Contributions au contrôle et au dimensionnement des micro-réseaux par apprentissage par renforcement : application aux systèmes avec production renouvelable et stockage hybride batterie-hydrogène

Valentin Père

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

IMT Mines Albi-Carmaux
École Mines-Télécom

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

délivré par
IMT – École Nationale Supérieure des Mines d'Albi-Carmaux

---

présentée et soutenue par

Valentin PÈRE

le 7 décembre 2023

Contributions au contrôle et au dimensionnement des micro-réseaux par apprentissage par renforcement : application aux systèmes avec production renouvelable et stockage hybride batterie-hydrogène

---

École doctorale et discipline ou spécialité :
MEGEP : Énergétique et Transferts

Unité de recherche :
Centre RAPSODEE, UMR CNRS 5302, IMT Mines Albi

Direction de thèse :
Jean-Louis DIRION, Maître assistant HDR, IMT Mines Albi

Autres membres du jury :
Stéphane GRIEU, Professeur, Université de Perpignan, Rapporteur
Robin ROCHE, Professeur, Université de Franche-Comté, Rapporteur
Hossam AFIFI, Professeur, Telecom SudParis, Examinateur
Catherine AZZARO-PANTEL, Professeure, INP ENSIACET, Présidente
Vincent DEBUSSCHERE, Maître de conférence HDR, Université Grenoble Alpes, Examinateur
Fabien BAILLON, Maître-assistant, IMT Mines Albi, Co-encadrant
Mathieu MILHÉ, Maître-assistant, IMT Mines Albi, Co-encadrant, Invité
Rachid OUARET, Enseignant-chercheur, INP ENSIACET, Invité

Remerciements

Je tiens à remercier Jean-Louis Dirion, mon directeur de thèse. Je le remercie d'avoir placé sa confiance en moi depuis le début. Il a su me guider en m'empêchant de partir dans de mauvaises directions. Un directeur attentif à ce que je raconte, tout en ayant un franc-parler pour donner son avis, je n'aurais pas pu rêver mieux ! Merci Mathieu Milhé pour sa force de proposition et son humour au quotidien, des idées, du rire, je n'aurais pas pu m'en passer. Merci également à Fabien Baillon pour m'avoir fait profiter de son recul et son expertise technique sur divers problèmes qui gravitent autours de mes recherches. J'aurais été souvent au point mort sans lui. Ces trois personnes m'ont laissé parcourir mon bout de chemin avec bienveillance, en se rendant disponibles et en me réorientant au moindre égarement. On formait une super équipe et vous allez me manquer.

Je tiens également à remercier Claire Billet, qui m'a donné l'opportunité de connaître la recherche en IA et en énergie. Cette thèse en fut la suite logique. Merci à Rachid Ouaret pour les échanges mutuellement enrichissants, les pistes évoquées. Merci également à Robin Roche et à Stéphane Grieu qui m'ont fait l'honneur d'être rapporteurs de ma thèse. Merci aux membres du jury, Hossam Afifi, Catherine Azzaro-Pantel et Vincent Debusschere pour cette série de questions plus intéressantes les unes que les autres.

Si cette thèse me rend si fier aujourd'hui, c'est parce que sa construction a été une bataille au long terme, que l'on a fini par remporter. De l'appropriation du sujet à la correction de la conclusion générale, rien n'a été lisse. Mon encadrement a eu l'oeil, ils ont su déceler que je travaille mieux dans un environnement de confiance, et ont tout mis en œuvre pour que cela se déroule harmonieusement. Cependant, l'aventure de la thèse constitue une part importante de ma vie, qui dépasse le simple temps de travail. Cette confiance, je la dois aussi à mes proches, à ma famille, à mes amis. Ce bien être quotidien, ces zones de lumières dans laquelle je peux être moi-même et avancer avec assurance, vous avez su me les apporter aux moments où j'en avais le plus besoin.

Sommaire

Introduction générale

Selon Valérie Masson-Delmontte, ancienne co-présidente du groupe de travail 1 du GIEC, « Chaque degré compte, chaque année compte et chaque décision compte : ne pas agir aujourd'hui c'est ajouter au fardeau des générations futures. Limiter le réchauffement à 1,5 °C n'est pas impossible mais nécessite une politique forte et immédiate. ». La réduction des gaz à effet de serre est un challenge à relever à toutes les échelles. En France, l'industrie de l'énergie représente 11 % des émissions de CO_2 en 2022 (Citepa, 2023). Le secteur du transport routier est le plus émetteur, avec 32,3 % des émissions totales. La décision du Parlement Européen d'interdire la vente de voitures neuves à essence et au diesel à partir de 2035 (*Règlement Du Parlement Européen et Du Conseil* 2023) va fortement accélérer l'industrialisation des véhicules électriques. Le secteur de l'énergie doit s'adapter rapidement sans augmenter son impact carbone.

Les énergies renouvelables sont les sources d'énergie peu carbonées ayant le plus grand potentiel de déploiement pour leur prix et leur facilité à s'intégrer sur le territoire (H. Lee et al., 2023). En revanche, ces énergies ont des potentiels de production variables et peu contrôlables, c'est le cas des énergies éolienne et solaire photovoltaïque. La topologie du réseau électrique français permet de les intégrer de manière décentralisée au réseau de distribution. Avec l'augmentation des sources de production d'énergie renouvelable, le réseau de distribution doit s'adapter à des systèmes de production décentralisés (Lebrouhi et al., 2022). Les points de charge des véhicules électriques vont fortement augmenter la demande sur ces réseaux. Le gestionnaire du réseau de distribution est le garant de la sécurité et de la qualité de fourniture de l'électricité, et il doit faire face à des challenges de flexibilité et de résilience.

Dans une répartition distribuée des installations de production, le stockage d'électricité permet de lisser les intermittences. Grâce au stockage électrique, l'électricité produite peut être conservée puis utilisée à des moments où la production est faible ou inexistante. Le micro-réseau électrique permet de regrouper localement stockages, points de consommation et unités de production (au sein d'une résidence, d'un quartier ou d'une zone commerciale économique par exemple). Ces systèmes permettent d'atteindre l'équilibre entre la demande et la production lorsqu'elle est intermittente. Leur adaptabilité leur permet de les déconnecter du réseau de distribution en cas de besoin. Les conditions pour garantir autonomie et viabilité économique reposent sur un dimensionnement adapté et une gestion intelligente des unités de stockage. L'opération des micro-réseaux repose sur des variables aléatoires, comme les variables météorologiques ou la consommation électrique à chaque instant. Ainsi, l'équilibre du réseau local dépend d'unités pilotables et de variables incontrôlables. Le contrôle de micro-réseaux électriques s'effectue en adaptant les consignes de charge des systèmes de stockage, en décalant l'utilisation de certains appareils et en modifiant des puissances en consigne d'appareils contrôlables par exemple. Ces consignes sont prises par des systèmes de gestion de l'énergie ou *Energy Management Systems* (EMS). La modélisation analytique des

différentes unités qui constituent le micro-réseau affine la précision des consignes de contrôle de l'EMS en intégrant les dynamiques et les interactions entre les différentes composantes. De plus, un modèle basé sur des données est essentiel afin qu'il puisse mieux prendre en considération les variables aléatoires.

