Apprentissage par renforcement appliqué à la commande des systèmes électriques

Jing Dai, Yannick Phulpin, Jean-Claude Vannier, Damien Ernst

Conférence EF 2009
UTC, Compiègne, 24-25 Septembre 2009

APPRENTISSAGE PAR RENFORCEMENT APPLIQUE A LA COMMANDE DES SYSTEMES ELECTRIQUES

DAI Jing, PHULPIN Yannick*, VANNIER Jean-Claude* et ERNST Damien°*
**SUPELEC, 3 rue Joliot Curie, 91192 Gif sur Yvette, France*
°FNRS, Université de Liège, Grande Traverse, 10, Sart-Tilman, B-4000 Liège, Belgique
{jing.dai, yannick.phulpin, jean-claude.vannier}@supelec.fr, dernst@ulg.ac.be

Résumé

Cet article propose une revue de littérature concernant les applications de l'apprentissage par renforcement à la commande des systèmes électriques. L'apprentissage par renforcement a pour caractéristique principale de résoudre des problèmes de commande optimale à partir de la seule observation des trajectoires du système. Il présente l'intérêt de ne pas requérir de connaissance à priori sur la dynamique du système à commander et convient ainsi aux problèmes de commande des systèmes complexes. Dans un premier temps, l'article détaille les caractéristiques des problèmes auxquels l'apprentissage par renforcement s'applique, puis cette technique est décrite. Ensuite, deux exemples classiques d'application aux systèmes électriques sont présentés.

Mots Clés. Apprentissage par renforcement, Systèmes électriques, Réseaux, Commande.

Dans un contexte global de mutation des systèmes électriques, la commande de leurs différents éléments représente un enjeu de taille. En effet, il est d'autant plus difficile de déterminer des lois de commande performantes et robustes que les systèmes électriques sont des systèmes complexes : ils sont difficiles à modéliser du fait du grand nombre d'éléments interconnectés qui ont un comportement généralement non-linéaire et parfois discontinu, et ils sont sujets à des variations soudaines [1]. Les enjeux techniques associés aux systèmes électriques ont rapidement requis la mise en œuvre de schémas de commande afin de s'assurer de l'alimentation permanente des charges [2]. Ceux-ci ont généralement été synthétisés de manière empirique avant de s'orienter vers des schémas de commande basés sur modèles physiques. Or la modélisation des éléments d'un réseau implique une simplification qui peut nuire, dans le cas d'un système complexe, à la performance du système. Par exemple, pour modéliser un dispositif muni d'électronique de puissance, il est courant d'utiliser un modèle moyen de ses convertisseurs au risque de négliger certaines discontinuités qui pourraient avoir un impact sur le système [3]. Il est donc nécessaire que les lois de commande soient suffisamment robustes face aux erreurs introduites lors de la modélisation.

Alors que l'interconnexion progressive des systèmes électriques et la diversification des éléments qui y sont connectés atteignent un stade qui pourrait mettre en cause les modes de commande de ces systèmes [4], cet article propose de s'interroger sur le potentiel de nouveaux outils de commande basés sur une modélisation du système indépendante des niveaux physiques. Cette technique nommée apprentissage par renforcement (on utilisera dans cet article l'abréviation RL pour *reinforcement learning*) a pour but d'optimiser itérativement les lois de commande du système en vue de maximiser un indice de performance cumulée à long terme [5]. Cela revient à apprendre certaines caractéristiques macroscopiques du système en interagissant avec celui-ci, puis prendre des décisions en fonction de son état actuel. L'apprentissage par renforcement a déjà été appliqué avec succès aux systèmes électriques [6-8] ainsi que dans d'autres domaines tels que par exemple la robotique [9], la médecine [10] ou encore les télécommunications [11]. Dans cet article, nous allons donner un aperçu général des bénéfices potentiels de son application aux systèmes électriques.

