Vers un contrôle par apprentissage par renforcement de l'alimentation électrique dans un avion hybride

Aubin Delaveau[1,2], Olivier Buffet[2] Florent Teichteil-Königsbuch[1],
[1] Airbus Central Research & Technology, F-31000, Toulouse
[2] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy

prenom.nom@airbus.com, prenom.nom@loria.fr

Résumé

Nous nous intéressons à la gestion et l'optimisation de la consommation électrique dans les avions hybrides. Face aux dynamiques non-linéaires coûteuses à simuler, les approches de contrôle par PID et MPC ne sont pas adaptées aux systèmes exigeant une grande réactivité. Nous proposons ici une formulation MDP du problème, présentons des résultats expérimentaux préliminaires obtenus avec une approche myope, et discutons des pistes envisageables dans le cadre de l'apprentissage par renforcement.

Mots-clés

Avion hybride, apprentissage par renforcement, stratégie myope.

1 Introduction

La gestion optimale de la consommation d'énergie est un objectif technologique majeur pour le développement des avions hybrides pour lesquels l'énergie électrique est produite à la fois par des batteries et les moteurs. Elle doit répondre à la variabilité des phases de vol (décollage, vol de croisière, atterrissage) et des états internes de l'avion.

Cependant, le contrôle du système électrique est régi par des équations différentielles ordinaires (EDO) non linéaires. Ces dynamiques rendent l'usage des méthodes de contrôle classiques, basées sur le calcul de gradients, particulièrement ardues, voire impraticables en raison de la nature "boîte noire" de la simulation et du bruit numérique inhérent à la résolution des EDO.

Travaux antérieurs Le contrôle de systèmes électriques repose classiquement sur des approches de type PID ou des méthodes de contrôle prédictif (MPC). Les correcteurs PID souffrent d'un manque de méthode de paramétrage générique face aux fortes non-linéarités des systèmes hybrides. Les approches MPC standard, elles, bien que performantes pour gérer des contraintes explicites, nécessitent généralement une linéarisation locale du modèle dynamique. Cette simplification entraîne souvent une perte de précision, tandis que la résolution directe du problème non linéaire est trop coûteuse pour du temps réel.

Face à ces limites, nous proposons une formalisation comme un processus de décision markovien (MDP), l'évolution de l'état du système se faisant à temps discret en simulant numériquement l'EDO sous-jacente. On va ainsi pouvoir considérer des approches de résolution "boîte noire" ne reposant pas sur des calculs de dérivés.

Plan La section suivante présente le formalisme MDP. Nous formalisons ensuite notre scénario comme un MDP et proposons une première stratégie de contrôle myope, c'est-à-dire optimale sur un pas de temps. De premiers résultats expérimentaux sont enfin présentés avant de discuter des limitations de notre approche et des pistes envisagées.

2 Processus de décision markoviens

Le cadre des processus de décision markoviens (MDP) fournit une formalisation mathématique standard pour les problèmes de contrôle séquentiel [3]. Un MDP, ici déterministe, est défini par un quadruplet $\langle S, A, T, R\rangle$ où S est l'espace des *états* ; A est celui des *actions* possibles ; $s' = T(s,a)$, la *fonction de transition*, indique l'état s' atteint quand l'action a est effectuée dans l'état s ; et $R(s,a,s')$, la *fonction de récompense*, associe une récompense scalaire immédiate à chaque transition.

La stratégie de contrôle est dictée par une *politique* $\pi : S \to A$, qui associe à chaque état une action. L'objectif est de trouver une politique optimale π^*, c'est-à-dire maximisant en tout s la fonction de valeur $V_\pi(s)$, définie comme l'espérance de la somme cumulée des récompenses actualisées : $V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s\right]$, où $\gamma \in [0,1[$ est le facteur d'actualisation. Ce formalisme permet d'aborder le problème de la commande optimale comme une recherche de politique π dans un espace continu, ici résolue par optimisation sans dérivée.

3 Approche proposée

Notre scénario Le système étudié est un réseau électrique d'avion hybride composé de deux générateurs associés aux moteurs (HP et LP), d'une batterie, et d'une charge utile (CPL).

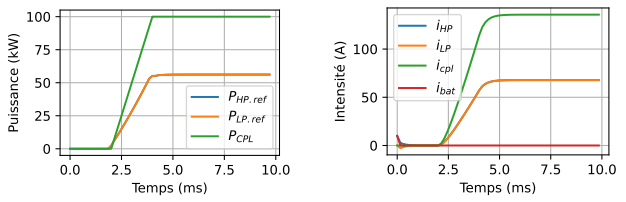
La dynamique du système est gouvernée par un système d'équations différentielles ordinaires (EDO) non linéaires : $\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}, P_{CPL}(t))$, où $\mathbf{x}$ représente le vecteur des variables d'état internes (ex : tensions des bus, courant batterie) ; $\mathbf{u}$ est le vecteur des variables de contrôle (puissances fournies par les générateurs) ; et $P_{CPL}(t)$ est la puissance consommée par la charge, entrée exogène connue a priori.

