce the computational time, for example by using GPUs Garcia et al. [2008] or by using tree search algorithms Arya et al. [1998].

2.4.2 Random forests

Random forests were introduced by Breiman Breiman [2001] for the estimation of conditional expectations. They have been used successfully for classification and regression, especially with problems where the number of variables is much larger than the number of observations Díaz-Uriarte and De Andres [2006].

Overview

The basic element of random forests is the *regression tree* T, a simple regressor built via a binary recursive partitioning process. Starting with all data in the same partition *i.e*

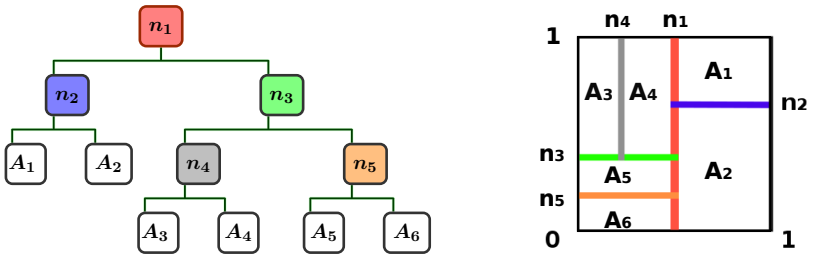


Figure 2.2: Left: a partitioning tree T. The nodes n_i $(1 \leqslant i \leqslant 5)$ represent the splitting points, the A_i's $(1 \leqslant i \leqslant 6)$ represent the leaves. Right: $\mathcal{X} = [0,1]^2$ as partitioned by T. The regression tree prediction is constant on each leaf A_i.

$\mathcal{X}$, the following sequential process is applied. At each step, the data is split into two, so that $\mathcal{X}$ is partitioned in a way that it can be represented by a tree as it is presented Figure 2.2.

Several splitting criteria can be chosen (see Ishwaran [2015]). In Meinshausen [2006], the splitting point x_S is the data point that minimizes

$$C(x_s) = \sum_{x_i \leqslant x_s} (y_i - \bar{Y}_L)^2 + \sum_{x_j > x_s} (y_j - \bar{Y}_R)^2, \tag{2.7}$$

where $\bar{Y}_L$ and $\bar{Y}_R$ are the mean of the left and right sub-populations, respectively. Equation (2.7) applies when the x's are real-valued. In the multidimensional case, the dimension d_S in which the split is performed has to be selected. The split then goes through x_S and perpendicularly to the direction d_S. There are several rules to stop the expansion of T. For instance, the process can be stopped when the population of each cell is inferior to a minimal size *nodesize*: then, each node becomes a terminal node or leaf. The result of the process is a partition of the input space into hyperrectangles $R(T)$. Like the KN method, the tree-based estimator is constant on each neighborhood. The hope is that the regression trees automatically build neighborhoods from the data that should be adapted to each problem.

Despite their simplicity of construction and interpretation, regression trees are known to suffer from a certain rigidity and a high variance (see Breiman [1996] for more details). To overcome this drawback, regression trees can be used with ensemble methods like bagging. Instead of using only one tree, bagging creates a set of tree $\mathcal{T}_N = \{T^1, .., T^N\}$ based on a bootstrap version $\mathcal{D}_{N,n} = \{((x_{1_t}, y_{1_t}), ..., (x_{n_t}, y_{n_t}))\}_{t=1}^N$ of $\mathcal{D}_n$. Then the final model is created by averaging the results among all the trees.

Bagging reduces the variance of the predictor, as the splitting criterion has to be optimized over all the input dimensions, but computing (2.7) for each possible split is costly when the dimension is large. The random forest algorithm, a variant of bagging, constructs an ensemble of weak learners based on $\mathcal{D}_{N,n}$ and aggregates them. Unlike plain bagging, at each node evaluation, the algorithm uses only a subset of $\widetilde{d}$ covariables

for the choice of the split dimensions. Because the $\widetilde{d}$ covariables are randomly chosen, the result of the process is a random partition $R(t)$ of $\mathcal{X}$ constructed by the random tree T^t.

Quantile prediction

We present the extension proposed in Meinshausen [2006] for conditional quantile regression. Let us define $\ell(x_*, t)$ the leaf obtained from the tree t containing a query point x_* and

$$\omega_i(x_*, t) = \frac{\mathbb{1}_{\{x_i \in \ell(x_*,t)\}}}{\#\{j : x_j \in \ell(x_*, t)\}}, \quad i = 1, ..., n$$

$$\bar{\omega}_i(x_*) = \frac{1}{N} \sum_{t=1}^{N} \omega_i(x_*, t) \ .$$

The $\bar{\omega}_i(x_*)$'s represent the weights illustrating the "proximity" between x_* and x_i. In the classical regression case, the estimator of the expectation is:

$$\widehat{\mu}(x_*) = \sum_{i=1}^{n} \bar{\omega}_i(x_*) y_i. \tag{2.8}$$

In Meinshausen [2006] the conditional quantile of order τ is defined as in (2.3) with the CDF estimator defined as

$$\widehat{F}(y|X = x_*) = \sum_{i=1}^{n} \bar{\omega}_i(x_*) \mathbb{1}_{\{y_i \leqslant y\}}. \tag{2.9}$$

Algorithm 3 details the implementation of the RF method.

Computational complexity

Assuming that the value of (2.7) can be computed sequentially for consecutive thresholds, the RF computation burden lies in the search of the splitting point that implies sorting the data. Sorting n variables has a complexity in $O(n \log n)$. Thus, at each node the algorithm finds the best splitting points considering only $\widetilde{d} \leqslant D$ covariables (classically $\widetilde{d} = D/3$). This implies a complexity of $O\big(\widetilde{d} n \log n\big)$ per node. In addition, the depth of a tree is generally upper bounded by $\log n$. Then the computational cost of building a forest containing N trees under the criterion (2.7) is

$$O\big(N \widetilde{d} n \log^2(n)\big)$$

Louppe [2014], Witten et al. [2016]. One may observe that RF are easy to parallelize and that contrary to KN the prediction time is very small once the forest is built.

Algorithm 3: Random forest

Training:
Input: $\mathcal{D}_n$, N, $\widetilde{d}$, m_s
for *each of the* N *trees* **do**
  Uniformly sample with replacement n points in $\mathcal{D}_n$ to create $\mathcal{D}_{t,n}$;
  Consider the cell $R = \mathcal{X}$;
  while *any cell of the tree contains more that* m_s *observations* **do**
    for *the cells containing more than* m_s *observations* **do**
      Uniformly sample without remplacement $\widetilde{d}$ covariables in $1, ..., D$;
      Compute the cell point among the $\widetilde{d}$ covariables that minimizes (2.7);
      Split the cell at this point perpendicularly to the selected covariable;
    end
  end
end
Prediction:
Input: $\mathcal{X}_{\text{test}}$, τ
for *each point in* $x_* \in \mathcal{X}_{\text{test}}$ **do**
  Compute $\bar{\omega}_i(x_*), \quad i = 1 \ldots, n$
  $\widehat{F}(y|X = x_*) = \sum_{i=1}^{n} \bar{\omega}_i(x_*) \mathbb{1}_{\{y_i \leqslant y\}}$
  $\widehat{q}_\tau(x_*) = \inf \left\{ y_i : \widehat{F}(y_i|X = x_*) \geqslant \tau \right\}$
end

2.5 Approaches based on functional analysis

Functional methods search directly for the function mapping the input to the output in a space fixed beforehand by the user. With this framework, estimating any functional S of the conditional distribution implies selecting a loss l (associated to S) and a function space $\mathcal{H}$. Thus, the estimator $\widehat{S} \in \mathcal{H}$ is obtained as the minimizer of the empirical risk $\mathcal{R}_e$ associated to l, *i.e.*

$$\widehat{S} \in \arg\min_{s \in \mathcal{H}} \mathcal{R}_e[s] = \arg\min_{s \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l\big(y_i - s(x_i)\big). \tag{2.10}$$

The functional space $\mathcal{H}$ must be chosen flexible enough to extract some signal from the data. In addition, $\mathcal{H}$ needs to have enough structure to make the optimization procedure feasible (at least numerically). In the literature, several formalisms such as linear regression Seber and Lee [2012], spline regression Marsh and Cormier [2001], support vector machine Vapnik [2013], neural networks Bishop [1995] or deep neural networks Schmidt-Hieber [2017] use structured functional spaces with different levels of flexibility.

However, using a too large $\mathcal{H}$ can lead to overfitting, *i.e.* return predictors that are good only on the training set and generalize poorly. Overcoming overfitting requires some *regularization* (see Schölkopf [2001], Zhao and Yu [2006], Zou and Hastie [2005] for instance), defining for example the regularized risk

$$\mathcal{R}_{r,e}[s] = \frac{1}{n} \sum_{i=1}^{n} l(y_i - s(x_i)) + \lambda \|s\|^{\beta}, \tag{2.11}$$

where $\lambda \in \mathbb{R}^+$ is a penalization factor, $\beta \in \mathbb{R}^+$ and $\|.\|$ is either a norm for some methods (Section 2.5.2) or a measure of variability for others (Section 2.5.1). The parameter λ plays a major role, as it allows to tune the balance between bias and variance.

Classically, squared loss is used: it is perfectly suited to the estimation of the conditional expectation. Using the pinball loss (Eq. 2.2) instead allows to estimate quantiles. In this section we present two approaches based on Equation (2.11) with the pinball loss. The first one is regression using artificial neural networks (NN), a rich and versatile class of functions that has shown a high efficiency in several fields. The second approach is the generalized linear regression in reproducing kernel Hilbert spaces (RK). RK is a non-parametric regression method that has been much studied in the last decades (see Steinwart and Christmann [2008]) since it appeared in the core of learning theory in the 1990's Schölkopf [2001], Vapnik [2013].

2.5.1 Neural Networks

Artificial neural networks have been successfully used for a large variety of tasks such as classification, computer vision, music generation, and regression Bishop [1995]. In the regression setting, feed-forward neural networks have shown outstanding achievements. Here we present quantile regression neural network Cannon [2011] which is an adaptation of the traditional feed-forward neural network.

Overview

A feed-forward neural network is defined by its number of hidden layers H, its numbers of neurons per layer $J_h, 1 \leqslant h \leqslant H$, and its activation functions g_h, $h = 1, \ldots, H$. Given an input vector $x \in \mathbb{R}^D$ the information is fed to the hidden layer 1 composed of a fixed number of neurons J_1. For each neuron $N_i^{(1)}$, $i = 1, .., J_1$, a scalar product (noted $\langle . \, , . \rangle$) is computed between the input vector $x = (x_1, ..., x_D) \in \mathbb{R}^D$ and the weights $w_i^{(1)} = (w_{i,1}^{(1)}, ..., w_{i,D}^{(1)}) \in \mathbb{R}^D$ of the $N_i^{(1)}$ neurons. Then a bias term $b_i^{(1)} \in \mathbb{R}$ is added to the result of the scalar product. The result is composed with the activation function g_1 (linear or non-linear) which is typically the sigmoid or the ReLu function Schmidt-Hieber [2017] and the result is given to the next layer where the same operation is processed until the information comes out from the outpout layer. For example, the output of a 3-layers NN at x_* is given by

$$s(x_*) = g_3 \left(\sum_{j=1}^{J_2} g_2 \Big(\sum_{i=1}^{J_1} g_1 \big(\langle w_i^{(1)}, x_* \rangle + b_i^{(1)} \big) w_{j,i}^{(2)} + b_j^{(2)} \Big) w_{1,j}^{(3)} + b^{(3)} \right). \tag{2.12}$$

The corresponding architecture can be found in Figure 2.3.

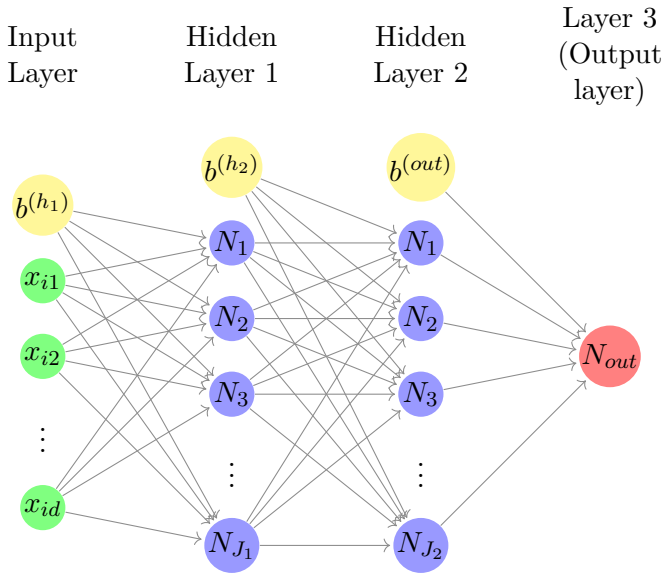


Figure 2.3: Architecture of 3-layer feedforward neural network.

The architecture of the NN defines $\mathcal{H}$. Finding the right architecture is a very difficult problem which will not be treated in this paper. However a classical implementation procedure consists of creating a network large enough (able to overfit) and then using techniques such as early stopping, dropout, bootstrapping or risk regularization to avoid overfitting Srivastava et al. [2014]. In Cannon [2011], the following regularized risk is

used:

$$\mathcal{R}_{r,e}[s] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i - s(x_i)\big) + \lambda \sum_{j=1}^{H}\sum_{z=1}^{J_j} \left\| w_z^{(j)} \right\|^2 . \tag{2.13}$$

Quantile regression

Minimizing Equation (2.13) (with respect to all the weights and biases) is in general challenging, as $\mathcal{R}_{r,e}$ is a highly multimodal function. It is mostly tackled using derivative-based algorithms and multi-starts (*i.e* launching the optimization procedure M_s times with different starting points). In the case of quantile estimation, the loss function is non-differentiable at the origin, which may cause problems to some numerical optimization procedures. To address this issue, Cannon [2011] introduced a smooth version of the pinball loss function, defined as:

$$l_\tau^\eta(\xi) = h^\eta(\xi)(\tau - \mathbb{1}_{\xi<0}),$$

where

$$h^\eta(\xi) = \begin{cases} \dfrac{\xi^2}{2\eta} & \text{if } \ 0 \leqslant |\xi| \leqslant \eta \\ |\xi| - \dfrac{\eta}{2} & \text{if } \ |\xi| \geqslant \eta. \end{cases} \tag{2.14}$$

Note that if the optimizer is based on a first order method such as Kingma and Ba [2014], then the transfer function does not require continuous derivatives. But using a second order method as it is done in the original paper implies the loss function to be twice differentiable with respect to the weights of the neural network. Then transfer functions such as logistic or hyperbolic tangent functions should be used over piecewise linear ones such as the ReLu or the PReLU functions Ramachandran et al. [2017].

Let us define $\boldsymbol{w}$ the list containing the weights and bias of the network. To find $\boldsymbol{w}^*$, a minimizer of $\mathcal{R}_{r,e}$, the idea is to solve a series of problems using the smoothed loss instead of the pinball one with a sequence E_K corresponding to K decreasing values of η. The process begins with the optimization with the larger value η_1. Once the optimization converges, the optimal weights are used as the initialization for the optimization with η_2, and so on. The process stops when the weights based on $l_\tau^{\eta_K}$ are obtained. Finally, $\widehat{q}_\tau(x_*)$ is given by the evaluation of the optimal network at x_*. Algorithm 4 details the implementation of the NN method.

Computational complexity

In Cannon [2011] the optimization is based on a Newton method. Thus the procedure needs to inverse a Hessian matrix. Without sophistications, its cost is $O(s_{\text{pb}}^3)$ with s_{pb} the size of the problem *i.e* the number of parameters to optimize. Note that using a high order method makes sense here because NN has few parameters (in contrast to deep learning methods). Moreover providing an upper bound on the number of iterations needed to reach an optimal point may be really hard in practice because of

Algorithm 4: Neural network

Training:
Input: $\mathcal{D}_n$, τ, λ, H, $(J_1, \ldots, J_H)$, $(g_1, \ldots, g_H)$, E_K
Initialize: Fix $\boldsymbol{w}_0$ as the list containing the initial weights and biases;
for *$t = 1$ to K* **do**
  $\varepsilon \leftarrow E_K[t]$;
  Starting the optimization procedure with $\boldsymbol{w}_0$ and define;

$$\boldsymbol{w}_\tau^* = \arg\min_{\boldsymbol{w}} \frac{1}{n} \sum_{t=1}^{n} l_\tau^\varepsilon(y_i - \widehat{q}^{\boldsymbol{w}}(x_i)) + \lambda \sum_{j=1}^{H} \sum_{i=1}^{J} \left\| w_i^{(j)} \right\|^2$$

  with $\widehat{q}^{\boldsymbol{w}}(\cdot)$ the output of the network with the weights $\boldsymbol{w}$;
  $\boldsymbol{w}_0 \leftarrow \boldsymbol{w}_\tau^*$;
end
Prediction
Input: $\mathcal{X}_{\text{test}}$, $\boldsymbol{w}_\tau^*$, λ, H, $(J_1, \ldots, J_H)$, $(g_1, \ldots, g_H)$
for *each point in $x_* \in \mathcal{X}_{\text{test}}$* **do**
  $\widehat{q}_\tau(x_*) = \widehat{q}^{\boldsymbol{w}_\tau^*}(x_*)$.
end

the non convexity of (2.13). In the non-convex case, there is no optimality guaranty and the optimization could be stuck in a local minima. However, it can be shown that the convergence near a local optimal point is at least super linear (see Boyd and Vandenberghe [2004] Eq. (9.33)) and may be quadratic (if the gradient is small). It implies, for each η, the number of iterations until $\mathcal{R}_{r,e}^\eta(\boldsymbol{w}) - \mathcal{R}_{r,e}^\eta(\boldsymbol{w}^*) \leqslant \varepsilon$ is bounded above by

$$\frac{\mathcal{R}_{r,e}^\eta(\boldsymbol{w}_0) - \mathcal{R}_{r,e}^\eta(\boldsymbol{w}^*)}{\gamma} + \log_2 \log_2(\varepsilon_0/\varepsilon),$$

with γ the minimal decreasing rate, $\varepsilon_0 = 2M_\eta^3/L_\eta^2$, M_η the strong convexity constant of $\mathcal{R}_{r,e}^\eta$ near $\boldsymbol{w}^*$ and L_η the Hessian Lipschitz constant (see Boyd and Vandenberghe [2004] page 489). As $\log_2 \log_2(\varepsilon_0/\varepsilon)$ increases very slowly with respect to ε, it is possible to bound the number of iterations N typically by

$$\frac{\mathcal{R}_{r,e}^\eta(\boldsymbol{w}_0) - \mathcal{R}_{r,e}^\eta(\boldsymbol{w}^*)}{\gamma} + 6.$$

That means, near an optimal point, the complexity is $O(L_\eta n(JD)^3)$, with J the total number of neurons. Then using a multistart procedure implies a complexity of

$$O(M_s L_{\eta^*} n(JD)^3),$$

with $L_{\eta^*} = \max_{\eta_1,\ldots,\eta_K} L_\eta$.

2.5.2 Generalized linear regression

Regression in RKHS was introduced for classification via Support Vector Machine by Cortes and Vapnik [1995], Hearst et al. [1998], and has been naturally extended for the estimation of the conditional expectation Drucker et al. [1997], Rosipal and Trejo [2001]. Since, many applications have been developed (see Steinwart and Christmann [2008], Schölkopf [2001] for some examples), here we present the quantile regression in RKHS Takeuchi et al. [2006], Sangnier et al. [2016].

RKHS introduction and formalism

Under the linear regression framework, S is assumed to be under the form $S(x) = x^T\alpha$, with α in $\mathbb{R}^D$. To stay in the same vein while creating non-linear responses, one can map the input space $\mathcal{X}$ to a space of higher dimension $\mathcal{H}$ (named the feature space), thanks to a feature map Φ. For example the feature space could be a polynomial space, in that case we are working with the spline framework Marsh and Cormier [2001]. For a large flexibility and few parameters, the feature space can even be chosen as an infinite dimensional space. In the following, $\Phi = (\varphi_1, \varphi_2, \varphi_3, ...)$ defines a feature map from $\mathcal{X}$ to $\mathcal{H}$, where $\mathcal{H}$ is the $\mathbb{R}$-Hilbert functional space defined as

$$\mathcal{H} = \left\{ s,\ s(x) = \sum_{j\in\mathbb{N}^*} \alpha_j\varphi_j(x),\ \text{s.t.}\ \ \|s\|_{\mathcal{H}} < +\infty \right\},$$

$$\text{with}\ \ \|s\|_{\mathcal{H}} := \sqrt{\langle s, s\rangle_{\mathcal{H}}},$$

where J is the cardinality of a basis of $\mathcal{H}$. Under the hypothesis that S belongs to $\mathcal{H}$, S can be written as

$$S(x) = \sum_{j\in\mathbb{N}^*} \alpha_j\varphi_j(x). \tag{2.15}$$

Notice that without more hypothesis on $\mathcal{H}$, estimating S is difficult. In fact it is impossible to compute (2.15) directly because of the infinite sum. Thus, using the sample $\mathcal{D}_n$, a solution of (2.10) is not known and cannot be computed.

However, this issue can be tackled by the introduction of the regularized empirical risk

$$\mathcal{R}_{r,e}[s] = \frac{1}{n}\sum_{i=1}^{n} l(y_i - s(x_i)) + \frac{\lambda}{2}\|s\|^2_{\mathcal{H}}\,, \tag{2.16}$$

and the utilization of the RKHS formalism that is based on the *representer theorem* and the so-called *kernel trick* (see Schölkopf [2001] for instance).

Let us first introduce the symmetric definite positive function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ such that:

$$k(x, x') = \big\langle \Phi(x'), \Phi(x)\big\rangle_{\mathcal{H}}\,. \tag{2.17}$$

Under this setting, $\mathcal{H}$ is a RKHS with the reproducing kernel k, that means for all $x \in \mathcal{X}$ $\Phi(x) = k(.,x) \in \mathcal{H}$ and the reproducing property

$$s(x) = \langle s, k(.,x)\rangle_{\mathcal{H}} \tag{2.18}$$

holds for all $s \in \mathcal{H}$ and all $x \in \mathcal{X}$. It can be shown that working with a fixed kernel k is equivalent to working with its associated functional Hilbert space. Note that the kernel choice is based on kernel properties or assumptions made on the functional space. See for instance Steinwart and Christmann [2008], chapter 4, for some kernel definitions and properties. In the following, $\mathcal{H}_\theta$ and k_θ denote respectively a RKHS and its kernel associated to the hyperparameters vector θ. $K^\theta_{x,x} \in \mathbb{R}^{n\times n}$ is the kernel matrix obtained via $K^\theta_{x,x}(i,j) = k_\theta(x_i, x_j)$.

From a theoretical point of view, the *representer theorem* implies that the minimizer $\widehat{S}$ of (2.16) lives in $\mathcal{H}^\theta_{|X} = \text{span}\{\Phi(x_i) : i = 1, ..., n\}$ with

$$\|s\|^2_{\mathcal{H}_{|X}} = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_\theta(x_j, x_i).$$

Combining this result to the definition (2.17) and the reproducing property (2.18), it is possible to rewrite $\widehat{S}$ as:

$$\widehat{S}(x) = \sum_{i=1}^n \alpha_i k_\theta(x, x_i).$$

Hence, the original infinite dimensional problem associated to the formalism (2.15) becomes an optimization problem over n coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n) \in \mathbb{R}^n$. More precisely, finding $\widehat{S}$ is equivalent to minimize in $\boldsymbol{\alpha}$ the quantity

$$\frac{1}{n}\sum_{i=1}^n l\Big(y_i - \big(\sum_{j=1}^n \alpha_j k_\theta(x_i, x_j)\big)\Big) + \frac{\lambda}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k_\theta(x_j, x_i). \tag{2.19}$$

Quantile regression

Quantile regression in RKHS was introduced by Takeuchi et al. [2006], followed by several authors Li et al. [2007], Steinwart et al. [2011], Christmann and Steinwart [2008a,b], Sangnier et al. [2016]. Quantile regression has two specificities compared to the general case. Firstly the loss l is defined as the pinball. Secondly, to ensure the quantile property, the intercept is not regularized. More precisely, we assume that

$$q_\tau(x) = g(x) + b \quad \text{with} \quad g \in \mathcal{H}_\theta \quad \text{and} \quad b \in \mathbb{R}.$$

and we consider the empirical regularized risk

$$\mathcal{R}_{r,e}[q] := \frac{1}{n}\sum_{i=1}^n l_\tau\big(y_i - q(x_i)\big) + \frac{\lambda}{2}\|g\|^2_{\mathcal{H}_\theta}. \tag{2.20}$$

Thus the *representer theorem* implies that $\widehat{q}_\tau$ can be written under the form

$$\widehat{q}_\tau(x_*) = \sum_{i=1}^{n} \alpha_i k_\theta(x_*, x_i) + b,$$

for a new query point x_*. Since (2.20) cannot be minimized analytically, a numerical minimization procedure is used. Cortes and Vapnik [1995] followed by Takeuchi et al. [2006] introduced nonnegative variables $\xi^{(*)} \in \mathbb{R}^+$ to transform the original problem into

$$\mathcal{R}_{r,e}[q] := \frac{1}{n}\sum_{i=1}^{n} \tau\xi_i + (1-\tau)\xi_i^* + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j k_\theta(x_j, x_i),$$

subject to

$$y_i - \left(\sum_{j=1}^{n} \alpha_j k_\theta(x_i, x_j) + b\right) \leqslant \xi_i$$

and

$$\sum_{j=1}^{n} \alpha_j k_\theta(x_i, x_j) + b - y_i \leqslant \xi_i^*, \text{ where } \xi_i^*,\ \xi_i \geqslant 0.$$

Using a Lagrangian formulation, it can be shown (see Steinwart and Christmann [2008] for instance) that minimizing $\mathcal{R}_{r,e}$ is equivalent to the problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & \frac{1}{2}\boldsymbol{\alpha}^T K_{x,x}^\theta \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{y} \\ s.t. \quad & \frac{1}{\lambda n}(\tau - 1) \leqslant \alpha_i \leqslant \frac{1}{\lambda n}\tau,\ \forall\ 1 \leqslant i \leqslant n \\ \text{and} \quad & \sum_{i=1}^{n} \alpha_i = 0. \end{aligned} \tag{2.21}$$

It is a quadratic optimization problem under linear constraint, for which many efficient solvers exist.

The value of b may be obtained from the Karush-Kuhn-Tucker slackness condition or fixed independently of the problem. A simple way to do so is to choose b as the τ-quantile of $\left(y_i - \sum_{j=1}^{n} \alpha_j k_\theta(x_i, x_j)\right)_{1\leqslant i\leqslant n}$. Algorithm 5 details the implementation of the RK method.

Computational complexity

Let us notice two things. Firstly, the minimal upper bound complexity for solving (2.21) is $O(n^3)$. Indeed solving (2.21) without the constraints is easier and it needs $O(n^3)$. Secondly the optimization problem (2.21) is convex, thus the optimum is global.

There are two main approaches for solving (2.21), the interior point method Boyd and Vandenberghe [2004] and the iterative methods like libSVM Chang and Lin [2011].

Algorithm 5: RKHS regression

Training:
Input: $\mathcal{D}_n$,τ, λ, k_θ
Initialize: Compute the $n \times n$ matrix $K^\theta_{x,x}$;
Optimization: Select $\boldsymbol{\alpha}^*$ as;

$$\begin{aligned} \boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & \frac{1}{2}\boldsymbol{\alpha}^T K^\theta_{x,x}\boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{y} \\ s.t \quad & \frac{1}{\lambda n}(\tau - 1) \leqslant \alpha_i \leqslant \frac{1}{\lambda n}\tau, \ \forall\ 1 \leqslant i \leqslant n \\ \text{and} \quad & \sum_{i=1}^n \alpha_i = 0 \end{aligned}$$

Define b as the τ-quantile of $\left(y_i - \sum_{j=1}^n \alpha_j k_\theta(x_i, x_j)\right)_{1\leqslant i\leqslant n}$;
Prediction: **Input:** $\mathcal{X}_{test}$, $\boldsymbol{\alpha}^*$, k_θ
for *each point in* $x_* \in \mathcal{X}_{\text{test}}$ **do**
  compute $K^\theta_{x_*,x}$;
  $\widehat{q}_\tau(x_{\text{test}}) = K^\theta_{x_*,x}\boldsymbol{\alpha}^* + b$;
end

The interior point method is based on the Newton algorithm, one method is the barrier method (see Boyd and Vandenberghe [2004] page 590). It is shown that the number of iterations N for reaching a solution with precision ε is $O(\sqrt{n}\log(\frac{n}{\varepsilon}))$. Moreover each iteration of a Newton type algorithm costs $O(n^3)$ because it needs to inverse a Hessian. Thus, the complexity of an interior point method for finding a global solution with precision ε is

$$O\left(n^{7/2}\log(n/\varepsilon)\right).$$

On another hand, iterative methods like libSVM transform the main problem into a smaller one. At each iteration the algorithm solves a subproblem in $O(n)$. Contrary to the interior point methods, the number of iterations depends explicitly on the matrix $K^\theta_{x,x}$. List and Simon [2009] shows that the number of iterations is

$$O\left(n^2\kappa(K^\theta_{x,x})\log(1/\varepsilon)\right),$$

where $\kappa(K^\theta_{x,x}) = \lambda_{\max}(K^\theta_{x,x})/\lambda_{\min}(K^\theta_{x,x})$. Note that $\kappa(K^\theta_{x,x})$ depends on the type of the kernel, it evolves in $O(n^{s'})$ with $s' > 1$ an increasing value of the regularity of k_θ Chang and Ha [1999]. For more information about the eigenvalues of $K^\theta_{x,x}$ one can consult Braun [2006].

To summarize, it implies that the complexity of the libSVM method has an upper bound higher than the interior point algorithm. However, these algorithms are known

to converge pretty fast. In practice, the upper bound is almost never reached, and thus the most important factor is the cost per iteration, rather than the number of iterations needed. This is the reason why libSVM is popular in this setting.

2.6 Bayesian approaches

Bayesian formalism has been used for a wide class of problems such as classification and regression Rasmussen and Williams [2006], model averaging Box and Tiao [2011] and model selection Raftery [1995].

The first Bayesian quantile regression framework was introduced in Yu and Moyeed [2001] where the authors worked under a linear framework and improper uniform priors. Nevertheless the linear hypothesis may be too restrictive to treat the stochastic black box setting. Taddy and Kottas [2010] introduced a mixture modeling framework called Dirichlet process to perform nonlinear quantile regression. However the inference is performed with MCMC methods (see Gilks et al. [1995], Gamerman and Lopes [2006] for instance), a procedure that is often costly. A possible alternative is the use of Gaussian process (GP). GPs are powerful in a Bayesian context because of their flexibility and their tractability (GPs are only characterized by their mean m and covariance k_θ). Using GPs, a possible approach is to use a joint modeling of the mean and variance, assuming that the distribution is Gaussian everywhere, and then to extract the quantiles of interest as it is done in Kersting et al. [2007], Lázaro-Gredilla and Titsias [2011]. Nevertheless this strategy may introduce a high bias when the true output distribution is not Gaussian.

In this section we present QK and VB, two approaches that use GP as a prior for q_τ. Contrary to the joint modelling approach, here the GP prior for q_τ does not imply any structure on the output distribution, which allows the creation of very flexible quantile models.

2.6.1 Quantile kriging

Kriging takes its origins in geostatistics and spatial data interpolation Cressie [1990], Stein [2012]. Since the 2000's, kriging drew attention of the machine learning community (see Rasmussen and Williams [2006] for some applications). In this section we present a very intuitive method that gives flexible quantile estimators based on data containing repetition and GPs Plumlee and Tuo [2014].

Kriging introduction

Kriging is based on the hypothesis that

$$S(x) \sim \mathcal{GP}\big(m(x), k_\theta(x, x')\big) \ . \tag{2.22}$$

Which means for every finite set $(x_1, \cdots, x_T)$, the output $\big(S(x_1), \cdots, S(x_T)\big)$ is multivariate Gaussian. Here m is the mean of the process and k_θ is a kernel function also known as the covariance function. Note that in the sequel we take $m = 0$ in order to

simplify the computations and notations. The covariance function conveys many properties of the process, so its choice should depend on the assumptions made on S. A first classical assumption is that the GP is stationary, *i.e* the correlation between two inputs does not depends on the location but only on the distance between the points. Then, different class of stationary kernels produce GPs with different regularities. The class of Matérn kernels is very convenient because it depends on a regularity hyperparameter that enables the end user to adapt its prior to his regularity assumptions (see Rasmussen and Williams [2006] for more details). For example, the Matérn 5/2 kernel is defined as

$$k_\theta(x,x') = \rho^2\big(1+\sqrt{5}\left\|x-x'\right\|_\theta + \frac{5}{3}\left\|x-x'\right\|_\theta^2\big)\exp(-\sqrt{5}\left\|x-x'\right\|_\theta), \tag{2.23}$$

where $\rho > 0$ and

$$\left\|x-x'\right\|_\theta^2 = (x-x')^T\Lambda_\theta(x-x'),$$

with Λ_θ a diagonal matrix with diagonal terms the inverses of the D squared length scales θ_i, $i = 1,\cdots,D$.

In addition to the assumption 2.22, let us assume y_i is observed with noise such that

$$y_i = S(x_i) + \varepsilon_i \tag{2.24}$$

with $\varepsilon_i \sim \mathcal{N}(0,\sigma_i^2)$. As a consequence the associated likelihood is Gaussian, *i.e*.

$$p(\mathcal{Y}_n|\mathcal{X}_n) = \mathcal{N}\big(\mathbf{0}, K^\theta_{x,x} + \mathrm{diag}(\boldsymbol{\sigma}^2)\big), \tag{2.25}$$

with $\boldsymbol{\sigma} = (\sigma_1,\cdots,\sigma_n)$. Because $\big(\mathcal{Y}_n, S(x_*)\big)^T$ is a Gaussian vector of zero mean and covariance

$$\boldsymbol{K} = \begin{pmatrix} K^\theta_{x,x} + \mathrm{diag}(\boldsymbol{\sigma}^2) & K^\theta_{x,x_*} \\ K^\theta_{x_*,x} & K^\theta_{x_*,x_*} \end{pmatrix}, \tag{2.26}$$

the distribution of $S(x_*)$ knowing $\mathcal{Y}_n$ is still Gaussian (see Rasmussen and Williams [2006] appendix A.2 for more details). Thus it is possible to provide the distribution a posteriori for Kriging regression model as

$$S(x_*) \sim \mathcal{N}\big(\bar{S}(x_*), \mathbb{V}_S(x_*)\big) \tag{2.27}$$

with

$$\bar{S}(x_*) = K^\theta_{x_*,x}(K^\theta_{x,x} + \mathrm{diag}(\boldsymbol{\sigma}^2))^{-1}\mathcal{Y}_n,$$
$$\mathbb{V}_S(x_*) = k_\theta(x_*,x_*) - K^\theta_{x_*,x}(K^\theta_{x,x} + \mathrm{diag}(\boldsymbol{\sigma}^2))^{-1}K^\theta_{x,x_*}.$$

As in Section 2.5.2, the covariance functions are usually chosen among a set of predefined ones (for example the Matérn 5/2 see Eq.2.23), that depend on a set of hyperparameters $\boldsymbol{\theta} \in \mathbb{R}^{D+1}$. The best hyperparameter $\boldsymbol{\theta}^*$ can be selected as the maximizer of the marginal likelihood. More precisely it follows (see Rasmussen and Williams [2006] for instance) that

$$\begin{aligned} p(\mathcal{Y}_n|\mathcal{X}_n,\boldsymbol{\theta},\boldsymbol{\sigma}) = &-\frac{1}{2}\mathcal{Y}_n^T(K^\theta_{x,x}+B)^{-1}\mathcal{Y}_n \\ &-\frac{1}{2}\log|(K^\theta_{x,x}+B)| - \frac{n}{2}\log(2\pi), \end{aligned} \tag{2.28}$$

where $|K|$ is the determinant of the matrix K. Maximizing this likelihood with respect to $\boldsymbol{\theta}$ is usually done using derivative-based algorithms, although the problem is non-convex and known to have several local maxima.

Different estimators of S may be extracted based on (2.27). Here $\widehat{S}$ is fixed as $\bar{S}_{\theta^*}$. Note that this classical choice is made because the maximum a posteriori of a Gaussian distribution coincides with its mean.