L'objectif de cette thèse est de développer une méthodologie de dimensionnement et de contrôle d'un micro-réseau comportant une source d'énergie photovoltaïque, une batterie électrochimique et un stockage hydrogène adaptés aux demandes et productions électriques. La modélisation du micro-réseau inclut des modèles dynamiques des unités et des données pour simuler les phénomènes aléatoires. La stratégie de contrôle est construite à partir de l'apprentissage par renforcement profond, qui permet de prendre des décisions séquentielles dans un environnement soumis à de fortes incertitudes. Le dimensionnement est effectué grâce à une méthode d'optimisation bi-niveau avec une méta-heuristique employée pour l'exploration de l'espace de solution en boucle externe et la stratégie de contrôle du micro-réseau développée en boucle interne. Une attention particulière est attribuée à la réduction du temps de calcul pour l'apprentissage des politiques de contrôle. Cet apprentissage peut s'avérer long et est répété dans le processus de dimensionnement bi-niveau.

Les verroux scientifiques suivants seront levés dans ce travail de thèse :

- L'étude de l'influence des paramètres de l'apprentissage par renforcement sur la politique de contrôle du stockage hydrogène dans un micro-réseau n'est pas explicite dans la littérature. En particulier, comment ajuster la valeur des paramètres pour assurer une stratégie de contrôle pertinente quand l'environnement de simulation se complexifie ?
- La simulation du micro-réseau inclut des modèles dynamiques des unités et des données pour simuler des phénomènes aléatoires. Son contrôle doit être mené à long terme en tenant compte de la dégradation de la batterie que cela implique. Comment construire, évaluer et analyser une politique de contrôle dans une simulation à horizon temporel long comprenant l'usure du stockage à court terme ?
- Le dimensionnement du micro-réseau comprend des coûts d'investissement, de maintenance et d'opération. Les coûts d'opération doivent être calculés selon des objectifs d'autonomie ou de rentabilité en tenant compte des variables aléatoires. Comment garantir l'inclusion d'incertitudes liées à la différence entre l'environnement de développement d'une stratégie de contrôle et son application réelle, dans la méthodologie de dimensionnement d'un micro-réseau, tout en garantissant des résultats optimaux ?
- Le temps d'entraînement des agents d'apprentissage par renforcement est conséquent. La volonté de considérer les coûts d'opération dans le processus de dimensionnement implique l'élaboration d'une stratégie de contrôle pour chaque dimensionnement considéré, ce qui accroît le temps de calcul. Quelle solution intégrer à la méthodologie de dimensionnement et de contrôle pour réduire ce temps de calcul ?

Ce manuscrit de thèse est organisé en deux parties. La première définit le contexte et le cadre scientifique et la seconde présente les contributions scientifiques apportées par ce travail. Elles sont décomposées en trois chapitres, chacune de la manière suivante :

Partie 1 : Contexte et Cadre Scientifique

1 Le Chapitre 1 définit les micro-réseaux électriques et leur intérêt dans le contexte actuel. Leur topologie et modes d'opération sont introduits. Le contrôle est classé en différentes catégories et les objectifs du dimensionnement sont présentés.

2 Les principes de l'apprentissage par renforcement sont introduits dans le Chapitre 2. Ce chapitre permet d'appréhender le fonctionnement des algorithmes et leur évolution

avec notamment l'intégration de l'apprentissage profond pour gérer des problèmes plus complexes. Une revue des différents algorithmes est proposée, mettant en lumière leurs caractéristiques distinctives afin de pouvoir sélectionner l'algorithme pertinent pour le travail proposé.

3 Le Chapitre 3 présente une analyse exhaustive et méthodique des travaux antérieurs sur l'application de l'apprentissage par renforcement au contrôle des micro-réseaux. Cet état de l'art est décomposé selon la nature des problèmes de contrôle résolus par les algorithmes d'apprentissage par renforcement.

Partie 2 : Contributions scientifiques

4 La modélisation du micro-réseau choisie est détaillée dans le Chapitre 4. Il expose la méthode de contrôle et son implémentation, ainsi que la formulation et le cadrage du problème de dimensionnement du micro-réseau. Ce chapitre présente l'interdépendance entre le contrôle et le dimensionnement et justifie le développement d'une méthodologie de dimensionnement couplée au contrôle à long terme du micro-réseau.

5 Le Chapitre 5 expose les travaux réalisés sur le contrôle du micro-réseau. Les différentes études de sensibilité sur l'apprentissage des modèles, leur adaptation à l'introduction de la dégradation des systèmes de stockage et l'analyse des stratégies développées par l'EMS sont présentées. Les paramètres de l'apprentissage par renforcement sont réajustés afin de développer une politique généralisable à des horizons temporels de simulation plus longs que ceux sur laquelle elle a été construite. Une analyse technico-économique est menée sur les différentes stratégies développées.

6 Le Chapitre 6 est consacré au dimensionnement du micro-réseau. La méthode de transfert de politique par apprentissage hors-ligne est d'abord introduite, suivie d'une explication détaillée sur le choix de l'algorithme d'optimisation bi-niveau. Le chapitre se conclut par la présentation des résultats. Des indicateurs permettent de mener une analyse multi-critère sur la fiabilité du micro-réseau tel qu'il est dimensionné par la méhode d'optimisation proposée. La performance et la fiabilité de la méthodologie sont analysées selon le temps de calcul et sa capacité à converger efficacement vers un optimum global.

Enfin, les conclusions tirées de ce travail de thèse sont présentées, suivies des perspectives et des recommandations pour des recherches futures.

I

Contexte et cadre scientifique

1

Les micro-réseaux électriques

Le rapport d'analyse de la politique énergétique de la France en 2021 par l'International Energy Agency (International Energy Agency, 2021) recommande au gouvernement d'accélerer le déploiement des énergies renouvelables décentralisées afin de tenir ses objectifs de neutralité carbone fixés pour 2050. Les ressources fossiles sont progressivement abandonnées en raison de préoccupations environnementales, et les structures nucléaires du parc français vieillissent, nécessitant souvent des maintenances et rénovations (Transport d'Électricité, 2023 et France, 2022). Par conséquent, la maîtrise de sources d'énergies décarbonnées et résilientes devient indispensable. La production centralisée d'énergie renouvelable ne constitue pas une solution viable pour compenser cette diminution de production (Cany et al., 2016), car elle est fortement affectée par les fluctuations intermittentes de la disponibilité des ressources. La décentralisation de la production d'énergie offre des opportunités pour déployer les énergies renouvelables localement à travers le territoire. Compatible avec la production centralisée d'énergie, elle favorise l'accès à des énergies propres à des communautés éloignées des centrales de productions. De ce fait, une gestion efficace et adaptable de ces ressources devient d'autant plus cruciale. D'après International Energy Agency et al., 2021, l'intégration d'une part importante d'énergies renouvelables nécessite une utilisation optimale de toutes les ressources de flexibilité, qu'elles soient en génération, en stockage ou en demande énergétique.
Par l'offre de solutions innovantes pour l'orchestration de ces ressources flexibles, les micro-réseaux peuvent être envisagés comme une réponse aux besoins changeants du système énergétique. Ils facilitent la gestion décentralisée des énergies intermittentes en rapprochant le producteur d'énergie du consommateur.
La définition d'un micro-réseau et son intérêt dans ce contexte seront présentés à la section 1.1. Le fonctionnement d'un micro-réseau et les méthodologies développées pour son contrôle sont décrits dans la section 1.2. Enfin, des méthodologies de dimensionnement des micro-réseaux sont mises en lumière à la section 1.3.