L'article est organisé comme suit. Dans un premier temps, nous allons formaliser certains types de problèmes qui peuvent être traités par apprentissage par renforcement. Puis, le principe de fonctionnement de cette méthode sera présenté. Ensuite, deux exemples classiques d'application de RL à la commande des systèmes électriques seront détaillés. Enfin, la conclusion fournira quelques commentaires sur la performance de cette approche et sur les nouvelles applications envisagées.

ISBN 978-2-913923-30-0

1. Problèmes de commande optimale pour des systèmes à temps discret

Le contrôle RL s'applique typiquement à des processus de décision markoviens. Dans cet article, nous nous concentrons sur les problèmes déterministes, où les équations qui régissent la dynamique n'impliquent pas de variables aléatoires. A tout instant t, le système est représenté par un état x_t. Une action u_t est alors appliquée et le système évolue vers l'état x_{t+1} à l'instant $t+1$. Cette transition peut s'exprimer sous la forme de l'équation suivante :

$$x_{t+1} = f(x_t, u_t). \quad (1)$$

A tout instant t, la performance du système est représentée par une récompense immédiate r_t, qui peut être déterminée comme suit:

$$r_t = \rho(x_t, u_t). \quad (2)$$

Si l'on considère un état de départ x_0 et une séquence d'actions de commande $u = (u_0, u_1, ..., u_{T-1})$ sur un horizon de temps T, on peut effectuer une somme pondérée des indices de performances du système de sorte à déterminer une performance cumulée:

$$J_T^u(x_0) = \sum_{t=0}^{T-1} \gamma^t r_t, \quad (3)$$

où γ est le facteur de dévaluation qui permet de prendre en compte les récompenses à plus ou moins loin long terme.

Dans ce contexte, la stratégie de commande à mettre en place doit mener à une séquence de d'actions de commande optimale $u^* = (u_0^*, u_1^*, ..., u_{T-1}^*)$ telle que:

$$u^* \in \max_u J_T^u(x_0). \quad (4)$$

2. Description générale de l'approche RL

L'approche RL est dite « agent-based ». Elle suppose en fait qu'un agent intelligent, qui ignore les fonctions f et ρ, est capable d'observer l'état du système x_t et la valeur de l'indice de performance r_t et détermine la valeur de u_t à tout instant t.

On peut identifier deux grandes familles de techniques en RL: celles qui apprennent un modèle et puis utilisent ce modèle pour en déduire une loi de commande (quasi-)optimale et celles qui déterminent une loi de commande (quasi-) optimale sans passer par l'apprentissage d'un modèle. Dans cet article, nous nous intéressons aux techniques d'apprentissage par renforcement qui ne nécessitent pas de devoir apprendre un modèle. Souvent, ces techniques reposent sur le calcul d'une fonction de valeur (nommée fonction V) qui dépend de l'état initial du système. Elle représente la performance cumulée maximale que l'on peut obtenir si le système évolue à partir de cet état initial en étant soumis à une séquence d'actions de commande optimale. Alternativement, on peut aussi définir cette fonction de valeur pour chaque paire état-action (x,u) (nommée fonction Q dans ce cas) qui représente la performance cumulée maximale que l'on peut obtenir quand on applique cette action à partir de cet état initial et que le système reçoit une commande optimale par la suite. L'agent intelligent cherche à estimer cette fonction de valeur en interagissant avec le système. Après une période d'entraînement, on considère qu'il en a fait une bonne estimation et il peut alors choisir, pour un état donné, l'action dont l'espérance de performance cumulée est la plus grande. Il existe différents algorithmes pour estimer la fonction de valeur (ou la fonction Q) sans estimer de modèle. Parmi ceux-ci on peut citer les algorithmes basés sur le concept de différence temporelle (e.g., SARSA [5] et Q-learning [12,13]) ou encore les algorithmes de la classe fitted-Q iteration qui sont basés sur l'algorithme d'itération sur les valeurs [14].

3. Exemples d'application

Certains problèmes classiques en commande des systèmes électriques ont été formalisés comme des problèmes de commande optimale à temps discret. Cette section présente deux types d'applications : en temps-réel, et en mode off-line.