Quantile kriging

Algorithm 6: Quantile kriging

Training:
Input: $\mathcal{D}_{n',r}$,τ, k_θ
Initialize Compute the $n' \times n'$ matrix $K^\theta_{x,x}$;
for *i = 1 to n'* **do**
  Define the local estimator of the τ-quantile: $\widehat{q}_\tau(x_i) = y_{i,([r\tau])}$;
  Estimate σ_i by bootstrap
end
Define $B = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_{n'}^2)$ and compute $(K^\theta_{x,x} + B)^{-1}$;
Define the kernel hyperparameters θ^* as

$$\theta^* = \arg\max_\theta -\frac{1}{2}\mathcal{Q}_{n'}^T (K^\theta_{x,x} + B)^{-1} \mathcal{Q}_{n'} - \frac{1}{2}\log|K^\theta_{x,x} + B| - \frac{n}{2}\log(2\pi)$$

Input: $\mathcal{X}_{test}$, $\mathcal{Q}_{n'}$, θ^*, B
for *each point in* $x_* \in \mathcal{X}_{\text{test}}$ **do**
  compute $K^{\theta^*}_{x_*,x}$
  $\widehat{q}^{\theta^*}_\tau(x_*) = K^{\theta^*}_{x_*,x}(K^{\theta^*}_{x,x} + B)^{-1}\widehat{q}_\tau$
end

As q_τ is a latent quantity, the solution proposed in Plumlee and Tuo [2014] is to consider the sample $\mathcal{D}_{n',r}$ corresponding to a design of experiments with n' different points that are repeated r times in order to obtain quantile observations. For each $x_i \in \mathcal{X}$, $1 \leqslant i \leqslant n'$, let us define:

$$y_{i,r} = (y_{i,1}, .., y_{i,r})$$

and

$$\mathcal{D}_{n',\tau,r} = \Big(\big(x_1, \widehat{q}_\tau(x_1)\big), ..., \big(x_n, \widehat{q}_\tau(x_n)\big)\Big), \text{with } \widehat{q}_\tau(x_i) = y_{i,([r\tau])}.$$

Following Plumlee and Tuo [2014], let us assume that

$$\widehat{q}_\tau(x_i) = q_\tau(x_i) + \varepsilon_i, \text{ with } \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2). \tag{2.29}$$

Note that from a statistical point of view Assumption (2.29) is wrong because the distribution is asymetric around the quantile but asymptotically consistent as illustrated

by the central limit theorem for sample quantiles. The resulting estimator is

$$\widehat{q}_\tau(x_*) = K^\theta_{x_*,x}(K^\theta_{x,x} + B)^{-1}\mathcal{Q}_{n'}, \tag{2.30}$$

with $\mathcal{Q}_{n'} = (\widehat{q}_\tau(x_1), \cdots, \widehat{q}_\tau(x_{n'}))$ and $B = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n'^2)$.

There are several possibilities to evaluate the noise variances σ_i^2. Here we choose to use a bootstrap technique (that is, generate bootstrapped samples of $y_{i,r}$, compute the corresponding $\widehat{q}_\tau(x_i)$ values and take the variance over those values as the noise variance) but it is possible to use the central limit theorem as it is presented in Bachoc [2013]. The hyperparameters are selected based on (2.28) changing $\mathcal{Y}_n$ by $\mathcal{Q}_{n'}$. Algorithm 6 details the implementation of the QK method.

Computational complexity

If $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D, \rho)$, optimizing (2.28) with a Newton type algorithm implies to inverse a $(D+1)\times(D+1)$ matrix. In addition, for each component of $\boldsymbol{\theta}$, obtaining the partial derivatives of (2.28) requires the computation of $(K^\theta_{x,x} + B)^{-1}$ Rasmussen and Williams [2006]. Thus at each step of the algorithm, the complexity is $O(n'^3 + D^3)$. Assuming the starting point θ_{start} is close to an optimal θ^*, based on the same analysis as in section 2.5.1, the complexity to find θ^* is

$$O\big(L(d^3 + n'^3)\big),$$

with L the Hessian Lipschitz constant.

Finally, obtaining $\widehat{q}_\tau^{\theta^*}$ from (2.30) implies inversing the matrix $K^\theta_{x,x} + B$ that is in $O(n'^3)$. So the whole complexity is

$$O\big(L(d^3 + n'^3) + n'^3\big).$$

2.6.2 Bayesian variational regression

Quantile kriging requires repeated observations to obtain direct observations of the quantile and make the hypothesis of Gaussian errors acceptable. Variational approaches allow us to remove this critical constraint, while setting a more realistic statistical hypothesis on ε. Starting from the decomposition of Eq.2.24, $\varepsilon(x)$ is now assumed to follow a Laplace asymmetric distribution Yu and Zhang [2005], Lum et al. [2012], implying:

$$p\big(y|q,\tau,\sigma,x\big) = \frac{\tau(1-\tau)}{\sigma}\exp\left(-\frac{l_\tau(y-q(x))}{\sigma}\right), \tag{2.31}$$

with the priors on q and σ that has to be fixed.

Such assumption may be justified by the fact that minimizing the empirical risk associated to the pinball loss is equivalent to maximizing the asymmetric Laplace likelihood, which is given by

$$p\big(\mathcal{Y}_n|q_\tau,\mathcal{X}_n,\theta\big) = \prod_{i=1}^{n}\frac{\tau(1-\tau)}{\sigma}\exp\left(-\frac{l_\tau(y_i-q_\tau(x_i))}{\sigma}\right). \tag{2.32}$$

According to the Bayes formula, the posterior can be written as

$$p(q_\tau|\mathcal{D}_n) = \frac{p(\mathcal{Y}_n|\mathcal{X}_n, q_\tau)\ p(q_\tau)}{p(\mathcal{Y}_n|\mathcal{X}_n)}.$$

As the normalizing constant is independent of q_τ, considering only the likelihood and the prior is enough. We obtain

$$p(q_\tau|\mathcal{D}_n) \propto p(\mathcal{Y}_n|\mathcal{X}_n, q_\tau)\ p(q_\tau). \tag{2.33}$$

Because of the Laplace asymetric likelihood, contrary to the classical kriging model, here the posterior distribution (2.33) is not Gaussian anymore. Thus it is not possible to provide an analytical expression for the regression model. To overcome this problem, Boukouvalas et al. [2012] used a variational approach with an expectation-propagation (EP) algorithm Minka [2001], while Abeywardana and Ramos [2015] used a variational expectation maximization (EM) algorithm which was found to perform slightly better.

Variational EM algorithm

The EM algorithm was introduced in Dempster et al. [1977] to compute maximum-likelihood estimates. Since then, it has been widely used in a large variety of fields (see McLachlan and Krishnan [2007] for more details). Classically, the purpose of the EM algorithm is to find ζ a vector of parameters that define the model and that maximizes $p(\mathcal{Y}_n|\zeta)$ thanks to the introduction of the hidden variables $z = (z_1, \cdots, z_M)$. However dealing with this classical formalism implies that $p(z|\mathcal{Y}_n, \zeta)$ is known or some sufficient statistics can be computed (see Tzikas et al. [2008], Robert and Casella [2013] for more details), which is not always possible. Using the variational EM framework is a possibility to bypass this requirement Tzikas et al. [2008], by approximating $p(z|\mathcal{Y}_n, \zeta)$ by a probability distribution $\widetilde{p}$ that factorizes under the form

$$\widetilde{p}(z) = \prod_{j=1}^{M} \widetilde{p}_j(z_j).$$

Starting from the log-likelihood $\log\big(p(\mathcal{Y}_n|\zeta)\big)$, thanks to Jensen's inequality, it is possible to show that:

$$\log\big(p(\mathcal{Y}_n|\zeta)\big) \geqslant \mathcal{L}(\widetilde{p}, \zeta) + \mathrm{kl}(\widetilde{p}||p),$$

where

$$\mathcal{L}(\widetilde{p}, \zeta) = \int \widetilde{p}(z) \log\left(\frac{p(\mathcal{Y}_n, z|\zeta)}{\widetilde{p}(z)}\right) dz,$$

and kl is the Kullback-Leibler divergence:

$$\mathrm{kl}(\widetilde{p}||p) = -\int \widetilde{p}(z) \log(\frac{p(z|\mathcal{Y}_n, \zeta)}{\widetilde{p}(z)}) dz.$$

As presented on Figure 2.6.2, the EM algorithm can be viewed as a two-step optimization technique. The lower bound $\mathcal{L}$ is first maximized with respect to $\widetilde{p}$ (E-step) so that to minimize $\mathrm{kl}(\widetilde{p}||p)$. Next the likelihood is directly maximized with respect to the parameter ζ (M-step).

Classically and in the following, the E-step optimization problem is analytically solved (see Tzikas et al. [2008] for details about the computation). In this particular case, the quantities $\widetilde{p}_j$ are limited to conditionally conjugate exponential families. But note that different strategies have been developed to relax this assumption (generally it comes with an higher computational complexity) so that creating more flexible models (see the review Zhang et al. for instance).

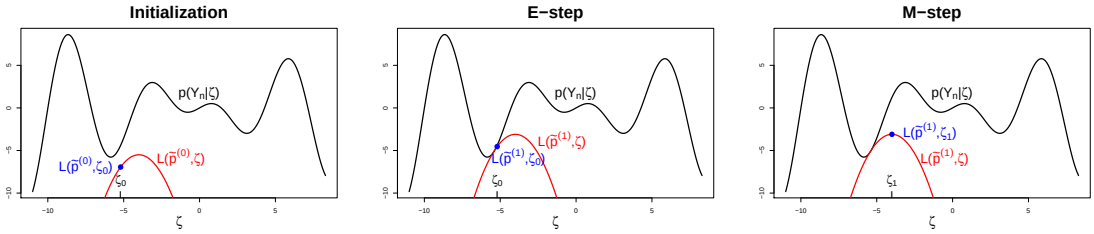


Figure 2.4: First steps of the variational EM algorithm.

Variational EM applied to quantile regression

Following Abeywardana and Ramos [2015], let us suppose that

$$q_\tau(x) \sim \mathcal{GP}\big(m(x), k_\theta(x, x')\big)$$

$$\sigma \sim \mathrm{IG}(10^{-6}, 10^{-6}),$$

with IG defining the inverse gamma distribution andfor sake of simplicity, $m(.) = 0$. Note that contrary to the formalism introduce with QK, here σ is taken as a random variable. To allow analytical computation, let us introduce an alternative definition of the Laplace distribution Lum et al. [2012], Kotz et al. [2012]:

$$p(y_i|q_\tau, x_i, \sigma, \tau) = \int \mathcal{N}(y_i|\mu_i, \sigma_{y_i}) \exp(-w_i) dw, \tag{2.34}$$

where $\mu_{y_i} = q_\tau(x_i) + \frac{1-2\tau}{\tau(1-\tau)}\sigma w_i$, $\sigma_{y_i} = \frac{2}{\tau(1-\tau)}\sigma^2 w_i$ and w_i is distributed according to an exponential law of parameter 1.

The distribution of q_τ at a new point x_* is given by averaging the output of all Gaussian models with respect to the posterior $p(q_\tau, \sigma, w|x, Y)$:

$$p\big(q_\tau(x_*)|, \mathcal{D}_n\big) = \int p\big(q_\tau(x_*)|q_\tau, \sigma, w, \mathcal{D}_n\big)\ p\big(q_\tau, \sigma, w|\mathcal{D}_n\big) dq\, d\sigma\, dw. \tag{2.35}$$

Here the crux is to compute the posterior

$$p\big(q_\tau, \sigma, w|\mathcal{D}_n\big) \propto p\big(\mathcal{Y}_n|q_\tau, \sigma, w, \mathcal{X}_n\big)\, p\big(q_\tau, \sigma, w\big).$$

To do so, in Abeywardana and Ramos [2015] the authors use $z = (q_\tau(\mathcal{X}_n), w, \sigma)$ as hidden variables and $\zeta = \theta \in \mathbb{R}^{D+1}$ as parameters and the variational factorization approximation

$$p(q_\tau, \sigma, w|\mathcal{D}_n) \approx \widetilde{p}(q_\tau, w, \sigma|\mathcal{D}_n) = \widetilde{p}(q_\tau|\mathcal{D}_n)\, \widetilde{p}(w|\mathcal{D}_n)\, \widetilde{p}(\sigma|\mathcal{D}_n). \tag{2.36}$$

The EM algorithm provides a nice formalism here. Although the goal is to find θ such that $p\big(\mathcal{Y}_n|\mathcal{X}_n, \theta\big)$ is maximal, the algorithm estimates the underlying GP (i.e. $p\big(q_\tau, \sigma, w|\mathcal{D}_n\big)$) that is able to have a likelihood as large as possible. Then the estimated value $p(q_\tau, \sigma, w|\mathcal{D}_n)$ is plugged into (2.35) to provide the final quantity of interest.

E-step. Because the posteriors are conjugated, it is possible to obtain an analytical expression of the optimal distribution $\widetilde{p}(q_\tau)$:

$$\widetilde{p}(q_\tau|\mathcal{D}_n) \sim \mathcal{N}\left(\mu_\theta, \Sigma_\theta\right),$$

where

$$\mu_\theta = \Sigma_\theta \left(\boldsymbol{D}\mathcal{Y}_n - \frac{1-2\tau}{2}\mathbb{E}\Big(\frac{1}{\sigma}\Big)\mathbf{1}\right)$$

and

$$\Sigma_\theta = \left(\boldsymbol{D} + {K^\theta_{x,x}}^{-1}\right)^{-1},$$

with

$$\boldsymbol{D} = \frac{\tau(1-\tau)}{2}\mathbb{E}\Big(\frac{1}{\sigma^2}\Big)\,\mathrm{diag}\left(\mathbb{E}\Big(\frac{1}{w_i}\Big)\right)_{i=1,..,n}.$$

The posterior on w_i is a Generalized Inverse Gaussian $\mathbf{GIG}(1/2, \alpha_i, \beta_i)$ with :

$$\alpha_i = \left(\frac{(1-2\tau)^2}{2\tau(1-\tau)} + 2\right)$$

and

$$\beta_i = \frac{\tau(1-\tau)}{2}\mathbb{E}\Big(\frac{1}{\sigma^2}\Big)\Big(y_i^2 - 2y_i\mathbb{E}\big(q_\tau(x_i)\big) + \mathbb{E}\big(q_\tau(x_i)^2\big)\Big).$$

Due to numerical problems, in Abeywardana and Ramos [2015] the authors use the restriction $\widetilde{p}(\sigma) = IG(a, b)$. Finding the best a, b is done numerically. Then finding the best a, b is equivalent to maximizing:

$$\begin{aligned} J(a,b) &= (a - N - 10^{-6})\log(b - \psi(a)) \\ &+ (b-\gamma)\frac{a}{b} - \delta\frac{a(a+1)}{b^2} - a\log(b) + \log\Gamma(a), \end{aligned}$$

with

$$\gamma = -\frac{1-2\tau}{2}\sum_{i=1}^{n} y_i - \mathbb{E}\big(q_\tau(x_i)\big)$$

and

$$\delta = \frac{\alpha(1-\tau)}{4}\sum_{i=1}^{n}\mathbb{E}\Big(\frac{1}{w_i}\Big)\Big(y_i^2 - 2y_i\mathbb{E}\big(q_\tau(x_i)\big) + \mathbb{E}\big(q_\tau(x_i)^2\big)\Big).$$

M-step. Ignoring terms that do not depend on θ, we obtain the lower bound:

$$\begin{aligned}\widetilde{\mathcal{L}}(\theta) &= \int \widetilde{p}(q_\tau|\theta)\,\widetilde{p}(w)\,\widetilde{p}(\sigma)\log p(y|q_\tau,w,\sigma)p(q_\tau|\theta)d\sigma dw dq_\tau \\ &- \int \widetilde{p}(q_\tau|\theta)\log\widetilde{p}(q_\tau|\theta)dq_\tau \\ &= \frac{1}{2}\Big(\mu_\theta^T\Sigma_\theta^{-1}\mu_\theta - \log|\boldsymbol{D}^{-1} + K^\theta_{x,x}|\Big). \end{aligned} \tag{2.37}$$

The optimization of $\widetilde{\mathcal{L}}$ with respect to θ is done using a numerical optimizer.

Recalling the goal is to compute (2.35), thanks to (2.36), we make the approximation:

$$p(q_\tau|x_*,\mathcal{D}_n) \approx \int p\big(q_\tau(x_*)|q_\tau,\sigma,\mathcal{D}_n\big)\,\widetilde{p}(q_\tau)\,\widetilde{p}(\sigma)\,\widetilde{p}(w)\,dq_\tau dw d\sigma.$$

Then we obtain

$$p(q_\tau|x_*,\mathcal{D}_n) \approx \mathcal{N}\big(\bar{q}_\tau(x_*),\mathbb{V}_q(x_*)\big),$$

where

$$\begin{aligned}\bar{q}_\tau(x_*) &= K^\theta_{x_*,x}{K^\theta_{x,x}}^{-1}\mu_\theta \text{ and} \\ \mathbb{V}_q(x_*) &= k_\theta(x_*,x_*) - K^\theta_{x_*,x}{K^\theta_{x,x}}^{-1}{K^\theta_{x_*,x}}^T + K^\theta_{x_*,x}{K^\theta_{x,x}}^{-1}\Sigma_\theta{K^\theta_{x,x}}^{-1}{K^\theta_{x_*,x}}^T.\end{aligned}$$

Finally, as explained in section 2.6.1, the quantile estimator $\widehat{q}_\tau$ is selected as $\bar{q}_\tau$. Algorithm 7 details the implementation of the VB method.

Computational complexity

E-step. The complexity of this step is in

$$O(n^3).$$

In fact the algorithm computes Σ_θ that implies inversing a matrix of size $n \times n$.

M-step. Optimizing $\widetilde{\mathcal{L}}$ with a Newton type algorithm costs $O(n^3+D^3)$ at each iteration (for details refer to the optimization description of (2.28)). Assuming the starting point θ_{start} is close to an optimal θ^*, based on the same analysis as in section 2.5.1, the whole complexity is in

$$O\big(L(D^3+n^3)\big).$$

Algorithm 7: Bayesian variational regression

Training:
Input: $\mathcal{D}_n$,τ, k_{θ_0}
Initialize Compute the $n \times n$ matrix $K^{\theta}_{x,x}$ and ${K^{\theta}_{x,x}}^{-1}$
$\theta = \theta_0$;
for *t = 1 to* n_{it} **do**
  E-step
  Compute Σ_θ, μ_θ, α_i, β_i, w_i, (a, b);
  M-step
  $\theta = \arg\max \frac{1}{2}\Big(\mu_\theta^T \Sigma_\theta^{-1} \mu_\theta - \log|\boldsymbol{D}^{-1} + K^{\theta}_{x,x}|\Big)$;
end
Prediction: **Input:** $\mathcal{X}_{\text{test}}$, $\theta^* = \theta$, μ_{θ^*}
for *each point in* $x_* \in \mathcal{X}_{test}$ **do**
  $\widehat{q}_\tau(x_*) = K^{\theta^*}_{x_*,x} {K^{\theta^*}_{x,x}}^{-1} \mu_{\theta^*}$
end

Overall complexity. At each iteration of the EM algorithm, the computation cost is $O(L(D^3 + n^3) + n^3)$. The final complexity is obtained by multiplying by the number of iterations n_{it} of the EM algorithm. Thus, the overall complexity is in

$$O\big(n_{\text{it}}(L(D^3 + n^3) + n^3)\big).$$

2.7 Metamodel summary and implementation

In this section we detail our implementation procedure. After providing a summary of the six metamodels in Table 2.1, we present the packages we used and the hyperparameters we chose (which hyperparameters we set and which hyperparameters we optimized). We then describe the procedure we used to optimize the hyperparameters (optimization strategies and evaluation metrics).

2.7.1 Summary of the models

Table 2.1 lists the analytical expressions of the six metamodels, along with the associated underlying quantity.

2.7.2 Packages and hyperparameter choices

Each method depends on many parameters that can be tuned to improve performance, for example the choice of the kernel function and the value of its parameters for RK, QK and VB or the penalization factor for RK and NN. Here, to limit the computational burden, we chose to optimize only the most critical ones. When possible, for the other parameters, we applied the arbitrary choices and values made by the authors of the

Method	Definition of $\widehat{q}_\tau(x_*)$	Related quantity	Complexity
KN	$y_{([K\tau])}(x_*)$	The K-nearest points from x_*	$O\big(nN_{\text{new}}(D+\log n)\big)$
RF	$\inf\{y_i : \widehat{F}(y_i\|X=x_*) \geqslant \tau\}$	$\widehat{F}(y_i\|X=x_*) = \sum_{i=1}^{n} \bar{\omega}_i(x_*)\mathbb{1}_{\{y_i \leqslant y\}}$	$O(N\widetilde{d}n\log^2 n)$
NN	For a 3 layer NN $g_3(\sum_{j=1}^{J_2} g_2(\sum_{i=1}^{J_1} g_1(\langle w_i^{(h_1)}, x_* \rangle$ $+b_i^{(h_1)})w_j^{(h_2)} + b_j^{(h_2)})w^{(h_3)} + b^{(h_3)})$	With $w_i^{(h_1)}$, $w_j^{(h_2)}$, $b_i^{(h_1)}$, $b_j^{(h_2)}$, $w^{(h_3)}$, $b^{(h_3)}$, $1 \leqslant i \leqslant J_1$, $1 \leqslant j \leqslant J_2$, minimizing $\frac{1}{n}\sum_{t=1}^{n} l_\tau(y_i - \widehat{q}_\tau(x_i)) + \sum_{j,i} \frac{\lambda}{J_j} \left\| w_i^{(h_j)} \right\|^2$	$O\big(M_s L_{\eta^*} n (JD)^3\big)$
RK	$\sum_{i=1}^{n} \alpha_i k_\theta(x_*, x_i) + b$	With $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ minimizing $\frac{1}{2}\boldsymbol{\alpha}^T K_{x,x}^\theta \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{y}$ s.t $\frac{\tau-1}{\lambda n} \leqslant \alpha_i \leqslant \frac{\tau}{\lambda n}$, $\forall\ 1 \leqslant i \leqslant n$ and $\sum_{i=1}^{n} \alpha_i = 0$ and b the τ-quantile of $(y_i - \sum_{j=1}^{n} \alpha_j k_\theta(x_i, x_j))_{1 \leqslant i \leqslant n}$	$O\big(n^{7/2}\log(\frac{n}{\varepsilon})\big)$
QK	$K_{x_*,x}^\theta (K_{x,x}^\theta + B)^{-1}\widehat{q}_\tau$	Maximizing the likelihood: $p(\mathcal{Q}_{n'}\|\mathcal{X}_{n'})$ $\widehat{q}_\tau(x_i) = q_\tau(x_i) + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0, B_{ii})$	$O\big(L(D^3 + n'^3) + n'^3\big)$
VB	$K_{x_*,x}^\theta {K_{x,x}^\theta}^{-1}\mu_\theta$	Approached solution that maximize: $p(\mathcal{Y}_n\|\mathcal{X}_n)$ $y_i = q_\tau(x_i) + \varepsilon \quad \varepsilon \sim \text{ALP}(0, \sigma)$	$O\Big(n_{\text{it}}\big(L(D^3 + n^3) + n^3\big)\Big)$

Table 2.1: Summary of the metamodels, included the definition of the estimators, the associated numerical quantity and the related computation complexity. In this table L and L_{η^*} are Hessian Lipschitz constants, M_s the number of multistarts (on the weights at hyperparameters fixed), D the input dimension, N_{new} the size of the prediction set, n_{it} the number of iterations for the EM algorithm, $k_\theta(.,.)$ the kernel function, B a diagonal matrix, J the total number of neurons, $\widetilde{d}$ the number of covariables considered to find the best splitting point.

original papers. Most changes were made to improve robustness. Below, we describe our experimental settings, also summarized in Table 2.2.

Nearest Neighbors. We set $d()$ as the Euclidean distance and optimized only the size K of the neighborhood.

Random forest. In this case, the only hyperparameter we optimized was the maximum size of the leaves $m_s \in \mathbb{N}^*$. Regarding the number of trees, we noticed that the metamodel needs many more trees than are needed for the estimation of the expectation. In some problems, the metamodel needs up to 5,000 trees to stabilize the variance. Thus, in our experiments we set the number of trees at 10,000 in all cases. We set the number of dimensions considered for the split at the default choice $D/3$. The depth of the tree is not constrained and the splitting rule is based on Eq. 2.7. We used the R package *QuantregForest* Meinshausen and Meinshausen [2007].

Neural network. Based on Cannon [2011], we set the number of hidden layers at one and the transfer function as the sigmoid. The optimization algorithm is a Newton method, we set E_K at $1/2^K$ with $K = 1, 2, 5, 10, 15, 20, 25, 30, 35$ and the number of multistarts to optimize the empirical risk at five. We optimized the number of neurons J_1 in the hidden unit and the regularization parameter $\lambda \in \mathbb{R}_+$. The metamodel is generated using the R package qrnn Cannon [2011].

Regression in RKHS. The kernel was set as a Matérn 5/2. We optimized the length scale parameters $\theta \in \mathbb{R}_+^D$ and the regularization hyperparameter $\lambda \in \mathbb{R}_+$. Optimization (2.21) is done with the quadratic optimizer quadprog Turlach and Weingessel [2007].

Quantile Kriging. The kernel was set as a Matérn 5/2. The number of repetitions was set according to the total number of observations (see Table 2.4). We optimized the length scale hyperparameter $\theta \in \mathbb{R}_+^D$ and variance hyperparameter $\rho \in \mathbb{R}_+$. QK is implemented in the R package *DiceKriging* Roustant et al. [2012].

Method	Hyperparameters
KN	number of neighbors $K \in \mathbb{N}^*$
RF	maximum size of the leaves $m_s \in \mathbb{N}^*$
NN	regularization $\lambda \in \mathbb{R}_+$, $J_1 \in \mathbb{N}^*$
RK	regularization $\lambda \in \mathbb{R}_+$, lengthscales $\theta \in {\mathbb{R}_+}^D$
QK	length scale and variance $\theta \in {\mathbb{R}_+}^{D+1}$
VB	length scale and variance $\theta \in {\mathbb{R}_+}^{D+1}$

Table 2.2: Hyperparameters optimized on our benchmark.

Variational regression. The kernel was set as a Matérn 5/2, the number of EM iterations n_{it} at 50. We optimized the length scale hyperparameter $\theta \in \mathbb{R}_+^D$ and variance hyperparameter $\rho \in \mathbb{R}_+$. The implementation is based on the Matlab code provided in Abeywardana and Ramos [2015].

2.7.3 Tuning the hyperparameters

In the previous section, we defined the hyperparameters we wanted to optimize for each method. In fact, once the type of metamodel is chosen, the quantile estimator is given by a function $\widehat{q}_\Theta : \mathcal{X} \to \mathbb{R}$ where $\Theta \in \mathbb{R}^v$ are called hyperparameters and v is metamodel dependent. Hyperparameter optimization (also known as model selection) is an essential procedure when dealing with non-parametric estimators. Although $\widehat{q}_\Theta$ may be very efficient on $\mathcal{D}_n$, the prediction may perform very poorly on an independent dataset $\mathcal{D}'_p$. The goal is to find the Θ that provides the best possible generalized estimator. In the following, we present the validation metric used to optimize the hyperparameter values and detail the hyperparameters optimization procedure associated with each method.

Metrics

In the standard conditional expectation estimation, the validation and performance metrics are both based on $\|\widehat{m}_\Theta - m\|_{L^2}$, where $\widehat{m}_\Theta$ is the estimator and m the targeted value. With the quantile estimation procedure the two metrics are no longer the same. The goal is to find $\widehat{q}_\Theta$ such that $\|\widehat{q}_\Theta - q\|_{L^2}$ is as small as possible. However, q is unobserved so the validation metric cannot be based on the L^2 norm. Here we present two metrics able to measure the generalization capacity of a quantile metamodel.

Bayesian metamodels (QK and VB) have their own validation metric, this is the likelihood function that can be maximized with respect to Θ. For quantile kriging, we use (2.28) while in the variational approach we use (2.37). The optimal hyperparameters are then:

$$\Theta^*_{mv} = \arg\max_\Theta p\big(\mathcal{Y}_n|\mathcal{X}_n, \Theta\big). \tag{2.38}$$

The second metric available for all metamodels is k-fold cross-validation associated with the pinball loss. The metric can be computed as follows. First, the data are split

into k equal parts, then the model is trained on $\mathcal{D}_{-j}$, the training set without the j-th part. The performance is evaluated on the remaining part $\mathcal{D}_j$. As the quantile minimizes the pinball loss (on $\mathcal{D}_j$), the evaluation metric is

$$E_{cv}(\widehat{q}_\tau^{\Theta}) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n'_j}\sum_{i=1}^{n'_j} l_\tau\big((y_i - \widehat{q}_\tau^{\Theta}(x_i)\big), \tag{2.39}$$

where n'_j is the number of observations in each fold. The optimal cross-validation hyperparameters are then:

$$\Theta^*_{cv} \in \arg\min_{\Theta} E_{cv}(\widehat{q}_\tau^{\Theta}).$$

In our experiments, we chose $k = 5$ to limit the computational cost. However, we observed empirically that choosing a larger k did not substantially modify the performances of the metamodels. Although cross-validation is available for QK, we chose to stay in the spirit of the methods and to only use likelihood to select hyperparameters. Our choice is supported by Bachoc [2013] that does not show a clear improvement using cross validation instead of maximum likelihood techniques.

Hyperparameters optimization procedure

Both likelihood functions come with analytical derivatives, enabling the use of gradient-based algorithms. However, since both functions are multi-modal, multi-start techniques are necessary (and generally very efficient, see Hansen [2009]) to avoid being trapped in local optima. To account for the increasing difficulty of the optimization task with the dimension while limiting the computational cost, we ran $n_{\text{start}} = 20D$ optimization procedures from different starting points θ_{start} and chose the set of starting points based on a *maximin* Latin hypercube design.

For QK, the BFGS algorithm is used to optimize (2.28). For VB, two derivative-based optimizers are used alternately for the E- and M-steps. Since each step may lead the algorithm toward a local minimum, we chose to apply the multi-start strategy in the entire EM procedure.

Optimization of the cross-validation metric (2.39) is done under the black box framework, since no structural, derivative or even regularity information is available. Hence, all optimizations are carried out using the branch-and-bound algorithm named Simultaneous Optimistic Optimization (SOO) Munos [2011]. SOO is a global optimizer, hence robust to local minima.

We used Tange [2018] to parallelize the computations.

Oracle metamodels

Each method presented in this paper is a trade-off between power and the difficulty of finding good hyperparameters. A good method should be powerful (i.e. provide flexible fits) but easy to tune. In order to assess the strengths and weaknesses of the hyperparameter tuning methods in addition to standard metamodels, we provide what

we call the *oracle metamodels* for each problem. Instead of using the cross-validation or likelihood metric, the oracle tunings are directly based on the evaluation metric

$$E_{L^2}(\widehat{q}) = \sum_{i=1}^{n_{\text{test}}} \left(\widehat{q}^{\Theta}(x_i) - q(x_i)\right)^2, \tag{2.40}$$

where n_{test} is the size of the test set. In a sense, they provide a upper bound on the performance of each method. This allows us to show which metamodels have the potential to tackle the problems and which are intrinsically too rigid or make poor use of information. In addition, this allows us to directly assess the quality of the validation procedure.

2.8 Benchmark design and experimental setting

Many factors can affect the efficiency of methods to estimate the right quantile. For our benchmark system, we considered five models or test cases to evaluate the performance of the six metamodels. We decided to focus primarily on the dimensionality of the problem, the number of training points available, the signal-to-noise ratio defined as

$$\text{SNR} = \frac{\mathbb{V}_X\big(\mathbb{E}(Y|X)\big)}{\mathbb{E}_X\big(\mathbb{V}(Y|X)\big)},$$

and the pdf value at the targeted quantile for test cases in which the distribution shape and the distribution spread (i.e. level of heteroscedasticity) can vary significantly. Our two objectives were to:

1. discover if there is a single best method for all factors variations considered or specific choices depending on the configuration at hand, and

2. assess the performance of the quantile regression, and in particular, the configurations for which the current state-of-the-art is insufficient.

A full 3D factorial experimental design was used to analyze the efficiency of the metamodels, the 3 factors being the test case (4 test cases), the number of training points (4 levels) and the quantile order (0.1, 0.5 and 0.9). We used part of this complete design to focus our analyses on the characteristics of the test cases (dimension, pdf shape and heteroscedasticity).

2.8.1 Test cases and numerical experiments

Test case 1 is a $1D$ toy problem on $[-1, 1]$ defined as

$$Y_x = 5\sin(8x) + (0.2 + 3x^3)\xi,$$

with $\xi = \eta \mathbb{1}_{\eta \leqslant 0} + 7\eta \mathbb{1}_{\eta > 0}$ where $\eta \sim \mathcal{N}(0, 1)$.

The signal-to-noise ratio is SNR ≈ 0.5, it is consider as small. The pdf value $\widetilde{f}(x, q_\tau)$ varies substantially according to x for all q_τ. Indeed, on the interval $[-0.5, 0.3]$, for all values of q_τ, the pdf is very large (almost equals to $+\infty$) because the variance of the distribution on this interval is very small. In contrast, for $\tau = 0.9$ (resp. $\tau = 0.1$) the pdf is very small in the interval $[0.6, 1]$ (resp. $[-1, -0.6]$). In $[0.6, 1]$ the pdf of the 0.9-quantile is equal to the pdf of the 0.9-quantile of a normal distribution of variance $49(0.2 + 3x^3)^2$ that is, for example, approximately equal to 0.01 for $x = 0.9$. Because of this important variations, we consider the values of the pdf for all quantiles of interest as variable.

Test case 2 is a $2D$ toy problem on $[-5, 5] \times [-3, 3]$ based on the Griewank function Dixon and Szego [1978], defined as

$$Y_x = G(x)\xi,$$

with

$$G(x) = \left[\sum_{i=1}^{2} \frac{x_i^2}{4000} - \prod_{i=1}^{2} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1\right]$$

and $\xi = \eta \mathbb{1}_{\eta \leqslant 0} + 5\eta \mathbb{1}_{\eta > 0}$ where $\eta \sim \mathcal{N}(0, 1)$.

The signal-to-noise ratio is SNR $= 0$, it is consider as small. The pdf value $\widetilde{f}(x, q_\tau)$ varies substantially according to q_τ. Indeed, for $\tau = 0.1$ (resp. $\tau = 0.9$) the pdf is small (resp. very small), more precisely the pdf of the 0.1-quantile (resp. 0.9-quantile) is equivalent to the pdf of the 0.1-quantile (resp. 0.9-quantile) of a normal distribution of variance G^2 (resp. $25G^2$) with G that takes values in $[0, 2]$. That implies $\widetilde{f}(x, q_\tau)$ varies with respect to x as well. Note that close to $x = (0, 0)$ the pdf is very large because of the very small variance of the associated distribution. The pdf at the 0.5-quantile is equal to the pdf at the median of a normal distribution of variance G^2, that we consider as large.

Test case 3 is a $1D$ toy problem based on the Michalewicz function Dixon and Szego [1978] on $[0, 4]$, defined as

$$Y_x = -2\sin(x)\sin^{30}\left(\frac{x^2}{\pi}\right) - \frac{0.1\cos(\pi x/10)^3}{\left|-\sin(x)\sin^{30}\left(\frac{x^2}{\pi}\right) + 2\right|}\xi^2,$$

with $\xi = 3\eta \mathbb{1}_{\eta \leqslant 0} + 6\eta \mathbb{1}_{\eta > 0}$ where $\eta \sim \mathcal{N}(0, 1)$.

The signal-to-noise ratio is SNR ≈ 0.04, we consider it as small. The pdf value $\widetilde{f}(x, q_\tau)$ varies substantially according to q_τ and x. The conditional distribution of this problem is not a classical one but it is close to the distribution of $-\mathcal{X}^2$ with one degree of freedom. It implies at x fixed, the pdf value increases with τ. For the 0.1-quantile (resp. 0.9-quantile) the mean of the pdf in the interval $[0, 2.5]$ is approximately 0.05

(resp. 9.1). According to x the pdf varies as well. For $x \in [3.5, 4]$ the variance of the conditional distribution is very small thus the pdf near the quantiles of interest is very large. The value of the pdf at the 0.5-quantile is between the value of the pdf at the 0.1 and 0.9 quantile. Thus we consider the pdf at $q_{0.1}$ as globally small and at $q_{0.5}$ and at $q_{0.9}$ as globally large

Test case 4 is a $9D$ toy problem based on the Ackley function Ackley [2012] on $[-1, -0.7] \times [0, 1] \times [-0.7, -0.3] \times [0.5, 1] \times [-1, -0.5] \times [-3, -2.6] \times [-0.1, 0] \times [0, 0.1] \times [0, 0.8]$, defined as a function

$$Y_x = 30 \times A(x) + R(x) \times \xi$$

with

$$A(x) = a \exp\left(- b \sqrt{\frac{1}{9} \sum_{i=1}^{9} x_i^2} \right) - \exp\left(\frac{1}{9} \sum_{i=1}^{9} \cos(c x_i) \right) + a + \exp(1), \tag{2.41}$$

and

$$R(x) = 3A(x_2, x_3, \cdots, x_9, x_1), \tag{2.42}$$

with $a = 10$, $b = 2 \times 10^{-4}$, $c = 0.9\pi$ and ξ follows a log-normal distribution of parameters $(0, 1)$.

