Métamodèles et Optimisation Coûteuse Avec des variables mixtes catégorielles

Sujet de thèse

Rodolphe Le Riche, CNRS & EMSE (Directeur) ; Sanaa Zannane, Julien Pelamatti, Merlin Keller, EDF (Encadrants)
En collaboration avec Safran Tech, IFPEN, CEA, polytechnique Montréal

Contexte de la thèse :

L'ANR SAMOURAI est un projet ambitieux de recherche financé en partie par l'Agence Nationale de la Recherche. Ce projet débute en 2021 pour une durée de quatre ans. Il regroupe plusieurs industriels (CEA, EDF, l'IFP Energies Nouvelles et SAFRAN) et académiques (CentraleSupélec, l'Ecole Nationale des Mines de Saint-Etienne et Polytechnique Montréal), autour des thèmes de l'optimisation, de l'analyse d'incertitudes et de la fiabilité, basées sur des simulations et des méta-modèles. La thèse proposée s'inscrit dans le cadre de ce projet, et bénéficiera à ce titre d'un environnement scientifique riche et stimulant.

Problématique générale :

La conception de systèmes industriels complexes donne souvent lieu à un arbitrage entre les performances techniques espérées du système (ex. : puissance d'un système de production d'énergie, efficacité d'un réseau de distribution) et le coût économique associé (ex. : coûts de l'investissement initial, coûts de maintenance, etc.), tout en s'astreignant au respect des contraintes du système (ex : compatibilité entre composants mis en œuvre, plages de fonctionnement, etc.). Ce problème se formalise naturellement par un problème d'optimisation sous contrainte G, mettant en œuvre des variables de décision X et une fonction objectif F.

$$x^{\square} \epsilon arg \min_{xtqG(x)<0} F(x) \quad (1)$$

Lorsque la fonction objectif F est coûteuse à calculer (par ex. lorsqu'elle est la résultante d'un code de calcul complexe) et qu'on ne dispose pas de son expression analytique, recourir aux méthodes classiques d'optimisation mixte (typiquement les méthodes évolutives ou les méthodes de séparation- évaluation) peut s'avérer prohibitif ou inadapté.

De plus, il est courant en pratique que les variables d'optimisation (ou plus largement de décision), regroupées sous la notation X, soient de nature mixte, continue pour les unes et discrète, voire catégorielle, pour les autres. Par exemple, la conception d'un réseau de neurones implique d'ajuster des poids continus, de choisir un nombre de neurones par couches cachées, sans parler des choix catégoriels de la forme des fonctions d'activation ou de l'architecture générale du réseau… parce que l'optimisation mixte correspond à une formulation très générale de ce type de problèmes, on peut en trouver des exemples dans des disciplines en apparence aussi diverses que l'ingénierie, les sciences physiques ou l'apprentissage statistique.

Un tel mélange de variables continues et catégorielles pose de réelles difficultés aux techniques d'optimisation à base de méta-modèles, que ce soit dans la construction de ce dernier, ou dans le choix d'une stratégie d'optimisation efficace, les deux problèmes étant intimement liés. En outre, une difficulté d'ordre plus pratique réside dans le fait que les problèmes d'optimisation mixtes sont en général traités par des familles distinctes d'approches qui correspondent à des communautés scientifiques également distinctes. C'est pourquoi l'un des objectifs fondateurs de cette thèse est d'aboutir à une formulation suffisamment générique du problème d'optimisation mixte, pour pouvoir unifier les approches existantes en combinant leurs forces.

Objectifs de la thèse :

L'ambition principale de la thèse est d'aboutir à une solution la plus générique possible aux problèmes d'optimisation mixte coûteuse, en surmontant notamment les difficultés suivantes :

- **Explosion combinatoire et coût calculatoire** : la présence de variables discrètes (ordinales ou nominales) en l'absence de toute notion de convexité mène à un nombre de combinaisons possibles pour les variables discrètes qui augmente exponentiellement avec leur nombre. C'est particulièrement problématique lorsque les fonctions associées (objectifs et contraintes) sont coûteuses à évaluer. Le développement de métamodèles de type processus gaussiens et de stratégies d'enrichissement de plans d'expérience numériques adaptées à des variables mixtes semble à l'heure actuelle une perspective très prometteuse ;
- **Généricité** : Les problèmes d'optimisation mixte ont été étudiés depuis longtemps par la communauté de la recherche opérationnelle, et a donné lieu à la création d'un grand nombre d'approches spécialisées, adaptées à divers cas. L'apparition de métamodèles mixtes et de critères d'enrichissement adaptés ouvre une avenue pour développer des méthodes plus génériques. La démonstration de cette généricité passe notamment par la possibilité de tester les nouvelles méthodes sur différents cas d'applications industriels ; c'est pourquoi quatre cas-tests principaux sont envisagés pour cette thèse, issus de différents secteurs industriels : conception de centrales éoliennes, de turbo-machines, de flotteurs d'éolienne offshore, et dimensionnement d'un réseau électrique (voir détails plus bas) ;