Formalisation MDP Le système est modélisé comme un MDP comme suit : on discrétise le temps, en supposant ici un pas $\delta t = 1$ pour pouvoir noter t le temps dans le modèle physique comme dans le modèle MDP ; l'**état** est défini par le vecteur $s_t = [\mathbf{x}_t, t]^T$, où $\mathbf{x}_t$ représente les variables internes de l'avion et t l'instant courant ; l'**action** $a_t \in A$ correspond aux puissances de commande $\mathbf{u}_t$ appliquées aux générateurs pendant δt, en imposant un certain ratio entre ces puissances ; la **transition** $s_{t+1} = T(s_t, a_t)$ s'obtient en résolvant une équation aux dérivées ordinaires (EDO) sur un pas de temps ; pour ici minimiser l'emploi de la batterie (donc le courant la traversant), la **récompense** instantanée correspond à $-\|i_{bat}(s_{t+1})\|^2$. On notera que la dynamique de P_{CPL} est une entrée exogène connue (déterministe), indépendante de l'action a_t, et intégrée dans l'EDO.

Algorithme de contrôle Les espaces d'états et d'actions étant continus, un tel MDP serait typiquement abordé par un algorithme d'apprentissage par renforcement profond tel que *proximal policy optimization* [6]. Dans un premier temps, afin d'obtenir rapidement une première solution pragmatique à notre problème, nous nous limitons à mettre en œuvre une stratégie myope. À chaque pas de temps, étant donné s_t, on cherche l'action a_t maximisant la récompense immédiate. On doit donc optimiser une fonction $J(a) \stackrel{\text{def}}{=} -\|I_{bat}(T(s_t, a))\|^2$ non dérivable, mais simulable par l'EDO Φ, dans un espace continu. Nous employons à cette fin l'algorithme COBYQA (*Constrained Optimization BY Quadratic Approximation*) [5].

4 Résultats expérimentaux

L'algorithme de contrôle proposé a été évalué sur un premier circuit type, permettant de satisfaire la demande de la charge tout en assurant la stabilité du système. La figure 1a illustre la répartition des puissances entre les générateurs (courbes superposées), démontrant la capacité du solveur à gérer la dynamique non linéaire. L'intensité batterie i_{bat} est maintenue proche de zéro (figure 1b), confirmant que les générateurs couvrent la demande de la charge.
Réduire le pas de temps conduit toutefois à de fortes oscillations de i_{bat}, illustrant les limites d'un contrôle myope dans certaines situations.



(a) Puissances de commande (b) Intensités sources et charge

FIGURE 1 – Évolution de diverses grandeurs

5 Discussion et Perspectives

Discussion La stratégie proposée permet d'atteindre des trajectoires de contrôle stables avec une intensité batterie maintenue proche de zéro dans des situations simples. Toutefois, cette approche comporte des verrous techniques : 1. une vision "myope" qui ne garantit pas une optimalité globale sur l'horizon temporel, en particulier pour des dynamiques rapides de $P_{CPL}(t)$, 2. une latence computationnelle élevée (30–60 s) incompatible avec le temps réel, et 3. une dépendance à une connaissance immédiate des états.

Perspectives Pour lever ces verrous, nous envisageons plusieurs axes de recherche dans le contexte de l'apprentissage par renforcement : se tourner vers le *Deep RL* en espaces d'états et d'actions continus (par ex. avec PPO [6]), et éventuellement guider un *apprentissage par imitation* avec notre stratégie myope [4], [7] ; exploiter les *Physics-Informed Neural Networks* [2] pour intégrer la connaissance du modèle physique dans l'apprentissage ; intégrer la présence de *délais d'observation* dans la boucle de contrôle [1].
Cette approche hybride, combinant la robustesse de l'optimisation sans dérivée et la capacité de généralisation des modèles neuronaux, constitue une voie prometteuse pour le contrôle embarqué haute performance des avions hybrides.

Références

[1] M. AGARWAL et V. AGGARWAL, "Blind decision making : Reinforcement learning with delayed observations," *Pattern Recognition Letters*, t. 150, p. 176-182, 2021.

[2] C. BANERJEE, K. NGUYEN, C. FOOKES et M. RAISSI, "A survey on physics informed reinforcement learning : Review and open problems," *Expert Systems with Applications*, t. 287, p. 128 166, 2025.

[3] R. BELLMAN, "A Markovian Decision Process," *Journal of Mathematics and Mechanics*, t. 6, n° 5, p. 679-684, 1957.

[4] G. LIBARDI, G. D. FABRITIIS et S. DITTERT, "Guided Exploration with Proximal Policy Optimization using a Single Demonstration," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. MEILA et T. ZHANG, éd., sér. Proceedings of Machine Learning Research, t. 139, PMLR, 2021, p. 6611-6620.

[5] T. M. RAGONNEAU, "Model-based derivative-free optimization methods and software," thèse de doct., Hong Kong Polytechnic University, 2023.

[6] J. SCHULMAN, F. WOLSKI, P. DHARIWAL, A. RADFORD et O. KLIMOV, "Proximal policy optimization algorithms," *arXiv preprint arXiv :1707.06347*, 2017.

[7] M. ZARE, P. M. KEBRIA, A. KHOSRAVI et S. NAHAVANDI, "A survey of imitation learning : Algorithms, recent developments, and challenges," *IEEE Transactions on Cybernetics*, t. 54, n° 12, p. 7173-7186, 2024.