The signal-to-noise ratio is SNR ≈ 2.3, it is consider as large. The pdf value $\widetilde{f}(x, q_\tau)$ varies substantially according to q_τ and x. Indeed, for $\tau = 0.1$ (resp. $\tau = 0.9$) the pdf is large (resp. very small). At the conditional distribution of this test case is a log-normal distribution of parameters $\big(\log R(x), 1\big)$. The expectations of the pdf value at different values of τ are $\mathbb{E}(\widetilde{f}(., q_{0.1})) = 0.1$, $\mathbb{E}(\widetilde{f}(., q_{0.5})) = 0.04$ and $\mathbb{E}(\widetilde{f}(., q_{0.9})) = 0.005$. Thus the pdf at $q_{0.1}$ is consider as large, the pdf at $q_{0.5}$ and $q_{0.9}$ are consider as small and very small.

To provide a better intuition about this problem, we plotted what we call the marginals. For all dimensions except the j-th, the values of the input are fixed to $x_{-j} \in \mathbb{R}^8$ and the j-th dimension varies. Figure 2.6 represents the evolution of the quantiles w.r.t. the j-th dimension for two different x_{-j} and for $j = 1, \cdots, 9$. In particular it shows that the difference between the 0.1 and 0.9-quantile depends significantly on x.

Note that on those four toy problems, the random term ξ is defined such that the resulting distribution of Y would be strongly asymmetric. As ξ is also multiplied by a factor that depends on x, the distribution of Y is also heteroscedastic. The first three toy problems are represented in Figure 2.5 and some illustrations of the fourth test case are available Figure 2.6.

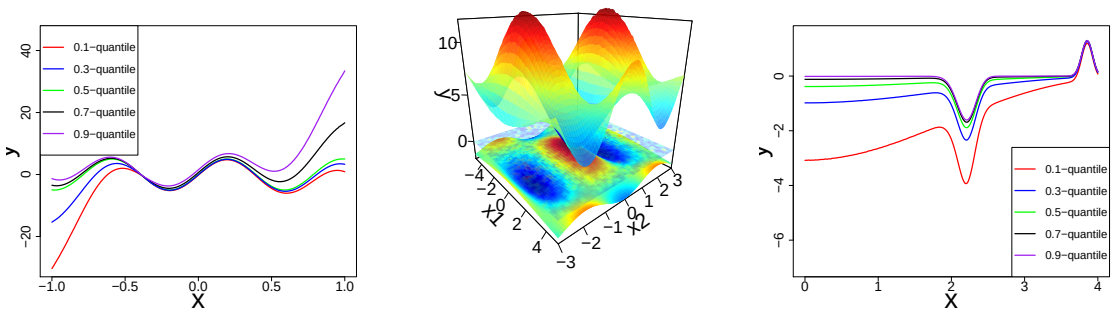


Figure 2.5: Illustration of the test cases (left: test case 1, center: test case 2, right: test case 3). For test case 1 and 3 the $0.1, 0.3, 0.5, 0.7, 0.9$ -quantiles are represented. For test case 2, only the $0.1, 0.5, 0.9$-quantiles are represented.

Test case 5 is based on the agronomical model SUNFLO, a process-based model which was developed to simulate sunflower grain yield (in tons per hectare) as a function of climatic time series, environment (soil and climate), management practices and genetic diversity. The full description of the model is available in Casadebaig et al. [2011]. In the regression model we consider $\mathcal{X}$ corresponding to nine macroscopic traits that characterize the sunflower variety. Although SUNFLO is a deterministic model, for each simulation the climatic time series are randomly chosen within a database containing 190 years of weather records, which makes the output stochastic (see also Picheny et al. [2017] for more details).

The signal-to-noise ratio is SNR ≈ 0.1. The pdf value $\widetilde{f}(x, q_\tau)$ varies substantially according to q_τ, more precisely $\mathbb{E}(\widetilde{f}(., q_{0.1})) = 0.17$, $\mathbb{E}(\widetilde{f}(., q_{0.5})) = 0.15$ and $\mathbb{E}(\widetilde{f}(., q_{0.9})) = 0.05$. Thus the pdf at $q_{0.1}$ and at $q_{0.5}$ are consider as large, and the pdf at $q_{0.9}$ is consider as small.

In addition, the shape of the distribution varies significantly with x.

Numerical experiments. On all problems, we consider four sample sizes. Those sizes depend on the dimension and are empirically chosen so that the smallest size corresponds to the minimal information required by the metamodels to work and the largest size is chosen keeping in mind the potentially high cost of simulators. Besides, our focus is on computer experiments, where data sizes rarely exceed thousands of points. For the $1D$ problems, the points are generated on a uniform grid. For the $2D$ and $9D$ problems, the observations are taken on a Latin hypercube design optimized for a *maximin* criterion to ensure space-filling Fang et al. [2005]. The same samples are used by all methods except QK, as it requires repetitions. For QK, the number of distinct points and number of repetitions depends on the budget. The different sample sizes are reported in Table 2.4. Finally, for each sample size and problem, 10 samples are drawn in order to assess

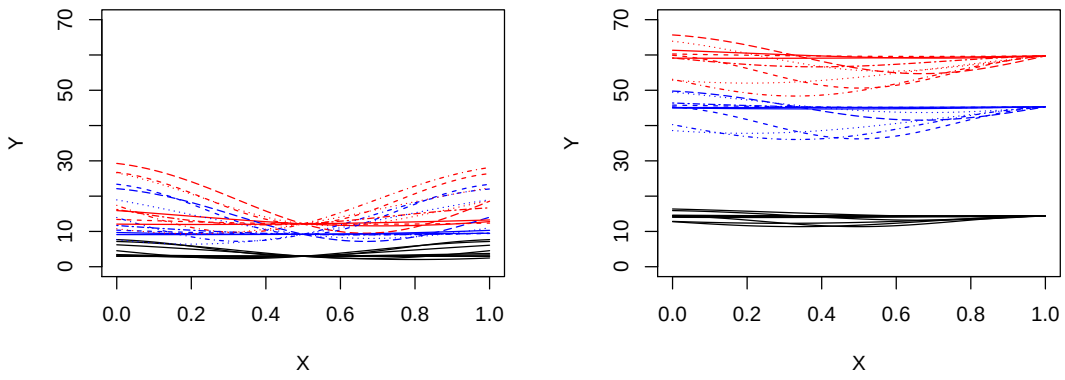


Figure 2.6: Illustration of some marginals of Test case 5. The conditional quantiles of order 0.9 (resp. 0.1) are represented in red (resp. in blue). The black curves represent the difference between the 0.1-conditional quantiles and the 0.9-conditional quantiles so that to measure the level of heteroscedasticity. To the right the noise level is low *i.e* between 0 and 10 while to the left the noise level is higher *i.e* between 10 and 20.

robustness with respect to the data.

2.8.2 Structuration between the questions and the numerical setting

Factors. Three factors are explicit in our benchmark system: the number of training points, the problem dimension and the quantile level. The other factors depend on the characteristics of the problem concerned: shape variation, pdf value at the quantile, level of heteroscedasticity, signal-to-noise ratio. For all four test cases, we consider three quantile levels: 0.1, 0.5 and 0.9. Note that due to the asymmetry of the problems, learning for the 0.1 and 0.9 quantiles is not equivalent in terms of difficulty. Indeed, when the response is heteroscedastic (a variance/spread depending on x) and/or when

		Test case 1	Test case 2	Test case 3	Test case 4	Sunflo
Dimension		1	2	1	9	9
Heteroscedasticity		very strong	very strong	very strong	strong	weak
Shape variation		very strong	weak	weak	weak	strong
pdf value near the τ-quantile	$\tau = 0.1$	variable	small	globally very small	large	large
	$\tau = 0.5$	variable	large	small	small	large
	$\tau = 0.9$	variable	very small	very large	very small	small

Table 2.3: Summary of the characteristic of the problems.

Dimension	Data size (no repetitions)				Data size (with repetitions)			
1	40	80	160	320	5 (8)	10 (8)	10 (16)	16 (20)
2	100	200	400	800	10 (10)	20 (10)	25 (16)	40 (20)
9	250	500	1000	2000	25 (10)	50 (10)	100 (10)	100 (20)

Table 2.4: Data sizes for the different problems. The number in parentheses are the number of repetitions for QK.

the shape of $\mathbb{P}_x$ varies in x, the pdf $\widetilde{f}(x, q_\tau)$ may vary in x as well. Intuitively, quantiles with large pdf values are easier to learn, as the data points may be closer to them. Figure 2.7 illustrates this effect. Table 2.3 summarizes the characteristics of our design concerning the number of training points with respect to the dimension of the test case. To make our results easier to analyze, we divided the problems into groups that allow us to focus on subsets of factors.

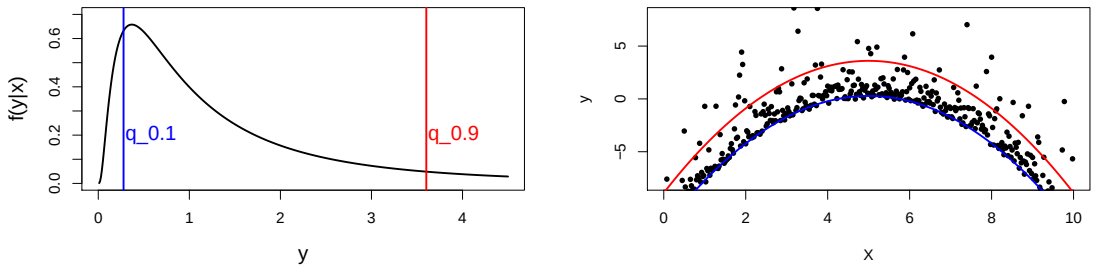


Figure 2.7: Left: log-normal density fonction with $\mu = 0$ and $\sigma = 1$. Right: a sample generated by the function $f(x) = \xi - (x-5)^2/2$, with ξ following the density represented on the left. The 0.9- (resp. the 0.1-) quantile is represented in red (resp. in blue). One can notice that more information is available in areas with large pdf (i.e. for the 0.1-quantile) than areas with small pdf.

Focus 1: is there a universal winner? To provide a universal ranking of the methods, we use all test cases, training points and quantile levels. As highlighted in Table 2.3, we created a set of different problems representative of a large number of characteristics that could be met dealing with any quantile regression problem. Note that our benchmark system is slightly biased towards small-dimensional problems, since only three-fifths of the cases have a dimension higher than two.

Focus 2: what behavior with respect to the dimension, the number of training points, signal-to-noise ratio and pdf value? To analyze the effects of these factors

on the performance of the methods, we combine toy problems 1, 2, 3, 4 and the SUNFLO model. Note that once the pdf value is taking as a explanation variable, toy problem 1 is excluded from the group because the pdf value near all the studied quantiles cannot be classified as large or small.

2.8.3 Performance evaluation and comparison metrics

Assessing the performance of quantile regression is not an easy task when only limited data are available. Here, since we are considering toy problems (exept for SUNFLO, in that case the true quantile values are taken as the quantiles of the 190 years of weather records), the true quantile values can be approximated with precision, so we can evaluate the L^2 error for each emulator. The value of the criterion is privided by (2.40).

We chose $n_{\text{test}} = 250$ for the $1D$ problems and $n_{\text{test}} = 4000$ for the others. Now, since the problems vary in difficulty and in their response range (Figure 2.5), $E_{L^2}(\widehat{q})$ cannot be aggregated directly over several problems or configurations. To do so, we normalize this error by the error obtained by a constant model (the constant being taken as the quantile of the sample):

$$E_{cq}(\widehat{q}) = \sqrt{\frac{E_{L2}(\widehat{q})}{E_{L2}(\text{CQ})}} \times 100, \tag{2.43}$$

where CQ stands for constant quantile.

As an alternative criterion, we consider the ranks of the metamodels based on their L^2 error. Although ranks do not provide information regarding the range of errors, they are insensitive to scaling issues, which makes aggregation between configurations more sensible. They allow us to assess whereas any method consistently outperforms others, regardless of overall performance.

2.9 Results

2.9.1 Focus 1: overall performance and ranks

First, we consider the overall performance and ranks, integrated over all runs. We have considered 5 test cases, for each test case we have generated 4 sizes of training sets, for each test case and sample size 10 samples are drawn and for each of this occurrences we have estimated 3 different conditional quantiles. Thus we have $5 \times 4 \times 10 \times 3 = 600$ experiments. They are shown as boxplots in Figure 2.8. Based on these ranks (Figure 2.8, left), VB appears to be the best solution since it is ranked either 1st or 2nd in 50% of the problems. RK is in second position, its median is the same as VB but it is generally ranked between 2nd an 3rd. In addition RK seems slightly less risky than VB in the sens that it is almost never rank 5th or 6th. KN is the worst since its median rank is equal to five. However, all boxplots range between 1 and 6, indicating that no method is outperformed by another on all problems. This finding is reinforced by the performance boxplots (Figure 2.8, right), where all median performances are similar (VB and RK being the best and QK the worst which means QK may be very bad sometimes), and

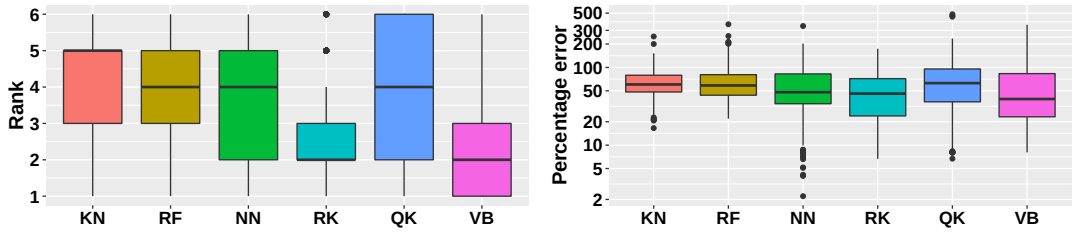


Figure 2.8: Boxplots of ranks and E_{cq} error over the entire benchmark. Note that for clarity, the right boxplots do not contain the error of the median for the toy problem 2.

the variance is very large. Indeed, the errors range from 2% for NN (of the error achieved by a constant metamodel) to 500% for QK, all methods experiencing cases with more than 100% error (i.e. situations where they are worse than the constant metamodel).

2.9.2 Focus 2: dimension, number of training points and pdf value

Performance according to the constant quantile

In this section, we analyze the performance of the methods with respect to the pdf value and the number of points.

Sample size: Figure 2.9 shows the performances of the methods grouped according to the size of the sample. As expected, the performances increase with the size of the sample. For size 1 ($n \approx 50D$), the distribution of E_{CQ} of all the metamodels is almost centered around 100%, implying that these correspond to limit cases for quantile regression since the metamodels do not outperform the constant metamodel (although in some cases the error is as small as 40%). For size 4 ($n \approx 300D$), the median performance is roughly 50% (twice as accurate as the constant metamodel). BV, RK and especially NN experience situations with very accurate models. However, all the methods also experience bad performances (error greater than 100%) in the large sample regime. Unfortunately, from Figure 2.9 we can conclude that no method is sufficiently robust in all cases.

Signal-to-noise ratio. Figure 2.11 groups performance with respect to the signal-to-noise ratio. According to the figure the performance depends to a great extent on the signal-to-noise ratio. Dealing with a high signal and considering the performance, there is a clear difference between the statistical methods (KN, RF) and the four others (RK, NN, VB, QK) while this difference is not visible in the small signal setting. The impact of the signal-to-noise ratio is strong on the performance. Dealing with a high signal, the median of the performance for NN, RK, QK, VB is close to 20% with a small variance while if the signal is small, the error is larger (the median is above 50% for all methods) and the variance is larger.

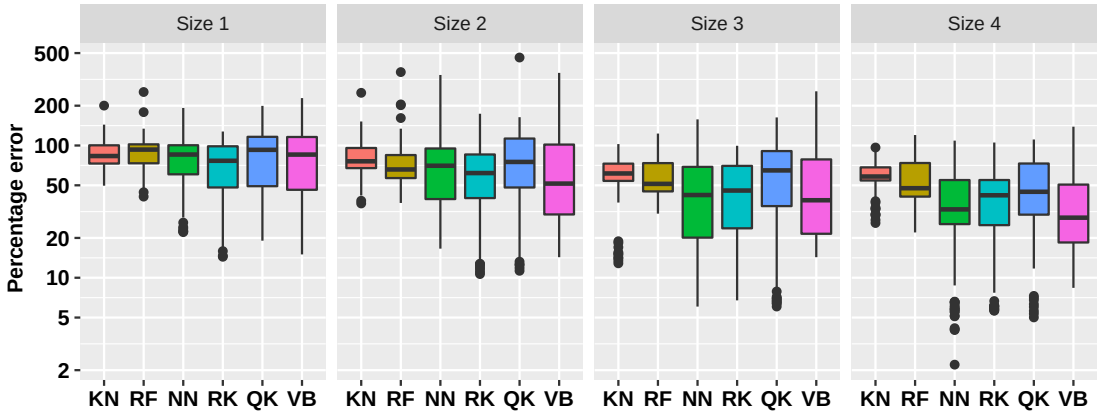


Figure 2.9: Error according to the sample size

In the following we focus our interest only on cases with small signal-to-noise ratio. Indeed as the performance of RK, NN, QK and VB are close to each others, we think we do not have enough experiments to extract patterns.

Pdf value with small signal-to-noise ratio. Figure 2.11 groups performance with respect to sample size (either small, i.e. level 1 and 2 or large, i.e. level 3 and 4) and pdf value (according to Table 2.3). According to the Figure 2.10, the performance depends to a great extent on the pdf value in the neighborhood of the targeted quantile. With a small n, the pdf value has no significant impact on the median of the performance (except for VB) but it does have an impact on the lower bound of the error. More precisely, the median of the error does not depend on the pdf value in the case of a small pdf but sometimes the metamodel errors are sensibly smaller when the pdf is large. With large samples, both the median and the lower bound of the error depend on the pdf value. Metamodels may be very good when the pdf is large, for example 20 times better than the constant metamodel for NN whereas the error appears to have a lower bound when the pdf is small even with large n. In addition, for a problem with a small n and a large pdf, the performance is similar to the performance for problems with a large n and a small pdf (Figure 2.11, see the two columns in the center).

Rank in the context of a low signal-to-noise ratio

Pdf value. Figure 2.12 shows clearly that when the pdf is large, VB is the best model while when the pdf is small (and the problem is heteroscedastic), VB is less good than RN, RK and KN. This observation is supported by Figure 2.11 which reveals a strong contrast between the performance of the VB method. QK is poor in both cases, whereas RK performs comparatively better with small pdf.

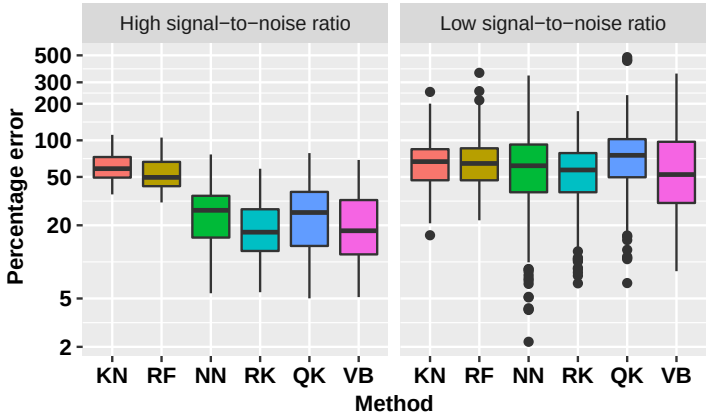


Figure 2.10: Error according to the signal-to-noise ratio.

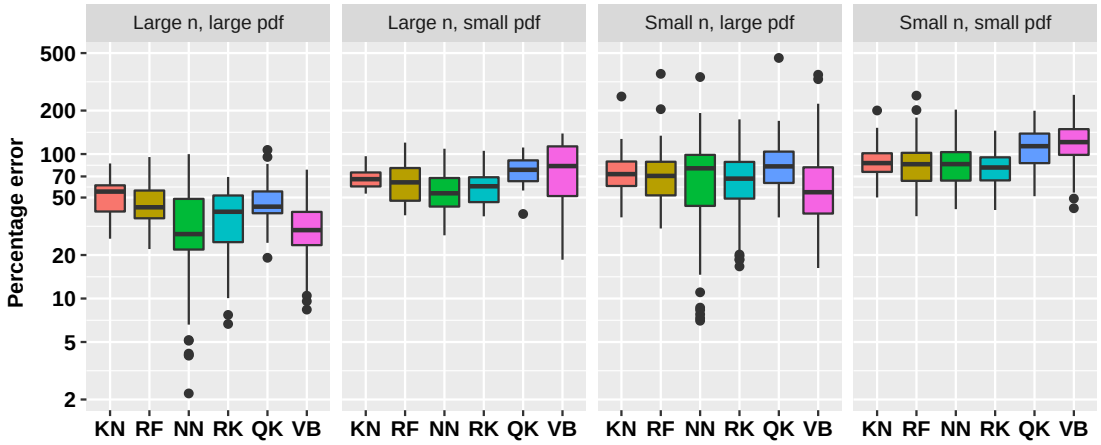


Figure 2.11: Error according to the size of the training set and the pdf value.

Sample size. Figure 2.13 shows that the number of points has a major impact on the ranking of some methods. The ranking of QK and VB is relatively insensitive to the size of the sample. The other methods are less distinguishable when the sample size is small than when it is large. With a small sample KN, RF, NN and RK are comparable, whereas when the sample size increases, NN and RK clearly outperform KN and RF. For the largest size, NN is slightly better than VB.

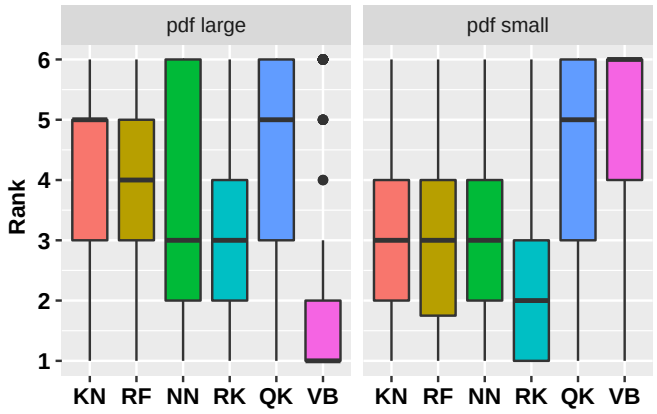


Figure 2.12: Rank according to the pdf value

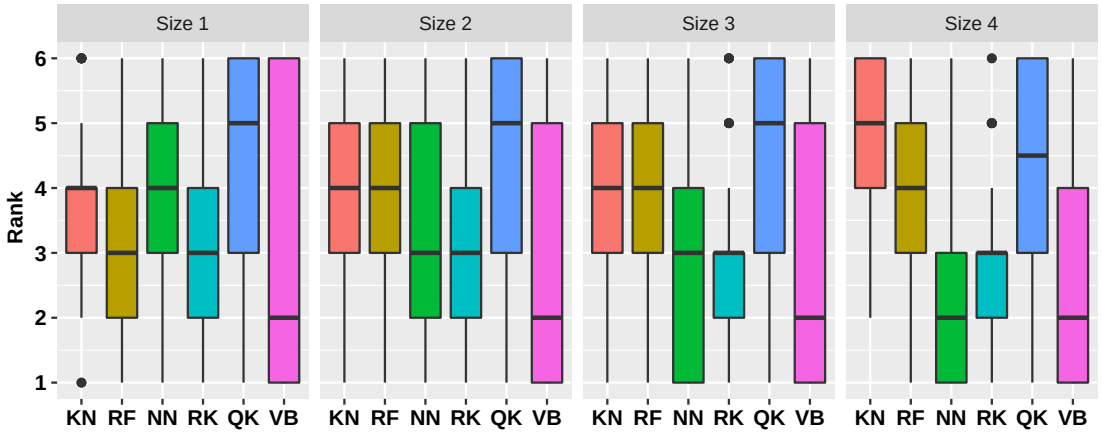


Figure 2.13: Rank according to the size of the sample

Dimension. Figure 2.14 groups performance based on dimension. The first contrast is the permutation between RK and NN. With a small dimension, RK is better than NN but the relative performance of NN increases w.r.t. the dimension. With small dimensions, RF and KN are comparable, but with high dimensions, RF outperforms KN.

High dimension, small pdf. Figure 2.15 shows an extreme case in which the pdf is small but the dimension is high. With a small n, the best method is clearly RF followed by RK and KN. VB and QK are not well ranked. With a larger n, as mentioned above, NN and VB are better but with large variance, while RF, RK and KN rank less well.

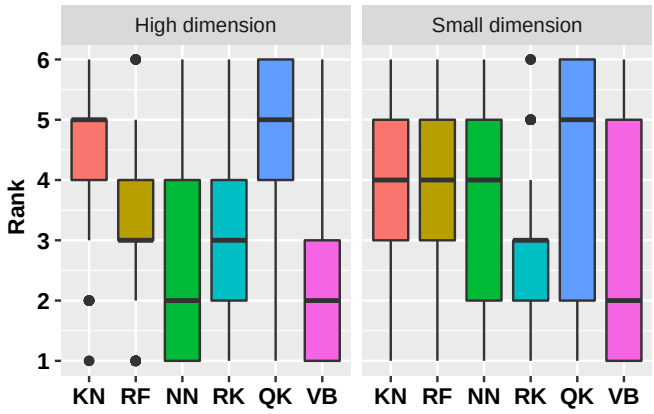


Figure 2.14: Rank according to the dimension

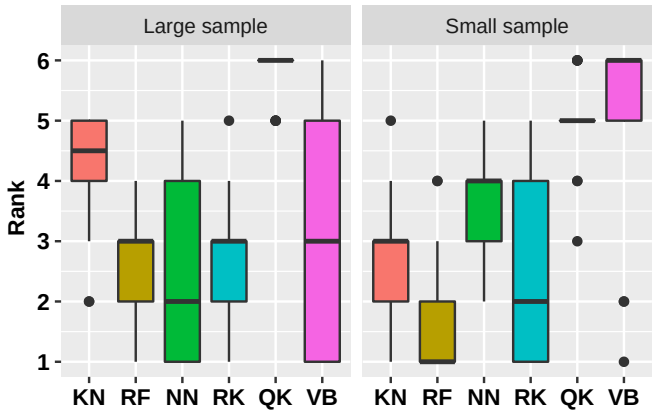


Figure 2.15: Rank associated to the case where little information is available, *i.e* high dimension and small pdf

2.10 Extensions and open questions

2.10.1 Effect of hyperparameter tuning

In the following we define

$$\Delta E(\widehat{q}_\tau^{\Theta}) = E_{cq}(\widehat{q}_\tau^{\Theta}) - E_{cq}(\widehat{q}_\tau^{\Theta^*}),$$

the performance gap between the regular metamodel and its oracle performance (the loss in performance between actual hyperparameter tuning and oracle tuning). Figure 2.16 gives the average values of ΔE aggregated respectively over all problems and only aggregated over the problems with a large pdf, and considering the effect of dimension and sample size. In high dimension, the easiest methods to tune are KN, NN, and RF. In our study, KN and RF have a single hyperparameter to tune regardless of the

dimension, and NN has two. This is clearly an advantage in terms of robustness in high dimension. The other methods are kernel-based and require the tuning of at least $D+1$ hyperparameters. This consistently affects RK and QK, but affects VB only in the case of small pdf, while it is the most stable method in the other cases.

With a small dimension, all the methods have roughly the same number of hyperparameters. The most noticeable change compared to the case of a high dimension is the good performance of RK, while NN becomes comparatively the most difficult method to train.

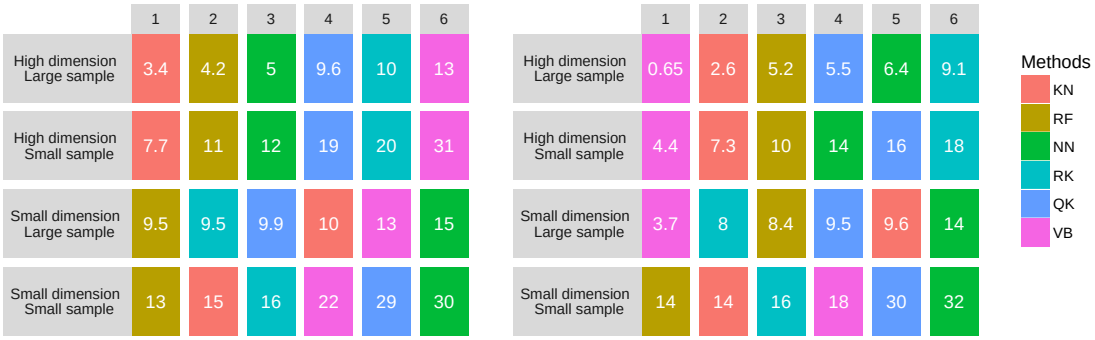


Figure 2.16: Average ΔE aggregated over all problems (left) and over the problems with a large pdf only (right), arranged in increasing order. For each method the rank is provide on the top of each figure.

2.10.2 On the methods' behavior

Statistical order methods. As presented on Figure 2.10 this methods are not relevant on high signal-to-noise regime. A possible explanation is the difficulty to fit smooth variations with high amplitudes with a piecewise constant model.

Concerning the low signal-to-noise regime, it is clear from Figure 2.14 that KN performs poorly in a high dimension. This may be due to the irrelevance of the Euclidean distance when there is a significant increase in dimension. RF clearly outperforms KN in this situation, as it is able to produce better neighborhoods than the Euclidean distance. Overall, (compared with the other methods), RF performance increases with dimension. This may be due to the fact that it has fewer hyperparameters to tune.

Functional methods. As presented on Figure 2.10 this methods are relevant on high signal-to-noise regime.

Concerning the low signal-to-noise regime, Figure 2.13 shows that NN works poorly in a small sample setting, but it is one of the best methods when the sample is large. This result reflects the high flexibility of NN. Too much flexibility leads to overfitting when the sample is small. In contrast, when the number of points is large, NNs are able to fit

the data very well (e.g. Figure 2.9, Size 4). According to Figures 2.9 and 2.13, RK is a robust method. Its robustness in both small and large data settings can be attributed in part to the selected kernel. If the selected kernel is sufficiently smooth (here continuous and derivable), the resulting metamodel cannot produce instable results. However, it seems (Figure 2.9, Size 3 and 4) that this lack of flexibility may affect the performance with an increase in the size of the data set. In this case, more flexible methods like NN may outperform RK. The contrast between RK and NN shown in Figure 2.14 can be explained by the level of difficulty associated with each method involved in finding good hyperparameters (as explained above).

Bayesian models. As presented on Figure 2.10 this methods are relevant on high signal-to-noise regime.

Dealing with low signal-to-noise ratio the QK method under-performs comparing to others. One possible explanation is the erroneous assumption in Equation (2.29) that the noise is centered, which is more critical for extreme quantiles. Another possible explanation lies in the small number of replica. The local inference (that uses statistical orders) is biased and in a low signal-to-noise regime it has high variance. In addition, the increasingly bad performance of QK with an increase in dimension (Figure 2.9) is likely a consequence of the fact that empty areas become larger in high dimensions.

VB is one of the best methods presented in our paper. Figures 2.12 and 2.16 show that VB is also the most dependent on the pdf value. When the pdf is large, it is the best method whereas when the pdf is small (and the shape of the distribution and/or the variance depend on the input), it may be the worst. The explanation lies in the philosophy of the model. In the case of NN and RK, the model complexity (i.e. smoothness) is almost entirely related to the regularity of parameter λ that is selected by cross-validation. Hence, the model cannot excessively overfit and cannot perform very poorly. With Bayesian methods, the regularization is included in the model hypothesis: in our setting, the quantile is assumed to be a Gaussian process with covariance function $k_\theta(.,.)$, so θ performs the regularization. We observed that if the local quantity of information (roughly the product of the number of points times the pdf value in the neighborhood of the quantile) is too small comparing to the information needed to fit well the quantile, the metamodel tends to interpolate the available data. When sufficient information is available, the optimization of the marginal likelihood provides θ values that allow a good trade-off between flexibility and smoothness. This is likely the reason why VB is easily beaten by RN and RK when the pdf is small.

2.10.3 Varying shape and heteroscedasticity.

If the shape of $\mathbb{P}_x(Y)$ or the variance of Y_x (heteroscedasticity) vary w.r.t. x, then $\widetilde{f}(x, q_\tau)$ may vary in x. Figure 2.17 illustrates the ability of RF, RK and VB to estimate quantiles of a distribution with a strongly varying shape. In this problem, as depicted

in Figure 2.17 (top row), the quantiles are not perfectly estimated but the metamodels provide good indications about the shape of the true distribution. However, as can be seen in Figure 2.17 (bottom row), the methods can present strong instabilities. Here, for a sample virtually indistinguishable from the one leading to accurate estimates, the median estimates largely overestimate the true values for large x values. Such instabilities can be partly imputed to the difficulty of the task. However, this is also because no method is actually designed to deal with strongly varying pdf, as we explain below.

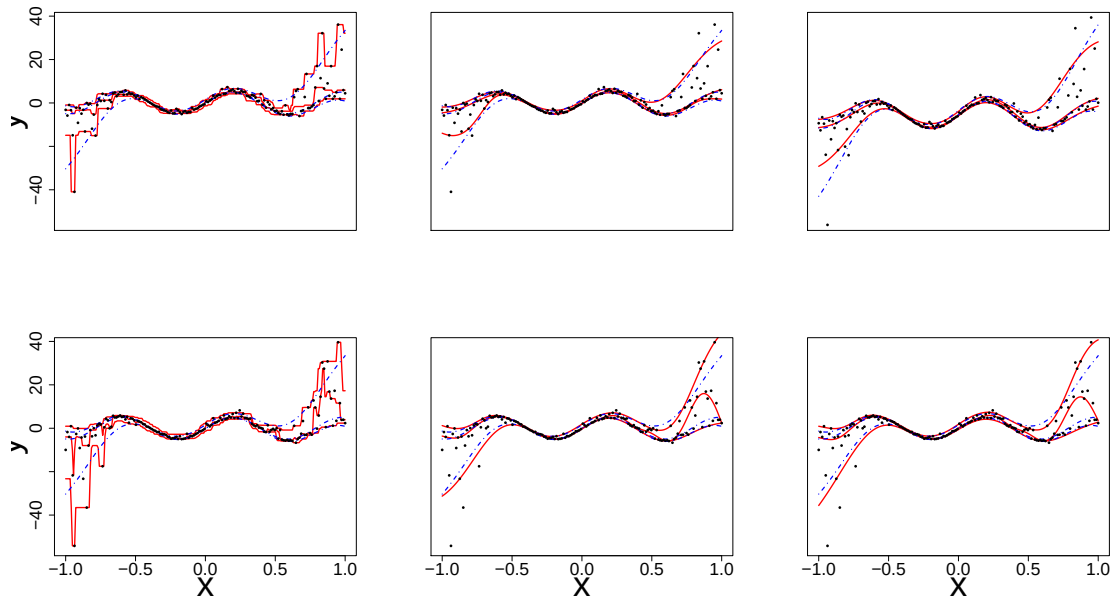


Figure 2.17: Quantiles estimates using RF (left), RK (middle), VB (right) for two 160-point samples (top and bottom rows, resp.) of the toy problem 1. Dots: observations; plain red lines: metamodels for the $0.1, 0.5, 0.9$ quantile estimates; dotted blue lines: actual quantiles.

An ideal method would be almost interpolant for a very large pdf but only loosely fit the data when the pdf is small. Indeed, if the pdf is very large then the output is almost deterministic, thus the metamodel should be as closer as possible to the data. In the small pdf case, a point does not provide a lot of information. Information should be extracted from a group of points, that means the metamodel must not interpolate the data. However, most of the methods presented here rely on a single hyperparameter to tune the trade-off between data fitting and generalization: the number of neighbors for KN, the maximum size of the leafs for RF and the penalization factor for NN and RK. As a result, the selected hyperparameters are the ones that are best on average. Theoretically, this is not the case for the Bayesian approaches: QK accounts for it *via* the error variance σ_i^2 computed by bootstrap, and the weights w_i (Eq. 2.34) allow VB to attribute different "confidence levels" to the observations. However in practice, both

methods fail to tune the values accurately, as we illustrate below. Figure 2.18 shows the three quantiles of toy problem 3 and their corresponding RF, RK and VB estimates. For $\tau = 0.1$ in particular, the pdf ranges from very small (x close to 0) to very large (x close to 4). Here, RF and RK use a trade-off that globally captures the trend of the quantile, but cannot capture the small hill in the case of large x. Inversely, VB perfectly fits this region but dramatic overfitting occurs on the rest of the domain.

Finally, Figure 2.19 illustrates that this is not an issue of hyperparameter tuning. For each method, we show the oracle estimate, a tuning that tends to underfit and another that tends to overfit. One can see that no tuning is entirely satisfactory, since capturing the region with high pdf leads to overfitting on the rest of the domain and vice-versa.

We believe that further research is necessary to obtain estimators that intrinsically account for strong heteroscedasticity and varying shape. One possible direction is the use of stacking, in the spirit of Sill et al. [2009]. Under the stacking framework the final estimator could be

$$\widehat{q}(x) = \sum_{i=1}^{N} g_i(x)\widehat{q}_{\theta_i}(x),$$

where $\{\widehat{q}_{\theta_i}\}_{1\leqslant i\leqslant N}$ is a set of metamodel and $\{g_i(x)\}_{1\leqslant i\leqslant N}$ is a set of weight functions. Choosing $\{\widehat{q}_{\theta_i}\}_{1\leqslant i\leqslant N}$ such that they correspond to different pdf values might provide more flexible estimates.