1.1 Définition et intérêts des micro-réseaux électriques

1.1.1 Besoin émergeant de micro-réseaux électriques

Un micro-réseau électrique est un réseau local regroupant la production et la demande d'électricité (Schwaegerl et al., 2013). En générant de l'électricité près de la consommation, les pertes de charges dues au transport (résistance des câbles) sont évitées. Cela a aussi pour effet de réduire la dépendance de la demande aux infrastructures centralisées. Ainsi, la génération d'électricité locale peut compléter ou remplacer ces structures, ce qui améliore la résilience du système dans son ensemble. Les unités de productions présentes dans le micro-réseau produisent de l'électricité directement utilisable dans le réseau local. Prenant la forme de panneaux photovoltaïques (PV), micro-turbines, piles à combustibles (PAC), générateurs diesels ou à gaz et éoliennes, ces unités produisent de l'électricité à faible puissance comparées aux générateurs centralisés. Elles sont appelées micro-générations.
L'intégration progressive de ces ressources énergétiques distribuées exerce une influence sur les réseaux de distribution électrique (Lopes et al., 2019). Les micro-générations rendent la demande aux points de consommation variable selon l'énergie produite du point de vue du gestionnaire du réseau de distribution (GRD). Les points de couplage comportant des micro-générations sont à considérer dans la distribution comme une demande flexible ou une production d'électricité, ce qui introduit des défis de gestion. Le flux traditionnel de l'électricité depuis la génération jusqu'à la consommation passant par le transport et la distribution est alors à revoir (Mourshed et al., 2015). Pour cela, le paradigme de *Smart Grid* est introduit. La gestion du réseau de distribution devient active, disposant de points d'injection en plus des points de soutirage traditionnels.
Un management à tous les niveaux (production, transport, distribution, consommation) doit s'organiser autour de la micro-génération et des demandes flexibles. Les micro-réseaux sont des systèmes particulièrement adaptés en raison de leur capacité à réagir aux fluctuations et incertitudes du réseau central auquel ils sont connectés, telles que des déséquilibres de fréquences et des risques de pannes généralisées. Les points de demande de consommation électrique ne doivent plus être vus comme des demandes passives, mais sont actifs et sensibles à la politique de distribution de l'électricité.

1.1.2 Modes d'opération

Une caractéristique notable du micro-réseau est qu'il est capable d'opérer en mode connecté comme îloté. S'il est connecté, il échange avec le réseau central de distribution ; s'il est îloté, il est isolé de tout autre réseau électrique. Un micro-réseau comprend un système de gestion de l'énergie ou *Energy Management System* (EMS) qui communique avec le réseau de distribution lorsque c'est possible. L'EMS orchestre les décisions de contrôle des unités dans le micro-réseau en optimisant leur fonctionnement selon des conditions changeantes. Les micro-réseaux peuvent être urbains (quartier, immeuble), ruraux (maisons), privés (résidences, espace commercial, industrie), publics (hôpital, campus, collectivités) ou totalement isolés du réseau de distribution (camp militaire, zone non-desservie). L'objectif du déploiement d'un micro-réseau prend différentes natures selon le contexte. S'il est totalement isolé, il permet d'électrifier une zone tout en réduisant les coûts de carburant, généralement liés à l'utilisation d'un générateur contrôlable, avec un contrôle intelligent. Si connecté au réseau principal, le micro-réseau favorise une intégration plus efficace des sources d'énergie intermittentes, optimise la gestion des pannes, réduit les coûts énergétiques et accroît l'efficacité énergétique des infrastructures concernées.
Le micro-réseau électrique constitue un réseau de basse tension fonctionnant le plus souvent

en courant alternatif (AC), afin d'assurer une compatibilité avec les exigences des charges locales et le réseau de distribution. Cette configuration permet une gestion adaptative des unités en réponse à divers aléas, tels que la production intermittente, l'état du réseau de distribution et les variations de la demande (A. Hirsch et al., 2018). En revanche, elle requiert de nombreux convertisseurs électroniques de puissance. De plus, lors du passage du mode îloté au mode connecté, appelé resynchronisation, l'alignement de la tension, de la fréquence et de la phase au Point de Couplage Commun (PCC) est un processus technique et complexe (Kandari et al., 2022). Il existe des micro-réseaux fonctionnant en courant continu (DC). Ils sont plus efficaces énergétiquement, car le besoin de convertisseurs électroniques est moindre (ils sont compatibles avec de nombreuses sources de production électrique). Les micro-réseaux DC sont moins courants, principalement parce que les infrastructures historiques ont été construites en utilisant du courant alternatif. L'AC a été favorisé pour le réseau de distribution, car il minimise les pertes sur de grandes distances. Toutefois, avec l'augmentation de l'utilisation des panneaux PV, des batteries et des bornes de recharge pour véhicules électriques (EV), qui sont plus aisément intégrés dans les micro-réseaux DC, ces derniers deviennent une option de plus en plus pertinente dans les nouvelles structures. Les applications spéciales de micro-réseau comme celles pour les navires, les véhicules électriques et les systèmes de communication fonctionnent souvent en DC (Justo et al., 2013).

1.1.3 Gestion et stockage de l'énergie

En intégrant les énergies renouvelables comme unique système de production électrique, les micro-réseaux s'exposent à des fluctuations importantes de génération, ce qui déséquilibre l'offre et la demande locale. Le stockage électrique est une solution efficace qui permet de stocker l'électricité excédentaire produite en période de forte production pour la restituer en période de faible production ou de forte demande (Rohit et al., 2017). Ainsi, le micro-réseau est plus stable et l'intégration des énergies renouvelables est plus efficace. Le stockage électrique réduit aussi la dépendance du micro-réseau aux générateurs contrôlables le cas échéant. Les technologies les plus utilisées dans les micro-réseaux sont les batteries (Lithium-ion, flux redox, nickel, plomb-acide), le stockage chimique (hydrogène), le stockage électrochimique (supercondensateurs) et le stockage mécanique (volant d'inertie) (Kandari et al., 2022).
La combinaison de plusieurs technologies est courante pour assumer différents rôles (Amrouche et al., 2015). Dans un micro-réseau avec production PV, il est courant que l'hydrogène assume le rôle de stockage énergétique à long terme (en stockant le surplus énergétique de l'été pour l'hiver) et soit combiné à une batterie qui stocke la production de jour pour la nuit.

Catégorie	**Type**	**Densité Énergétique**	**Densité de Puissance**	**Rendement**	**Durée de vie**	**Coût**
Batterie	Lithium-ion	Élevée	Moyenne	Élevé	Moyenne	Élevé
	Flux Redox	Faible	Moyenne	Moyen	Élevée	Moyen
	Nickel	Moyenne	Moyenne	Moyen	Moyenne	Moyen
	Plomb-acide	Faible	Moyenne	Moyen	Faible	Faible
Stockage Chimique	Hydrogène	Élevée	Moyenne	Faible	Moyenne	Élevé
Électrochimique	Super-condensateurs	Faible	Élevée	Élevé	Élevée	Élevé
Mécanique	Volant d'inertie	Faible	Élevée	Élevé	Moyenne	Élevé

Table 1.1 – Comparaison des caractéristiques des technologies de stockage d'énergie selon les valeurs relevées par diverses sources.