ISBN 978-2-913923-30-0

L'approche RL en temps-réel consiste à laisser l'agent intelligent interagir directement avec le système, sans construire de modèle mathématique du système, comme indiqué dans la Figure 1. Dans [6], Ernst *et al.* appliquent cette approche à la commande d'un TCSC (condensateurs série contrôlés par thyristor) pour amortir les oscillations d'un système à dix nœuds avec quatre générateurs et deux charges. Un autre exemple est abordé dans [7], où l'approche RL en temps-réel est utilisée pour résoudre le problème de la commande de génération automatique dans un système à deux régions liées par une ligne d'interconnexion.

Avec l'approche RL off-line (Figure 2), l'agent est soumis à un ensemble de scénarii afin d'apprendre un certain nombre de caractéristiques du système. Il peut ensuite extraire une loi de commande à appliquer au système ou comparer les résultats obtenus avec différentes lois de commande génériques. Ainsi, dans [6], un contrôleur de frein résistif est conçu pour améliorer la stabilité transitoire d'un réseau à dix nœuds. Un autre exemple est la commande d'un TCSC dans un simple système à un générateur et une charge dans [8].

Les deux sections suivantes détaillent les applications à la commande d'un frein résistif et d'un TCSC présentées en [6] et [8].

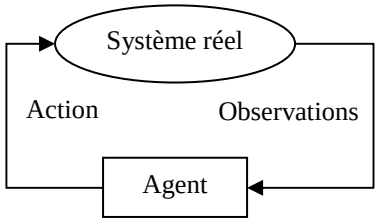

Figure 1.Commande RL temps-réel [6]

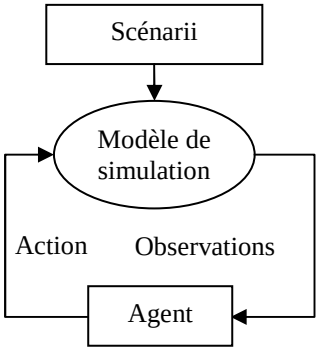

Figure 2. Commande RL off-line [6]

3.1 Conception d'un contrôleur dynamique de frein

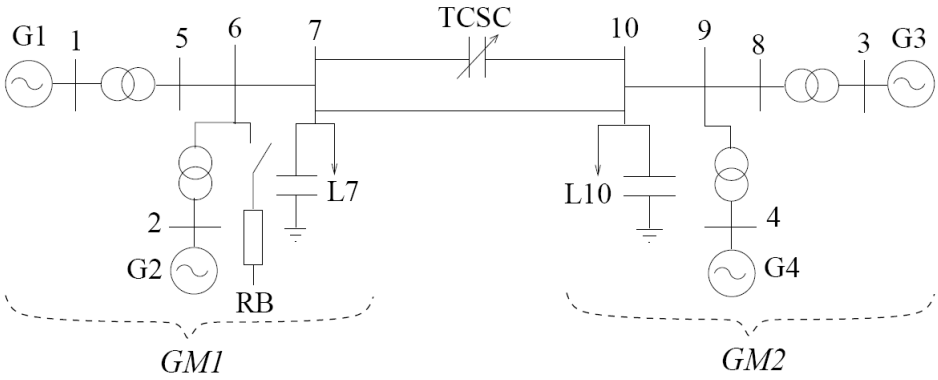

Figure 3. Système électrique à dix nœuds et deux régions [6]

Le système utilisé pour cette application compte deux régions comportant chacune deux générateurs et une charge (Figure 3). Suite à un défaut franc triphasé dans le système, on peut activer le frein résistif (RB) au nœud 6 afin d'éviter la perte de synchronisme entre les deux zones. La loi de commande à synthétiser doit donc déterminer à chaque instant si le frein résistif doit être activé.