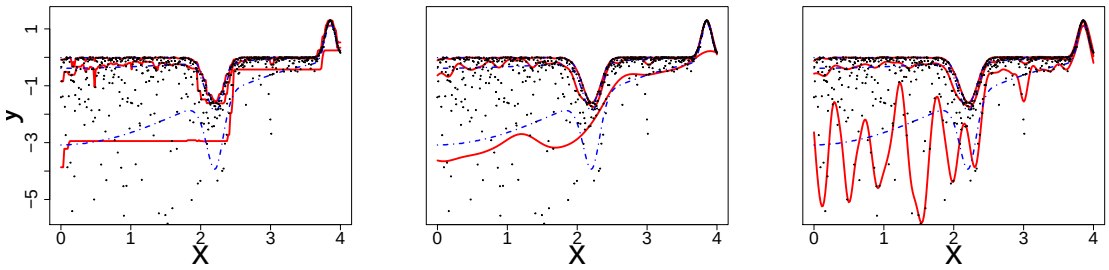


Figure 2.18: Quantiles estimates using RF (left), RK (middle), VB (right) for a 640-point sample of the toy problem 3. Dots: observations; plain red lines: metamodels for the $0.1, 0.5, 0.9$ quantile estimates; dotted blue lines: actual quantiles.

2.10.4 On the non-crossing of the quantile functions

While the quantile functions (for different quantile levels) may obviously never cross, unfortunately, their estimators may not always satisfy this property. This is a well-known issue against which none of the methods presented here is immune.

The neighborhood approaches first estimate the CDF, then extract the quantiles. If the hyperparameters are the same for all quantiles, crossing is impossible. However in our setting, different neighborhood sizes were used for different quantiles.

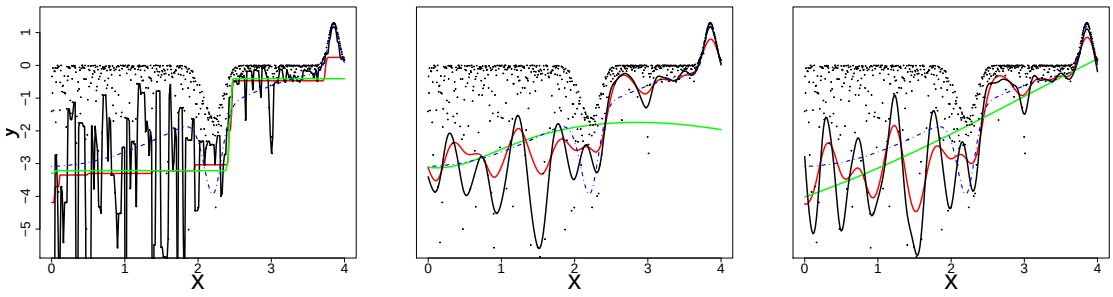


Figure 2.19: Metamodel responses for toy problem 3 and $\tau = 0.1$ with 640 training sample (left: RF, center: RK, right: VB) for different values of hyperparameters. The true 0.1-quantile is presented in dotted blue lines. In green and black two extreme metamodels associated to two extreme hyperparameter values, while in red the oracle metamodels are represented.

With functional analysis approaches, crossing may happen even if each quantile is built with the same hyperparameters. In the literature, authors have produced methods to address this issue. It could be reduced by the introduction of additional constraints in the model Takeuchi et al. [2006] or by the construction of a new model that intrinsically produces non-crossing curves Sangnier et al. [2016]. However in both cases, the dimension of the optimization problem then increases significantly.

Finally the stochastic process approaches estimate each quantile in independent Gaussian processes, so crossing may occur. While the number of training points is larger than what we consider here (10^4 repetitions for each input point), Browne et al. [2016] use GP and takes into account all the quantile orders at the same time and thus ensures non-crossing.

Another approach available for all the methods presented here is related to the rearrangement of curves or isotonic regression Belloni et al. [2011], Abrevaya [2005]. The idea is to perform many quantile regressions with a large number of different values of τ or a large set of bootstrapped versions of $\mathcal{D}_n$ and then to rearrange the curves, thereby obtaining the whole distribution and then extracting the quantiles that by definition do not cross.

In theory, adding non-crossing constraints and predicting several quantiles simultaneously could improve the quality of the estimates (in particular as it might add some robustness). However, in practice, it also makes the model more rigid (i.e. a single regularization hyperparameter for all quantiles), and preliminary experiments have shown no gain in accuracy compared to independent predictions, despite a considerably higher computational cost. Hence, multi-quantile predictors were not considered in our study.

2.10.5 Assessment of prediction accuracy

To assess the accuracy of the results and the confidence that we can have in the estimation it could be useful to provide confidence intervals for the predictor. From this point of view the methods are not equal. With Bayesian approaches, theoretical confidence intervals are provided with the models. More precisely as the output model is Gaussian and as it returns the mean and the variance, confidence intervals can be created. For example the 0.9-confidence interval is provided by

$$\mathrm{CI}(x) = \widehat{q}_\tau(x) \pm 1.96\sqrt{\mathbb{V}_q(x)}.$$

However as presented on Figure 2.20, while the confidence intervals obtained from QK seem useful, the VB model is clearly overconfident.

The statistical order methods consider that inside each neighborhood the samples are i.i.d. Based on that, it is possible to use Wilks' formula (see Reiss and Ruschendorf [1976] for instance) or deviation inequality as presented in Torossian et al. [2019a] to extract confidence intervals. For example, using Chernoff's inequality, for any $\eta > 0$, the confidence interval of order $1-\eta$ is as $\mathrm{CI}(x) = [L_K(x), U_K(x)]$ with

$$U_K(x) = \min\left\{q, \quad \widehat{F}^K(q|X=x) \geqslant \tau \text{ and } n\,\mathrm{kl}(\widehat{F}^K(q|X=x), \tau) \geqslant \log(2/\eta)\right\},$$

and

$$L_K(x) = \max\left\{q, \quad \widehat{F}^K(q|X=x) \leqslant \tau \text{ and } n\,\mathrm{kl}(\widehat{F}^K(q|X=x), \tau) \geqslant \log(2/\eta)\right\}.$$

Using this method enables the confidence intervals to be data dependent. For example in Figure 2.20 the confidence intervals are not symmetric. Note that while the confidence intervals obtained with this technique come with theoretical guarantees, they are very conservative and they depend a lot on the size of the training set.

Finally for all methods it is possible to use a bootstrap technique to create different regression models and then to aggregate them in order to create empirical confidence intervals. For example Figure 2.20 shows confidence intervals using a bootstrap technique for RK. The main drawback of such method is its computational cost.

To the best of our knowledge there are no other methods that sensibly improve this results. Thus there is still a room for improvement concerning quantile regression model assessment.

2.11 Summary and perspectives

2.11.1 General recommendations

In this presentation we have introduced six metamodels for quantile regression. In the first part of the paper we have provided a full description of the six metamodels, first focusing on their theoretical basis, then discussing their implementation procedure.

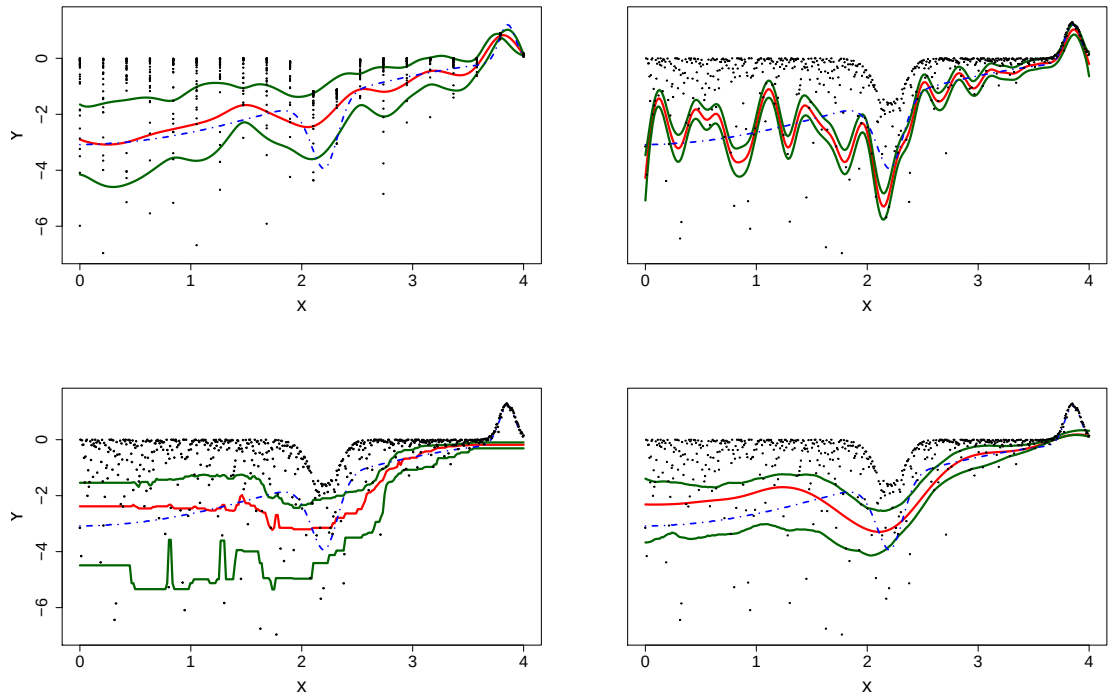


Figure 2.20: In red quantile metamodel, in blue the true 0.1-quantile and in green the 0.9-confidence intervals. Top left QK, top right VB, bottom left KN, bottom right RK.

This part of the paper have enabled us to highlight the similarities and differences of the methods so that providing critical perspectives on the state of the art. The second part of the paper focused on performance comparison according to the dimension of the problem, the size of the learning set, the signal-to-noise ratio and the value of the pdf at the targeted quantile. We have compared the presented methods on 4 toy problems in dimension 1, 2, 9 and on an agronomic model in dimension 9.

Figure 2.21 summarizes our findings. In a nutshell, when the signal-to-noise ratio is high RK, VB, QK and NN shows good results in our experimental setting but as soon as the signal-to-noise ratio decreases, quantile regression requires larger budgets and comparing the methods seems to be more complicated. Indeed, while the rule-of-thumb for computer experiments is a budget (i.e. number of experiments) 10 times the dimension (see Loeppky et al. [2009] for instance), as we work on problems with low signal-to-noise ratio, we found that no method was able to provide a relevant quantile estimate with a number of observations less than 50 times the dimension. For larger budgets, no method works uniformly better than any other. NN and VB are best when the budget is large. When the budget is smaller, RF, RK, KN are best when the pdf is small in the neighborhood of the quantile, in other words, when little information is

available. However, VB outperforms all the other methods when more information is available, that is, when the pdf is large in the neighborhood of the quantile.

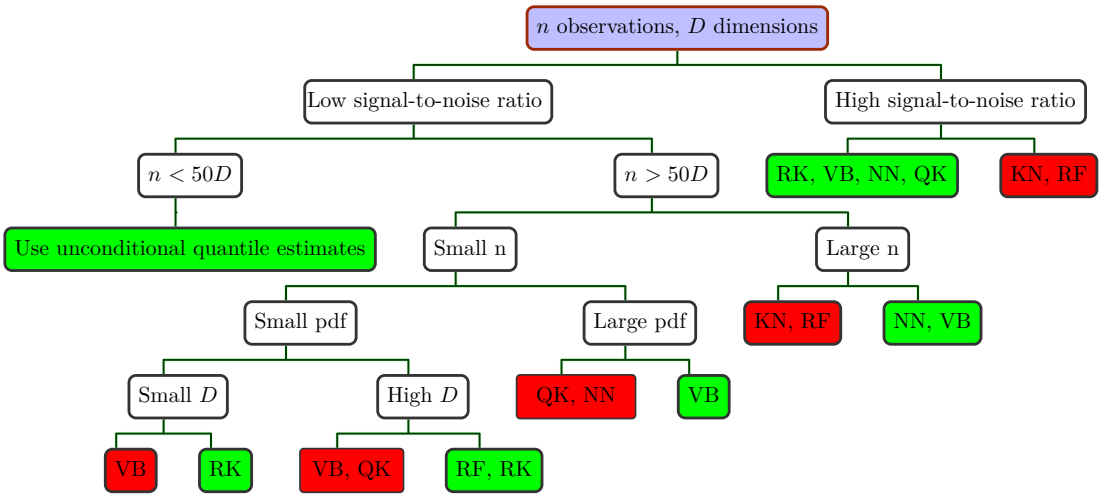


Figure 2.21: Method recommendation depending on the problem at hand (green: recommended methods, red: methods to avoid). KN: nearest-neighbors, RF: random forests, NN: neural networks, RK: RKHS regression, QK: quantile kriging, VB: variational Bayesian, unconditional quantile estimates: KN with $K = n$.

2.11.2 Possible ways of improvement

In our benchmark, we generally followed the approaches as presented by their authors. However, most of them could be improved. The optimization scheme of NN is the computational bottleneck of the method, which makes it the most expensive method in our benchmark system. One possible improvement would be using the BFGS algorithm (see Lewis and Overton [2009] for details about the BFGS algorithm applied to non-smooth functions) or the ADAM algorithm Kingma and Ba [2014] to optimize directly the empirical risk associated to the pinball loss. A faster scheme would allow more restarts, and hence improve robustness.

Another improvement concerns the splitting criterion (2.7) of RF, which is not designed for the quantile but for the expectation. This could lead to poor estimates for problems where quantiles are weakly correlated with expectations. Defining an appropriate splitting criterion could significantly improve the performance of this method.

In our experiments, QK used a predefined number of sampling points that were heuristically defined as a trade-off between space-filling and pointwise quantile estimation accuracy. The performance of QK could be significantly improved by optimally tuning the ratio between the number of points and repetitions, in the spirit of Binois et al. [2018].

The KN method can naturally be extended to a variant that uses the whole sample instead of the K nearest points. The weights associated to each point of the sample could be based on Gaussian or triangular kernel for example. This idea has been developed in Yu et al. [2002] to estimate the conditional expectation but we think that it could be possible to adapt this approach to the estimation of the conditional quantile.

Finally, in practice, finding the best hyperparameters was the most difficult part of the proposed benchmark system. While this aspect is often toned down by authors, we believe hyperparameters tuning is a key practical aspect that remains a challenging problem in quantile regression.

Chapter 3

$\mathcal{X}$-Armed Bandits: Optimizing Quantiles, CVaR and other Risks

Contents

3.1 Résumé

Ce chapitre reprend l'article Torossian et al. [2019a] publié à la conférence ACML 2019. Nous proposons et analysons l'algorithme StoROO qui est un algorithme d'optimisation de mesures de risque de fonctions boîte noire stochastique. Cet algorithme est une adaptation de l'algorithme StoOO Munos [2014]. Nous proposons une analyse générique du regret simple de StoROO et illustrons son applicabilité sur deux exemples : l'optimisation de quantiles et de CVaR. Inspiré par la littérature bandit et les optimiseurs de la moyenne conditionnelle d'une fonction boîte noire, StoROO construit des intervalles de confiance sur la fonction cible grâce à des échantillons de taille aléatoire. Nous détaillons la mise en place de tels intervalles, d'abord en utilisant des inégalités sous-optimales mais explicites permettant d'obtenir des bornes non-asymptotiques sur le regret simple. Puis nous utilisons des bornes plus fines mais moins explicites dont l'impact est illustrée numériquement.

Ce travail a été réalisé en collaboration avec Aurélien Garivier et Victor Picheny.

3.2 Introduction

We consider an unknown function $\Psi : \mathcal{X} \times \Omega \to [0,1] \subset \mathbb{R}$, where $\mathcal{X} \subset [0,1]^D$ and Ω denotes the probability space representing some uncontrollable variables. For any fixed $x \in \mathcal{X}, Y_x = \Psi(x, \cdot)$ is a random variable of distribution $\mathbb{P}_x$ and we consider $g(x) = \rho(\mathbb{P}_x)$ with ρ a real-valued functional defined on probability measures. We assume that there exists at least one $x^* \in \mathcal{X}$ such that $g(x^*) = \sup_{x \in \mathcal{X}} g(x)$. Using a set of sequential observations $(\Psi(x_1, \omega_1), \cdots, \Psi(x_T, \omega_T))$, our goal is to minimize the simple regret $r_T = g(x^*) - g(x_T)$, with x_T the value returned after using a budget T.

Different families of algorithms have been developed to treat this problem. Some are for example of Bayesian flavor [see Shahriari et al., 2016, for instance], some are inspired by the bandit literature. Here we focus our interest on the bandit framework.

In the classical $\mathcal{X}$-armed bandit problem, a forecaster selects repeatedly a point x in the input space $\mathcal{X} \in [0,1]^D$ and receives a reward distributed according to an unknown distribution $\mathbb{P}_x$. Historically, the main goal was to minimize the *cumulative regret*, i.e. the sum of the difference between his collected rewards and the ones that would have been brought by optimal actions. In the last decade, other works focused on the simple regret. These can be divided in two: algorithms that optimize an unknown function with the knowledge of the smoothness, for example StoOO [Munos, 2014], HOO [Bubeck et al., 2011] or Zooming [Kleinberg et al., 2008] and others focusing on the optimization of unknown functions without the knowledge of the smoothness, such as POO [Grill et al., 2015], StroquOOL [Bartlett et al., 2018], GPO [Xuedong et al., 2019], StoSOO [Valko et al., 2013] or Locatelli and Carpentier [2018].

Those algorithms focus on the optimization of the conditional *expectation* of $\mathbb{P}_x$. This choice is questionable in some situations. For example if the shape and variance of the reward distribution depend on the input, a forecaster may be interested in different aspects of the unknown distribution in order to modulate its risk exposure. In the liter-

ature, some measures of risk have been proposed to replace the expectation: for instance quantiles [also referred to as Value-at-Risk, see Artzner et al., 1999], the Conditional Value-at-Risk [CVaR also referred as Superquantile or Expected Shortfall, Rockafellar et al., 2000] or expectiles [Bellini and Di Bernardino, 2017]. The purpose of this paper is to present a risk optimization framework of an unknown stochastic function with the knowledge of the smoothness using only pointwise sequential observations and a finite budget T.

$\mathcal{X}$-armed bandit algorithms rely on *optimistic strategies* that associate with each point of the space an upper confidence bound (UCB), that is, an *optimistic* prediction of the outcome. Adapting the classical setting to the optimization of risk measures implies being able to create high-probability confidence bounds for that particular measure. This problem has been tackled in the multi-armed bandit setting (*i.e.* when the input space is discrete and finite). For instance, Audibert et al. [2009], Sani et al. [2012] focused on the empirical variance, Galichet et al. [2013], Kolla et al. [2019], Hepworth [2017] on the CVaR while in David and Shimkin [2016], Szorenyi et al. [2015] the authors based their policies on the quantile. However, the literature is scarce in the continuous input space case.

In this paper we provide a new version of the Stochastic Optimistic Optimization (StoOO) algorithm [Munos, 2014], named StoROO (Stochastic Risk Optimistic Optimization), which is designed to handle any functional ρ. In a first part, we provide an analysis of the simple regret from a generic point of view. We then particularize our analysis in two important illustrative cases: conditional quantiles and CVaR. In the case of quantiles, assuming that the output distribution has a continuous, strictly increasing cumulative density function, we first propose an upper bound on the simple regret using Hoeffding's inequality, then, we derive tighter confidence intervals that take into account the order of the quantile respectively based on Bernstein's and Chernoff's inequalities. In the case of the CVaR, we first derive an upper bound on the regret using the deviation inequality of Brown [2007], then using the work of Thomas and Learned-Miller [2019] we derived tighter confidence bounds. Finally, we present numerical experiments that illustrate the ability of our method to optimize conditional quantiles and CVaR of a black-box function and the relevance of using tight deviation bounds.

3.3 Problem setup

3.3.1 Hierarchical partitioning

The upper confidence bounds on which optimistic algorithms are based are surrogate functions $U : \mathcal{X} \to \mathbb{R}$ larger than the objective (in a sense detailed below) with high probability. At each round t, the point $X(t)$ having the highest UCB is sampled and a reward $Y_X(t)$ is collected.

In the classical multi-armed bandit problem, computing and sorting the UCB can be done without major issues. But dealing with continuous input spaces implies maximizing a UCB function over a continuous space, which can be both computational intensive

and algorithmically challenging. For example, Piyavskii's algorithm [see Bouttier, 2017, and references therein] defines U using a global Lipschitz assumption on the targeted function. Because of the Lipschitz hypothesis, the UCB maximizer is at an intersection of hyperplanes, i.e. where the UCB is non-differentiable. Thus a gradient-based algorithm cannot be used, implying that finding the point with the highest UCB is a very hard problem to solve.

To overcome the computational difficulties, a popular alternative is to rely on hierarchical partitions (see Bubeck et al. [2011], Munos [2014] for instance), $\mathcal{P} = \{\mathcal{P}_{h,j}\}_{h,j}$ of $\mathcal{X}$ such that

$$\mathcal{P}_{0,1} = \mathcal{X}, \quad \mathcal{P}_{h,j} = \bigcup_{i=0}^{K-1} \mathcal{P}_{h+1,Kj-i} \ ,$$

with K the number of sub-regions obtained after expanding a cell and $\mathcal{P}_{h,j}$ the j-th cell at depth h. In the following we assume that:

Assumption 1: There exists a decreasing sequence $\delta(h)$, such that for any $h \geqslant 0$ and for any cell $\mathcal{P}_{h,j}$, $\sup_{x \in \mathcal{P}_{h,j}} \|x - x_{h,j}\|_\infty \leqslant \delta(h)$, with $x_{h,j}$ the center of $\mathcal{P}_{h,j}$.

Assumption 2: There exists $\nu > 0$ such that every cell of depth h contains a ball of radius $\nu\delta(h)$.

Starting with $\mathcal{P}_{0,1}$ and following an optimistic strategy, at time t the algorithm has expanded some cells and the result is a tree $\mathcal{T}_t$ that is a subset of $\mathcal{P}$ and a partition of $\mathcal{X}$. In this setting U is taken as a piecewise constant function. Indeed for any $(\mathcal{P}_{h,j})_{h,j \in \mathcal{T}_t}$ we define $\bar{U}_{h,j}$ such that for all $x \in \mathcal{P}_{h,j}$, $U(x) = \bar{U}_{h,j}$.

In the literature of $\mathcal{X}$-armed bandits there are two ways to select a cell of $\mathcal{T}_t$ at each round. In Bubeck et al. [2011], the algorithm follows an *optimistic path* from the root to the leaves. In Munos [2014], StoOO selects the cell having the highest UCB among all the cells of $\mathcal{T}_t$ that have not been expanded, *i.e.* the set $\mathcal{L}_t$ of leaves of $\mathcal{T}_t$. We consider here this second alternative. Hence, to find the maximizer of U at time t, we only need to evaluate and sort a finite number of values $(\bar{U}_{h,j})_{(h,j) \in \mathcal{L}_t}$.

3.3.2 Regularity assumptions, noise and bias

Even in the absence of noise, optimization from finite samples requires some regularity of the objective. Following Munos [2014], we assume the following smoothness property:

$$\forall x \in \mathcal{X}, \quad g(x) \geqslant g(x^*) - \beta ||x - x^*||^\gamma \text{ with } \gamma, \ \beta \ > 0 \ . \tag{3.1}$$

Note that this condition is less restrictive than a global Hölder condition. In particular, the objective may be very irregular (even possibly discontinuous) except in the neighborhood of global maxima.

At first glance, in our stochastic setting, it may not be easy to asses that g satisfies (3.1). Sufficient conditions can be derived from the continuity of the conditional distribution $\mathbb{P}_x$ with respect to x. The relevant metric on the space of distributions actually depends on the chosen risk. For conditional quantiles, the natural assumption is that $x \mapsto F_x^{-1}(\tau)$ satisfies (3.1), and a sufficient condition is that $\|F_x^{-1} - F_y^{-1}\|_\infty \leqslant \beta \|x - y\|^\gamma$. In the case of the the conditional expectation and for the CVaR (or more generally for a

large class of Optimized Certainty Equivalent Ben-Tal and Teboulle [2007]), the natural metric involved is the *Wasserstein distance* $\mathcal{W}_1$, as explained in Section 3.9.1.

To create confidence bounds for $(\mathcal{P}_{h,j})_{(h,j)\in\mathcal{L}_t}$, StoOO samples the leafs at their centers $(x_{h,j})_{(h,j)\in\mathcal{L}_t}$. Then using that all observed values are independent, *deviation inequalities* are used to create $(U_{h,j})_{(h,j)\in\mathcal{L}_t}$, a UCB for $(g(x_{h,j}))_{(h,j)\in\mathcal{L}_t}$. Finally to create $(\bar{U}_{h,j})_{(h,j)\in\mathcal{L}_t}$, a UCB over the cells, a *bias term* is added that takes into account how g can potentially increase from the center of the cell to its edges. Because the convergence of StoOO (and StoROO) only needs $\bar{U}_{h,j}$ to be a UCB of $\max_{x\in\mathcal{P}_{h,j}} g(x)$ for the cell containing x^* (see the proof of Proposition 3.4.2 (see also Munos [2014]), it is enough to use the condition (3.1) to define a UCB as

$$\bar{U}_{h,j} = U_{h,j} + B_{h,j}, \text{ with } B_{h,j} = \widehat{\beta}\delta(h)^{\widehat{\gamma}} ,$$

and $\beta \leqslant \widehat{\beta}$, $\gamma \geqslant \widehat{\gamma}$. The algorithm also needs a quantity that bounds g from below in order to provide guaranties on the value of g over each cell. We thus construct a lower confidence bound, termed $L_{h,j}$, for $g(x_{h,j})$, and use it as a LCB for the maximum of g on $\mathcal{P}_{h,j}$. In particular, on the cell $\mathcal{P}_{h^*,j^*}$ containing the optimum x^*, it holds that

$$L_{h^*,j^*} \leqslant g(x^*) \leqslant U_{h^*,j^*} + \widehat{\beta}\delta(h^*)^{\widehat{\gamma}}$$

with high probability. To summarize, the estimation of $g(x^*)$ is altered by two sources of error: the local estimation error $E_{h^*,j^*} = U_{h^*,j^*} - L_{h^*,j^*}$ made at the center of the cell, and the bias term B_{h^*,j^*}. Balancing those two terms naturally provides a trade-off between exploration and exploitation.

3.4 Stochastic Risk Optimistic Optimization

3.4.1 The StoROO algorithm

StoROO starts by sampling one time each K sub-region of the root node. Then, at each time $1 \leqslant t \leqslant T$ the algorithm selects $\mathcal{P}_{h_t,j_t} \in (\mathcal{P}_{h,j})_{(h,j)\in\mathcal{L}_t}$ having the highest UCB. To reduce the estimation error, StoROO can either get more samples from $\mathcal{P}_{h_t,j_t}$ (to reduce the variance), or split the cell in order to reduce its diameter (to reduce the bias). The good balance between these two options is found by dividing a cell as soon as the local estimation error is smaller than the bias, that is when

$$U_{h_t,j_t} - L_{h_t,j_t} \leqslant \widehat{\beta}\delta(h_t)^{\widehat{\gamma}} . \tag{3.2}$$

If Condition (3.2) is satisfied, StoROO expands $\mathcal{P}_{h_t,j_t}$ and requires a new sample at the center of each sub-region. If Condition (3.2) is not satisfied, then StoROO requires a new sample at the center x_{h_t,j_t} which is used to update U_{h_t,j_t} and L_{h_t,j_t}.

When the budget is exhausted, several choices are possible for the return value: they have the same theoretical guarantees. Following Munos [2014], one can return the deepest node among those that have been expanded. Here we propose a different, more conservative choice. Denoting by $\mathscr{L}_T$ the set of nodes having the highest LCB among

those that have been expanded after a budget T, StoROO returns the node with the highest value $\widehat{g}$ (an estimator of g) among the deepest nodes of $\mathscr{L}_T$. The pseudo-code of the full algorithm is given in Algorithm 8. It requires the parameters $\widehat{\beta}$ and $\widehat{\gamma}$ that satisfy Condition (3.1), but of course the inequality do not have to be tight.

Algorithm 8: StoROO

Input: error probability $\eta > 0$; number of children K; time horizon T; $\widehat{\beta} > 0$; $\widehat{\gamma} > 0$;
Define: UCB and LCB
Initialization $n = 1$; $t = 1$;
Expand into K sub-regions the root node $(0, 0)$ and sample one time each child;
while $n \leqslant T$ **do**
  foreach $(h, j) \in \mathcal{L}_t$ **do**
    compute $\bar{U}_{h,j}(t)$;
  end
  Select $(\widetilde{h}, \widetilde{j}) = \arg\max_{(h,j)\in\mathcal{L}_t} \bar{U}_{h,j}(t)$;
  Compute the LCB $L_{\widetilde{h},\widetilde{j}}(t)$;
  if $U_{\widetilde{h},\widetilde{j}}(t) - L_{\widetilde{h},\widetilde{j}}(t) \leqslant \widehat{\beta}\delta(h)^{\widehat{\gamma}}$ **then**
    expand the node, remove $(\widetilde{h}, \widetilde{j})$ from $\mathcal{L}_t$, add to $\mathcal{L}_t$ the K sub-cells of $\mathcal{P}_{\widetilde{h},\widetilde{j}}$ and sample each new node once,
    $n = n + K$, $t = t + 1$;
  else
    Sample the state $x_t = x_{\widetilde{h},\widetilde{j}}$ and collect the observation $Y_{x_{h_t,j_t}}$, $n = n + K$, $t = t + 1$
  end
end
Return the node according to the returning rule.;

3.4.2 Analysis of the algorithm

In this section we provide a theoretical analysis of StoROO. It is inspired by Munos [2014], but differs most notably by the fact that the analysis is suited for any g and not only for the conditional expectation. The analysis relies on the possibility to construct, for any $\eta > 0$, upper- and lower-confidence bounds $U^{\eta}_{h,j}(t)$ and $L^{\eta}_{h,j}(t)$ such that the event

$$\mathcal{A}_\eta = \bigcap_{T\geqslant t\geqslant 1} \bigcap_{\mathcal{P}_{h,j}\in\mathcal{T}_t} \left\{ U^{\eta}_{h,j}(t) \geqslant g(x_{h,j}),\ L^{\eta}_{h,j}(t) \leqslant g(x_{h,j}) \right\}$$

has probability $\mathbb{P}(\mathcal{A}_\eta)$ at least $1 - \eta$. We defer to Section 3.5 their specific expression for the cases of the quantile and CVaR. Especially Section 3.5 shows that in our setting the size of the confidence interval associated to each node is not always explicit, by opposition of the classical case. We thus need to introduce the following definition to

quantify how many times a node needs to be sampled before satisfying the expansion condition (Eq. 3.2).

Definition 3.4.1. Let

$$m_{\eta,h}(\theta,\kappa,\alpha) = \log(\theta T^2/\eta)\left(\frac{\kappa}{\widehat{\beta}\delta(h)^{\widehat{\gamma}}}\right)^{\alpha}$$

and $N_{h,j}(t) = \sum_{s=1}^{t} \mathbb{1}_{X(s)\in\mathcal{P}_{h,j}}$, a *vector of safe constants* $v = (\theta,\kappa,\alpha)$ is composed of constants $\theta > 0$, $\kappa > 0$, and $\alpha > 0$ such that the event

$$\mathcal{B}_\eta = \bigcap_{T\geqslant t\geqslant 1} \bigcap_{N_{h,j}\geqslant m_{\eta,h}(\theta,\kappa,\alpha)} \bigcap_{\mathcal{P}_{h,j}\in\mathcal{T}_t} \left\{U^{\eta}_{h,j}(t) - L^{\eta}_{h,j}(t) \leqslant \widehat{\beta}\delta(h)^{\widehat{\gamma}}\right\}$$

has probability at least $1-\eta$.

For example, in the case of the conditional expectation a direct consequence of Hoeffding's inequality provides $\theta = 2$, $\alpha = 2$ and $\kappa = \sqrt{1/2}$ (see Munos [2014]).

To ensure the convergence of StoROO, we first prove (Proposition 3.4.2) that any point at the center of an expanded cell of depth h belongs to

$$J_h = \{\ x_{h,j} \text{ such that } g(x_{h,j}) + 2\widehat{\beta}\delta(h)^{\widehat{\gamma}} \geqslant g^*\}\ . \tag{3.3}$$

Next, Proposition 3.4.3 shows that using a budget T, the tree $\mathcal{T}_T$ reaches at least a depth $H^*_\eta(T)$. This implies the point returned by the algorithm belongs to $J_{H^*_\eta(T)}$ (Proposition 3.4.4). Finally, using an assumption on the size of J_h that can be formalized by the so-call *near-optimality dimension*, we provide an upper bound on the regret (Theorem 3.4.7).

Proposition 3.4.2. *Conditionally on $\mathcal{A}_\eta$, StoROO only expands cells $\mathcal{P}_{h,j}$ such that $x_{h,j} \in J_h$.*

Given the safe constants v and the total budget T, the deeper the algorithm builds the tree, the better are the guarantees on the final point returned. So the goal of the following proposition is to provide a lower bound on the depth of $\mathcal{T}_T$.

Proposition 3.4.3. *Define $n_{\eta,h} = m_{\eta,h}(v)$ and define H_η the largest $h \in \mathbb{N}$ such that*

$$S_h = K\sum_{h'\leqslant h} n_{\eta,h'+1}|J_{h'}| \leqslant T, \quad \textit{with} \quad |J_{h'}| \textit{the cardinal of } J_{h'}\ .$$

The deepest node H^*_η expanded by StoROO is such that $H^*_\eta \geqslant H_\eta$.

Intuitively, S_h is the budget needed to expand all the nodes in J_h for all $h' \leqslant h$. It may be that some of this nodes will not be visited, but in the worst case they are and they need to be considered in order to obtain a valid bound. Putting Propositions 3.4.2 and 3.4.3 together, yields a first upper bound on the simple regret:

Proposition 3.4.4. *Running StoROO with budget T, with probability $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta)$ the regret is bounded as*

$$r_T \leqslant 2\widehat{\beta}\delta\big(H_\eta^*(T)\big)^{\widehat{\gamma}} .$$

A more explicit bound for the regret can be obtained by quantifying the volume of $\mathcal{X}_\varepsilon = \{x \in \mathcal{X},\ g(x) \leqslant g^* - \varepsilon\}$ for small values of ε. Introducing the Holderian semi-metric

$$\ell_{\beta,\gamma}(x, x') = \beta \left\|x - x'\right\|^\gamma ,$$

that is associated with its regularity constants β and γ, the *near-optimality* dimension of the function is defined as follows, (see Munos [2014], Bubeck et al. [2011] for more details).

Definition 3.4.5. The ν-near optimality dimension is the smallest $d \geqslant 0$ such that for all $\varepsilon \geqslant 0$, there exists $C \geqslant 0$ such that the maximal number of disjoint $\ell_{\widehat{\beta},\widehat{\gamma}}$-balls of radius $\nu\varepsilon$ with center in $\mathcal{X}_\varepsilon$ is less than $C\varepsilon^{-d}$.

In order to evaluate H_η^*, we need to bound $|J_h|$ for all $h \geqslant 0$. The following proposition makes the link between the near optimality dimension and $|J_h|$.

Proposition 3.4.6. *Let d be the $\frac{\nu^{\widehat{\gamma}}}{2}$-near-optimality dimension, and C the corresponding constant. Then*

$$|J_h| \leqslant \frac{C}{\big(2\widehat{\beta}\delta(h)^{\widehat{\gamma}}\big)^d} .$$

Finally, combining Propositions 3.4.4 and 3.4.6 with an hypothesis on the decreasing sequence $\delta(h)$, it is possible to provide the speed of convergence of r_T.

Theorem 3.4.7. *Assume that $\delta(h) = c\rho^h$ for some $c \geqslant 0$ and $\rho < 1$, and assume that $v = (\theta, \kappa, \alpha)$. Thus with probability $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta)$, the regret of StoOO is bounded as*

$$r_T \leqslant c_1 \left[\frac{\log(\theta T^2/\eta)}{T}\right]^{\frac{1}{d+\alpha}} \quad \text{with} \quad c_1 = 2\widehat{\beta}\left[\frac{KC\kappa^\alpha [2\widehat{\beta}]^{-d}}{(1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\alpha})}\right]^{\frac{1}{d+\alpha}},$$

where d is the near optimality dimension and C the corresponding near optimality constant.

If g is the conditional expectation, a vector of *safe constants* is $(\theta = 2, \alpha = 2, \kappa = \sqrt{1/2})$ (based on Hoeffding's inequality). Thus if we plug it into the quantity defined in Theorem 3.4.7 we obtain

$$r_T \leqslant c_1 \left[\frac{\log(2T^2/\eta)}{T}\right]^{\frac{1}{d+2}} \quad \text{with} \quad c_1 = 2\widehat{\beta}\left[\frac{KC[2\widehat{\beta}]^{-d}}{2(1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\alpha})}\right]^{\frac{1}{d+2}},$$

that is equivalent to what it is obtained in Munos [2014].