Des caractéristiques spécifiques à chaque technologie guident le choix des systèmes de stockage d'un micro-réseau. La table 1.1 présente une comparaison de ces caractéristiques entre les technologies de stockage citées. Les intervalles précis de valeurs pour les différentes caractéristiques ne sont pas exposés, car les valeurs diffèrent d'une source à l'autre. La densité énergétique, exprimée en Wh/kg ou Wh/L, définit la quantité maximale d'énergie qu'une technologie de stockage peut contenir par unité de masse ou de volume. En contraste, la densité de puissance, mesurée en W/kg ou W/L, caractérise la vitesse avec laquelle cette technologie peut libérer ou absorber de l'énergie par unité de masse et de volume. Les valeurs de densité énergétique et de puissance peuvent varier considérablement pour un même moyen de stockage, selon qu'elles sont exprimées par unité de masse ou de volume. Les technologies de batterie à flux redox ont une densité énergétique faible, la batterie lithium-ion a une densité de puissance forte par rapport aux autres batteries (Kebede et al., 2022, Espinar et al., 2011 et Townsend et al., 2022). La densité énergétique des batteries et du stockage hydrogène permet d'intégrer les énergies renouvelables au réseau, en lissant les pics de production et de demande. Les supercondensateurs et le volant d'inertie ont une densité de puissance élevée, ils peuvent absorber et restituer la puissance électrique facilement. Leur intérêt est donc de régler des perturbations pour des durées très courtes sur le réseau (Espinar et al., 2011). Peu flexibles à cause de leur faible densité énergétique, leur emploi seul comme système de stockage à moyen et long terme n'est pas idéal.
Le rendement des systèmes de stockage, exprimé en pourcentage, est le taux de conservation de l'énergie sous forme d'électricité après le processus de stockage et de restitution. Dans le contexte d'un micro-réseau cherchant à intégrer les énergies renouvelables, un rendement élevé permet d'optimiser l'utilisation des ressources énergétiques fournies par les énergies renouvelables, minimisant ainsi les pertes et contribuant à une gestion énergétique plus efficace. Enfin, le coût peut contraindre le choix des technologies impliquées dans un micro-réseau. Une batterie plomb-acide peut être privilégiée à une batterie lithium-ion pour son prix et sa durée de vie, malgré une densité énergétique, une densité de puissance et un rendement plus faibles (Kebede et al., 2022 et Townsend et al., 2022). La durée de vie de l'unité doit être prise en considération parallèlement à son coût, car un remplacement sera nécessaire lorsque l'usure atteindra un certain seuil. D'autres caractéristiques sont prises en compte selon l'usage. Par exemple, le taux d'autodécharge n'est pas à négliger sur un système de stockage long terme. La précision avec laquelle est évalué l'état de charge d'une unité et la fréquence avec laquelle il est mesurable pendant l'opération sont importantes pour une régulation plus fine de l'énergie. Les densités énergétiques et de puissance de la batterie lithium-ion la rendent intéressante pour les usages mobiles. En revanche, sa puissance dépend de sa capacité maximale qui décroît à mesure de son vieillissement. Il est estimé qu'une telle batterie n'est plus utilisable pour un usage mobile lorsque sa capacité est dégradée à plus de 20% (Hesse et al., 2017). L'usure de la batterie lithium-ion étant fonction de son utilisation, une surveillance régulière et fréquente de divers indicateurs techniques est essentielle. Cela permet un contrôle précis de l'état de la batterie, contribuant ainsi à minimiser les coûts de remplacement en prévenant une dégradation accélérée. Les indicateurs pertinents sont collectés par le système de gestion de la batterie, connu sous l'acronyme BMS (pour *Battery Management System*), qui fonctionne en liaison avec l'EMS. Ce système supervise et contrôle la batterie à court terme. La Figure 1.1 illustre les différentes typologies de micro-réseaux présentées dans cette section.

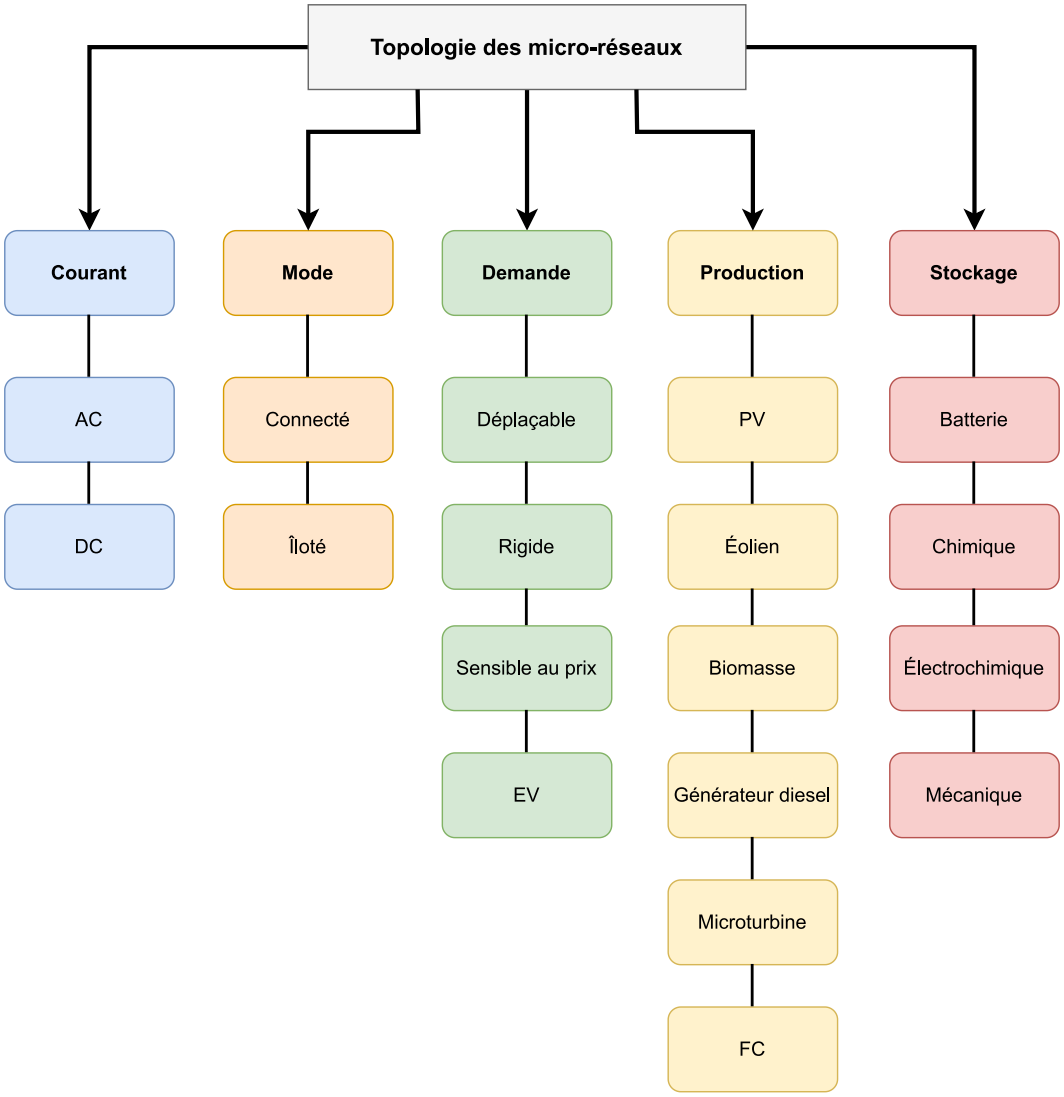

Figure 1.1 – Diagramme résumant les classifications des typologies de micro-réseaux présentées.