Les générateurs sont modélisés par un modèle détaillé, avec une excitation à courant continu, un régulateur automatique de tension et un régulateur de vitesse, alors que les charges sont modélisées comme un courant constant et une impédance constante. Pour accélérer la résolution du problème, on réduit le modèle initial à 60 dimensions à un modèle à deux variables d'états, à savoir l'angle (δ_t) et la vitesse (ω_t) relatifs entre les deux zones, qui fournissent l'information nécessaire tant pour décider des actions de commande que pour mesurer la performance. La fonction de récompense est définie comme suit en fonction de la stabilité du système

$$r(x_t, u_t) = \begin{cases} -|\omega_t| - c \cdot u_t & \text{if } |\delta_t| < \pi \text{ rad} \\ -1000 & \text{if } |\delta_t| < \pi \text{ rad} \end{cases} \quad (5)$$

où $u_t \in \{0,1\}$ indique si le frein est activé à l'instant t et c détermine le degré auquel on pénalise l'activation du frein.

ISBN 978-2-913923-30-0

La phase d'apprentissage consiste en différents scenarii, où le système est soumis après 10 s de régime statique à un défaut franc triphasé au nœud 10 dont la durée est déterminée aléatoirement entre 250 et 350 ms. Le système entre ensuite en régime transitoire jusqu'à ce qu'une instabilité soit détectée ou que 60 s se soient écoulées. 1000 scénarii sont générés pour l'apprentissage, dont 163 instables. Pendant l'apprentissage, on adopte un facteur d'avidité (ε) de 0.1, c'est-à-dire, l'agent applique avec une probabilité de 90% une action « optimale » basée sur sa connaissance déjà obtenue, ou essaie une action au hasard pour explorer.

Le résultat des simulations montre qu'après un apprentissage de 1000 scénarii, on a obtenu une commande structurée sur toute l'espace d'état. La police de commande ainsi obtenue permet au système de subir sans perte de stabilité un défaut franc d'une durée de 350 ms, alors que sans le frein, le défaut ne peut être plus long que 215 ms avant que le système perde le synchronisme. La robustesse de ce mode de commande est aussi testée pour un défaut franc se trouvant au nœud 7, pour lequel on arrive à augmenter cette durée maximale du défaut de 141 ms, sans frein, à 252 ms.

3.2 Apprentissage on-line pour commander un TCSC

Le deuxième exemple porte sur la commande d'un TCSC pour amortir l'oscillation de puissance électrique. Il s'appuie également sur le modèle de système électrique représenté en Figure 3, où un TCSC est placé entre les deux zones. Les paramètres de l'AVR des générateurs sont modifiés pour que le système soit initialement amorti négativement.

Contrairement au premier exemple, où on dispose d'un modèle de simulation pour exercer l'agent, on adopte dans cet exemple le mode on-line, où l'agent commande le système tout en apprenant de nouvelles trajectoires. Pour ce faire, les grandeurs utilisées par l'agent doivent être définies à partir de mesures en temps réel. Pour concevoir une police de commande locale, il faut aussi choisir une grandeur mesurable proche de l'emplacement du TCSC. Ainsi, la puissance (P_e) transférée dans la ligne à laquelle le TCSC est connecté est supposée mesurée. A travers cette grandeur uniquement, le système est partiellement observable. Pour compenser ce problème, on définit un pseudo-état qui contient les observations et les actions prises à certains instants précédents. A l'occurrence, ce pseudo-état à instant t est défini comme les puissances injectées aux instants t, $t-1$ et $t-2$, ainsi que les actions prises à instants $t-1$ et $t-2$. Quant à la variable de commande, elle est définie comme la réactance du TCSC, qui prend 5 valeurs discrètes entre sa zéro et sa capacité maximale. La fonction de récompense dépend de l'état de stabilité du système, et est défini comme suit :

$$r(x,u) = \begin{cases} -\left|P_e - \overline{P_e}\right| & \text{if } \left|P_e\right| \le 250\text{MW} \\ -1000 & \text{if } \left|P_e\right| > 250\text{MW} \end{cases}$$

où $\overline{P_e}$ représente la valeur correspondante en régime permanent, qui est estimée comme la moyenne au cours de 60 dernières secondes.