Remark: In the particular case where each cell is a hypercube and the sub-regions are created by the division of the parent-cell into $K = 2^D$ sub-regions of equal size, then $K = 2^D$, c is equal to $\sqrt{D}$ and ρ is equal to $\frac{1}{2}$.

3.5 Optimizing Quantiles

In this section, we focus on the optimization of *quantiles*, which are well-established tools in (risk-averse) decision theory [see Rostek, 2010, for instance]. In particular, they benefit from interesting robustness properties, with respect to outliers or heavy tails. Let

$$g(x) = q_x(\tau) = \inf\left\{q \in \mathbb{R} : F_x(q) \geqslant \tau\right\},$$

be the τ-quantile of Y_x, where F_x is the cumulative distribution function (CDF) of $\mathbb{P}_x$. Here we detail how to construct the UCB and LCB for quantiles. First, we provide bounds based on Hoeffding's inequality and we use them to adapt the regret bounds of Theorem 3.4.7. Then we provide two more refined bounds that take into account the order τ of the quantile based respectively on the Bernstein's inequality and on the Kullback-Leibler divergence.

Let us first introduce some notations. For all $1 \leqslant t \leqslant T$, $1 \leqslant h \leqslant t$, $1 \leqslant j \leqslant K^h$ and $q \in \mathbb{R}$ we denote

$$\widehat{F}^t_{h,j}(q) = \frac{\sum_{s=1}^t \mathbb{1}_{Y(t)\leqslant q}\mathbb{1}_{X(t)\in\mathcal{P}_{h,j}}}{N_{h,j}(t)}$$

the empirical CDF of the reward inside the cell $\mathcal{P}_{h,j}$, where $N_{h,j}(t)$ is the (random) number of times the cell was sampled up to time t (see Definition 3.4.1). The *generalized inverse* $\widehat{F}^{t\ -}_{h,j}$ of the piecewise constant function $\widehat{F}^t_{h,j}$ is defined as

$$\widehat{q}_{h,j}(\tau) = \inf\left\{q \in \mathbb{R} :\ \widehat{F}^t_{h,j}(q) \geqslant \tau\right\},$$

that is the $\lceil N_{h,j}(t) \times \tau \rceil$ order statistic of the sample that has been collected from the node $x_{h,j}$ until time t.

To define confidence bounds on the conditional quantile we proceed in two steps. First we propose confidence bounds on $\widehat{F}_{h,j}(q_\tau)$. To do so, we simply use deviation bounds for Bernoulli distributions, since for all $x \in \mathcal{X}$, for all $1 \leqslant n \leqslant T$, the random variables $\left(\mathbb{1}_{Y_x(\xi_s)\leqslant q_x(\tau)}\right)_{s=1,\cdots,n}$ are independent and identically distributed with a Bernoulli law of parameter τ, if ξ_s denotes the time when the node x has been sampled for the s-th time. Then we use the properties

$$\forall\ \varepsilon > 0 \text{ such that } \tau + \varepsilon < 1,\ \widehat{F}^t_{h,j}(q_{h,j}(\tau)) \geqslant \tau + \varepsilon \quad \Leftrightarrow \quad q_{h,j}(\tau) \geqslant \widehat{F}^{t\ -}_{h,j}(\tau + \varepsilon) \tag{3.4}$$

$$\forall\ \varepsilon > 0 \text{ such that } \tau + \varepsilon > 0,\ \widehat{F}^t_{h,j}(q_{h,j}(\tau)) < \tau - \varepsilon \quad \Leftrightarrow \quad q_{h,j}(\tau) \leqslant \widehat{F}^{t\ -}_{h,j}(\tau - \varepsilon) \tag{3.5}$$

to create confidence bounds on $q_{h,j}(\tau)$ using bounds on $\widehat{F}^t_{h,j}(q_\tau)$. Note that here we just assume that the output distribution has a continuous, stricly increasing cumulative density function. It is not necessary to assume something else, such as bounded support or bounded moments because here we refer to Bernouilli distributions.

3.5.1 Hoeffding's bound and regret analysis

Let $\varepsilon^{\eta,T}_{N_{h,j}(t)} = \sqrt{\dfrac{\log(2T^2/\eta)}{2N_{h,j}(t)}}$, and let

$$U^{\eta}_{h,j}(t) = \begin{cases} \min\left\{q,\ \widehat{F}^{t}_{h,j}(q) \geqslant \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)} < 1 \\ +\infty & \text{otherwise,} \end{cases} \tag{3.6}$$

$$L^{\eta}_{h,j}(t) = \begin{cases} \max\left\{q,\ \widehat{F}^{t}_{h,j}(q) \leqslant \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)} > 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{3.7}$$

The next proposition motivates the choice of the above quantities as a UCB and a LCB for the quantile of order τ at the points $(x_{h,j})_{(h,j)\in\mathcal{T}_t}$.

Proposition 3.5.1. *Assume that for all $x \in \mathcal{X}$, $\mathbb{P}_x$ has a continuous, stricly increasing cumulative density function then for any $\eta > 0$, for all $h \geqslant 0$, for all $0 \leqslant j \leqslant K^h$ and for all $1 \leqslant t \leqslant T$, if $L^{\eta}_{h,j}(t)$ and $U^{\eta}_{h,j}(t)$ are defined according to (3.7) and (3.6), respectively, then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.*

Now, analyzing the regret requires a high probability bound on the number of time a node is sampled before being expanded:

Proposition 3.5.2. *Under the conditions required by Proposition 3.5.1, define f_x as the density of $\mathbb{P}_x$ and define*

$$\bar{f}(x) = \min_{\tau' \in [\tau - 2\varepsilon^{\eta,T}_{M_\tau}, \tau + 2\varepsilon^{\eta,T}_{M_\tau}]} f_x \circ F_x^{-1}(\tau')$$

with

$$M_\tau = 2m_\tau^{-2}\log(2T^2/\eta) \quad \textit{and} \quad \textit{and}m_\tau = \min(\tau, 1-\tau).$$

If $U^{\eta}_{h,j}(t)$ and $L^{\eta}_{h,j}(t)$ are defined according to (3.6) and (3.7), respectively, then for any $\eta > 0$, $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta) \geqslant 1-\eta$ and a vector of safe constants is given as

$$v = \left(2, \frac{\sqrt{8m_\tau^2 + 4\left(\widehat{\beta}\operatorname{diam}(\mathcal{X})^{\widehat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)\right)^2}}{m_\tau \min_{x\in\mathcal{X}} \bar{f}(x)}, 2\right).$$

According to the previous proposition, if we have sampled a node at depth h more than

$$n_{\eta,h} = \log(2T^2/\eta)\left(\frac{8m_\tau^2 + 4\left(\widehat{\beta}\operatorname{diam}(\mathcal{X})^{\widehat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)\right)^2}{\left(\min_{x\in\mathcal{X}} \bar{f}(x) m_\tau \widehat{\beta}\delta(h)^{\widehat{\gamma}}\right)^2}\right) \tag{3.8}$$

times, then with probability $1-\eta$, Condition (3.2) is satisfied and thus the node is expanded.

Equality (3.8) reflects two dependencies. The smaller the minimum of the density over a neighborhood of the quantile and the closer τ from 0 or 1, the larger the upper

bound on the number of samples needed before being expanded. Indeed a small density value in a neighborhood of the targeted quantile will produce samples with few observations close to the quantile, hence the estimation error will be large. In addition from Proposition (3.5.1), to obtain non trivial UCB and LCB, the value $N_{h,j}$ has to be large enough to ensure $\tau \pm \varepsilon^{\eta,T}_{N_{h,j}} \in [0,1]$ and this value increases as τ comes close from 0 or 1. Thus a more precise way to understand the behaviour of StoROO is that the number of time a node needs to be sampled before expansion depends on the pdf value in a neighborhood (of decreasing size with $N_{h,j}$) of the targeted quantile.

To obtain an upper bound on the simple regret, we now just need to combine Theorem 3.4.7 with Proposition 3.5.2 so as to obtain the following theorem.

Theorem 3.5.3. *Under the conditions required by Proposition 3.5.1 and 3.5.2, if $\delta(h) = c\rho^h$ for some $c \geqslant 0$ and $\rho < 1$, then with probability $1-\eta$, the regret of StoROO for maximizing the quantile is bounded as*

$$r_T \leqslant c_2 \left[\frac{\log(2T^2/\eta)}{T}\right]^{\frac{1}{d+2}} \text{ with } c_2^{d+2} = KC\widehat{\beta}^2 \frac{16m_\tau^2 + 8\left(\widehat{\beta}\,\mathrm{diam}(\mathcal{X})^{\widehat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)\right)^2}{\left(m_\tau \min_{x\in\mathcal{X}} \bar{f}(x)\right)^2 (1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\alpha})},$$

with d the near-optimality dimension and C the near-optimality corresponding constant.

Note that the speed of convergence is the same as the one obtained in the conditional expectation optimization setting; only the constant varies.

3.5.2 Tighter bounds

Using Hoeffding's inequality is convenient because it leads to explicit lower and upper confidence bounds, which simplifies the deriviation of bounds on the regret. However, it implicitly upper-bounds the variance of all $[0,1]$-valued random variables by $1/4$, which is overly pessimistic when the inequality is applied to variables whose expectations are far from $1/2$. This is in particular the case for quantile estimation, when the quantile is of order close to 0 or 1. To take into account the order of the quantile, following David and Shimkin [2016], a first possibility is to derive confidence intervals from Bernstein's inequality as presented in the following proposition.

Proposition 3.5.4. *For any $\eta > 0$, for all $1 \leqslant t \leqslant T$, $1 \leqslant h \leqslant t$ and $1 \leqslant j \leqslant K^h$, define*

$$U^\eta_{h,j}(t) = \begin{cases} \min\left\{q,\ \widehat{F}^t_{h,j}(q) \geqslant \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)} < 1 \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$L^\eta_{h,j}(t) = \begin{cases} \max\left\{q,\ \widehat{F}^t_{h,j}(q) \geqslant \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)}\right\} & \text{if } \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)} > 0 \\ -\infty & \text{otherwise,} \end{cases}$$

with

$$\varepsilon^{\eta,T}_{N_{h,j}(t)} = \frac{\log(2T^2/\eta)}{3N_{h,j}(t)}\left(1 + \sqrt{1 + \frac{18N_{h,j}(t)\tau(1-\tau)}{\log(2T^2/\eta)}}\right).$$

If g is the conditional quantile of order τ then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

Although Bernstein's inequality takes into account the order of the quantile, it is possible to do something better. In order to create tighter confidence bounds, we thus go back to Chernoff's inequality and derive less explicit, but more accurate upper- and lower- confidence bounds on the τ-quantiles. We follow here Garivier and Cappé [2011], but a close inspection at the proofs shows however a difference in the order of the marginals of the KL functions. Recall that the binary relative entropy is defined for $(p, q) \in [0, 1]^2$ as:

$$\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \,,$$

with by convention, $0 \log 0 = 0$, $\log 0/0 = 0$ and $x \log x/0 = +\infty$ for $x > 0$.

Proposition 3.5.5. *For any $\eta > 0$, for all $1 \leqslant t \leqslant T$, $1 \leqslant h \leqslant t$ and $1 \leqslant j \leqslant K^h$, define*

$$U^{\eta}_{h,j}(t) = \min \left\{ q,\ \widehat{F}^{n}_{h,j}(q) \geqslant \tau \text{ and } \mathrm{kl}(\widehat{F}^{t}_{h,j}(q), \tau) \geqslant \frac{\log(2T^2/\eta)}{N_{h,j}(t)} \right\} \text{ if } \mathrm{kl}(1, \tau) > \frac{\log(2T^2/\eta)}{N_{h,j}(t)}$$

and $+\infty$ otherwise. Define

$$L^{\eta}_{h,j}(t) = \max \left\{ q,\ \widehat{F}^{t}_{h,j}(q) \leqslant \tau \text{ and } \mathrm{kl}(\widehat{F}^{t}_{h,j}(q), \tau) \geqslant \frac{\log(2T^2/\eta)}{N_{h,j}(t)} \right\} \text{ if } \mathrm{kl}(0, \tau) > \frac{\log(2T^2/\eta)}{N_{h,j}(t)}$$

and $-\infty$ otherwise. Then the event $\mathcal{A}_\eta$ has probability at least $1 - \eta$.

Contrary to Bernstein's inequality, Chernoff's bound

$$\mathbb{P}\big(\widehat{F}^n(q(\tau)) \geqslant x\big) \leqslant \exp(-n\mathrm{kl}(x, \tau))$$

is always tighter than Hoeffding's inequality

$$\mathbb{P}(\widehat{F}^n\big(q(\tau)\big) \geqslant x\big) \leqslant \exp\big(-2n(\tau - x)^2\big),$$

which follows from Pinsker's inequality [see e.g. Garivier et al., 2018]:

$$\forall\ 0 \leqslant p < q \leqslant 1,\ \mathrm{kl}(p, q) \geqslant \frac{1}{2 \max_{x \in [p,q]} x(1 - x)} (p - q)^2 \geqslant 2(p - q)^2 \,.$$

For example, given $\tau > 0.5$ and an i.i.d. sample of size n, one can see that

$$U^{\mathrm{kl}}_n \leqslant \widehat{q}_n \left(\tau + \sqrt{\frac{2\tau(1 - \tau)\log(2/\eta)}{n}} \right) < \widehat{q}_n \left(\tau + \sqrt{\frac{\log(2/\eta)}{2n}} \right) = U^{\mathrm{H}}_n,$$

with U^{kl} (resp. U^{H}) the UCB associated to Chernoff's inequality (resp. Hoeffding's inequality). Bernstein's inequality is tighter than Hoeffding's when τ is different from $1/2$ and n sufficiently large, but always looser than Chernoff. It follows in particular that the regret of StoROO using confidence bounds derived from Chernoff's inequality has, at least, the guarantees presented in Theorem 3.5.3.

The online setting we consider in this article induces that, after t steps, the set of nodes and the number of observations in each node are random. To cope with this, we thus need deviation bounds for random size samples. The most simple way to obtain such inequalities is to use a union bound on the possible number of observations in each node, as presented above. Tighter results can be obtained from a more thorough analysis (sometimes called *peeling trick*): this is what is presented below.

Proposition 3.5.6. *For any $\eta \in (0,1)$ let $\delta_\eta(T) = \inf\left\{\delta > 0 :\ Te\lceil \delta \log(T)\rceil \exp(-\delta) \leqslant \eta/2\right\}$, and define*

$$U^\eta_{h,j}(t) = \min\left\{q,\ \widehat{F}^n_{h,j}(q) \geqslant \tau \text{ and } N_{h,j}(t)\mathrm{kl}(\widehat{F}^t_{h,j}(q),\tau) \geqslant \delta_\eta(T)\right\} \text{ if } \mathrm{kl}(1,\tau) > \frac{\delta_\eta(T)}{N_{h,j}(t)}$$

and $+\infty$ otherwise. Define

$$L^\eta_{h,j}(t) = \max\left\{q,\ \widehat{F}^t_{h,j}(q) \leqslant \tau \text{ and } N_{h,j}(t)\mathrm{kl}(\widehat{F}^n_{h,j}(q),\tau) \geqslant \delta_\eta(T)\right\} \text{ if } \mathrm{kl}(0,\tau) > \frac{\delta_\eta(T)}{N_{h,j}(t)}$$

and $-\infty$ otherwise. Then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

Note that for every $0 < \delta \leqslant \log(2/\eta)$, $\lceil \delta \log(T)\rceil \geqslant 1$ and thus $Te\lceil \delta \log(T)\rceil \exp(-\delta) > \eta/2$; hence, $\delta_\eta(T) > \log(2/\eta)$.

3.6 Optimizing CVaR

We now detail how StoROO can be applied to the optimization of another important notion of risk: the CVaR. CVaR has raised a great interest in recent years, notably because it is a *coherent* risk indicator (see Ben-Tal and Teboulle [2007] for instance). For $\tau \in [0,1)$ the condition value at risk at level τ of a continuous random variable Y is defined as

$$\mathrm{CVaR}_\tau(Y) = \inf_{z\in\mathbb{R}}\left\{z + \frac{1}{(1-\tau)}\mathbb{E}[(Y-z)^+]\right\} = \mathbb{E}\Big(Y | Y \geqslant q(\tau)\Big)\ ,$$

with $(z)^+ = \max(0,z)$. Following Brown [2007], it can be estimated by

$$\begin{aligned}\widehat{\mathrm{CVaR}_\tau}^{\,n} &= \inf_{z\in\mathbb{R}}\left\{z + \frac{1}{(1-\tau)n}\sum_{i=1}^n (Y_i - z)^+\right\} \\ &= Y_{(\lfloor n\tau\rfloor)} + \frac{1}{(1-\tau)n}\sum_{i=1}^n (Y_i - Y_{(\lfloor n\tau\rfloor)})^+\ .\end{aligned}$$

Note that the second equality can be demonstrated using the fact that $\widehat{\mathrm{CVaR}_\tau}^{\,n}$ is piecewise convex and that the slope is negative for $z < Y_{(\lfloor n\tau\rfloor)}$ and positive for $z > Y_{(\lfloor n\tau\rfloor)}$.

Since Y often stands for a loss, the CVaR is usually to be minimized. In order to stay consistent with the rest of the paper, we choose in the following to maximize $g = -\mathrm{CVaR}_\tau$. Assuming the random variables are bounded in an interval $[a,b]$, the next proposition adapts the deviation inequalities presented in Brown [2007] to our sequential setting.

Proposition 3.6.1. *For any $\eta > 0$, for all $h \geqslant 0$, for all $0 \leqslant j \leqslant K^h$ and for all $1 \leqslant t \leqslant T$, define*

$$U^{\eta}_{h,j}(t) = -\widehat{\mathrm{CVaR}}^{\mathrm{t}}_{\tau}(h,j) + \frac{b-a}{1-\tau}\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}},$$

and

$$L^{\eta}_{h,j}(t) = -\widehat{\mathrm{CVaR}}^{\mathrm{t}}_{\tau}(h,j) - (b-a)\sqrt{\frac{5\log(6T^2/\eta)}{(1-\tau)N_{h,j}(t)}}.$$

with

$$\widehat{\mathrm{CVaR}}^{\mathrm{t}}_{\tau}(h,j) = Y^{h,j}_{(\lfloor N_{h,j}(t)\tau \rfloor)} + \frac{1}{(1-\tau)N_{h,j}(t)}\sum_{i=1}^{t} \mathbb{1}_{X(i)\in\mathcal{P}_{h,j}}(Y_i - Y^{h,j}_{(\lfloor N_{h,j}(t)\tau \rfloor)})^+,$$

where $Y^{h,j}_{(k)}$ represents the value of $Y_{(k)}$ for the node (h,j).

If the random variables Y_x are bounded in $[a,b]$ for all $x \in \mathcal{X}$ and have continuous distribution functions, then the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

Note that *deviation inequalities* can be established for CVaR in sub-Gaussian or light-tailed cases (see Kolla et al. [2019] for instance) but an assumption has to be made on the value of the pdf in a neighborhood of the τ-quantile.

From Proposition (3.6.1), one can see that whenever a node has been played more than

$$m_{\eta,h} = \log(6T^2/\eta)(b-a)^2\left(\frac{1+\sqrt{10(1-\tau)}}{\sqrt{2}(1-\tau)\widehat{\beta}\delta(h)^{\widehat{\gamma}}}\right)^2$$

times, it has been expanded. Thus a possible associated vector of *safe constants* is

$$v = \left(6, (b-a)\left(\frac{1+\sqrt{10(1-\tau)}}{\sqrt{2}(1-\tau)\widehat{\beta}\delta^{\widehat{\gamma}}}\right), 2\right).$$

Combining v with Theorem 3.4.7 provides the following upper bound on the regret.

Theorem 3.6.2. *Under the conditions required by Proposition 3.6.1, if $\delta(h) = c\rho^h$ for some $c \geqslant 0$ and $\rho < 1$, then with probability $1-\eta$, the regret of StoROO for minimizing CVaR_τ is bounded as*

$$r_T \leqslant c_3\left[\frac{\log(6T^2/\eta)}{T}\right]^{\frac{1}{d+2}} \quad \text{with} \quad c_3 = 2\widehat{\beta}\left[\frac{\left(1+\sqrt{10(1-\tau)}\right)^2 KC(b-a)^2[2\widehat{\beta}]^{-d}}{2(1-\tau)^2(1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\alpha})}\right]^{\frac{1}{d+2}},$$

with d the near-optimality dimension and C the near-optimality corresponding constant.

The inequalities obtained in Proposition 3.6.1 are convenient because they lead to explicit lower and upper confidence bounds, which simplifies the derivation of bounds on the regret. However, as they are based on Hoeffding's inequality, they can be over-conservative. To obtain better bounds, Thomas and Learned-Miller [2019] propose data-dependent inequalities derived from the *Dvoretzky-Kiefer-Wolfowitz* inequality. The following proposition provides the UCB and LCB based on these inequalities.

Proposition 3.6.3. *Assume for all $x \in \mathcal{X}$, Y_x is bounded by $(a,b) \in \mathbb{R}^2$. For any $\eta \in (0, 0.5]$, for all $1 \leqslant t \leqslant T$, $1 \leqslant h \leqslant t$ and $1 \leqslant j \leqslant K^h$, define*

$$L^{\eta}_{h,j}(t) = \frac{1}{1-\tau} \sum_{i=1}^{N_{h,j}(t)} (Y^{h,j}_{i+1} - Y^{h,j}_{i}) \Big(\frac{i}{N_{h,j}(t)} - \sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}} - \tau \Big)^+ - Y_{n+1}$$

and

$$U^{\eta}_{h,j}(t) = \frac{1}{1-\tau} \sum_{i=0}^{N_{h,j}(t)-1} (Y^{h,j}_{i+1} - Y^{h,j}_{i}) \Big(\min\Big\{1, \frac{i}{N_{h,j}(t)} + \sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}}\Big\} - \tau \Big)^+ - Y^{h,j}_{n},$$

with $Y^{h,j}_0 = a$ and $Y_{n+1} = b$. Then if $g = -\operatorname{CVaR}_\tau$, the event $\mathcal{A}_\eta$ has probability at least $1-\eta$.

Although we do not propose an analysis of the regret based on this bounds, it is immediate to state that the upper bound on the regret is always smaller than the bound obtained in Theorem 3.6.2 because the inequalities of Thomas and Learned-Miller [2019] are strictly tighter than Brown's inequalities. In the following section, we numerically highlight the relevance of using these tight bounds.

3.7 Experiments

We empirically highlight the capacity of StoROO to optimize the conditional quantile and CVaR of a black-box function. Four versions of StoROO are compared for both cases.

For the conditional quantile we compare StoROO using confidence bounds repectively derived from Hoeffding's, Bernstein's, Chernoff's inequalities (resp. denoted StoROO_{H}, StoROO_{B} and $\text{StoROO}_{\text{kl}}$) and Chernoff's inequality and the *peeling trick* ($\text{StoROO}_{\text{kl-p}}$).

For the optimization of the conditional CVaR, we compare the use of confidence bounds derived from Brown's inequality and from Thomas and Learned-Miller [2019]. To use these inequalities we have to provide $(a,b) \in \mathbb{R}^2$ that bound the output. Hence, we compare two cases: one where we provide conservative bounds for (a,b) (here $(a,b) = (0,1)$), and one where we provide their actual values ($a_x = \min \operatorname{supp}(Y_x)$ and $b_x = \max \operatorname{supp}(Y_x)$, *i.e.* the minimum and the maximum of the support of the conditional distribution). We denote the four variants $\text{StoROO}_{\text{Br}}$ (from Brown's inequality), StoROO_{T} (from Thomas and Learned-Miller [2019]), and $\text{StoROO}_{\text{Br-o}}$ and $\text{StoROO}_{\text{T-o}}$ for their variants with oracle bounds.

As a test-case, we chose two functions with heteroscedastic noise and local extrema. The first is

$$\Psi_1(x, \cdot) = 0.18(\sin(3x)\sin(13x) + 1.3) + 0.062\zeta(\cdot)\big(\cos(8x-2) + 1.2\big),$$

where ζ is a log-normal random variable of parameters 0 and 1 truncated at its 0.95-quantile (the truncated mass is uniformly reallozcated between $q(0.91)$ and $q(0.95)$).

Note that to initialise StoROO not too close from a global optimum, we optimize the quantiles of Ψ_1 on $[-0.1, 0.9]$ and the CVaR on $[0, 1]$. Figure 3.1 (left) shows the shape of the 0.1 and 0.9 -quantiles and -CVaR of Ψ_1, while Figure 3.1 (right) shows samples of the 0.1-quantile. The second test-case is

$$\Psi_2(x, \cdot) = \mathrm{Cr}(x) + \zeta(\cdot)|\,\mathrm{Cr}(x) + 1.5\sqrt{x_1^2 + x_2^2}|,$$

on $[-0.5, 1]^2$ with

$$\mathrm{Cr}(x) = 0.1\Big(\Big|\sin(x_1)\sin(x_2)\exp\Big(\Big|3 - (\sqrt{x_1^2 + x_2^2}/\pi)\Big|\Big)\Big| + 1\Big)^{1.4}$$

and ζ a random variable that follows a Cauchy distribution of parameters $(0, 0.75)$. Note that for all $x \in \mathcal{X}$, $\Psi_2(x, \cdot)$ is unbounded and it has unbounded moments. Thus we can only apply quantile optimization on Ψ_2 based on the strategies developed in the past sections. Figure 3.2 (left) shows the shape of the 0.1-quantile of Ψ_2. The performance of each version of StoROO is evaluated for different values of τ and quantified according to the simple regret. In our experiments we fix the values $\beta = 12$ and $\gamma = 1.4$ (resp. $\beta = 2$, $\gamma = 0.5$ and $\beta = 2$, $\gamma = 0.7$) for the optimization of the quantiles (resp. the CVaR of order 0.1 and 0.9) of Ψ_1 and $\beta = 13$ and $\gamma = 1$ for the optimization of the 0.1-quantile of Ψ_2. Note that these values underestimate the regularity conditions at optimum so that satisfying the condition (3.1). In addition we fix $K = 3^D$ and we choose to expand the nodes into sub-region of equal sizes.

Figure 3.1 and 3.2 report the average of the simple regret over 100 runs. For both values of τ all the variants of StoROO have a regret that decreases with the budget. However from our experiments a ranking can be created.

For the optimization of the quantile, the less efficient method is $\mathrm{StoROO_H}$. For $\tau = 0.9$ its simple regret decreases slower than the three others methods and for $\tau = 0.1$ $\mathrm{StoROO_H}$ does not reach the performance of the others variants. To reach a fixed accuracy, $\mathrm{StoROO_H}$ sometimes needs a much larger budget than others variants. For example, on Ψ_1, taking $\tau = 0.9$, $\mathrm{StoROO_H}$ needs a budget of $15,000$ to reach a simple regret of order 10^{-4}, while $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$ need a budget equal to $5,000$. Second-to-last is $\mathrm{StoROO_B}$. Using the maximal budget, on both experiments on Ψ_1, this variant reaches the same accuracy as $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$ but its simple regret decreases slower. For some levels of performance $\mathrm{StoROO_B}$ needs a much larger budget than $\mathrm{StoROO_{kl}}$. For example, taking $\tau = 0.1$, to reach the value $r_T = 10^{-4}$ $\mathrm{StoROO_B}$ needs a budget of $T = 15,000$ while $T = 10,000$ is enough for $\mathrm{StoROO_{kl}}$. Finally, the most efficient methods are clearly $\mathrm{StoROO_{kl}}$ and $\mathrm{StoROO_{kl\text{-}p}}$. The use of a peeling argument (instead of a plain union bound) in $\mathrm{StoROO_{kl\text{-}p}}$ provides some additional gain over $\mathrm{StoROO_{kl}}$ on Ψ_1 but the effect is negligible on Ψ_2.

For the optimization of the CVaR, the variant based on tighter bounds is almost always better than the other and it is independent of the use of oracle bounds. The use of oracle bounds always improves the performance of StoROO and this effect is stronger if the confidence intervals are created with the inequalities of Thomas and Learned-Miller [2019]. Of course, in a real problem the oracle bounds are not known. Nevertheless this

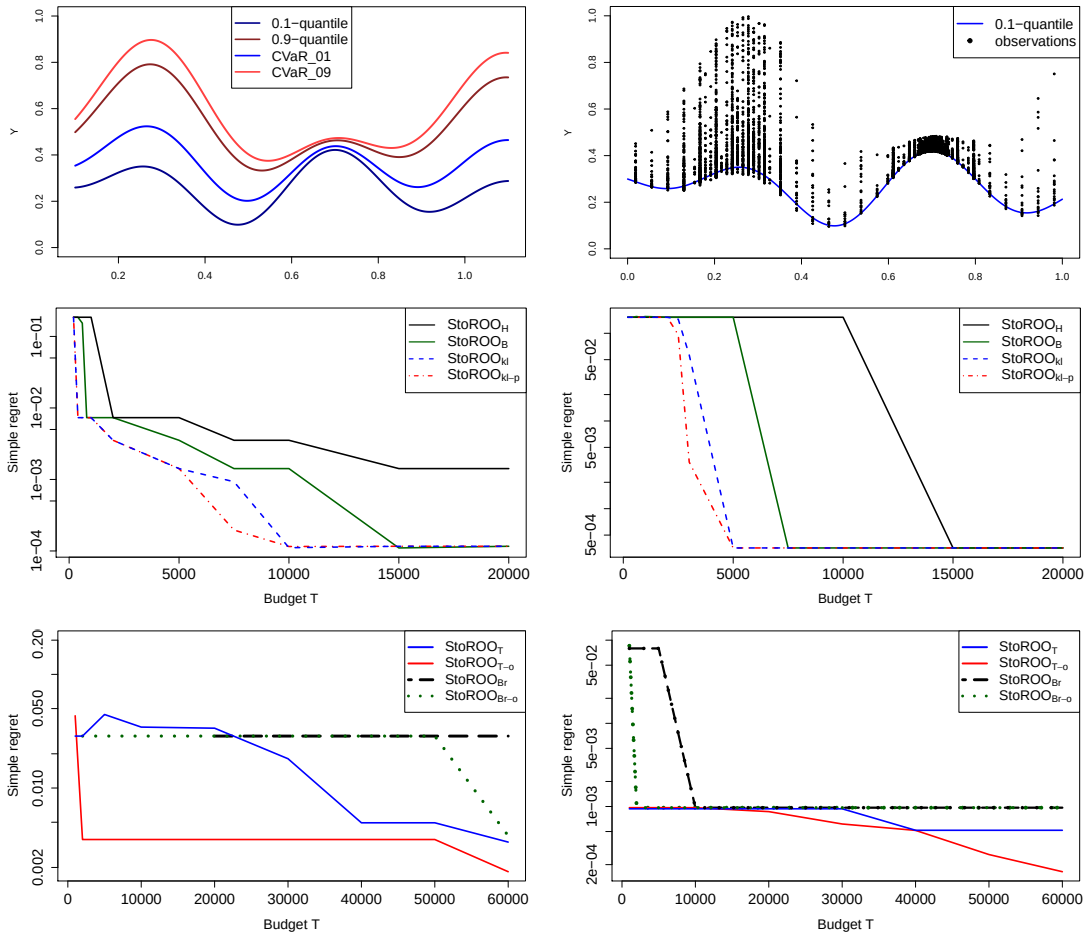


Figure 3.1: Results for the Ψ_1 test function. Top left: conditional quantiles and CVaR of Ψ_1. Top right: one run of StoROO$_{\text{kl}}$ for the 0.1-quantile with $T = 5,000$, $\beta = 12$ and $\gamma = 1.4$. Middle: evolution of the simple regret for the optimization of the quantile of order 0.1 (left) and 0.9 (right). Bottom: evolution of the simple regret for the optimization of the CVaR of order 0.1 (left) and 0.9 (right).

result motivates the use of estimators of the minimum and the maximum to estimate the conditional support so that to accelerate convergence.

3.8 Conclusion

In this work, we extended StoOO to a generic algorithm applicable to any functional of the reward distribution. We proposed a tailored application to the problem of quantile optimization, with four variants: one based on the classical Hoeffding's inequality, one based on Bernstein's inequality, and two others based on Chernoff's inequality. We

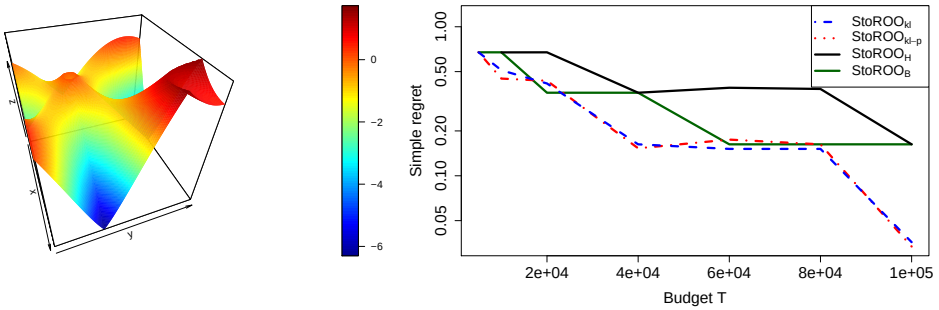


Figure 3.2: Results for the Ψ_2 test function. Left: Conditional quantile of order 0.1 of Ψ_2, right: Simple regret for the optimization of the conditional quantile presented to the left.

showed that using Chernoff's inequality to build confidence intervals resulted in a dramatic improvement, both in theory and practice. We also illustrated the ability of StoROO to optimize the CVaR and compared numerically four variants.

For simplicity, we assumed that the local regularity (or at least, an upper bound) of the target function at the optimum was known to the user. However, we believe that it might be possible to combine our results to the procedure defined in Grill et al. [2015], Xuedong et al. [2019] so as to propose an algorithm able to optimize g without the knowledge of the smoothness near an optimal point: this is left for future work. A second possible extension is to leverage the results proposed here to design an algorithm for the cumulative regret, in the spirit of HOO Bubeck et al. [2011] for example.

3.9 Appendix

3.9.1 Details about the regularity hypothesis

In the classical setting the Optimized Certainty Equivalent is defined as

$$S_u(Y) = \sup_z \Big\{ z + \mathbb{E}\big(u(Y - z)\big) \Big\},$$

with u a concave function. Here we assume u is concave and k-lipschitzian ($k - \mathrm{Lip}$). Let us consider two random variables Y_{x_1} and Y_{x_2}, then

$$\begin{aligned} |S_u(Y_{x_1}) - S_u(Y_{x_2})| &= \Big| \sup_z \Big\{ z + \mathbb{E}\big(u(Y_{x_1} - z)\big) \Big\} - \sup_z \Big\{ z + \mathbb{E}\big(u(Y_{x_2} - z)\big) \Big\} \Big| \\ &\leqslant \sup_z \Big\{ \big| \mathbb{E}\big(u(Y_{x_1} - z)\big) - \mathbb{E}\big(u(Y_{x_2} - z)\big) \big| \Big\}. \end{aligned}$$

Using the *Kantorovich-Rubinstein* representation one obtains

$$\begin{aligned} \sup_z \Big\{ \big| \mathbb{E}\big(u(Y_{x_1} - z)\big) - \mathbb{E}\big(u(Y_{x_2} - z)\big) \big| \Big\} &\leqslant k \times \mathcal{W}_1(Y_{x_1} - z, Y_{x_2} - z) \\ &= k \times \mathcal{W}_1(Y_{x_1}, Y_{x_2}) \end{aligned}$$

with $\mathcal{W}_1$ the Wasserstein distance associated with $p = 1$. Thus if $g = S_u$, then a sufficient condition to satisfied (1) is $\mathcal{W}_1(Y_{x*}, Y_x) \leqslant \frac{\beta}{k}\|x^* - x\|^\gamma$, for all $x \in \mathcal{X}$.

To treat the case of the CVaR_τ, we use the fact that if $u(z) = \dfrac{\min(z, 0)}{1 - \tau}$ then we have the equality $S_u = -\mathrm{CVaR}_\tau$.