1.2 Contrôle des micro-réseaux

1.2.1 Catégories de contrôle

Le contrôle des micro-réseaux peut être catégorisé en trois niveaux selon la résolution temporelle des décisions de contrôle, ou l'horizon temporel considéré lors de la prise de décision (De Brabandere et al., 2007, Shahgholian, 2021, Salehi et al., 2022) :

— Niveau primaire : Régulation réactive de la puissance entre différents onduleurs pour le contrôle de la tension et de la fréquence à une échelle de temps inférieure à la seconde.

— Niveau secondaire : Contrôle en régime permanent pour corriger les écarts de tension et de fréquence en programmant l'énergie dans la production et le stockage.

— Niveau tertiaire : Planification à long terme pour un fonctionnement optimal du micro-réseau en termes de coûts.

Les trois niveaux de contrôle dans un micro-réseau ont des fonctions distinctes et sont interconnectés (F. Gao et al., 2019). Le niveau tertiaire et le niveau secondaire sont gérés par l'EMS, qui prend des décisions basées sur la connaissance des états des différentes unités du micro-réseau. Ces niveaux se concentrent sur l'optimisation à long terme et la coordination des ressources, respectivement. Le niveau de contrôle primaire, en revanche, est propre à chaque unité et fonctionne sans communication avec les autres, en se concentrant sur la régulation instantanée et l'équilibre local. C'est un mode de contrôle décentralisé. Le niveau de contrôle secondaire peut être centralisé ou décentralisé (Bharath et al., 2019, Ouramdane et al., 2021). Le contrôle est considéré comme centralisé lorsque l'EMS détermine conjointement les opérations de toutes les unités, à partir d'informations transmises par celles-ci et sur des données exogènes telles que l'ensoleillement et le prix de l'électricité. Bien que la solution avancée par l'EMS puisse satisfaire l'ensemble des contraintes (qui lui sont accessibles), la collecte de données augmente les risques d'erreur, d'autant plus que de multiples variables sont intégrées dans les équations. Le contrôle est décentralisé s'il inclut une solution de contrôle calculée localement depuis chaque unité. Les unités reçoivent des informations via leurs capteurs et des consignes de l'EMS. Les décisions de contrôle sont prises selon les contraintes de chaque unité. L'EMS tranche sur les décisions finales retenues en considérant les contraintes globales du micro-réseau. Cela requiert la présence d'outils rapides de communication et de calcul.

Cette structure hiérarchique avec trois niveaux de contrôle permet une gestion flexible et efficace du micro-réseau. Bien que les trois niveaux de contrôle soient largement étudiés et présentent des défis et opportunités variés (Vandoorn et al., 2013, F. Gao et al., 2019, Rodriguez-Martinez et al., 2023), l'effort visant à intégrer les énergies renouvelables afin de rendre le système plus résilient, économiquement viable et respectueux de l'environnement, est intrinsèquement lié à la planification à long terme. Cela se situe principalement au niveau de contrôle tertiaire. Par conséquent, les travaux menés et présentés se focaliseront sur ce niveau spécifique. Il est défini ici comme la prise de décision sur les flux d'énergie dans un micro-réseau en fonctionnement stationnaire (Rios et al., 2022), qu'il soit complètement isolé ou connecté au réseau central de distribution, avec une résolution temporelle minimale de 15 minutes.

Le contrôle tertiaire ou de haut-niveau des micro-réseaux peut être classé en trois catégories (Battula et al., 2021, Sarwar et al., 2022, Allwyn et al., 2023) :

— Planification des systèmes de stockage et engagement des unités
— Gestion de la demande
— Partage d'énergie entre plusieurs micro-réseaux

Il y a une dépendance entre le type de contrôle et la configuration du micro-réseau, caractérisée par plusieurs facteurs. Parmi eux, la faisabilité de chaque type de contrôle est influencée par des éléments tels que la flexibilité de la consommation, la présence de systèmes de stockage ou d'unités contrôlables de production électrique, et la possibilité de mettre en liaison plusieurs micro-réseaux. Ces facteurs déterminent les modalités de contrôle dans le cadre de l'exploitation d'un micro-réseau. La résolution de certains problèmes demande une combinaison de plusieurs catégories de contrôle.

La planification des systèmes de stockage et l'engagement des unités consistent à contrôler les systèmes de stockage et de production d'électricité (K. Gao et al., 2021, Battula et al., 2021). Le stockage absorbe les déséquilibres entre demande et production électrique en se chargeant et en se déchargeant. Ainsi, l'électricité produite en excès par des sources

de production renouvelables à certaines heures peut satisfaire des demandes à des heures ou la production est moins abondante. La gestion d'un générateur contrôlable augmente l'autonomie du micro-réseau en répondant aux demandes électriques avec une ressource supplémentaire. L'intérêt est plus marqué lorsque les ressources se font rares (en hiver pour un micro-réseau à production PV par exemple). Cependant, les unités de production d'énergie contrôlables sont souvent des unités qui émettent des gaz polluants. Leur utilisation est liée à un coût environnemental et à un coût d'approvisionnement en combustible. La consommation du micro-réseau peut être déplacée ou effacée. Lors du fonctionnement du micro-réseau en mode îloté, la capacité à subvenir de manière autonome aux demandes énergétiques devient primordiale. Dans ce cas, le pilotage efficace des systèmes de stockage et des unités de production contrôlables est essentiel.

La gestion de la demande permet de gérer les déséquilibres du réseau en pilotant directement la demande énergétique (Kanakadhurga et al., 2022, Dahiru et al., 2023). La diminution de la demande peut stabiliser le réseau, mais dégrade la satisfaction des consommateurs. Un compromis entre la satisfaction des utilisateurs et la stabilité du réseau est recherché. Le délestage, le déplacement de charge et l'effacement énergétique sont des techniques de gestion orientées sur la demande. Le chauffage et la purification de l'air sont des consommations flexibles, dont l'alimentation peut être programmée. Les horaires d'utilisation des lave-linges et des lave-vaisselles sont déplaçables : ils peuvent être décalés pendant les heures creuses. Certains contrôleurs utilisent un prix virtuel de l'électricité au sein du micro-réseau afin d'influencer les comportements des unités de production et des consommateurs. Il dépend généralement de la différence entre la génération et la demande d'énergie ainsi que du prix d'achat de l'électricité extérieure au réseau (dans le cas où le micro-réseau n'est pas îloté) qui varie avec le temps.

Le partage d'énergie entre plusieurs micro-réseaux améliore l'autonomie de plusieurs micro-réseaux proches géographiquement (Rashidi et al., 2021). Puisque plusieurs micro-réseaux peuvent être connectés les uns aux autres, un micro-réseau, avec des problèmes de stabilité énergétique à un instant, peut être approvisionné par d'autres. Les micro-réseaux qui surproduisent peuvent également échanger l'énergie excédentaire à un micro-réseau proche. Des politiques de partage efficaces sont établies, notamment grâce au développement d'algorithmes qui optimisent un objectif commun préalablement convenu. Pour cela, les évolutions du prix de l'électricité, de la production, de la consommation et de la capacité de stockage de chacun doivent être anticipées. Un système de tarification est souvent utilisé pour donner la priorité à la transmission d'énergie entre les micro-réseaux plutôt qu'à l'échange avec un réseau extérieur. Le prix de l'électricité fixé entre les micro-réseaux, qu'il soit virtuel ou non, doit être plus intéressant que l'offre d'un réseau externe.