Les simulations avec une charge constante montrent qu'au bout de 10 heures après la mise en œuvre de l'agent, il arrive à déduire une loi de commande qui enlève l'oscillation initialement présente dans le système non compensé. Les petites fluctuations restantes sont dues à la discrétisation. La loi de commande a aussi été testée avec une charge périodique, dont le résultat montre que la police est convient également aux conditions stochastiques.

4. Conclusion

Dans cet article, nous avons mis en évidence le potentiel d'application de l'apprentissage par renforcement dans la commande de systèmes électriques en nous appuyant sur deux exemples pratiques. Pour ces deux applications possédant un nombre relativement faible de variables d'états, des algorithmes d'apprentissage par renforcement assez simples se sont avérés performants

En revanche, pour des problèmes impliquant plusieurs centaines ou milliers de variables d'états, comme c'est souvent le cas pour un grand système électrique, il faudra appliquer des algorithmes plus sophistiqués de sorte à décomposer le problème et/ou réduire l'espace d'état. Les approches de RL hiérarchique, l'agrégation d'état, etc. nous semblent particulièrement prometteuses à cet égard. Cette perspective s'annonce comme un défi scientifique, car la recherche dans le domaine de RL s'applique le plus souvent à des problèmes dont la solution peut-être trouvée en

ISBN 978-2-913923-30-0

grande partie en se basant sur l'intuition humaine alors que ce n'est souvent pas le cas lorsque l'on traite des problèmes de commande liés aux grands réseaux électriques.

REFERENCES

[1] JENNINGS N.R et BUSSMANN S., Agent-based control systems: Why are they suited to engineering complex systems?**,** IEEE Control Systems Magazine, Vol. 23, No. 3, p. 61-73 (Juin 2003).

[2] BOUNEAU C., DERDEVET M. et PERCEBOIS J., Les Réseaux Electriques au Cœur de la Civilisation Industrielle, Timée-éditions, (2007).

[3] USMAN IFTIKHAR M., GODOY E., LEFRANC P., SADARNAC D. et KARIMI C., A control strategy to stabilize PWM DC-DC converters with input filters using state-feedback and pole-placement, Proc. de INTELEC, p. 1-5, (Septembre 2008).

[4] BIALEK J.W., Why has it happen again ? Comparison between the UCTE blackout in 2006 and the blackouts of 2003, Proc. de Powertech, Lausanne, Suisse, p. 51-56, (Juillet 2007).

[5] SUTTON R.S. et BARTO A. G., Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, (1998).

[6] ERNST D., GLAVIC M., et WEHENKEL L., Power systems stability control: reinforcement learning framework, IEEE Transactions on Power Systems, Vol. 19, p. 427-435, (Février 2004).

[7] IMATHIAS AHAMED T. P., NAGENDRA RAO P. S. et SASTRY P. S., A reinforcement learning approach to automatic generation control, Electric Power Systems Research 63 9/26, (2004).

[8] ERNST D., GLAVIC M., CAPITANESCU F. et WEHENKEL L., Learning versus model predictive control: a comparison on a power system problem, IEEE Transactions on Systems, Man and Cybernetics - Part B, (2008).

[9] HAFNER R. et RIEDMILLER M., Neural reinforcement learning controllers for a real robot application. ICRA 2007, p. 2098-2103:

[10] MURPHY S. A., Optimal dynamic treatment regimes, Journal of the Royal Statistical Society, Series B 65 (2): 331-366, 2003

[11] SINGH S. et BERTSEKAS D., Reinforcement learning for dynamic channel allocation in cellular telephone systems, Advances in Neural Information Processing Systems 9 (NIPS), p. 974-980, 1997.

[12] WATKINS C., Learning From Delayed Rewards, Thèse de Cambridge University, (1989).

[13] WATKINS C. et DAYAN P., Q-learning, Machine Learning, Vol. 8, p. 279-292 (1992).

[14] ERNST D., GEURTS P., et WEHENKEL L., Tree-based batch mode reinforcement learning," Journal of Machine Learning Research, Vol. 6, p. 503-556, April 2005.

ISBN 978-2-913923-30-0