In the case of the conditional expectation the same kind of condition can be sufficient. Indeed we have

$$|\mathbb{E}\big(Y_{x_1}\big) - \mathbb{E}\big(Y_{x_2}\big)| \leqslant \sup_{\|f\| \in 1-\mathrm{Lip}} \Big\{ \big|\mathbb{E}\big(f(Y_{x_1})\big) - \mathbb{E}\big(f(Y_{x_2})\big)\big| \Big\} = \mathcal{W}_1(Y_{x_1}, Y_{x_2}).$$

3.9.2 Proofs related to the generic analysis of StoROO

Proof. of Proposition 3.4.2

Let us define $\mathcal{P}_{h^*,j^*}$ the partition containing x^*. Assume that the partition $\mathcal{P}_{h,j}$ has been selected, thus

$$\bar{U}_\eta^{h,j}(t) \geqslant \bar{U}_\eta^{h^*,j^*}(t).$$

By definition $\bar{U}_\eta^{h^*,j^*}(t) \geqslant g^*$, thus $\bar{U}_\eta^{h,j}(t) \geqslant g^*$. Conditionally on $\mathcal{A}_\eta$, $L_\eta^{h,j}(t)) \leqslant g(x_{h,j}(t))$ that implies

$$g^* - g(x_{h,j}) \leqslant \bar{U}_\eta^{h,j}(t) - L_\eta^{h,j}(t) \leqslant U_\eta^{h,j}(t) + \widehat{\beta}\,\delta(h)^{\widehat{\gamma}} - L_\eta^{h,j}(t) \leqslant 2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}.$$

Note that the last inequality is obtained because the partition is expanded, which implies that

$$U(x_{h,j})(t) - L(x_{h,j})(t) \leqslant \widehat{\beta}\,\delta(h)^{\widehat{\gamma}}.$$

Finally:

$$g^* \leqslant g(x_{h,j}) + 2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}},$$

thus $x_{h,j}$ belongs to J_h. □

Proof. of Proposition 3.4.3

$$\begin{aligned}
T = \sum_{h,j \in \mathcal{T}_T} N_{h,j}(t) &\leqslant \sum_{h,j \in \mathcal{T}_T} n_{\eta,h} \quad \text{because } N_{h,j}(t) \leqslant n_{\eta,h} \\
&\leqslant \sum_{h'=0}^{\mathrm{depth}(\mathcal{T}_T)-1} K|\mathcal{T}_T \cap J_h| n_{\eta,h'+1} \quad \text{StoROO has not expanded all the sampled nodes} \\
&\leqslant \sum_{h'=0}^{\mathrm{depth}(\mathcal{T}_T)-1} K|J_h| n_{\eta,h'+1} = S_{\mathrm{depth}(\mathcal{T}_T)-1}.
\end{aligned}$$

Thus $S_{H_\eta} \leqslant S_{\mathrm{depth}(\mathcal{T}_T)-1} \leqslant S_{\mathrm{depth}(\mathcal{T}_T)}$ so $H_\eta \leqslant \mathrm{depth}(\mathcal{T}_T)$. There is at least an expanded node of depth $H_\eta^* \geqslant H_\eta$ after a budget T was used. □

Proof. of Proposition 3.4.4

Proposition 3.4.2 implies that the center of an expanded partition is in J_h. Proposition 3.4.3 implies that a partition of depth at least H^*_η has been expanded. Thus StoROO has expanded a node in $J_{H^*_\eta}$. At the end of the budget StoROO returns the node having the highest LCB among the nodes that have been expanded and not the deepest node among those that have been expanded. But

$$g^* - g(x_{h,j}) \leqslant \bar{U}_{H^*_\eta(T),j'} - L_{h,j} \leqslant \bar{U}_{H^*_\eta(T),j'} - L_{H^*_\eta(T),j'} \leqslant 2\,\widehat{\beta}\,\delta(H^*_\eta(T))^{\widehat{\gamma}}.$$

That ensure the node having the highest LCB has the same theoretical regret as the node of maximal depth among those that have been expanded. □

Proof. of Proposition 3.4.6

According to the assumption 2, each cell $\mathcal{P}_{h,j}$ contains a ball of radius $\nu\delta(h)$ centered in $x_{h,j}$ that is a $\ell_{\widehat{\beta},\widehat{\gamma}}$-ball of radius $\widehat{\beta}(\nu\delta(h))^{\widehat{\gamma}}$ centered in $x_{h,j}$. If d is the $\nu^{\widehat{\gamma}}/2$ near optimality dimension then there is at most $C[2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}]^{-d}$ disjoint $\ell_{\widehat{\beta},\widehat{\gamma}}$- balls of radius $\widehat{\beta}(\nu\delta(h))^{\widehat{\gamma}}$ inside $\mathcal{X}_{2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}}$. Thus if $|J_h| = |x_{h,j} \in \mathcal{X}_{2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}}| > C[\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}]^{-d}$ this implies there is more than $C[2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}]^{-d}$ disjoint $\ell_{\widehat{\beta},\widehat{\gamma}}$ balls of radius $\widehat{\beta}(\nu\delta(h))^{\widehat{\gamma}}$ with center in $\mathcal{X}_{2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}}$, that is a contradiction.

□

Proof. of Therorem 3.4.7

$$\begin{aligned}
T &\leqslant \sum_{h=0}^{H^*} K|J_h|n_{\eta,h+1} && \text{by definition of } H^* \\
&\leqslant \sum_{h=0}^{H^*} KC[2\,\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}]^{-d} n_{\eta,h+1} && \text{using Proposition 3.4.6} \\
&= \sum_{h=0}^{H^*} KC[2\,\widehat{\beta}(c\rho^h)^{\widehat{\gamma}}]^{-d} n_{\eta,h+1} && \text{using the exponential decay of the diameter of the cells} \\
&\leqslant \sum_{h=0}^{H^*} KC[2\,\widehat{\beta}(c\rho^h)^{\widehat{\gamma}}]^{-d} \times \kappa^\alpha \frac{\log(T^2/\eta)}{(\widehat{\beta}(c\rho^h)^{\widehat{\gamma}})^\alpha} && \text{using Definition 3.4.1} \\
&= \log(T^2/\eta)\frac{KC\kappa^\alpha[2\,\widehat{\beta}\,c^{\widehat{\gamma}}]^{-d}}{\widehat{\beta}\,c^{\widehat{\gamma}\,\alpha}} \sum_{h=0}^{H^*} \rho^{h(-d\widehat{\gamma}-\widehat{\gamma}\,\alpha)} \\
&= \log(T^2/\eta)\frac{KC\kappa^\alpha[2\,\widehat{\beta}\,c^{\widehat{\gamma}}]^{-d}}{\widehat{\beta}\,c^{\widehat{\gamma}\,\alpha}} \times \frac{\rho^{(H^*+1)(-d\widehat{\gamma}-\widehat{\gamma}\,\alpha)} - 1}{\rho^{-d\widehat{\gamma}-\widehat{\gamma}\,\alpha} - 1} && \text{rewriting the sum} \\
&\leqslant \frac{\log(T^2/\eta)}{(1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\,\alpha})}\frac{KC\kappa^\alpha[2\,\widehat{\beta}\,c^{\widehat{\gamma}}]^{-d}}{\widehat{\beta}\,c^{\widehat{\gamma}\,\alpha}} \times \rho^{H^*(-d\widehat{\gamma}-\widehat{\gamma}\,\alpha)} \\
&= \frac{\log(T^2/\eta)}{(1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\,\alpha})}\frac{KC\kappa^\alpha[2\,\widehat{\beta}]^{-d}}{\widehat{\beta}} \times \delta(H^*)^{-d\widehat{\gamma}-\widehat{\gamma}\,\alpha}.
\end{aligned}$$

Finally

$$\left[\frac{KC\kappa^{\alpha}[2\,\widehat{\beta}]^{-d}}{\widehat{\beta}(1-\rho^{d\widehat{\gamma}+\widehat{\gamma}\,\alpha})}\right]^{\frac{1}{d\widehat{\gamma}+\widehat{\gamma}\,\alpha}}\left[\frac{\log(T^2/\eta)}{T}\right]^{\frac{1}{d\widehat{\gamma}+\widehat{\gamma}\,\alpha}} \geqslant \delta(H^*).$$

Using Proposition 3.4.4 we obtain

$$r_T \leqslant c_1\left[\frac{\log(T^2/\eta)}{T}\right]^{\frac{1}{\alpha+d}}.$$

□

3.9.3 Proofs related to the section Optimizing quantiles

Proof. of Proposition 3.5.1

Let us consider the event

$$\begin{aligned}\xi_\eta \;=\; & \{\forall\; h\geqslant 0, \forall\; 0\leqslant j\leqslant K^h, \forall\; 1\leqslant t\leqslant T,\\ & \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)\geqslant \tau+\varepsilon^{\eta}_{N_{h,j}(t)} \text{ or } \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)<\tau-\varepsilon^{\eta}_{N_{h,j}(t)}\}.\end{aligned}$$

$$\begin{aligned}\mathbb{P}\big(\xi_\eta\big) &= \mathbb{P}\Big(\forall h\geqslant 0,\forall 0\leqslant j\leqslant K^h,\forall 1\leqslant t\leqslant T,\ \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)\geqslant \tau+\varepsilon^{\eta}_{N_{h,j}(t)} \text{ or },\\ &\qquad\qquad \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)<\tau-\varepsilon^{\eta}_{N_{h,j}(t)}\Big)\\ &\leqslant \mathbb{P}\Big(\forall h\leqslant 0,\forall 0\leqslant j\leqslant K^h,\forall 1\leqslant t\leqslant T,\ \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)\geqslant \tau+\varepsilon^{\eta}_{N_{h,j}(t)})\Big)\\ &\quad+\mathbb{P}\Big(\forall h\geqslant 0,\forall 0\leqslant j\leqslant K^h,\forall 1\leqslant t\leqslant T,\ \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)<\tau-\varepsilon^{\eta}_{N_{h,j}(t)}\Big)\end{aligned}$$

Define $m\leqslant T$ the number of nodes expanded throughout the algorithm, define for $1\leqslant w\leqslant m$, ζ^s_w as the time when the cell w has been selected for the s-th time and define $Y_w(\zeta^s_w)$ the reward obtained at that time at the point x_w. Then one can write

$$\mathbb{P}\Big(\widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)\geqslant \tau+\varepsilon^{\eta,T}_{N_{h,j}(t)}\Big)=\mathbb{P}\Big(\frac{1}{N_{h,j}(t)}\sum_{s=1}^{N_{h,j}(t)}\mathbb{1}_{Y_{h,j}(\zeta^s_{h,j})\leqslant q_{h,j}(\tau)}\geqslant \tau+\varepsilon^{\eta}_{N_{h,j}(t)}\Big).$$

Using this notation, we have:

$$\begin{aligned}&\mathbb{P}\Big(\forall h\geqslant 0,\forall 0\leqslant j\leqslant K^h,\forall 1\leqslant t\leqslant T,\ \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big)\geqslant \tau+\varepsilon^{\eta}_{N_{h,j}(t)}\Big)\\ \leqslant\;& \mathbb{P}\Big(\exists\ 1\leqslant w\leqslant T,\ \exists\ 1\leqslant u\leqslant T,\ \frac{1}{u}\sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leqslant q_w(\tau)}\geqslant \tau+\varepsilon^{\eta}_{u}\Big)\\ \leqslant\;& \sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(\frac{1}{u}\sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leqslant q_w(\tau)}\geqslant \tau+\varepsilon^{\eta}_{u}\Big)\end{aligned}$$

By Hoeffding's inequality, if

$$\varepsilon_u^\eta = \sqrt{\frac{\log(2T^2/\eta)}{2u}},$$

we obtain

$$\mathbb{P}\left(\forall h \leqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{F}_{h,j}^t\Big(q_{h,j}(\tau)\Big) \geqslant \tau + \varepsilon^\eta_{N_{h,j}(t)}\right) \leqslant \frac{\eta}{2}.$$

Now using Equation (3.4) we can express this inequality directly in terms of quantiles:

$$\mathbb{P}\left(\forall h \leqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ q_{h,j}(\tau) \geqslant U^\eta_{h,j}(t)\right) \leqslant \frac{\eta}{2}.$$

Using the same scheme of proof with Inequality (3.5), we obtain:

$$\mathbb{P}\left(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ q_{h,j}(\tau) \leqslant L^\eta_{h,j}(t)\right) \leqslant \frac{\eta}{2},$$

and hence $\mathbb{P}\big(\mathcal{A}_\eta\big) = 1 - \mathbb{P}\big(\xi_\eta\big) \geqslant 1 - \eta$. □

Proof. of Proposition 3.5.2

Without loss of generality let us assume $\tau > 0.5$. Assume the node $x_{h,j}$ has been sampled $N_{h,j} \geqslant M_\tau = \max(n_\tau, n_{1-\tau})$ times, with

$$n_\tau > \frac{2\log(2T^2/\eta)}{\tau^2} \quad \text{and} \quad n_{1-\tau} > \frac{2\log(2T^2/\eta)}{(1-\tau)^2}$$

thus

$$\tau + 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} < 1 \text{ and } \tau - 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} > 0.$$

That implies

$$q_{h,j}\left(\tau + 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}}\right) < +\infty \text{ and } q_{h,j}\left(\tau - 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}}\right) > -\infty,$$

and in particular

$$U^\eta_{h,j} < +\infty \text{ and } L^\eta_{h,j} > -\infty.$$

Then define the event

$$\mathcal{C}_\eta = \bigcap_{T \geqslant t \geqslant 1} \bigcap_{\mathcal{P}_{h,j} \in \mathcal{T}_t} \Big\{ q_{h,j}\big(\tau + 2\varepsilon^{\eta,T}_{N_{h,j}(t)}\big)\big) \geqslant U^\eta_{h,j}(t) \geqslant q_{h,j}(\tau) \geqslant L^\eta_{h,j}(t) \geqslant q_{h,j}\big(\tau - 2\varepsilon^{\eta,T}_{N_{h,j}(t)}\big)\Big\},$$

with

$$\varepsilon^{\eta,T}_{N_{h,j}(t)} = \sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}}.$$

Using equivalences (3.4) and (3.5), one can write:

$$q_{h,j}\big(\tau + 2\varepsilon^{\eta,T}_{N_{h,j}(t)}\big) \geqslant U^{\eta}_{h,j}(t) \geqslant q_{h,j}(\tau) \geqslant L^{\eta}_{h,j}(t) \geqslant q_{h,j}\big(\tau - 2\varepsilon^{\eta,T}_{N_{h,j}(t)}\big)$$
$$\Leftrightarrow \widehat{F}(q_{h,j}(\tau + 2\varepsilon^{\eta,T}_{N_{h,j}(t)})) \geqslant \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)} > \widehat{F}(q_{h,j}(\tau) \geqslant \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)} > \widehat{F}(q_{h,j}(\tau + 2\varepsilon^{\eta,T}_{N_{h,j}(t)})).$$

Thus

$$\begin{aligned}\mathbb{P}(\mathcal{C}_\eta) \geqslant \quad & 1 - \mathbb{P}(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T, \sup_{y = q_\tau, q_{\tau + \varepsilon^{\eta,T}_{N_{h,j}(t)}}} |F_{h,j}(y) - \widehat{F}^t_{h,j}(y)| \geqslant \varepsilon^{\eta,T}_{N_{h,j}(t)}\ ) \\ \geqslant \quad & 1 - \mathbb{P}(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T, \sup_{y \in [0,1]} |F_{h,j}(y) - \widehat{F}^t_{h,j}(y)| \geqslant \varepsilon^{\eta,T}_{N_{h,j}(t)}\ ).\end{aligned}$$

Using the same notation as in the proof of Proposition 3.5.1, one can write

$$\geqslant \quad 1 - \sum_{w=1}^{T}\sum_{u=1}^{T} \mathbb{P}\Big(\sup_{y\in[0,1]} |F_w(y) - \frac{1}{u}\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leqslant q_w(\tau)}| \geqslant \varepsilon^{\eta,T}_u\Big).$$

Now by applying the Massart's inequality to bound

$$\mathbb{P}\Big(\sup_{y\in[0,1]} |F_w(y) - \sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leqslant q_w(\tau)}| \geqslant \varepsilon^{\eta,T}_u\Big),$$

one obtain $\mathbb{P}(\mathcal{C}_\eta) \geqslant \quad 1 - \eta$. Thus with probability $1 - \eta$, we have:

$$U^{\eta}_{h,j}(t) - L^{\eta}_{h,j}(t) \leqslant q_{h,j}\left(\tau + 2\varepsilon^{\eta,T}_{N_{h,j}(t)}\right) - q_{h,j}\left(\tau - 2\varepsilon^{\eta,T}_{N_{h,j}(t)}\right). \tag{3.9}$$

Assuming that $q_{h,j}$ is differentiable in τ, by the mean value theorem, we deduce

$$q_{h,j}(\tau + 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}}) - q_{h,j}(\tau - 2\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}}) \leqslant 4\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} \max_{\tau' \in [\tau - 2\varepsilon^{\eta,T}_{n_\tau}, \tau + 2\varepsilon^{\eta,T}_{n_{1-\tau}}]} \frac{1}{f_{x_{h,j}} \circ F^{-1}_{x_{h,j}}(\tau')}.$$

Next, using (3.9) it is possible to write that with probability $1 - \eta$:

$$U^{\eta}_{h,j} - L^{\eta}_{h,j} \leqslant 4\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} \frac{1}{\bar{f}_{x_{h,j}}} \leqslant 4\sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}}} \frac{1}{\min_{x\in\mathcal{X}} \bar{f}(x)}.$$

We define $n'_{\eta,h}$ as the smallest n such that

$$4\sqrt{\frac{\log(2T^2/\eta)}{2n}} \frac{1}{\inf_{x\in\mathcal{X}} \bar{f}(x)} \leqslant \widehat{\beta}\, \delta(h)^{\widehat{\gamma}},$$

that is

$$n'_{\eta,h} = \log(2T^2/\eta)\left(\frac{2\sqrt{2}}{\widehat{\beta}\, \delta(h)^{\widehat{\gamma}} \min_{x\in\mathcal{X}} \bar{f}(x)}\right)^2.$$

A proper $n_{\eta,h}$ has to verify

$$n_{\eta,h} \geqslant M_\tau \text{ and } n_{\eta,h} \geqslant \log(2T^2/\eta)\left(\frac{2\sqrt{2}}{\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}\min_{x\in\mathcal{X}}\bar{f}(x)}\right)^2.$$

To satisfy this constraint we define

$$\begin{aligned} n_{\eta,h} &= \log(2T^2/\eta)\left(\frac{\sqrt{8\min(1-\tau,\tau)^2 + 4\big(\widehat{\beta}\,\mathrm{diam}(\mathcal{X})^{\widehat{\gamma}}\min_{x\in\mathcal{X}}\bar{f}(x)\big)^2}}{\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}\min_{x\in\mathcal{X}}\bar{f}(x)\min(1-\tau,\tau)}\right)^2 \\ &\geqslant \log(2T^2/\eta)\left(\left(\frac{2\sqrt{2}}{\widehat{\beta}\,\delta(h)^{\widehat{\gamma}}\min_{x\in\mathcal{X}}\bar{f}(x)}\right)^2 + \left(\frac{2}{\min(1-\tau,\tau)}\right)^2\right) \\ &= n'_{\eta,h} + M_\tau. \end{aligned}$$

To conclude the whole proof, since $\mathcal{C}_\eta \subset \mathcal{A}_\eta \cap \mathcal{B}_\eta$, we obtain $\mathbb{P}(\mathcal{A}_\eta \cap \mathcal{B}_\eta) \geqslant 1-\eta$. □

Proof. of Proposition 3.5.4

Let $Y_1, \cdots, Y_n$ be n *i.i.d.* random variables bounded by the interval $[0,1]$. Define $\widehat{F}^n(q(\tau)) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \leqslant q(\tau)}$. For $x > \tau$ the Bernstein's inequality gives

$$\mathbb{P}(|\widehat{F}^n(q(\tau)) - \tau| > \varepsilon) \leqslant 2\exp\left(\frac{n\varepsilon^2}{2\tau(1-\tau) + 2\varepsilon/3}\right).$$

Let us consider the event

$$\begin{aligned} \xi_\eta \;=\; & \{\forall\, h \geqslant 0, \forall\, 0 \leqslant j \leqslant K^h, \forall\, 1 \leqslant t \leqslant T, \\ & \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) \geqslant \tau + \varepsilon^{\eta,T}_{N_{h,j}(t)} \text{ or } \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) < \tau - \varepsilon^{\eta,T}_{N_{h,j}(t)}\}. \end{aligned}$$

Using the same lines as in the proof of Proposition 3.5.1 we have

$$\begin{aligned} \mathbb{P}(\xi_\eta) \leqslant\; & \sum_{w=1}^T\sum_{u=1}^T \mathbb{P}\Big(|\frac{1}{u}\sum_{s=1}^u \mathbb{1}_{Y_w(\zeta^s_w)\leqslant q_w(\tau)} - \tau| > \varepsilon^{\eta,T}_u\Big) \\ & \text{then applying the Bernstein's inequality we obtain} \\ \leqslant\; & \sum_{w=1}^T\sum_{u=1}^T 2\exp\Big(-\frac{u{\varepsilon^{\eta,T}_{N_{h,j}(t)}}^2}{2\tau(1-\tau) + 2\varepsilon^{\eta,T}_{N_{h,j}(t)}/3}\Big). \end{aligned} \tag{3.10}$$

By now the goal is to find $\varepsilon^{\eta,T}_{N_{h,j}(t)} > 0$ such that

$$\frac{u{\varepsilon^{\eta,T}_{N_{h,j}(t)}}^2}{2\tau(1-\tau) + 2\varepsilon^{\eta,T}_{N_{h,j}(t)}/3} = \log(2T^2/\eta).$$

Finding such $\varepsilon^{\eta,T}_{N_{h,j}(t)}$ can be easily done because it is a square of a second order polynomial. The result is

$$\varepsilon^{\eta,T}_{N_{h,j}(t)} = \frac{\log(2T^2/\eta)}{3u}\left(1+\sqrt{1+\frac{18u\tau(1-\tau)}{\log(2T^2/\eta)}}\right).$$

Plugging the value of $\varepsilon^{\eta,T}_{N_{h,j}(t)}$ inside (3.10) concludes the proof. □

Proof. of Proposition 3.5.5

Step 1: bounds on $\widehat{F}^n(q(\tau))$ for a i.i.d sample

Let $Y_1,\cdots,Y_n$ be n *i.i.d.* random variables bounded by the interval $[0,1]$. Define $\widehat{F}^n(q) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\leqslant q}$. For $x>\tau$ the Chernoff's inequality gives

$$\mathbb{P}(\widehat{F}^n(q(\tau)) \geqslant x) \leqslant \exp(-n\mathrm{kl}(x,\tau)).$$

Let $\tau^+ > \tau$ be the value such that $\mathrm{kl}(\tau^+,\tau) = \frac{\log(2/\eta)}{n}$, then for all $x \geqslant \tau^+$:

$$\mathbb{P}(\widehat{F}^n(q(\tau)) \geqslant x) \leqslant \mathbb{P}(\widehat{F}^n(q(\tau)) \geqslant \tau^+) \leqslant \exp(n\frac{\log(2/\eta)}{n}) = \frac{\eta}{2}.$$

Now let us define the candidate for the UCB of a i.i.d sample:

$$U(n) = \min\left\{q,\quad \widehat{F}^n(q) \geqslant \tau \text{ and } n\mathrm{kl}(\widehat{F}^n(q),\tau) \geqslant \log(2/\eta)\right\},$$

and let us remark that

$$\widehat{F}^n(U(n)) \leqslant \widehat{F}^n(q(\tau)) \Leftrightarrow \tau \leqslant \widehat{F}^n(q(\tau)) \quad\text{and}\quad \mathrm{kl}(\widehat{F}^n(q(\tau)),\tau) \geqslant \frac{\log(2/\eta)}{n}, \tag{3.11}$$

thus

$$\begin{aligned}\mathbb{P}(\widehat{F}^n(U(n)) \leqslant \widehat{F}^n(q(\tau))) =& \mathbb{P}(\tau \leqslant \widehat{F}^n(q(\tau)) \quad\text{and}\quad \mathrm{kl}(\widehat{F}^n(q(\tau)),\tau) \geqslant \frac{\log(2/\eta)}{n})\\ \leqslant& \mathbb{P}(\widehat{F}^n(q(\tau)) \geqslant \tau^+) \leqslant \frac{\eta}{2}.\end{aligned}$$

For $x<\tau$ let us introduce

$$L(n) = \max\left\{q,\quad \widehat{F}^n(q) \leqslant \tau \text{ and } n\mathrm{kl}(\widehat{F}^n(q),\tau) \geqslant \log(2/\eta)\right\},$$

one proves in the same way

$$\mathbb{P}(\widehat{F}^n(L(n)) > \widehat{F}^n(q(\tau))) \leqslant \frac{\eta}{2}.$$

Step 2: Double union bound

Let us consider the event

$$\begin{aligned}\xi_\eta \quad=\quad &\Big\{\forall\ h\geqslant 0, \forall\ 0\leqslant j\leqslant K^h, \forall\ 1\leqslant t\leqslant T,\\ &\widehat{F}^t_{h,j}\big(q_{h,j}(\tau)\big) \geqslant \widehat{F}^t_{h,j}(U^\eta_{h,j}) \text{ or } \widehat{F}^t_{h,j}\big(q_{h,j}(\tau)\big) < \widehat{F}^t_{h,j}(L^\eta_{h,j})\Big\}.\end{aligned}$$

$$\mathbb{P}(\xi_\eta) \leqslant \mathbb{P}\Big(\forall h \leqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{F}^t_{h,j}\big(q_{h,j}(\tau)\big) \geqslant \widehat{F}^t_{h,j}(U^\eta_{h,j})\Big)$$
$$+ \mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{F}^t_{h,j}\big(q_{h,j}(\tau)\big) < \widehat{F}^t_{h,j}(L^\eta_{h,j})\Big)$$

Following the notation of the proof of Proposition 3.5.1 we have

$$\mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{F}^t_{h,j}\big(q_{h,j}(\tau)\big) \geqslant \widehat{F}^t_{h,j}(U^\eta_{h,j})\Big)$$
$$\leqslant \mathbb{P}\Big(\exists\ 1 \leqslant w \leqslant T,\ \exists\ 1 \leqslant u \leqslant T,\ \sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leqslant q_w(\tau)} \geqslant \sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leqslant U^\eta_w}\Big)$$
$$\leqslant \sum_{w=1}^{T}\sum_{u=1}^{T} \mathbb{P}\Big(\sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leqslant q_w(\tau)} \geqslant \sum_{s=1}^{u} \mathbb{1}_{Y_w(\zeta^s_w) \leqslant U^\eta_w}\Big).$$

Using the equivalence (3.11), the probability can be reformulated as

$$= \sum_{w=1}^{T}\sum_{u=1}^{T} \mathbb{P}\Big(\tau \leqslant \widehat{F}^u(q(\tau)) \ \text{ and } \ \mathrm{kl}(\widehat{F}^u(q(\tau)), \tau) \geqslant \frac{\log(2T^2/\eta)}{u}\Big).$$

Now using Chernoff's inequality we obtain

$$\mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{F}^t_{h,j}\big(q_{h,j}(\tau)\big) \geqslant \widehat{F}^t_{h,j}(U^\eta_{h,j})\Big)$$
$$\leqslant \sum_{w=1}^{T}\sum_{u=1}^{T} \exp(-u\frac{\log(2T^2/\eta)}{u}) = \eta/2.$$

By equivalence (3.4) this implies that, $\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T$, with probability at least $\eta/2$, $U^\eta_{h,j}(t) \leqslant q_{h,j}(\tau)$. Using the same lines one can show

$$\mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{F}^t_{h,j}\Big(q_{h,j}(\tau)\Big) < \widehat{F}^t_{h,j}(L)\Big) \leqslant \eta/2,$$

By equivalence (3.5) this implies that, $\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T$, $L^\eta_{h,j}(t) > q_{h,j}(\tau)$ with probability at least $\eta/2$. Putting this two probabilities together prove the result. □

Proof. of Proposition 3.5.6

Define

$$\widetilde{S}^\tau_{h,j}(n) = \sum_{i=1}^{n} \mathbb{1}_{Y_{h,j}(i) \leqslant q_{h,j}(\tau)}.$$

Step 1: Martingale For every $\lambda \in \mathbb{R}$, let $\varphi_\tau(\lambda) = \log \mathbb{E}[\exp(\lambda \mathbb{1}_{Y_{h,j}(1) \leqslant q_{h,j}(\tau)})]$. Let $W^\lambda_0 = 1$ and for $n \geqslant 1$,

$$W^\lambda_n = \exp(\lambda \widetilde{S}^\tau_{h,j}(n) - n\varphi_\tau(\lambda)).$$

$(W_n^\lambda)_{n\geqslant 0}$ is a martingale relative to $(\mathcal{F}_n)_{n\geqslant 0}$. In fact,

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(\lambda\{\widetilde{S}^\tau_{h,j}(n+1)-\widetilde{S}^\tau_{h,j}(n)\}\Big)|\mathcal{F}_n\Big] &= \mathbb{E}\Big[\exp(\lambda X_{n+1})|\mathcal{F}_n\Big]\\ &= \exp\Big(\log\mathbb{E}[\exp(\lambda X_1]\Big)\\ &= \exp\Big(\{(n+1)-n\}\varphi_\mu(\lambda)\Big)\end{aligned}$$

That is equivalent to

$$\mathbb{E}\Big[\exp\Big(\lambda\{\widetilde{S}^\tau_{h,j}(n+1)-\widetilde{S}^\tau_{h,j}(n)\}\Big)|\mathcal{F}_n\Big] = \exp\Big(\lambda S_n - n\varphi_\mu(\lambda)\Big).$$

Step 2: Peeling Let us devide the interval $\{1,\cdots,T\}$ into *slices* $\{t_{k-1}+1,\cdots,t_k\}$ of geometric increasing size. We may assume that $\delta>1$, since otherwise the bound is trivial. Take $\xi=1/(1-\delta_\eta(T))$, let $t_0=0$ and for all $k\in\mathbb{N}^*$, let $t_k=\lfloor(1+\xi)^k\rfloor$.

$$\begin{aligned}&\mathbb{P}\Big(\forall h\geqslant 0,\forall 0\leqslant j\leqslant K^h,\forall\ 1\leqslant t\leqslant T,\ U^\eta_{h,j}(t)\leqslant q_{h,j}(\tau)\Big)\\ \leqslant\ &\mathbb{P}\Big(\exists\ h\geqslant 0,\exists\ 0\leqslant j\leqslant K^h,\ \exists\ 1\leqslant t\leqslant T,\ U^\eta_{h,j}(t)\leqslant q_{h,j}(\tau)\Big).\end{aligned}$$

Define $m\leqslant T$ the number of nodes expanded throughout the algorithm, thus for $1\leqslant w\leqslant m$, it is possible to rewrite the last probability as

$$\begin{aligned}&\mathbb{P}\Big(\exists\ 1\leqslant w\leqslant T,\ \exists\ 1\leqslant n\leqslant T,\ U^\eta_w(n)\leqslant q_w(\tau)\Big)\\ \leqslant\ &\sum_{w=1}^{T}\mathbb{P}\Big(\ \exists\ 1\leqslant k\leqslant D,\ \exists\ t_{k-1}<n\leqslant t_k\ \text{ and }\ U^\eta_w(n)\leqslant q_w(\tau)\Big)\quad\text{with}\ \ D=\frac{\log(T)}{\log(1+\eta)}\\ \leqslant\ &\sum_{w=1}^{T}\sum_{k=1}^{D}\mathbb{P}\Big(A_k\Big),\end{aligned}$$

with

$$A_k=\Big\{\ \exists\ t_{k-1}<n\leqslant t_k\ \text{ and }\ U^\eta_w(n)\leqslant q_w(\tau)\Big\}.$$

Observe that $U^\eta_w(n)\leqslant q_w(\tau)$ if and only if $\frac{1}{n}\sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leqslant U^\eta_w}\leqslant\frac{1}{n}\widetilde{S}^\tau_w(n)$ and

$$\frac{1}{n}\sum_{s=1}^{u}\mathbb{1}_{Y_w(\zeta^s_w)\leqslant U^\eta_w}\leqslant\frac{\widetilde{S}^\tau_w(n)}{n}\Leftrightarrow\tau\leqslant\frac{\widetilde{S}^\tau_w(n)}{n}\ \text{ and }\ \mathrm{kl}(\frac{\widetilde{S}^\tau_w(n)}{n},\tau)\geqslant\delta_\eta(T)+\frac{1}{n}.$$

Define $\delta=\delta_\eta(T)+1/n$, let s be the smallest integer such that $\delta/(s+1)\leqslant\mathrm{kl}(1,\tau)$; if $n\leqslant s$, then $n\mathrm{kl}(\frac{\widetilde{S}^\tau_w(n)}{n},\tau)\leqslant s\mathrm{kl}(\frac{\widetilde{S}^\tau_w(n)}{n},\tau)\leqslant s\mathrm{kl}(1,\tau)<\delta$ thus $\mathbb{P}(U(n)<q(\tau))=0$. Thus for all k such that $t_k\geqslant s$, we obtain $\mathbb{P}(A_k=0)$. For k such that $t_k>s$, let $\widetilde{t}_{k-1}=\max\{t_{k-1},s\}$. Let $x\in]\tau,1[$ be such that $\mathrm{kl}(x,\tau)=\delta/n$ and let $\lambda(x)=\log(x(1-\tau))-\log(\tau(1-x))>0$, so that $\mathrm{kl}(x,\tau)=\lambda(x)x-(1-\tau+\tau\exp(\lambda(x)))$. Consider z such that $z>\tau$ and $\mathrm{kl}(z,\tau)=\delta/(1+\xi)^k$.

Observe that

- if $n > \widetilde{t}_{k-1}$, then
$$\mathrm{kl}(z,\tau) = \frac{\delta}{(1+\xi)^k} \geqslant \frac{\delta}{(1+\xi)n};$$
- if $n \leqslant t_k$, then as
$$\mathrm{kl}\big(\frac{\widetilde{S}_w^\tau(n)}{n},\tau\big) > \frac{\delta}{n} > \frac{\delta}{(1+\xi)^k} = \mathrm{kl}(z,\tau),$$

it holds that:
$$\tau \leqslant \frac{\widetilde{S}_w^\tau(n)}{n} \quad \text{and} \quad \mathrm{kl}(\frac{\widetilde{S}_w^\tau(n)}{n},\tau) \geqslant \frac{\delta}{n} \Rightarrow \frac{\widetilde{S}_w^\tau(n)}{n} \geqslant z.$$
Hence on the event $\{\widetilde{t}_{k-1} < n < t_k\} \cap \{\tau \leqslant \frac{\widetilde{S}_w^\tau(n)}{n}\} \cap \{\mathrm{kl}(\frac{\widetilde{S}_w^\tau(n)}{n},\tau) \geqslant \frac{\delta}{n}\}$ it holds that
$$\lambda(z)\frac{\widetilde{S}_w^\tau(n)}{n} \geqslant \lambda(z)z - \varphi_\tau(\lambda(z)) = \mathrm{kl}(z,\tau) \geqslant \frac{\delta}{(1+\xi)n}.$$

Step 3: Putting everything together
$$\begin{aligned}
&\{\widetilde{t}_{k-1} < n < t_k\} \cap \{\tau \leqslant \frac{\widetilde{S}_w^\tau(n)}{n}\} \cap \{\mathrm{kl}(\frac{\widetilde{S}_w^\tau(n)}{n},\tau) \geqslant \frac{\delta}{n}\} \\
&\quad \subset \{\lambda(z)\frac{\widetilde{S}_w^\tau(n)}{n} - \varphi_\tau(\lambda(z)) \geqslant \frac{\delta}{n(1+\xi)}\} \\
&\quad \subset \{\lambda(z)S_w(n) - n\varphi_\tau(\lambda(z)) \geqslant \frac{\delta_\eta(T)}{(1+\xi)}\} \\
&\quad \subset \{W_n^{\lambda(z)} > \exp(\frac{\delta_\eta(T)}{(1+\xi)})\}.
\end{aligned}$$

As $(W_n^\lambda)_{n\geqslant 0}$ is a martingale, $\mathbb{E}[W_n^{\lambda(z)}] \leqslant \mathbb{E}[W_0^{\lambda(z)}] = 1$. Thus the Doob's inequality for martingales provides:
$$\mathbb{P}\left(\sup_{\widetilde{t}_{k-1}<n<t_k} W_n^{\lambda(z)} > \exp\Big(\frac{\delta_\eta(T)}{1+\xi}\Big)\right) \leqslant \exp\Big(-\frac{\delta_\eta(T)}{1+\xi}\Big)$$

Finally
$$\sum_{w=1}^{T}\sum_{k=1}^{D} \mathbb{P}\Big(\exists\, t_{k-1} < n \leqslant t_k \ \text{ and } \ U_w^\eta(n) \leqslant q_w(\tau)\Big) \leqslant TD\exp(-\frac{\delta_\eta(T)}{(1+\xi)}).$$

But as $\xi = 1/(\delta_\eta(T)-1)$, $D = \Big\lceil \dfrac{\log(T)}{\log(1+1/(\delta_\eta(T)+1))} \Big\rceil$ and as long as
$$\log(1+1/(\delta_\eta(T)-1)) \geqslant 1/\delta_\eta(T),$$
we obtain:
$$\mathbb{P}(\mathcal{A}^c) \leqslant T\Big\lceil \frac{\log(T)}{\log(1+1/(\delta_\eta(T)+1))} \Big\rceil \exp(-\delta_\eta(T)+1) \leqslant Te^{\lceil \delta_\eta(T)\log(T)\rceil}\exp(-\delta_\eta(T)) \leqslant \eta/2.$$
Using the same lines for the LCB concludes the proof. □

3.9.4 Proofs related to the section Optimizing CVaR

Proof. of Proposition 3.6.1

Let us consider the event

$$\begin{aligned}\xi_\eta \;=\; &\Big\{\forall\ h \geqslant 0, \forall\ 0 \leqslant j \leqslant K^h, \forall\ 1 \leqslant t \leqslant T,\\ &\widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \geqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) + \widetilde{\varepsilon}^{\eta}_{N_{h,j}(t)} \text{ or } \widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \leqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) - \varepsilon^{\eta}_{N_{h,j}(t)}\Big\}.\end{aligned}$$

$$\begin{aligned}\mathbb{P}\big(\xi_\eta\big) = \mathbb{P}\Big(&\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \geqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) + \widetilde{\varepsilon}^{\eta}_{N_{h,j}(t)} \text{ or },\\ &\widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \leqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) - \varepsilon^{\eta}_{N_{h,j}(t)}\Big)\\ \leqslant\ &\mathbb{P}\Big(\forall h \leqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \geqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) + \widetilde{\varepsilon}^{\eta}_{N_{h,j}(t)}\Big) \qquad (3.12)\\ &+ \mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \leqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) - \varepsilon^{\eta}_{N_{h,j}(t)}\Big) \qquad (3.13)\end{aligned}$$

First let us consider (3.12):

$$\begin{aligned}&\mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \geqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) + \widetilde{\varepsilon}^{\eta}_{N_{h,j}(t)}\Big)\\ \leqslant\ &\mathbb{P}\Big(\exists\ 1 \leqslant w \leqslant T,\ \exists\ 1 \leqslant u \leqslant T,\ \inf_{z\in\mathbb{R}}\{z + \frac{1}{u(1-\tau)}\sum_{s=1}^{u}(Y_w(\zeta_w^s) - z)^+\} \geqslant \mathrm{CVaR}_\tau(Y_{x_w}) + \widetilde{\varepsilon}^{\eta}_u\Big)\\ \leqslant\ &\sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(\inf_{z\in\mathbb{R}}\{z + \frac{1}{u(1-\tau)}\sum_{s=1}^{u}(Y_w(\zeta_w^s) - z)^+\} \geqslant \mathrm{CVaR}_\tau(Y_{x_w}) + \widetilde{\varepsilon}^{\eta}_u\Big).\end{aligned}$$

Thus by Brown's inequality

$$(3.12) < \sum_{w=1}^{T}\sum_{u=1}^{T}\exp(-2(\tau\widetilde{\varepsilon}^{\eta}_u/(b-a))^2 u).$$

Taking

$$\widetilde{\varepsilon}^{\eta}_u = \frac{(b-a)}{\tau}\sqrt{\frac{\log(2T^2/\eta)}{2u}}$$

provides the first part, *i.e* $(3.12) < \frac{\eta}{2}$.