1.2.2 Méthodologies de contrôle

Le contrôle du micro-réseau peut se faire en minimisant une fonction coût associée à un objectif (Ahmad Khan et al., 2016). Cet objectif peut être environnemental (émissions), économique (coûts de carburant, de maintenance, d'achat d'énergie), ou divers (mécontentement des utilisateurs, pannes, stockage). La stratégie de contrôle peut concilier des objectifs multiples. S'ils sont contradictoires, un compromis doit être trouvé.
Des algorithmes sont déployés dans l'EMS pour orienter le processus décisionnel. Dans la littérature, la fréquence décisionnelle de l'EMS sur le contrôle de l'énergie au niveau tertiaire varie généralement de toutes les heures à toutes les 24 heures (Ahmad et al., 2023). Les entrées de l'EMS incluent des informations détaillées sur le système à chaque instant, notamment les données captées par le BMS et les autres unités. Ces informations sont essentielles pour la prise de décision au sein de l'EMS. Dans les situations où la gestion du micro-réseau est effectuée en simulation plutôt qu'avec un système réel, une modélisation dynamique des différentes unités est nécessaire afin de reproduire fidèlement les conditions opérationnelles.

Des contraintes propres au système et à ses unités sont à considérer dans la modélisation dynamique du micro-réseau. Dans la littérature, les micro-réseaux sont contrôlés de différentes manières.

Les algorithmes basés sur des règles sont des algorithmes déterministes et ont l'avantage d'être simples à implémenter. Ces algorithmes opèrent en suivant des règles logiques pré-établies, où des conditions spécifiques sous la forme de « si [...], alors [...] » régissent les décisions. Bien que pertinents pour des scénarii où la clarté et la rapidité d'exécution sont prioritaires, ils peuvent manquer de flexibilité et ne s'appuient que sur des données mesurées directement (Kanwar et al., 2015).

L'optimisation hors ligne consiste à planifier les décisions à un horizon temporel fixé selon la modélisation physique du système et des prévisions des données aléatoires. Un algorithme d'optimisation est utilisé pour résoudre un problème de planification qui consiste à minimiser un coût sous contrainte. Ces algorithmes sont déterministes et ne mènent pas à une prise de décision optimale en temps réel mais permettent d'élaborer une stratégie décisionnelle prévisionnelle. Toute incorporation de données actualisées impliquerait un nouveau calcul dans son intégralité. Les résultats sont grandement affectés par la qualité des prédictions (Battula et al., 2021). Il est parfois bénéfique de combiner des modèles de prédiction à la modélisation analytique du système pour déterminer la valeur de certaines variables (Gamarra et al., 2015). Pour illustrer cette démarche, Mohamed et al., 2013 a utilisé une modélisation thermique d'une maison basée sur un circuit RC et l'a combinée à des modèles de données pour estimer les températures extérieures. Cette approche intégrée, en tenant compte du gain thermique, a permis de déduire les températures intérieures. Ces méthodologies mixtes sont nommées *boîtes grises*, car elles contiennent les caractéristiques des *boîtes noires* (modèles de données dont les mécanismes internes ne sont pas explicitement définis) et des *boîtes blanches* (modèles analytiques où les équations sont posées).

La Commande prédictive ou *Model Predictive Control* (MPC) (García et al., 1989) s'appuie aussi sur les caractéristiques physiques du modèle du micro-réseau pour minimiser un objectif selon des contraintes. La différence avec l'optimisation hors ligne est que cette méthodologie intègre les valeurs actualisées des variables à mesure du temps. Contrairement à l'optimisation hors ligne, le MPC incorpore les mises à jour en temps réel des variables. Cette approche utilise des algorithmes d'optimisation à plusieurs reprises en adoptant une fenêtre temporelle glissante (Kamal et al., 2022, Erazo-Caicedo et al., 2022). Cela permet de recalculer la planification en fonction des nouvelles conditions, adaptant ainsi la solution aux évolutions et écarts des prédictions initiales.

Il existe une grande variété d'algorithmes d'optimisation pouvant être utilisés pour le contrôle de micro-réseaux en optimisation hors ligne ou en MPC. Il n'est pas intéressant de les lister ici de manière exhaustive, mais leurs spécificités permettent de les catégoriser et de les choisir selon le type de problème à résoudre.
Certains algorithmes sont conçus pour optimiser des variables continues et d'autres des variables discrètes (ou combinatoires). Un problème d'optimisation intégrant des variables continues et discrètes est dit à variables mixtes. Si l'espace de valeur des variables est un ensemble dénombrable, le problème est combinatoire.
Un algorithme d'optimisation peut être déterministe ou stochastique. Alors que ces termes peuvent évoquer la certitude sur les données en jeu, comme évoqué précédemment pour distinguer l'optimisation hors ligne du MPC, ils peuvent également qualifier la méthodologie employée pour la résolution. Une méthode déterministe poursuit un cheminement précis pour trouver la solution au problème d'optimisation, tandis qu'une méthode stochastique incorpore des éléments aléatoires pour explorer l'espace des solutions. L'optimisation déterministe est adaptée à la résolution de problèmes représentables par un modèle dans leur totalité et permet de calculer un résultat optimal. En revanche, ces méthodologies ne sont pas adaptées aux problèmes partiellement modélisés. L'optimisation stochastique ne permet pas de parvenir à un résultat optimal avec certitude, mais elle est adaptée aux variables aléatoires et à la

résolution de problèmes non-convexes.
Parmi les algorithmes stochastiques, les algorithmes méta-heuristiques sont des algorithmes génériques, développés pour être adaptés à la résolution de nombreux problèmes complexes (Collet et al., 2007). Ils sont construits pour explorer l'espace des solutions de manière efficace tout en incorporant des processus aléatoires. Les algorithmes méta-heuristiques sont couramment inspirés de phénomènes naturels comme l'évolution (Katoch et al., 2021), le comportement en essaim (Kennedy et al., 1995) ou en meute (Mirjalili et al., 2014).

Enfin, certaines approches de contrôle de micro-réseau sont basées sur des agents. Chaque entité de contrôle prend des décisions de manière séquentielle et indépendante. Le contrôle basé sur des agents est adapté pour des décisions prises en temps réel. Si le contrôle est centralisé, un agent est utilisé pour prendre les décisions séquentiellement. La programmation dynamique (Bellman, 1958) et l'apprentissage par renforcement (Sutton et al., 1995) sont des catégories d'algorithmes basées sur un agent. Si le contrôle est décentralisé, on parle de systèmes multi-agents (Altin et al., 2023). Dans ce cas, les agents prennent des décisions distinctement et peuvent avoir des objectifs différents. Ces systèmes multi-agents permettent à chaque entité décisionnelle d'être indépendante et autonome. La théorie des jeux peut être utilisée pour coordonner la prise de décision des différents agents de manière à atteindre un équilibre à chaque instant (J.-W. Lee et al., 2021).

1.3 Méthodes de dimensionnement

Le dimensionnement d'un micro-réseau doit être optimisé pour que l'achat des équipements puisse être amorti tout en conservant l'atteinte des objectifs associés. La localisation du micro-réseau est la première décision de dimensionnement lorsqu'elle est réalisable. L'accès aux énergies renouvelables dépend fortement de la localisation du micro-réseau puisqu'une zone ensoleillée ou bénéficiant de vents réguliers permet de générer plus d'électricité à partir des ressources naturelles. Les équipements sont dimensionnés selon le productible renouvelable déterminé en fonction de la localisation et des habitudes de consommation des utilisateurs. Le stockage et les unités de production sont choisis pour que la demande soit satisfaite à chaque instant, en considérant les pics de consommation. Le choix des valeurs de dimensionnement dépend généralement d'autres objectifs à atteindre et se déterminent selon différents indicateurs présentés à la sous-section 1.3.1. Les différentes méthodologies utilisées pour le dimensionnement des micro-réseaux sont présentées à la sous-section 1.3.2.