Programme de travail :

Les pistes suivantes ont été identifiées comme semblant les plus prometteuses afin de répondre aux défis ci-dessus:

1. **Développement des noyaux de covariance pour métamodèles à variables mixtes** : Les travaux récents de [Hutter, 2011] et [Pelamatti, 2019], entre autres, ouvrent de grands champs d'exploration de nouvelles possibilités pour construire des métamodèles à entrées mixtes, pouvant inclure des variables à optimiser dont la dimension est variable, et fait elle-même partie des variables à optimiser ! De tels noyaux joints sont obtenus en combinant des choix de noyaux indépendants pour les variables continues, discrètes et dimensionnelles. Bien que de nombreux noyaux soient permis en théorie selon cette approche, seuls quelques cas particuliers ont été explorés jusqu'à présent. Cela est vrai pour le choix des noyaux de covariance continus, exponentiels-quadratiques dans [Pelamatti, 2019], Matérn dans [Muñoz & Sinoquet, 2020]. Et ce l'est encore plus pour les noyaux discrets, pour lesquels une large littérature existe déjà qui demande à être explorée, de la représentation par variables latentes dans [Zhang et al. 2019; Cuesta Ramirez et al. 2019] aux noyaux à dimension variable dans [Pelamatti, 2020]. C'est pourquoi, lors de la phase de construction du métamodèle, choisir les différents modèles de noyaux et la manière de les combiner, à partir de critères de sélection de modèle (AIC, BIC, facteur de Bayes, scores prédictifs par validation croisée, …), peut présenter en soit un défi si l'on considère l'explosion combinatoire du nombre de modèles de noyaux mixtes disponibles. Si bien qu'il s'agit en soit d'un problème d'optimisation mixte du critère de choix de modèle à optimiser en fonction du choix des variables catégorielles représentant les différents choix de noyaux continus, catégoriels, dimensionnels, et de leur mode de combinaison. L'approche par processus gaussien nous semble à l'heure actuelle la piste la plus prometteuse. Cependant, la comparaison avec d'autres classes de métamodèles utilisés dans les problèmes mixtes, tels que les bases de fonctions radiales (RBF) [Müller et al., 2013; Costa and Nannicini, 2018] ou les forêts aléatoires [Hutter et al, 2011], aideront à valider leur intérêt opérationnel.

2. **Choix d'une méthode d'optimisation globale pour variables mixtes** : L'optimisation bayésienne, à l'instar de l'algorithme EGO [Jones et al., 1998] semble une piste naturelle. Celui-ci dépend d'un plan d'expérience initial, ainsi que d'un critère d'enrichissement, typiquement l'amélioration espérée ou *expected improvement* (EI) ainsi que des variantes basées sur les moments du métamodèle. La perspective d'ajouter plusieurs points à la fois au plan d'expérience courant, en s'appuyant sur des architectures de calcul parallèle, pourrait dans une certaine mesure contenir

l'explosion du nombre de configurations. En complément, d'autres approches populaires dans la communauté optimisation pourront être envisagées, telles que la méthode de zone de confiance quadratique adaptative dans [Conn et al. 2009]. Une extension de cette méthode au cas de conception d'une turbomachine [Tran et al., 2020], dans lequel interviennent des variables mixtes-binaires, fournit naturellement une approche de référence sur celui-ci. Enfin, des approches hybrides, comme proposé dans [Régis, 2015] pour les variables mixtes, pourraient permettre de combiner la flexibilité des processus gaussiens au bon comportement observé des approches par région de confiance en grande dimension. Un exemple de telles approches est fourni dans [Audet et al., 2019], avec une implémentation dans le logiciel NOMAD [Le Digabel, 2011] qui pourra servir de référence pour le développement de nouveaux algorithmes.

3. **Développement des stratégies d'enrichissement pour l'optimisation bayésienne** : Optimiser le critère d'acquisition revient encore une fois à résoudre un problème d'optimisation avec les mêmes variables que le problème d'origine, mais cette fois-ci avec un critère gratuit à évaluer ou presque. La littérature très large sur le sujet fournit de nombreuses solutions qu'il conviendra de comparer sur les différents cas d'application : algorithme évolutifs hybrides [Cauwet et al., 2019], adaptation de l'algorithme MADS [Audet and Dennis, 2006] à des problèmes mixtes dans [Munoz Zuniga and Sinoquet, 2020], stratégies d'allocation budgétaire évolutives [Pelamatti, 2020], ou d'inspiration bayésienne dans [Bergstra et al., 2011], pour n'en citer que quelques-unes. Comparer toutes ces approches sur de multiples cas permettra de mieux comprendre quels avantages tirer des chacune.