We use the same scheme of proof to bound (3.13), the only difference comes from the fact that the inequality of deviation is different:

$$\begin{aligned}
&\mathbb{P}\Big(\forall h \geqslant 0, \forall 0 \leqslant j \leqslant K^h, \forall 1 \leqslant t \leqslant T,\ \widehat{\mathrm{CVaR}_\tau}^{\,t}(Y_{x_{h,j}}) \leqslant \mathrm{CVaR}_\tau(Y_{x_{h,j}}) - \varepsilon^\eta_{N_{h,j}(t)}\Big) \\
\leqslant\ &\mathbb{P}\Big(\exists\ 1 \leqslant w \leqslant T,\ \exists\ 1 \leqslant u \leqslant T,\ \inf_{z\in\mathbb{R}}\{z + \frac{1}{u(1-\tau)}\sum_{s=1}^{u}(Y_w(\zeta_w^s) - z)^+\} \leqslant \mathrm{CVaR}_\tau(Y_{x_w}) - \varepsilon^\eta_u\Big) \\
\leqslant\ &\sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(\inf_{z\in\mathbb{R}}\{z + \frac{1}{u(1-\tau)}\sum_{s=1}^{u}(Y_w(\zeta_w^s) - z)^+\} \leqslant \mathrm{CVaR}_\tau(x_w) - \varepsilon^\eta_u\Big).
\end{aligned}$$

By Brown's inequality

$$(3.13) < \sum_{w=1}^{T}\sum_{u=1}^{T} 3\exp\Big(-\frac{\tau}{5}\Big(\frac{\varepsilon^\eta_u}{b-a}\Big)^2 u\Big)$$

Taking

$$\widetilde{\varepsilon}^\eta_u = (b-a)\sqrt{\frac{5\log(6T^2/\eta)}{\tau u}}$$

provides $(3.13) < \frac{\eta}{2}$.

Finally putting (3.12) and (3.13) together provides $\mathbb{P}\big(\xi_\eta\big) < \eta$ and hence $\mathbb{P}(\xi_\eta^c) = \mathbb{P}(\mathcal{A}_\eta) = 1 - \eta$.

□

Proof. of Proposition 3.6.3 If $Y_1 \cdots, Y_n$ are i.i.d random variables bounded by (a, b) then Thomas-Learned-Miller's inequalities provide

$$\mathbb{P}\Big(-\mathrm{CVaR}_\tau < \frac{1}{1-\tau}\sum_{i=1}^{n}(Y_{i+1} - Y_i)\Big(\frac{i}{n} - \sqrt{\frac{\log(1/\eta)}{2n}} - \tau\Big)^+ - Y_{n+1}\Big) < \eta$$

and

$$\mathbb{P}\Big(-\mathrm{CVaR}_\tau > \frac{1}{1-\tau}\sum_{i=0}^{n-1}(Y_{i+1} - Y_i)\Big(\min\big\{1, \frac{i}{n} + \sqrt{\frac{\log(2T^2/\eta)}{2N_{h,j}(t)}}\big\} - \tau\Big)^+ - Y_n\Big) < \eta.$$

Define

$$\xi_{\eta,1} = \{\forall\ h \geqslant 0, \forall\ 0 \leqslant j \leqslant K^h, \forall\ 1 \leqslant t \leqslant T, -\mathrm{CVaR}_\tau(Y_{h,j}) < U^\eta_{N_{h,j}(t)}\},$$

and

$$\xi_{\eta,2} = \{\forall\ h \geqslant 0, \forall\ 0 \leqslant j \leqslant K^h, \forall\ 1 \leqslant t \leqslant T, -\mathrm{CVaR}_\tau(Y_{h,j}) > L^\eta_{N_{h,j}(t)}\},$$

To treat the sequential point of view, here we use a double union bound as it is done in the proof of Proposition 13, then it can be shown that

$$\mathbb{P}(\xi_{\eta,1}) < \sum_{w=1}^{T}\sum_{u=1}^{T}\mathbb{P}\Big(- \mathrm{CVaR}_\tau(Y_w^u) < U_u^\eta\Big).$$

Thus by defining

$$U_u^\eta = \frac{1}{1-\tau}\sum_{i=0}^{u-1}(Y_{i+1}-Y_i)\Big(\min\big\{1, \frac{i}{u} + \sqrt{\frac{\log(2T^2/\eta)}{2u}}\big\} - \tau\Big)^+ - Y_u$$

we obtain

$$\mathbb{P}(\xi_{\eta,1}) < \sum_{w=1}^{T}\sum_{u=1}^{T}\frac{\eta}{2T^2} = \frac{\eta}{2}.$$

Using the same scheme of proof with

$$L_u^\eta = \frac{1}{1-\tau}\sum_{i=1}^{u}(Y_{i+1}-Y_i)\Big(\frac{i}{u} - \sqrt{\frac{\log(2T^2/\eta)}{2u}} - \tau\Big)^+ - Y_{u+1}$$

provides

$$\mathbb{P}(\xi_{\eta,2}) < \frac{\eta}{2}.$$

Finally

$$\mathbb{P}(\xi_{\eta,1}\cup\xi_{\eta,1}) < \eta,$$

and hence $\mathbb{P}\Big((\xi_{\eta,1}\cup\xi_{\eta,1})^c\Big) = \mathbb{P}(\mathcal{A}_\eta) = 1-\eta.$ □

Chapter 4

Bayesian Quantile and Expectile Optimisation

Contents

4.1 Résumé

Ce chapitre reprend un article soumis à la conférence AISTAT 2020. Dans le chapitre précédent nous avons proposé un algorithme d'optimisation de mesures de risque sur lequel nous pouvons dériver une borne supérieure sur son regret. La recherche de garanties théorique pouvant rendre les algorithmes trop conservatifs, dans ce chapitre nous considérons un cadre plus heuristique basé sur l'optimisation à base de métamodèles gaussiens. L'objectif est la création d'un algorithme capable d'optimiser un quantile ou un expectile conditionnel dans un régime *small data*.

Ce travail est réalisé conjointement avec Victor Picheny et Nicolas Durrande.

4.2 Introduction

Let $\Psi : \mathcal{X} \times \Omega \to \mathbb{R}$ be an unknown function , where $\mathcal{X} \subset [0,1]^D$ and Ω denotes a probability space representing some uncontrolled variables. For any fixed $x \in \mathcal{X}$, $Y_x = \Psi(x, \cdot)$ is a random variable of distribution $\mathbb{P}_x$. We assume here a classical *black-box optimisation* framework: Ψ is available only through (costly) pointwise evaluations of Y_x, and no gradient or structural information are available. Typical examples may include stochastic simulators in physics or biology (see Skullerud [1968] for simulations of ion motion and Székely Jr and Burrage [2014] for simulations of heterogeneous natural systems), but Ψ can also represent the performance of a machine learning algorithm according to some hyperparameters (see Bergstra et al. [2011], Li et al. [2016] for instance). In the latter case, the randomness can come from the use of minibaching in the training procedure, the choice of a stochastic optimiser or the randomness in the optimisation initialisation.

Let $g(x) = \rho(\mathbb{P}_x)$ be the objective function we want to maximise, where ρ is a real-valued functional defined on probability measures. The canonical choice for ρ is the conditional expectation (i.e. conditional on x), which is sensible when the exposition to extreme values is not a significant aspect of the decision. However, in a large variety of fields such as agronomy, medicine or finance, the decision maker has an incentive to protect himself against extreme events which typically have little influence on the expectation but that can lead to severe consequences. To take these rare events into account, one should consider alternative choices for ρ that depend on the tails of $\mathbb{P}_x$, such as the quantile [Rostek, 2010], conditional value-at-risk (CVaR, see Rockafellar et al. [2000]) or expectile [Bellini and Di Bernardino, 2017]. In this paper we focus our interest on the conditional quantile and expectile.

Given an estimate of g based on available data, global optimization algorithms define a policy that finds a trade-off between exploration and intensification. More precisely, the algorithm has to explore the input space in order to avoid getting trapped in a local optimum, but it also has to concentrate its budget on input regions identified as having a high potential. The latter results in accurate estimates of g in the region of interest and allow the algorithm to return an optimal input value with high precision.

In the context of Bayesian optimisation (BO), such trade-offs have been initially

studied by Mockus et al. [1978] and Jones et al. [1998] in a noise-free setting. Their framework has latter been extended to optimisation of the conditional expectation of a stochastic black box (see e.g. Frazier et al. [2009], Srinivas et al. [2009] or Picheny et al. [2013] for a review). Although the literature is very scarce for quantile or expectile optimization under the BO framework, an algorithm based on Gaussian processes for quantile optimization is presented in Browne et al. [2016]. The approach they propose however requires many replications per input point and is not compatible with small-data settings.

Contributions The contributions of this paper are the following: 1) We propose a new metamodel based on variational inference to estimate either quantiles or expectiles without repetitions in the design of experiment and suited to the heteroscedastic case. 2) We propose a new Bayesian algorithm suited to optimise conditional quantiles or expectiles in a small data setting. Two batch-sequential acquisition strategies are designed to find a good trade-off between exploration and intensification. The ability of our algorithm to optimise quantiles and expectiles is illustrated on several test problems.

4.3 Bayesian metamodels for tails dependant measures

The conditional quantile of order $\tau \in (0,1)$ can be defined as:

$$q_\tau(x) = \underset{q\in\mathbb{R}}{\arg\min}\, \mathbb{E}\big[l_\tau(Y_x - q)\big], \tag{4.1}$$

with

$$l_\tau(\xi) = (\tau - \mathbb{1}_{(\xi<0)})\xi, \quad \xi \in \mathbb{R}, \tag{4.2}$$

the so-called pinball loss introduced by Koenker and Bassett Jr [1978]. Following this formalism, Newey and Powell [1987] introduced the square pinball loss defined as

$$l^e_\tau(\xi) = |\tau - \mathbb{1}_{(\xi<0)}|\xi^2, \quad \xi \in \mathbb{R}, \tag{4.3}$$

and the expectile of order τ as

$$e_\tau(x) = \underset{q\in\mathbb{R}}{\arg\min}\, \mathbb{E}\big[l^e_\tau(Y_x - q)\big]. \tag{4.4}$$

We detail in the next section how these losses can be used to get an estimate of $g(x)$ using a dataset $\mathcal{D}_n = \big((x_1, y_1)\cdots, (x_n, y_n)\big) = (\mathcal{X}_n, \mathcal{Y}_n)$ with $\mathcal{X}_n$ a $n \times D$ matrix.

4.3.1 Quantile and Expectile Metamodel

To estimate a conditional quantile of fixed order, different metamodels such as artificial neural networks [Cannon, 2011], random forest [Meinshausen, 2006] or nonparametric estimation in reproducing kernel Hilbert spaces [Takeuchi et al., 2006, Sangnier et al., 2016] have been proposed. While the literature on expectile regression is less extended,

neural network [Jiang et al., 2017] or SVM-like approaches [Farooq and Steinwart, 2017] have been developed as well. All the approaches cited above defined an estimator of g as the function that minimises (optionally with a regularization term)

$$\mathcal{R}_e[g] = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i - g(x_i)\big), \tag{4.5}$$

with $l = l_\tau$ for the quantile estimation and $l = l_\tau^e$ for the expectile. This framework makes sense because asymptotically minimising (4.5) is equivalent to minimising (4.4) or (4.1).

All these approaches however have a drawback: they do not quantify the uncertainty associated with each prediction. This is a significant problem in our settings since this knowledge is of paramount importance to define the exploration/intensification trade-off. Alternatively, using Bayesian models can overcome this issue as they provide a posterior distribution on g. To do so, Yu and Moyeed [2001] proposed the model:

$$y = q(x) + \varepsilon(x), \quad \text{with} \quad q(x) = x^T\alpha,$$

$\alpha \in \mathbb{R}^D$ with an improper uniform prior and ε a random variable that follows an asymmetric Laplace distribution, *i.e.*

$$p\big(y|q_\tau, \sigma, x\big) = \frac{\tau(1-\tau)}{\sigma}\exp\left(-\frac{l_\tau(y - q_\tau(x))}{\sigma}\right).$$

The associated likelihood is given by

$$p\big(\mathcal{Y}_n|q_\tau, \mathcal{X}_n, \sigma\big) = \prod_{i=1}^{n}\frac{\tau(1-\tau)}{\sigma}\exp\left(-\frac{l_\tau(y_i - q_\tau(x_i))}{\sigma}\right). \tag{4.6}$$

Then an estimator of q_τ is taken as the function that maximises this likelihood. This model is intuitive for two reasons. First ε is asymmetric, such that $\mathbb{P}(Y \leqslant q) = \tau$ and $\mathbb{P}(Y \geqslant q) = 1 - \tau$. Second, minimising the empirical risk associated to the pinball loss (4.5) is equivalent to maximising the asymmetric Laplace likelihood (4.6). To the best of our knowledge, there are no Bayesian expectile models in the literature. However, similarly to quantiles, it is possible to use the asymmetric Gaussian distribution defined as

$$p(y|e_\tau, \sigma, x) = C(\tau, \sigma)\exp\left(-\frac{l_e^\tau\big(y_i - g(x_i)\big)}{2\sigma^2}\right), \tag{4.7}$$

with

$$C(\tau, \sigma) = \frac{\sqrt{2\tau(1-\tau)}}{\sigma\sqrt{\pi}(\sqrt{\tau} + \sqrt{1-\tau})}.$$

To estimate these models, we can refer to the existing methods for the quantile: see for instance Boukouvalas et al. [2012], Abeywardana and Ramos [2015]. In the review of Torossian et al. [2019b], quantile estimation using variational approaches and a Gaussian process prior for g appears to be one of the most competitive approach on the

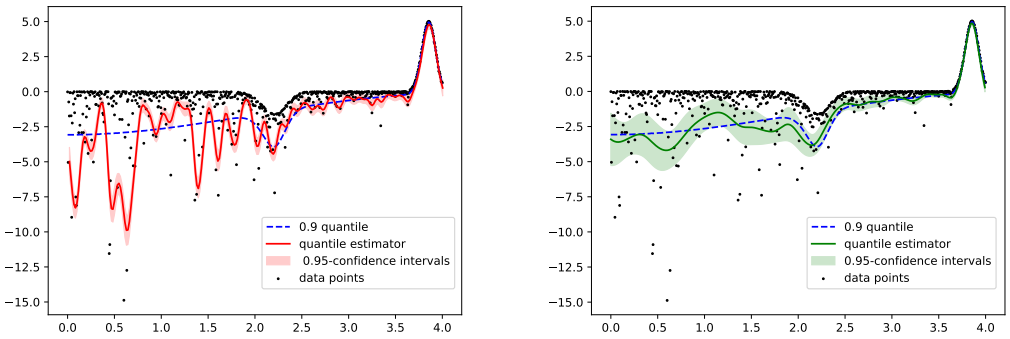


Figure 4.1: Original GP quantile model from Abeywardana and Ramos [2015] (left) and chained GP (right) on a very heteroscedastic model. The model on the left cannot compromise between very small observation variances around $x = 4$ and very large variances ($x \leq 2$) and largely overfits on half of the domain, while returning overconfident confidence intervals. The chained GP model captures the low variance region and the high variance one, while returning well-calibrated confidence intervals.

considered benchmark. However although GPs theoretically provide confidence intervals, the original metamodel of Abeywardana and Ramos [2015] seems to be overconfident in heteroscedastic problems as presented Figure 4.1. The main reason for this is the use (also present in the aforementioned papers) of a single spread parameter σ for the likelihood function. This amounts to considering the spread of Y_x as constant over $\mathcal{X}$, which is particularly harmful as quantile optimisation precisely aim at leveraging the varying spread of Y_x.

To overcome this issue and to propose a relevant model for heteroscedastic cases, it is necessary to add a degree of freedom to the spread of ε and make it dependent on the variance of Y_x. For both the asymmetric Laplace and asymmetric Gaussian distributions, it can be done simply by defining σ in equations 4.6 and 4.7 as a function of the input parameters. Intuitively, using a small σ creates a very confident model that tends to interpolate the data while using a large σ will add regularity in the model and produce a more robust estimate. To incorporate such flexibility, we propose a model with a GP prior on g and on the scale parameter σ, *i.e.*

$$g(x) \sim \mathcal{GP}\big(0, k_\theta^g(x, x')\big), \tag{4.8}$$

$$\sigma(x) = s_{\max}\big(r(x)\big) \text{ where } r(x) \sim \mathcal{GP}\big(0, k_\theta^\sigma(x, x')\big), \tag{4.9}$$

where $s_{\max}$ is the *softmax* function. It is used in order to ensure the positivity of σ. Estimating the parameters of this model may appear challenging, but the chained GP formalism introduced by Saul et al. [2016] provides the appropriate framework to do so.

4.3.2 Inference Procedure

Although our study is limited to the small data regime, quantile and expectile regression are much more challenging problems than classical regression. In the review on quantile regression of Torossian et al. [2019b], the typical budget to perform estimation is defined as 50 times the input dimension, while a classical rule-of-thumb for GP regression is 10 times the dimension [Loeppky et al., 2009]. As we wish to propose a scalable algorithm and as optimisation naturally needs more points than regression, we need a model able to train on large datasets (say, $n \geq 1,000$). In addition, the concentration of points which results of the intensification part of our optimisation scheme would produce instability during the computation of the covariance matrix. To handle these two potential issues, it is natural to use the classical inducing point approach. Following Snelson and Ghahramani [2006], Titsias [2009], Hensman et al. [2013] we introduce N 'pseudo inputs' (named inducing points) at location $\boldsymbol{z} = \{z_i\}_{i=1}^N$ and the corresponding output $\boldsymbol{u}_g = g(\boldsymbol{z})$ and $\boldsymbol{u}_r = r(\boldsymbol{z})$. The marginal likelihood is thus provided as

$$p(\mathcal{Y}_n) = \int p(\mathcal{Y}_n|g,\sigma)p(g,\sigma,\boldsymbol{u}_r,\boldsymbol{u}_g)dgd\sigma d\boldsymbol{u}_g d\boldsymbol{u}_r,$$

with $p(g,\sigma,\boldsymbol{u}_r,\boldsymbol{u}_g) = p(g,\sigma|\boldsymbol{u}_g,\boldsymbol{u}_\sigma)p(\boldsymbol{u}_g,\boldsymbol{u}_r)$. This later quantity is not analytically tractable because the likelihoods introduced to estimate quantiles and expectiles are not conjugated with the Gaussian likelihood related to the assumptions (4.8) and (4.9). Thus to estimate the parameters of the model we use a variational black box formalism with a stochastic optimisation scheme as introduced in Saul et al. [2016], Hensman et al. [2013]. Assuming the mean field approximation for g and σ implies

$$p(g,\sigma,\boldsymbol{u}_r,\boldsymbol{u}_g) = p(g|\boldsymbol{u}_g)\widetilde{p}(\boldsymbol{u}_g)p(\sigma|\boldsymbol{u}_r)\widetilde{p}(\boldsymbol{u}_r).$$

It results the following evidence lower bound (ELBO)

$$\log p(\mathcal{Y}_n) \geqslant \int \widetilde{p}(g)\widetilde{p}(\sigma)\log p(y|g,\sigma)dgd\sigma \\ - \mathrm{kl}\left(\widetilde{p}(\boldsymbol{u}_g)||p(\boldsymbol{u}_g)\right) - \mathrm{kl}\left(\widetilde{p}(\boldsymbol{u}_r)||p(\boldsymbol{u}_r)\right).$$

The posterior on $\boldsymbol{u}_g$ and $\boldsymbol{u}_r$ is assumed to be Gaussian,

$$\widetilde{p}(\boldsymbol{u}_g) \sim \mathcal{N}(\boldsymbol{u}_g|\mu_g, S_g) \quad \text{and} \quad \widetilde{p}(\boldsymbol{u}_r) \sim \mathcal{N}(\boldsymbol{u}_r|\mu_r, S_r),$$

with μ_r, μ_g in $\mathbb{R}^N$ and S_g, S_r in $\mathbb{R}^{N\times N}$ the variational quantities that are fully parametrised. Next, because the considered distributions follow Gaussian priors, we obtain

$$\begin{aligned} p(g|\boldsymbol{u}_g) &= \mathcal{N}(g|K_{g,\boldsymbol{u}_g}K^{-1}_{\boldsymbol{u}_g,\boldsymbol{u}_g}u_g, K_{g,g} - Q_g) \\ p(r|\boldsymbol{u}_r) &= \mathcal{N}(r|K_{g,\boldsymbol{u}_r}K^{-1}_{\boldsymbol{u}_r,\boldsymbol{u}_r}\boldsymbol{u}_r, K_{r,r} - Q_r), \end{aligned}$$

where for $j = (g,\ r)$, $Q_j = K_{j,\boldsymbol{u}_j}K^{-1}_{\boldsymbol{u}_j,\boldsymbol{u}_j}K_{\boldsymbol{u}_j,j}$.

Finally the approximation of the posterior is

$$\begin{aligned}\widetilde{p}(g) &= \mathcal{N}(g|K_{g,\boldsymbol{u}_g}K^{-1}_{\boldsymbol{u}_g,\boldsymbol{u}_g}\mu_g, K_{g,g}+\widehat{Q}_g)\\ \widetilde{p}(r) &= \mathcal{N}(r|K_{r,\boldsymbol{u}_r}K^{-1}_{\boldsymbol{u}_r,\boldsymbol{u}_r}\mu_r, K_{r,r}+\widehat{Q}_r),\end{aligned}$$

where $\widehat{Q}_j = K_{j,\boldsymbol{u}_j}K^{-1}_{\boldsymbol{u}_j,\boldsymbol{u}_j}(S_j - K_{\boldsymbol{u}_j,\boldsymbol{u}_j})K^{-1}_{\boldsymbol{u}_j,\boldsymbol{u}_j}K_{\boldsymbol{u}_j,j}$.

To compute the intractable approximation of the log-likelihood $\int \widetilde{p}(q)\widetilde{p}(\sigma)\log p(\mathcal{Y}_n|q,\sigma)dqd\sigma$, it is possible to take advantage of the factorized form of our likelihood across the data in order to optimize stochastically an equivalence of the ELBO provided by

$$\sum_{i=1}^{n}\int \widetilde{p}(g_i)\widetilde{p}(\sigma_i)\log p\big(y_i|g,\tau,\sigma,x_i\big) - \mathrm{kl}\left(\widetilde{p}(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\right) - \mathrm{kl}\left(\widetilde{p}(\boldsymbol{u}_\sigma)||p(\boldsymbol{u}_\sigma)\right).$$

Note that due to the non differentiability of the pinball loss at the origin, the lower bound is not differentiable everywhere. We thus use a first order optimizer (ADAM optimiser Kingma and Ba [2014]) as it does not need the objective function to have continuous derivative. To estimate $\widetilde{p}(q_i)$ and $\widetilde{p}(\sigma_i)$ for $i = 1, \cdots, n$ we use a quadrature approximation.

To make predictions at a query point x_*, following Section 1.8.2 it is possible to write

$$\begin{aligned}\widetilde{p}(g|x_*) &= \mathcal{N}(K_{x_*,\boldsymbol{u}_g}K^{-1}_{\boldsymbol{u}_g,\boldsymbol{u}_g}\mu_g, K_{x_*,x_*}+\widehat{Q}^*_g)\\ \widetilde{p}(r|x_*) &= \mathcal{N}(K_{x_*,\boldsymbol{u}_r}K^{-1}_{\boldsymbol{u}_r,\boldsymbol{u}_r}\mu_r, K_{x_*,x_*}+\widehat{Q}^*_r),\end{aligned}$$

where $\widehat{Q}^*_j = K_{x_*,\boldsymbol{u}_j}K^{-1}_{\boldsymbol{u}_j,\boldsymbol{u}_j}(S_j - K_{\boldsymbol{u}_j,\boldsymbol{u}_j})K^{-1}_{\boldsymbol{u}_j,\boldsymbol{u}_j}K_{\boldsymbol{u}_j,x_*}$.

4.4 Bayesian optimisation

Classical BO algorithms work as follow. First, a posterior distribution on g is inferred from an initial set of experiments $\mathcal{D}_n$ (typically obtained using a space-filling design). Then the next input point to evaluate is chosen as the maximiser of an *acquisition function*, computed from the g posterior. The objective function is sampled at the chosen input and the posterior on g is updated. These steps are repeated until the budget is exhausted.

The efficiency of such strategies depends on the relevance of the g posterior but also on the exploration/exploitation trade-off provided by the acquisition function. Many acquisition functions have been designed to fit this trade off, among them the *Expected improvement* [EI, Jones et al., 1998], *upper confidence bound* [UCB, Srinivas et al., 2009], *knowledge gradient* [KG Frazier et al., 2009] or *Entropy search* [PES Hernández-Lobato et al., 2014]. In the case of quantiles and expectiles, adding points one at a time may be impractical, as many points may be necessary to modify significantly the g posterior. One solution is to rely on replications, i.e. evaluating repeatedly Y a single input, as

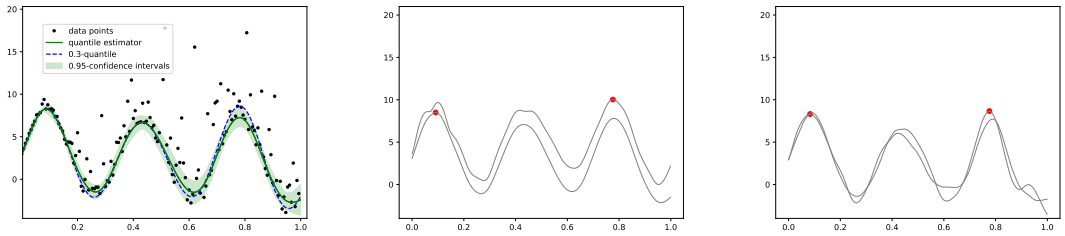


Figure 4.2: Left: estimator of the 0.3-quantile and the associated confidence intervals; middle: two UCB (with resp. $\beta = 5$ and $\beta = 1$) with different maximisers (red); right: two sample trajectories of g using RFF, with different maximisers.

in Browne et al. [2016], Wang et al. [2019]. However, in Torossian et al. [2019b] using replications was clearly found less efficient than using distributed observations.

All of the above-mentioned acquisition functions have been extended to batches of points: see for instance Ginsbourger et al. [2010], Marmin et al. [2015] for EI, [Wu and Frazier, 2016] for KG or Contal et al. [2013], Desautels et al. [2014] for UCB. However, none actually fit our settings for two main reasons. First, most parallel acquisitions make use of explicit update equations for the GP moments, which are not available for our model. Second, most are designed for small batches (say, ≤ 10) and become numerically intractable for larger batches (say, 100), which is our aim.

We propose in the following two alternatives, one based on a simple adaptation of UCB, the other based on the Thompson sampling approach proposed by Hernández-Lobato et al. [2017].

4.4.1 Batch GP-UCB via Multiple Optimism Levels

Assume the posterior on g is a GP of mean μ and covariance matrix $\sigma_c(x, x')$ then the classical UCB acquisition function is

$$\arg\max_{x \in \mathcal{X}} \widehat{g}(x) + \beta_t \sqrt{\sigma_c(x, x)}, \tag{4.10}$$

with β_t a positive hyperparameter that tunes the trade-off between exploration (large β_t, implying more weight on the variance) and exploitation (small β_t, implying more weight on the mean).

A simple way to parallelise this criterion consists in selecting different values of β_t at the same time. Denoting B the batch size, we choose $\boldsymbol{\beta}_t = (\beta_1, \cdots, \beta_B)$ as: $\beta_i = \Phi^{-1}(0.5 + \frac{i}{2(b+1)})$ $(1 \leq i \leq B)$, with Φ^{-1} the inverse of the cumulative distribution function of the standard Gaussian distribution. Intuitively, each batch of new inputs is based on a gradient of exploration / exploitation trade-offs. This idea is represented at the center of Figure 4.2. However such values of $\boldsymbol{\beta}$ are to small to guaranty the exploration thus we finally multiply it by $5D$. Algorithm 9 presents the pseudo-code for this strategy.

Contrary to Srinivas et al. [2009], Contal et al. [2013], Desautels et al. [2014], due to the chained GP framework our approach does not have theoretical guarantees. However, this might have a limited practical effect, as the theoretically sound values for β_t are known to be overly conservative and typical algorithms use constant β_t's.

Algorithm 9: Risk Parallel GP-UCB

Input: initial data $\mathcal{D}_I$; batch size B
for $t = 1$ **to** T **do**
  Compute the posterior $p(g|\mathcal{D}_{I(t)}) = \mathcal{GP}(\boldsymbol{\mu}, \boldsymbol{\sigma_c})$;
  for $b = 1$ **to** B **do**
    Select $\beta = \beta_b$;
    $x(b) \leftarrow \arg\max_{x \in \mathcal{X}} \mu(x) + \beta\sqrt{\sigma_c(x,x)}$;
    Observe $y_{x(b)}$ by sampling Ψ at $x(b)$;
  end
  $\mathcal{D}_{I(t+1)} = \mathcal{D}_{I(t)} \cup \{x(b), y_{x(b)}\}_{b=1}^{B}$;
end

4.4.2 Thompson Sampling

In this section we adapt the parallel Thompson Sampling strategy of Hernández-Lobato et al. [2017] to the case of the *Chained GPs* with a Matérn prior on the kernel.

Given the posterior on g, an intuitive approach is to sample Ψ according to the probability that $x = x^*$. However this distribution is usually intractable. Alternatively, one may achieve the same goal by sampling a trajectory from the posterior of g and selects the input that corresponds to its maximiser. Such approach directly extends to batches of inputs, by drawing several strategies and selecting all the maximisers. Algorithm 10 illustrates this strategy.

Algorithm 10: Risk Parallel Thompson Sampling

Input: initial data $\mathcal{D}_I$; batch size B
for $t = 1$ **to** T **do**
  Compute the posterior $p(g|\mathcal{D}_{I(t)}) = \mathcal{GP}(\boldsymbol{\mu}, \boldsymbol{\sigma_c})$;
  for $b = 1$ **to** B **do**
    Sample the trajectory tr_b according to $p(g|\mathcal{D}_{I(t)})$;
    $x(b) \leftarrow \arg\max_{x \in \mathcal{X}} \mathrm{tr}_b(x)$;
    Observe $y_{x(b)}$ by sampling Ψ at $x(b)$;
  end
  $\mathcal{D}_{I(t+1)} = \mathcal{D}_{I(t)} \cup \{x(b), y_{x(b)}\}_{b=1}^{B}$;
end

The main difficulty of this strategy lies in the creation of sample trajectories of g.

It is well-known that the values of trajectories from $\mathcal{GP}(\mu, \Sigma)$ can be obtained on any discrete set $\mathbb{X}$ of size M using

$$\mathrm{tr}(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \boldsymbol{\Sigma}^{1/2}\boldsymbol{N},$$

where $\boldsymbol{N} = (N_1, \cdots, N_M)$ is a vector of independent standard Gaussian samples, $\boldsymbol{\Sigma}^{1/2}$ is the lower triangular matrix of the Cholesky decomposition of Σ evaluated on $\mathbb{X}$. But this framework has two drawbacks. First the obtained trajectories are not continuous functions, the optimisation can only be made over the discrete set. Second, as $\boldsymbol{\Sigma}^{1/2}$ is obtained with a Cholesky decomposition, defining such trajectories has a $O(M^3)$ cost [Diggle et al., 1998]. So this approach quickly meets its limitations as the dimension of $\mathcal{X}$ increases and cannot be well represented by $\mathbb{X}$.

To overcome these drawbacks, a solution is to go back to the parametric formulation of g and to use the random Fourier features (RFF) to approximate the kernel k, as it is presented in Rahimi and Recht [2008]. Let us introduce Bochner's theorem that asserts the existence of a dual formulation for a large class of kernels and in particular the Matérn family.

Theorem 4.4.1. *A continuous, shift-invariant kernel is positive definite if and only if it is the Fourier transform of a non-negative, finite measure.*

Thus, giving a stationary kernel k, there exists an associated *spectral density* s such that

$$k(x, x') = \int \exp\big(-iw^T(x - x')\big)s(w)dw$$

with

$$s(w) = \frac{1}{(2\pi)^d}\int \exp(iw^T r)k(r, 0)dr.$$

Note that s is not a probability density function because it is not normalized. It is possible to define $p(w) = s(w)/\alpha$ where the normalizing constant is $\alpha = \int s(w)dw$. Using this formulation enables to write

$$\begin{aligned} k(x, x') &= \alpha\mathbb{E}_{p(\omega)}\big(\exp\big(-iw^T(x - x')\big) \\ &= 2\alpha\mathbb{E}_{p(w,b)}\big(\cos(\omega^T x + b)\cos(\omega^T x' + b)\big), \end{aligned}$$

with $p(b) = \mathcal{U}(0, 2\pi)$. RFFs then consists in approximating this expectation using a Monte-Carlo estimate:

$$k(x, x') \approx \varphi(x)^T\varphi(x') \tag{4.11}$$

with $\varphi(x)$ a m-dimensional feature such that $\varphi_i(x) = \sqrt{2\alpha/m}\cos(w_i^T x + b_i)$ where w_i and b_i are i.i.d. samples from $p(\omega)$ and $p(b)$.

Such methodology has been classically used to approximate the squared-exponential kernel because it is self conjugated [see Hernández-Lobato et al., 2014]. Here we present how to use RFFs to approximate anisotropic Matern kernels. Our goal is to determine the spectral density associated to the Matern kernel of variance parameter $\sigma_M \in \mathbb{R}^+$ and length scales $\theta \in \mathbb{R}^D$.