1.3.1 Objectifs et indicateurs

Le coût d'investissement lié à la conception du micro-réseau peut être important, mais il est nécessaire de prendre en compte des coûts liés à sa phase d'opération pour atteindre les objectifs associés. Ces objectifs sont de différentes natures (Upadhyay et al., 2014) : économiques, techniques et liés à l'utilisation d'énergie renouvelable. La réalisation ou non d'objectifs est vérifiée grâce à des indicateurs.

Indicateurs économiques

Parmi les objectifs économiques, l'indicateur le plus populaire est le coût nivelé de l'énergie, ou *Levelized cost of energy* (LCE), exprimé en €/kWh. Son expression est donnée par l'équation 1.1.

$$\mathrm{LCE} = \frac{\mathrm{CAPEX} + \mathrm{OPEX}}{\sum_{\mathrm{t}}^{\mathrm{T}} \mathrm{D(t)}} \tag{1.1}$$

Avec CAPEX les dépenses de capital, OPEX les dépenses d'exploitation, D(t) la demande à l'instant t et T le temps maximum de l'intervalle sur lequel l'indicateur est calculé. En

général, le LCE est comparé au prix de l'électricité du réseau central auquel le micro-réseau est connecté (François-Lavet, 2017), ou à une valeur cible à ne pas dépasser si le micro-réseau est isolé. Il est calculable sur des échelles de temps variées allant de la durée de vie du micro-réseau à la journée. La valeur net actualisée (VNA) est un autre indicateur économique (Chebabhi et al., 2023) calculé par la soustraction des coûts (comprenant l'investissement) aux flux de trésorerie actualisés du micro-réseau. Si elle est positive, alors le micro-réseau est rentable. Le taux de rendement interne (TRI) est un indicateur lié à la VNA puisqu'il s'agit du facteur d'actualisation nécessaire pour que la VNA soit nulle. Plus cette valeur est grande et plus le projet est rentable.

Indicateurs liés à l'utilisation d'énergie renouvelable

Les objectifs liés à l'utilisation d'énergie renouvelable prennent différentes formes selon la nature du micro-réseau. Les générateurs contrôlables sont généralement les unités émettant le plus de gaz à effet de serre parmi les unités de productions électriques décentralisées. Lorsqu'il y en a, leur utilisation est souvent associée à un coût d'émission. Le mix énergétique du réseau central de distribution est pris en compte dans le calcul des émissions d'un micro-réseau électrique opérant en mode connecté. Une fonction liée aux émissions est donc calculable (Upadhyay et al., 2014). L'intégration des énergies renouvelables est un objectif considérable dans la conception d'un micro-réseau électrique. La quantité d'énergie renouvelable produite en excès (c'est-à-dire ne pouvant être ni stockée ni utilisée pour les demandes locales) est à minimiser. Le ratio d'excès d'énergie (REE) (An et al., 2015) permet de comparer cette quantité sur différents horizons temporels. Son expression est donnée par l'équation 1.2.

$$\mathrm{REE} = \frac{\mathrm{E_{PE}}}{\mathrm{E_{PL}} + \mathrm{E_{SR}}} \tag{1.2}$$

Avec $\mathrm{E_{PE}}$, $\mathrm{E_{PL}}$ et $\mathrm{E_{SR}}$ respectivement l'énergie produite en excès, l'énergie produite localement et l'énergie soutirée au réseau central de distribution. La fraction d'énergie renouvelable (FER) est un rapport utilisé pour quantifier la part d'énergie d'origine renouvelable utilisée pour satisfaire la demande locale. L'équation 1.3 présente son expression.

$$\mathrm{FER} = \frac{\mathrm{E_{PR}}}{\mathrm{E_{PL}} + \mathrm{E_{SR}}} \tag{1.3}$$

Avec $\mathrm{E_{PR}}$ l'énergie produite d'origine renouvelable. Lors du dimensionnement, la minimisation du REE et la maximisation du FER sont recherchées. D'autres facteurs environnementaux comme l'analyse de cycle de vie peuvent être optimisés lors du dimensionnement (Mori et al., 2021).

Indicateurs techniques

Les objectifs techniques sont liés à la satisfaction des utilisateurs. La satisfaction de la demande des utilisateurs est un objectif primordial pour un micro-réseau, particulièrement lorsqu'il est isolé. Le facteur de perte équivalent ou *Equivalent loss factor* (ELF) est le ratio entre la demande non-approvisionnée et la demande totale (Ardakani et al., 2010). Cette relation est présentée dans l'équation 1.4.

$$\mathrm{ELF} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=0}^{\mathrm{T}} \frac{\mathrm{D_L(t)}}{\mathrm{D(t)}} \tag{1.4}$$

Avec $\mathrm{D_L(t)}$ la demande non-approvisionnée à l'instant t. Ce rapport est utile pour déterminer l'autonomie d'un micro-réseau isolé. Il s'applique à un micro-réseau connecté en ajoutant

l'énergie soutirée au réseau central au numérateur et les charges des unités de stockage au dénominateur. Du point de vue de l'utilisateur, il est nécessaire d'évaluer le nombre de coupures plutôt que leur amplitude. La probabilité de perte de charge ou *loss of power supply probability* (LPSP), dont l'expression est donnée par l'équation 1.5, permet de déterminer une probabilité de panne (H. Yang et al., 2008).

$$\mathrm{LPSP} = \sum_{t=0}^{T} \frac{t_c(t)}{T} \tag{1.5}$$

Avec $t_c(t)$ le temps de coupure sur le pas de temps t. L'acceptabilité sociale du projet est à considérer comme critère d'évaluation lorsque le projet peut susciter de la résistance. L'impact visuel, acoustique et les effets sur la biodiversité sont à inclure dans la conception (Stigka et al., 2014).

1.3.2 Méthodologies de dimensionnement

Les objectifs de dimensionnement d'un micro-réseau dépendent de la nature de son fonctionnement, donc des contraintes et de la modélisation dynamique des composants. Une simulation incluant la phase opératoire est nécessaire puisque des données d'opérations réelles ne peuvent exister avant la conception du micro-réseau.
Les méthodes de dimensionnement dépendent des outils et des données à disposition. Si les équations mathématiques qui régissent le système à dimensionner sont disponibles, alors un modèle analytique peut être employé (Upadhyay et al., 2014). Ces modèles permettent de décrire les contraintes du micro-réseau. Des algorithmes d'optimisation analytiques (Atia et al., 2016) ou méta-heuristiques (Abdel-hamed et al., 2019) permettent alors de trouver un dimensionnement optimal selon les données et la typologie du problème (Ren et al., 2023). Des outils de simulation et des logiciels spécialisés (HOMER, ODYSSEY, Simulink) permettent de modéliser et simuler les contraintes et les comportements dynamiques pré-établis lors du dimensionnement.
Lorsque les données ne sont pas accessibles, des scénarii hypothétiques sont simulés. Afin que le dimensionnement soit adapté à toute situation et pas uniquement aux scénarii typiques et attendus, des conditions extrêmes peuvent être simulées. Le dimensionnement optimal pour de telles conditions est appelé dimensionnement robuste (Pecenak et al., 2020, Gazijahani et al., 2018).
Lorsqu'un jeu de données est exploitable, le recours à l'optimisation déterministe comme décrite dans la sous-section 1.2.2 est possible (Roy et al., 2021). Si de nombreuses données sont disponibles, des modèles prédictifs permettent de les estimer ou de les reproduire avec une certaine marge d'erreur. L'adaptabilité de la méthodologie s'accroît si le modèle peut être appliqué à d'autres données en conservant une erreur de prediction limitée (B. Li et al., 2017). De cette manière, le dimensionnement s'établit avec des données différentes des données d'entraînement du modèle prédictif et incorpore les éventuelles erreurs de prédictions observables en conditions réelles.