Cas d'application :

Cette thèse qui ambitionne de proposer une solution générique aux problèmes à variables mixtes issus de l'apprentissage non-linéaire et de l'optimisation pour des modèles coûteux, a également l'ambition d'aborder une grande diversité de cas d'applications afin de stimuler le développement des approches les plus génériques.

Ainsi, deux cas d'applications concerneront des problèmes de conception, l'un d'une turbomachine et l'autre d'un flotteur d'éolienne. Ils impliqueront chacun des choix discrets de composantes (à travers les différents choix de technologies, de nombres de composantes installées, etc.).

Deux autres cas s'intéresseront à des problèmes à plus grande échelle, à travers la conception d'une ferme éolienne et la gestion d'un réseau d'électricité impliquant des choix discrets de dimensionnement (nombre et type d'éoliennes, d'unités de production) qui influent sur le nombre de variables d'optimisation continues (positionnement des éoliennes).

A ces quatre principaux cas d'application pourront s'ajouter d'autres, en fonction du temps disponible, afin de valoriser les développements méthodologiques de la thèse, comme par exemple l'optimisation des hyperparamètres d'un système d'intelligence artificielle (réseau de neurones) et conception de procédé innovant de détection de défauts par courants de Foucault.

Encadrement :

- La thèse sera dirigée par Rodolphe Le Riche, directeur de recherche au CNRS (laboratoire LIMOS à Mines St-Etienne), en collaboration avec EDF R&D, en les personnes de Sanaa Zannane, Julien Pelamatti et Merlin Keller, ingénieurs chercheurs au département PRISME à EDF Lab Chatou ;
- Cette thèse s'inscrit dans le contexte de l'ANR SAMOURAI (Simulation Analytics, Meta-model-based solutions for Optimization, Uncertainty and Reliability Analysis), qui démarre en mars 2021. Cette ANR regroupe 3 thèses (dont la présente) et 2 post-docs sur les différents axes de recherche de l'ANR : métamodèles en grande dimension, méthodes séquentielles en fiabilité, métamodèles et optimisation mixte, gestion de contraintes cachées. Il est d'ailleurs prévu un

appui de la part de l'IFPEN (D. Sinoquet) et du laboratoire GERAD (S. Le Digabel) pour des actions en lien avec la plateforme NOMAD ;
- Les quatre cas d'application envisagés pour la thèse sont fournis par les différents partenaires : EDF (conception de ferme éolienne), Safran Tech (conception de turbomachine), IFPEN (conception de flotteur pour éolienne offshore), CEA (gestion de réseau de production et de distribution d'électricité), ce qui favorisera la découverte de différents acteurs du monde industriel ;

Lieu :

- Mines Saint-Etienne ou EDF Chatou (Ile de France)

Dates, durée du contrat :

- 3 ans à compter de l'automne 2021.

Profil recherché :

- Etudiant probabilités/statistiques/recherche opérationnelle, en M2 ou en Grande Ecole
- Bases solides en apprentissage statistique et optimisation
- Aisance en informatique, maîtrise de R, Python

Contacts :

leriche@emse.fr
sanaa.zannane@edf.fr
julien.pelamatti@edf.fr
merlin.keller@edf.fr

Références bibliographiques :

[Hutter et al., 2011] doi:10.1007/978-3-642-25566-3_40
[Pelamatti et al., 2019] doi:10.1007/s10898-018-0715-1
[Müller et al., 2013] doi:10.1016/j.cor.2012.08.022
[Costa and Nannicini, 2018] doi:10.1007/s12532-018-0144-7
[Jones et al., 1998] doi:10.1023/A:1008306431147
[Conn et al., 2009] doi:10.1137/1.9780898718768
[Tran et al, 2020] doi:10.1007/978-3-030-38364-0
[Regis, 2015] doi:10.1080/0305215X.2015.1082350
[Audet et al., 2019] doi:10.1137/18M1175872
[Le Digabel, 2011] doi:10.1145/1916461.1916468
[Cauwet et al., 2019] hal-02170283
[Audet and Dennis, 2006] doi:10.1137/040603371
[Munoz Zuniga and Sinoquet, 2020] doi:10.1080/03155986.2020.1730677
[Pelamatti et al., 2020] arxiv:2003.03300
[Bergstra et al., 2011] hal-00642998