In its simplest form the Matérn kernel is provided by

$$k(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}||x-x'||_2^{\nu}\mathcal{K}_\nu(||x-x'||_2),$$

with $\mathcal{K}_\nu$ the modified Bessel function of the second kind with order ν, Γ the gamma function and its Fourier transform [see Rasmussen, 2003, for more details] is given by

$$s(w) = \frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)}\frac{(2\sqrt{\pi})^d}{(1+w^2)^{\frac{d}{2}+\nu}}.$$

If $||x-x'||_2 = \sqrt{(x-x')^T\Lambda^{-1}(x-x')}$, with $\Lambda = \text{diag}(\theta_1,\cdots,\theta_D)$ the diagonal matrix containing the length scale hyperparameters, then the Fourier transform is

$$s(w) = |\Lambda|^{1/2}\frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)}\frac{(2\sqrt{\pi})^d}{(1+w^T\Lambda w)^{\frac{d}{2}+\nu}}.$$

Now it is possible to use $\Lambda' = 2\nu \times \Lambda$ that provides

$$s(w) = |\Lambda'|^{1/2}\frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)\nu^{d/2}}\frac{\sqrt{2\pi}^d}{(1+\frac{1}{2\nu}w^T\Lambda' w)^{\frac{d}{2}+\nu}},$$

next if we define $\alpha = (\sqrt{2}\pi)^d$ we obtain $s(w) = \alpha p(w)$ with $p(w)$ its associated normalized probability density function that is the *multivariate t-distribution*:

$$p(w) = |\Lambda|^{1/2}\frac{\Gamma(\frac{d}{2}+\nu)}{\Gamma(\nu)\pi^{d/2}\nu^{d/2}}\frac{1}{(1+\frac{1}{2\nu}w^T\Lambda w)^{\frac{d}{2}+\nu}}.$$

As the Fourier transform is linear, we simply have to multiply by σ_M to obtain the normalizing constant, *i.e.* $\alpha = \sigma_M(\sqrt{2}\pi)^d$.

Now, to approximate the trajectories, we only need to rewrite g under the parametric form. Combining (4.11) with (4.10), we obtain

$$\widetilde{p}\big(g(x)\big) = \mathcal{N}\Big(\varphi(x)^T\Phi^{-1}\mu,$$
$$\varphi(x)^T\big[I_m + \Phi_{u_g}\Phi^{-1}(S_j - \Phi_{u_g}^T\Phi_{u_g})\Phi^{-1}\Phi_{u_g}^T\big]\varphi(x)\Big),$$

with $\Phi = \Phi_{u_g}^T\Phi_{u_g}$ and $\Phi_{u_g}^T = \big(\varphi(x_1),\cdots,\varphi(x_n)\big)$. Consequently it is possible to factorize by φ to obtain

$$g(x) \approx \varphi(x)^T\omega, \text{ with}$$

$$\omega \sim \mathcal{N}\Big(\Phi_{u_q}\Phi^{-1}\mu, I_m + \Phi_{u_q}\Phi^{-1}(S_j - \Phi_{u_q}^T\Phi_{u_q})\Phi^{-1}\Phi_{u_q}^T\Big).$$

With this sampling strategy, an analytic expression of the trajectory is known that enables its optimisation. In addition the cost to obtain a trajectory is $O(n^2m)$.

4.4.3 Adding Noise

Both algorithms presented above select sampling points that correspond to a potential reduction of the simple regret. They do not correspond necessarily to inputs that improve the accuracy of the model in the vicinity of the maximum [contrary to the approaches in Frazier et al., 2009, Hernández-Lobato et al., 2014, Picheny, 2014, for instance].

We observed empirically that focusing on the simple regret resulted in overly myopic strategies, as our model delivers much better local predictions using well-spread observations over a local region rather than highly concentrated points around a local optimum. In a sense, our acquisition functions point towards the right optima but do not propose an efficient sampling strategy to improve our model.

However, a simple way to correct this problem is to add a small centered multivariate Gaussian noise to the selected inputs, with a diagonal covariance matrix with terms $(\theta_1, \cdots, \theta_D)/4$.

4.5 Experiments

4.5.1 Test Cases Description

In this section we show the capacity of our algorithms to optimise a conditional quantile or expectile. To do so, we propose two challenging toy problems of dimension 2 and 7, respectively.

Test case 1 is a $2D$ toy problem on $[-4, 1] \times [2, 6]$ based on the Griewank function (see Dixon and Szego [1978]), defined as $Y_x = G(x) + R(x)\xi$, with

$$G(x) = \left[\sum_{i=1}^{2} \frac{x_i^2}{4000} - \prod_{i=1}^{2} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1 \right],$$

$R(x) = G(-3 - x_1, 8 - x_2)$ and $\xi = \eta \mathbb{1}_{\eta \leqslant 0} + \sqrt{3}\eta \mathbb{1}_{\eta > 0}$ where $\eta \sim \mathcal{N}(0, 1)$. The quantiles of order $\tau = 0.1,\ 0.5,\ 0.9$ are represented Figure 4.5.

Test case 2 is a $7D$ toy problem based on the Ackley function (see Ackley [2012]) on $[0, 1] \times [-0.7, -0.3] \times [0.5, 1] \times [-1, -0.5] \times [-0.1, 0] \times [0, 0.1] \times [0, 0.8]$, defined as a function

$$Y_x = 30 \times A(x) + R(x) \times \xi$$

with

$$A(x) = a \exp\left(- b \sqrt{\frac{1}{7} \sum_{i=1}^{7} x_i^2} \right) - \exp\left(\frac{1}{7} \sum_{i=1}^{7} \cos(c x_i) \right) + a + \exp(1)$$

, and $R(x) = 3A(x_2, x_3, \cdots, x_6, x_1)$, with $a = 10$, $b = 2 \times 10^{-4}$, $c = 0.9\pi$ and ξ follows a log-normal distribution of parameters $(0, 1)$.

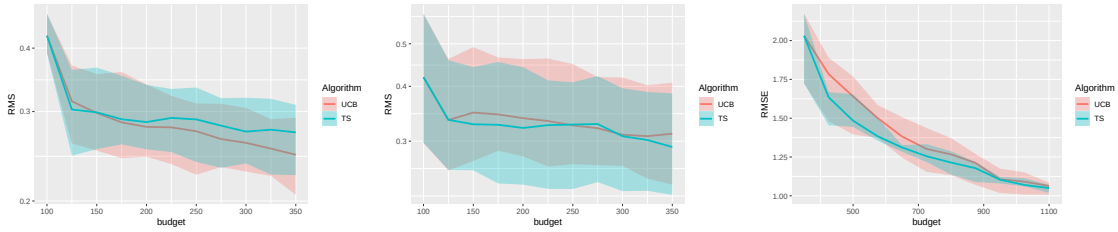


Figure 4.3: Evolution of the root mean square error during optimization: 2D, expectiles, $\tau = 0.1$ (left) and $\tau = 0.9$ (middle), 7D, quantile, $\tau = 0.3$ (right).

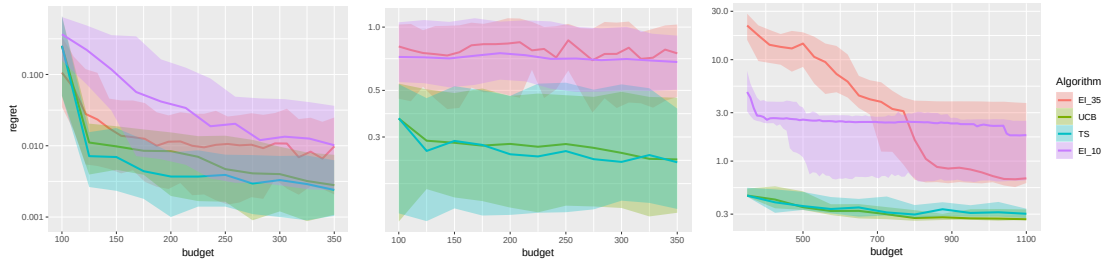


Figure 4.4: Regrets for: 2D, expectiles, $\tau = 0.1$ (left) and $\tau = 0.9$ (middle), 7D, quantile, $\tau = 0.3$ (right).

4.5.2 Quantile Kriging Baseline

Up to our knowledge, there exists no other BO algorithm to tackle quantile or expectile problems. A simple alternative is to use repetitions in the design of experiment to observe locally g and the observation noise σ (for instance by bootstrap). As direct observations are available, a standard GP inference can be used to provide a posterior on g [Plumlee and Tuo, 2014]. Next a BO procedure can be defined based on the EI criterion. As the number of repetitions is a potentially critical parameter of the method, in our experiments we use for different values: 10 and 20 in 2D, 10 and 35 in 7D. We refer to these algorithms as EI_s (for the smallest number of repetition) and EI_l (for the largest number of repetitions), which serve as baseline competitors.

4.5.3 Experimental Setting

Sequential strategy We created an initial design of size $50D$. At each update of the model we selected $10 \times D$ new inputs to be sampled and we added $2 \times D$ new points selected uniformly at random in $\mathcal{X}$. In $2D$ we used a budget equals to 350 while in $7D$ the budget is equals to 1100. At the end the point returned by the algorithm corresponds to the maximizer of our model.

On the hyperparameters of the model For the first test case we selected 100 inducing points at the location of the initial design of experiment. For the second test case we put 350 inducing points at the locations of the initial design of experiment and we add an inducing point at each corner of the input space which empirically helps to obtain relevant trajectories for TS. We trained the whole model on the initial design until convergence of the ELBO that took between 2000 and 3000 epochs with a leaning rate equals to 1×10^{-2}. To update the model we first trained only the variational parameters for 200 epochs with a learning rate equals to 5×10^{-3} then we optimised both the variationnal parameters and the kernel hyperparameters during 100 epochs with a leaning rate equals to 1×10^{-3}. Note that we did not optimise the inducing point location. To help the optimisation we used the whitening representation (see Hensman et al. [2015] for more details).

Metrics Each strategy is run 30 times with different initial conditions. As a primary performance metric, we consider the simple regret. In addition, we record the root mean square error of the models on $4,000$ randomly drawn test points over $\mathcal{X}$.

4.5.4 Results

Table 4.1 summaries our results. It is clear that for every of our test cases the strategy UCB and TS outperform the two versions of EI. In addition some problems are harder than others. For example the results of the 0.9-quantile are not as good as the results obtained for the 0.1 quantile. The reason for that lies in the high variance of the conditional distribution close to the maximum of $q^*_{0.9}$ while the variance is much more smaller close to the optimum $q^*_{0.1}$.

Figure 4.4 shows the regret curves for a subset of problems. We see that on the simplest 2D problem ($\tau = 0.1$), the baseline although not competitive with our approach, behaves correctly. On the much more difficult 2D problem ($\tau = 0.9$), the very high noise prevents the baseline from converging. While our approaches provide much better solutions from the start, the progress along iterations is limited. However, on average the model provides improved predictions (Figure 4.3, middle). On the 7D problem, our approaches directly start with a much better solution than the baseline and improve significantly over time. In addition, the overall prediction quality improves almost linearly (Figure 4.3, right).

Despite the stronger theoretical grounds of the Thompson sampling, the UCB approach offers comparable performances on our test problems for a smaller computational burden. This may be explained by the specificity of the problems at hand: the difficulty of the learning task results in large uncertainties in the model predictions, which reduces the influence of the sampling strategy. More significant differences may appear when more data is available, or on less demanding tasks such as low noise settings.

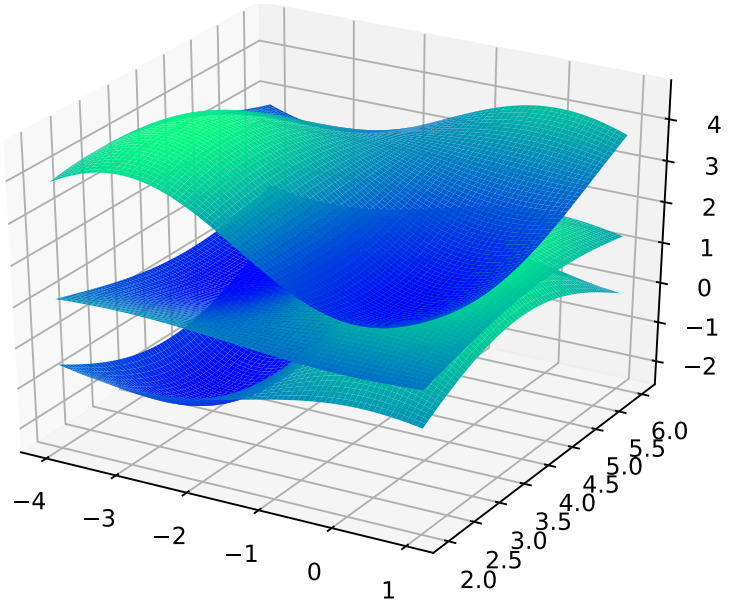


Figure 4.5: Quantiles of order 0.1, 0.5 and 0.9 of test case 1.

Quantile		EI_1	EI_2	UCB	TS
2D test case	$\tau = 0.1$	0.007	0.01	0.006	0.006
	$\tau = 0.9$	0.88	0.82	0.25	0.27
7D Test case	$\tau = 0.3$	2.1	0.73	0.29	0.31

Expectile		EI_1	EI_2	UCB	TS
2D test case	$\tau = 0.1$	0.01	0.01	0.003	0.003
	$\tau = 0.9$	0.65	0.61	0.27	0.27
7D Test case	$\tau = 0.3$	0.71	0.78	0.42	0.41

Table 4.1: Final values (median) of the simple regret with a budget of 350 points in 2D and 1100 points in 7D.

4.6 Conclusion

In this paper we have presented a new setting to estimate quantiles and expectiles of stochastic black box functions which is well suited to heteroscedastic cases. Then the proposed model has been used to create a Bayesian optimisation algorithm designed to optimise conditional quantiles and expectiles without repetitions in the experimental design. These algorithms showed good results on toy problems in dimension 2 and 7 and for different orders τ.

Chapter 5

Conclusion et perspectives

Contents

5.1 Conclusion

Métamodélisation : Dans cette thèse nous avons proposé une synthèse des différentes méthodes de régression pour le quantile, la CVaR et l'expectile, qui sont des indicateurs permettant de mesurer la probabilité d'observer des événements extrêmes. Six méthodes de régression quantile ont été étudiés, ce qui a permis l'identification de comportements propres à chaque approche et d'exposer leurs limites. Puis nous avons proposé une nouvelle méthode pensé spécialement pour faire de la régression dans des cas fortement hétéroscédastiques. Cette nouvelle méthode a été appliquée à l'estimation de quantiles et d'expectiles.

Optimisation bandit : Dans le Chapitre 3 nous avons proposé un algorithme d'optimisation inspiré de la littérature $\mathcal{X}$-armed bandits permettant l'optimisation d'une quelconque mesure de risque g sous réserve que des inégalités de déviation soient connues sur cette mesure. Notre approche a permis d'expliciter une borne supérieure générique sur le regret simple. Nous avons appliqué cet algorithme à l'optimisation d'un quantile conditionnel et d'une CVaR conditionnelle. D'un point de vue théorique nous avons montré que la borne supérieure sur le regret simple pour l'optimisation de ces deux mesures de risque était identique à celle obtenue pour l'optimisation de la moyenne conditionnelle modulo une constante. Pour le quantile, des inégalités de déviations plus fines que celle dérivés à partir de la propriété de Hoeffding ont été obtenues en utilisant l'inégalité de Bernstein et celle de Chernoff. Pour la CVaR des inégalités existantes ont été utilisées. Enfin nous avons illustré numériquement la capacité de notre algorithme à optimiser ces deux mesures de risque. Le résultat de ces expériences illustre qu'utiliser des inégalités de déviation plus fines améliore les vitesses de convergence.

Optimisation bayésienne : Dans le Chapitre 4 nous avons proposé deux algorithmes d'optimisations à base de métamodèles gaussiens pouvant optimiser une mesure de risque g sous réserve qu'une distribution a posteriori soit connue sur cette fonction. Deux routines d'optimisation ont été développées. La première est basée sur l'algorithme GP-UCB et la seconde sur une stratégie de type Thompson sampling. Les méthodes proposées ont été appliqués à l'optimisation d'un quantile conditionnel et d'un expectile conditionnel. Sur notre ensemble de fonctions testes ces approches ont montré de bons résultats dans le sens où elles sont meilleures que des méthodes existantes utilisant des répétitions dans le plan d'expérience. Les stratégies proposées semblent capables d'optimiser des mesures de risque jusqu'à des dimension de l'ordre de $D = 8$ à $D = 10$ mais des expériences complémentaires semblent nécessaires pour cerner plus précisément leur potentiel.

5.2 Perspectives

Régression multiquantile Il est possible que des métamodèles de quantile de différents ordres se croisent. Ce comportement est pathologique car par définition la fonc-

tion quantile est croissante en τ. Dans la littérature il y a un certain nombre de travaux traitant ce problème et apportant des solutions partielles Takeuchi et al. [2006], Sangnier et al. [2016]. Toutefois ces travaux se focalisent sur le métamodèle sans faire de liens entre des caractéristiques de la distribution ciblée et le phénomène de croisement. Or dans le Chapitre 2 nous avons pu constater que si la valeur de la densité est faible au voisinage du quantile estimé alors le métamodèle peut sur-apprendre et il en résultera des croisements potentiels. Une approche pour contrer ce phénomène de croisement et pour améliorer la qualité de la prédiction serait d'estimer séquentiellement différents quantiles. La stratégie s'initialiserait par l'estimation de quantiles associés à une valeur de densité élevée puis dans un second temps les quantiles associés à une densité plus faible seraient estimer non pas uniquement à partir des données mais en prenant en compte la valeur des premiers quantiles estimés. Cette stratégie demande de connaître la précision d'estimation de chaque quantile. Or l'approche par processus gaussiens développée dans le Chapitre 4 fournie une indication sur cette précision.

Améliorer le schéma d'exploration de la méthode $\mathcal{X}$-armed bandits et proposer de nouvelles inégalités de déviation Dans le Chapitre 3 nous avons uniquement échantillonné au centre des cellules. Cette approche permet de faire une hypothèse très faible sur la fonction visée tout en obtenant des garanties de convergence. Or il est possible d'échantillonner uniformément dans les cellules tout en obtenant des inégalités de déviation sur le quantile. Cependant il semble que pour obtenir des garanties de convergence il faille définir une hypothèse plus forte sur la régularité du quantile conditionnel. Par exemple une hypothèse suffisante serait que la fonction quantile soit globalement hölderienne.

En ce qui concerne le partitionnement hiérarchique de $\mathcal{X}$. Dans le chapitre 3 nous avons utilisé un partitionnement qui implique la division de chaque cellule en K^D nouvelles cellule. Or cette approche est trop gourmande quand la dimension augmente. Une possibilité pour améliorer l'utilisation de ces méthodes en grande dimension serait d'utiliser des découpage en K nouvelles cellules.

Pour définir une UCB et LCB, dans le Chapitre 3 nous avons besoin de connaître en partie la régularité de la fonction visée. Or dans la littérature bandit des approches ont été développées pour conduire l'optimisation sans cette connaissance. C'est le cas par exemple de Locatelli and Carpentier [2018], Shang et al. [2019], Bartlett et al. [2018]. Il serait donc possible de combiner notre approche à ces idées.

Enfin notre algorithme à besoin d'inégalités de déviation sur g pour conduire l'optimisation. Une première amélioration au travail de cette thèse serait de définir une inégalité de déviation sur l'expectile.

Optimisation bayésienne La première amélioration pour l'optimisation de mesures de risque par métamodèles gaussiens consisterait à proposer un critère robuste pour retourner le point final $\widehat{x}^*$.

La stratégie dérivée de l'algorithm GP-UCB semble prometteuse, elle pourrait être améliorée en proposant un vecteur $\boldsymbol{\beta}$ avec une approche plus théorique.

Le choix de la magnitude de la perturbation de chaque point sélectionné pourrait être améliorée. Par exemple la magnitude de la perturbation pourrait être une fonction décroissante du temps.

Enfin le modèle utilisé est certainement améliorable. Une première possibilité consisterait à supprimer l'hypothèse d'indépendance entre g et σ. Des indications pour estimer les paramètres d'un tel modèle peuvent êtres trouvées dans Adam [2017].

Optimisation hybride Les méthodes d'optimisation basées sur les métamodèles gaussiens semblent trouver les zones à fort potentiel relativement bien. En revanche estimer une mesure de risque avec une précision de l'ordre de 1×10^{-2} semble la limite de ces approches. Inversement, les approches $\mathcal{X}$-armed bandits ont plus de difficultés à trouver les zones à fort potentiel dès que la dimension augmente mais elles sont capables d'atteindre un très haut niveau de précision dans l'optimisation. Combiner les deux approches pourrait permettre d'améliorer les travaux présentés dans cette thèse.

Optimisation multi-objectif Dans cette thèse nous avons développé différents outils pour estimer et optimiser différentes mesures de risque indépendamment les unes des autres. Cependant la recherche d'optimum pour un problème multi-objectif où chaque objectif est une mesure de risque différente a du sens. On pourra notamment penser à l'optimisation conjointe d'un quantile bas et de la moyenne dans l'idée de ce qui est fait dans Picheny et al. [2017].

Bibliography

B. Abdous and B. Remillard. Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics*, 47(2):371–384, 1995.

S. Abeywardana and F. Ramos. Variational inference for nonparametric bayesian quantile regression. In *AAAI*, pages 1686–1692, 2015.

J. Abrevaya. Isotonic quantile regression: asymptotics and bootstrap. *Sankhyā: The Indian Journal of Statistics*, pages 187–199, 2005.

C. Acerbi and B. Szekely. Back-testing expected shortfall. *Risk*, 27(11):76–81, 2014.

D. Ackley. *A connectionist machine for genetic hillclimbing*, volume 28. Springer Science & Business Media, 2012.

V. Adam. Structured variational inference for coupled gaussian processes. *arXiv preprint arXiv:1711.01131*, 2017.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998.

J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. 2009.

J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

M. G. Azar, A. Lazaric, and E. Brunskill. Online stochastic optimization under correlated bandit feedback. In *ICML*, pages 1557–1565, 2014.

F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

F. Bachoc. *Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments.* PhD thesis, Université Paris-Diderot-Paris VII, 2013.

G. Barone Adesi. Var and cvar implied in option prices. *Journal of Risk and Financial Management*, 9(1):2, 2016.

P. L. Bartlett, V. Gabillon, and M. Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. *arXiv preprint arXiv:1810.00997*, 2018.

F. Bellini and E. Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506, 2017.

A. Belloni, V. Chernozhukov, D. Chetverikov, and I. Fernández-Val. Conditional quantile processes based on series or many regressors. *arXiv preprint arXiv:1105.6154*, 2011.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554, 2011.

P. K. Bhattacharya and A. K. Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

M. Binois, R. B. Gramacy, and M. Ludkovski. Practical heteroskedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, (just-accepted):1–41, 2018.

C. K. Birdsall and A. B. Langdon. *Plasma physics via computer simulation.* CRC press, 2018.

C. M. Bishop. *Neural networks for pattern recognition.* Oxford university press, 1995.

A. Boukouvalas, R. Barillec, and D. Cornford. Gaussian process quantile regression using expectation propagation. *arXiv preprint arXiv:1206.6391*, 2012.

C. Bouttier. Optimisation globale sous incertitudes: algorithmes stochastiques et bandits continus avec application à la planification de trajectoires d'avions. 2017.

G. E. Box and G. C. Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

M. L. Braun. Accurate error bounds for the eigenvalues of the kernel matrix. *Journal of Machine Learning Research*, 7(Nov):2303–2328, 2006.

L. Breiman. Bias, variance, and arcing classifiers. 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman. *Classification and regression trees*. Routledge, 2017.

D. B. Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

T. Browne, B. Iooss, L. L. Gratiet, J. Lonchampt, and E. Remy. Stochastic simulators based optimization by gaussian process metamodels–application to maintenance investments planning issues. *Quality and Reliability Engineering International*, 32(6): 2067–2080, 2016.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

A. J. Cannon. Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & geosciences*, 37(9):1277–1284, 2011.

A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141, 2015.

P. Casadebaig, L. Guilioni, J. Lecoeur, A. Christophe, L. Champolivier, and P. Debaeke. Sunflo, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. *Agricultural and forest meteorology*, 151(2):163–178, 2011.

M. Cavazzuti. Design of experiments. In *Optimization Methods*, pages 13–42. Springer, 2013.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

C.-H. Chang and C.-W. Ha. On eigenvalues of differentiable positive definite kernels. *Integral Equations and Operator Theory*, 33(1):1–7, 1999.

A. Christmann and I. Steinwart. Consistency of kernel-based quantile regression. *Applied Stochastic Models in Business and Industry*, 24(2):171–183, 2008a.

A. Christmann and I. Steinwart. How svms can estimate quantiles and the median. In *Advances in neural information processing systems*, pages 305–312, 2008b.

E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis. Parallel gaussian process optimization with upper confidence bound and pure exploration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 225–240. Springer, 2013.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

N. Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

Y. David and N. Shimkin. Pure exploration for max-quantile bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 556–571. Springer, 2016.

A. P. Dawid. Posterior expectations for large observations. *Biometrika*, 60(3):664–667, 1973.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *The Journal of Machine Learning Research*, 15(1):3873–3923, 2014.

P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of statistics*, pages 269–281, 1979.

R. Díaz-Uriarte and S. A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

P. J. Diggle, J. A. Tawn, and R. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

L. C. W. Dixon and G. P. Szego. The global optimization problem. an introduction. *Toward global optimization*, 2:1–15, 1978.

H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4):325–327, 1976.

S. Efromovich. Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association*, 105(490):761–774, 2010.

K.-T. Fang, R. Li, and A. Sudjianto. *Design and modeling for computer experiments*. Chapman and Hall/CRC, 2005.

M. Farooq and I. Steinwart. An svm-like approach for expectile regression. *Computational Statistics & Data Analysis*, 109:159–181, 2017.

R. P. Feynman. Mathematical formulation of the quantum theory of electromagnetic interaction. *Physical Review*, 80(3):440, 1950.

P. Frazier, W. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

V. Garcia, E. Debreuve, and M. Barlaud. Fast k nearest neighbor search using gpu. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

A. Garivier, P. Ménard, L. Rossi, and P. Menard. Thresholding bandit for dose-ranging: The impact of monotonicity. *arXiv preprint arXiv:1711.04454*, 2017.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 2018.

F. Ghaderinezhad and C. Ley. On the impact of the choice of the prior in bayesian statistics. In *Bayesian Inference*. IntechOpen, 2019.

E. Gijo and J. Scaria. Product design by application of taguchi's robust engineering using computer simulation. *International Journal of Computer Integrated Manufacturing*, 25(9):761–773, 2012.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

D. Ginsbourger, R. Le Riche, and L. Carraro. Kriging is well-suited to parallelize optimization. In *Computational intelligence in expensive optimization problems*, pages 131–162. Springer, 2010.

J.-B. Grill, M. Valko, and R. Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems*, pages 667–675, 2015.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

N. Hansen. Benchmarking the nelder-mead downhill simplex algorithm with many local restarts. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, pages 2403–2408. ACM, 2009.

W. Härdle. *Applied nonparametric regression*. Number 19. Cambridge university press, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

X. He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

J. Hensman, A. G. Matthews, M. Filippone, and Z. Ghahramani. Mcmc for variationally sparse gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1648–1656, 2015.

A. J. Hepworth. *A multi-armed bandit approach to superquantile selection*. PhD thesis, Monterey, California: Naval Postgraduate School, 2017.

J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.

J. M. Hernández-Lobato, J. Requeima, E. O. Pyzer-Knapp, and A. Aspuru-Guzik. Parallel and distributed thompson sampling for large-scale accelerated exploration of chemical space. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1470–1479. JMLR. org, 2017.

R. Herwig. Computational modeling of drug response with applications to neuroscience. *Dialogues in clinical neuroscience*, 16(4):465, 2014.

S. Isci, H. Dogan, C. Ozturk, and H. H. Otu. Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics*, 30(6):860–867, 2013.

H. Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015.

J. Janusevskis and R. Le Riche. Simultaneous kriging-based estimation and optimization of mean response. *Journal of Global Optimization*, 55(2):313–336, 2013.

C. Jiang, M. Jiang, Q. Xu, and X. Huang. Expectile regression neural network model with applications. *Neurocomputing*, 247:73–86, 2017.

D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

D. S. Jones. *The theory of electromagnetism*. Elsevier, 2013.

M. C. Jones. Expectiles and m-quantiles are quantiles. *Statistics & Probability Letters*, 20(2):149–153, 1994.

P. Jylänki, J. Vanhatalo, and A. Vehtari. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12(Nov):3227–3257, 2011.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of A/B testing. In *Conference on Learning Theory*, pages 461–481, 2014.

K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, pages 393–400. ACM, 2007.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM, 2008.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

R. K. Kolla, K. Jagannathan, et al. Risk-aware multi-armed bandits using conditional value-at-risk. *arXiv preprint arXiv:1901.00997*, 2019.

S. Kotz, T. Kozubowski, and K. Podgorski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Springer Science & Business Media, 2012.

H. Kozumi and G. Kobayashi. Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.

P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

C.-M. Kuan, J.-H. Yeh, and Y.-C. Hsu. Assessing value at risk with care, the conditional autoregressive expectile models. *Journal of Econometrics*, 150(2):261–270, 2009.

M. Lázaro-Gredilla and M. Titsias. Variational heteroscedastic gaussian process regression. 2011.

J. Lei. Stochastic modeling in systems biology. *J. Adv. Math. Appl*, 1(1):76–88, 2012.

A. S. Lewis and M. L. Overton. Nonsmooth optimization via bfgs. *Submitted to SIAM J. Optimiz*, pages 1–35, 2009.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*, 2016.

Y. Li, Y. Liu, and J. Zhu. Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

N. List and H. U. Simon. Svm-optimization and steepest-descent line search. In *Proceedings of the 22nd Annual Conference on Computational Learning Theory*, 2009.

A. Locatelli and A. Carpentier. Adaptivity to Smoothness in X-armed bandits. In *Conference on Learning Theory*, pages 1463–1492, 2018.

J. L. Loeppky, J. Sacks, and W. J. Welch. Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51(4):366–376, 2009.

D. Lopez-Martinez. Regularization approaches for support vector machines with applications to biomedical data. *arXiv preprint arXiv:1710.10600*, 2017.

G. Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

M. Ludkovski and J. Niemi. Optimal dynamic policies for influenza management. *Statistical Communications in Infectious Diseases*, 2(1), 2010.

K. Lum, A. E. Gelfand, et al. Spatial quantile multiple regression using the asymmetric laplace process. *Bayesian Analysis*, 7(2):235–258, 2012.

O.-A. Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

H. Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

S. Marmin, C. Chevalier, and D. Ginsbourger. Differentiating the multipoint expected improvement for optimal batch design. In *International Workshop on Machine Learning, Optimization and Big Data*, pages 37–48. Springer, 2015.

L. C. Marsh and D. R. Cormier. *Spline regression models*, volume 137. Sage, 2001.

M. A. McCARTHY and P. Masters. Profiting from prior information in bayesian analyses of ecological data. *Journal of Applied Ecology*, 42(6):1012–1019, 2005.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7 (Jun):983–999, 2006.

N. Meinshausen and M. N. Meinshausen. The quantregforest package. 2007.

T. C. Meng, S. Somani, and P. Dhar. Modeling and simulation of biological systems with stochasticity. *In silico biology*, 4(3):293–309, 2004.

T. P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

J. Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer, 1975.

J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

V. Moutoussamy, S. Nanty, and B. Pauwels. Emulators for stochastic simulation codes. *ESAIM: Proceedings and Surveys*, 48:116–155, 2015.

R. Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in neural information processing systems*, pages 783–791, 2011.

R. Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1): 1–129, 2014.

W. K. Newey and J. L. Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

A. O'Hagan. On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3):358–367, 1979.

G. C. Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic constrained optimization*, pages 272–281. Springer, 2000.

V. Picheny. A stepwise uncertainty reduction approach to constrained global optimization. In *Artificial Intelligence and Statistics*, pages 787–795, 2014.

V. Picheny, T. Wagner, and D. Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013.

V. Picheny, R. Trépos, and P. Casadebaig. Optimization of black-box models with uncertain climatic inputs-application to sunflower ideotype design. *PloS one*, 12(5): e0176815, 2017.

M. Plumlee and R. Tuo. Building accurate emulators for stochastic simulations via quantile kriging. *Technometrics*, 56(4):466–473, 2014.

Y.-H. Qian, D. d'Humières, and P. Lallemand. Lattice bgk models for navier-stokes equation. *EPL (Europhysics Letters)*, 17(6):479, 1992.

A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

C. E. Rasmussen and C. K. Williams. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 38:715–719, 2006.

R. D. Reiss and L. Ruschendorf. On wilks' distribution-free confidence intervals for quantile intervals. *Journal of the American Statistical Association*, 71(356):940–944, 1976.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

R. T. Rockafellar. Coherent approaches to risk in optimization under uncertainty. In *OR Tools and Applications: Glimpses of Future Technologies*, pages 38–61. Informs, 2007.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

R. T. Rockafellar, J. O. Royset, and S. I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014.

R. Rosipal and L. J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, 2(Dec):97–123, 2001.

M. Rostek. Quantile maximization in decision theory. *The Review of Economic Studies*, 77(1):339–371, 2010.

O. Roustant, D. Ginsbourger, and Y. Deville. Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. 2012.

M. Sangnier, O. Fercoq, and F. d'Alché Buc. Joint quantile regression in vector-valued rkhss. In *Advances in Neural Information Processing Systems*, pages 3693–3701, 2016.

A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

A. D. Saul, J. Hensman, A. Vehtari, and N. D. Lawrence. Chained gaussian processes. In *Artificial Intelligence and Statistics*, pages 1431–1440, 2016.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

B. Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

T. B. Schön and F. Lindsten. Manipulating the multivariate gaussian density. *Div. Automat. Control, Linköping Univ., Linkping, Sweden, Tech. Rep*, 2011.

E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016.

G. A. Seber and A. J. Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2015.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016.

X. Shang, E. Kaufmann, and M. Valko. General parallel optimization without a metric. In *30th International Conference on Algorithmic Learning Theory*, 2019.

J. Shim, C. Hwang, and K. H. Seok. Non-crossing quantile regression via doubly penalized kernel machine. *Computational Statistics*, 24(1):83–94, 2009.

J. Sill, G. Takács, L. Mackey, and D. Lin. Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*, 2009.

H. Skullerud. The stochastic computer simulation of ion motion in a gas subjected to a constant electric field. *Journal of Physics D: Applied Physics*, 1(11):1567, 1968.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

I. Steinwart, A. Christmann, et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

C. J. Stone. Nearest neighbor estimators of a nonlinear regression function. In *Computer Science and Statistics: 8th Annual Symposium on the Interface*, pages 413–418, 1975.

C. J. Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

C. Storlie and J. Helton. Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. *Reliability Engineering and System Safety*, 93:28–54, 2008.

T. Székely Jr and K. Burrage. Stochastic simulation in systems biology. *Computational and structural biotechnology journal*, 12(20-21):14–25, 2014.

B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. 2015.

M. A. Taddy and A. Kottas. A bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics*, 28(3):357–369, 2010.

I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *Journal of machine learning research*, 7(Jul):1231–1264, 2006.

O. Tange. Gnu parallel 2018. 2018.

R. Temam and A. Chorin. Navier stokes equations: Theory and numerical analysis, 1978.

P. Thomas and E. Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, pages 6225–6233, 2019.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

W. R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57(2):450–456, 1935.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

L. Torossian, A. Garivier, and V. Picheny. X-armed bandits: Optimizing quantiles and other risks. *arXiv preprint arXiv:1904.08205*, 2019a.

L. Torossian, V. Picheny, R. Faivre, and A. Garivier. A review on quantile regression for stochastic computer experiments. *arXiv preprint arXiv:1901.07874*, 2019b.

B. A. Turlach and A. Weingessel. quadprog: Functions to solve quadratic programming problems. *CRAN-Package quadprog*, 2007.

D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

M. Valko, A. Carpentier, and R. Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27, 2013.

A. Van Maanen and X.-M. Xu. Modelling plant disease epidemics. *European Journal of Plant Pathology*, 109(7):669–682, 2003.

J. Vanhatalo, P. Jylänki, and A. Vehtari. Gaussian process regression with student-t likelihood. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1910–1918. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3806-gaussian-process-regression-with-student-t-likelihood.pdf.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

G. Verdier and A. Ferreira. Adaptive mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 24(1):59–68, 2011.

N. Villa-Vialaneix, M. Follador, M. Ratto, and A. Leip. A comparison of eight meta-modeling techniques for the simulation of N_2O fluxes and N leaching from corn crops. *Environmental Modelling & Software*, 34:51–66, 2012.

S. Wang, S. H. Ng, and W. B. Haskell. A multi-level simulation optimization approach for quantile functions. *arXiv preprint arXiv:1901.05768*, 2019.

Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2009.

I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

J. Wu and P. Frazier. The parallel knowledge gradient method for batch bayesian optimization. In *Advances in Neural Information Processing Systems*, pages 3126–3134, 2016.

S. Xuedong, E. Kaufmann, and M. Valko. General parallel optimization a without metric. In *Algorithmic Learning Theory*, pages 762–787, 2019.

Q. Yao and H. Tong. Asymmetric least squares regression estimation: a nonparametric approach. *Journal of nonparametric statistics*, 6(2-3):273–292, 1996.

M. Ye and Y. Sun. Variable selection via penalized neural network: a drop-out-one loss approach. In *International Conference on Machine Learning*, pages 5616–5625, 2018.

K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.

K. Yu and J. Zhang. A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics-Theory and Methods*, 34(9-10):1867–1879, 2005.

K. Yu, L. Ji, and X. Zhang. Kernel nearest-neighbor algorithm. *Neural Processing Letters*, 15(2):147–156, 2002.

C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence, Volume: 41 , Issue: 8 , Aug. 1 2019*.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

J. F. Ziegel. Coherence and elicitability. *Mathematical Finance*, 26(4):901–918, 2016.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.