1.4 Conclusion

Le besoin de décentraliser la production des énergies renouvelables pour faciliter leur intégration tout en maintenant une stabilité acceptable des réseaux de distribution électriques a contribué au développement des micro-réseaux électriques. Ces réseaux regroupent localement la production, la demande et souvent le stockage de l'électricité, et augmentent l'autonomie des consommateurs vis-à-vis du réseau central de distribution. Ils peuvent opérer en mode connecté ou îloté.

À cause des fluctuations dans la production d'énergie d'origine renouvelable, leur autonomie repose sur la qualité du pilotage effectué par un EMS. Le contrôle des unités s'effectue à différents horizons temporels, entre des temps inférieurs à la seconde pour la sécurité des installations et jusqu'à plusieurs mois pour la planification à long terme des flux énergétiques. Le contrôle peut être centralisé ou décentralisé et s'applique à tout type d'unité : production, stockage et demande électrique. La politique d'échange d'électricité entre des micro-réseaux est contrôlable dans le cas où ils sont proches. Le contrôle d'un micro-réseau s'effectue avec un algorithme basé sur des règles, de l'optimisation hors ligne, de la commande prédictive ou avec une approche basée sur un agent. L'apprentissage par renforcement est une méthodologie stochastique avec un agent dont la stratégie est basée sur des données et une simulation de l'environnement. Cela rend la politique de contrôle particulièrement adaptable aux micro-réseaux à production d'électricité renouvelable avec une demande incertaine. Cette méthodologie sera présentée au chapitre 2 et son utilisation pour le contrôle des micro-réseaux au chapitre 3.
Le dimensionnement d'un micro-réseau est essentiel pour garantir sa viabilité et vérifier que ses objectifs sont atteints. Les méthodologies de dimensionnement dépendent des outils de simulation et des données à disposition.

2

L'apprentissage par renforcement

L'apprentissage par renforcement ou *Reinforcement Learning* (RL) est une forme d'apprentissage automatique, il n'est ni supervisé ni non-supervisé. C'est l'interaction entre un agent (apprenant) et son environnement qui produit les données d'apprentissage. Cette classe d'algorithme est particulièrement adaptée à la résolution de problèmes de Markov avec prises de décision séquentielles. Dans ce chapitre, les processus de décision de Markov et la programmation dynamique seront abordés dans la Section 2.1, avant d'introduire les méthodes usuelles d'apprentissage par renforcement. La Section 2.2 présentera les grandes classes d'algorithmes d'apprentissage par renforcement profond. Puis la Section 2.3 détaillera le fonctionnement de certains algorithmes hors ligne d'apprentissage par renforcement.

2.1 Des processus de décision de Markov à l'apprentissage par renforcement

2.1.1 Processus de décision de Markov

Un processus de décision de Markov ou *Markov Decision Process* (MDP) est la représentation mathématique de prises de décision dont les effets sont en partie déterministes et en partie aléatoires. La prise de décision dans un MDP est séquentielle selon le temps, les décisions sont prises les unes après les autres. Le but des MDP est de déterminer quelle suite de prises de décision a la meilleure espérance conformément à un objectif. Différentes entités définissent un MDP. L'**agent** est le preneur de décision. Il envoie des signaux d'action A_t à son **environnement** à chaque pas de temps. L'environnement est caractérisé à chaque instant par des états S_t, dont chacun est observable ou non par l'agent.

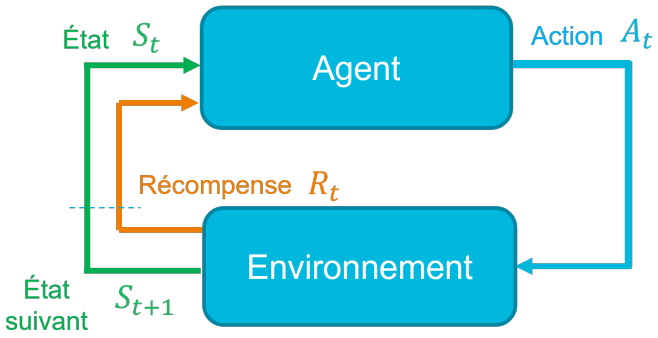

Figure 2.1 – Interaction entre agent et environnement dans un processus de décision de Markov

Dans de nombreuses représentations de problème, l'environnement inclut l'agent, et ses caractéristiques sont des états. L'espace d'états de l'environnement et l'espace d'actions de l'agent sont respectivement $\mathcal{S}$ et $\mathcal{A}$. Certains états de l'environnement sont affectés par l'action de l'agent prise à chaque instant. En plus des signaux d'états observables, l'agent reçoit un signal numérique appelé récompense $\mathrm{R_t} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ à chaque instant t. Ces signaux permettent de considérer la pertinence des états de l'environnement par rapport à un objectif. Lorsque l'agent a exécuté une action $\mathrm{A_t}$ depuis un état $\mathrm{S_t}$, l'environnement réagit : Il change d'état et renvoie le nouvel état $\mathrm{S_{t+1}}$ dans lequel il se situe et une récompense $\mathrm{R_t}$, positive, nulle ou négative indique la pertinence immédiate de l'action à l'agent. Le but de l'agent est donc de maximiser les récompenses perçues à long terme.
La Figure 2.1 illustre de manière schématique les échanges de signaux entre agent et environnement dans un MDP. L'action prise dans un état ne suffit pas à déterminer le prochain état de l'environnement, on parle de probabilité de transition $P^a_{ss'} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. La définition de cette probabilité est donnée avec les Équations 2.1 et 2.2.

$$P^a_{ss'} = \mathbb{P}\left(\mathrm{s'} \mid \mathrm{s}, \mathrm{a}\right) \tag{2.1}$$

$$= \mathbb{P}\left(\mathrm{S_{t+1}} = \mathrm{s'} \mid \mathrm{S_t} = \mathrm{s}, \mathrm{A_t} = \mathrm{a}\right) \tag{2.2}$$

$P^a_{ss'}$ est donc la probabilité pour l'environnement d'atteindre l'état s' à l'instant $\mathrm{t} + 1$ en sachant qu'il était dans l'état s et que l'agent a choisi l'action a à l'instant t. Les récompenses étant dépendantes des états de l'environnement, on peut donc les définir comme dépendantes de l'action prise par l'agent et de l'état dans lequel l'environnement était au moment de la prise de décision.

$$r(s, a) = \mathbb{E}\left[\mathrm{R_t} \mid \mathrm{S_t} = \mathrm{s}, \mathrm{A_t} = \mathrm{a}\right] \tag{2.3}$$

La probabilité de choisir chaque action selon l'état de l'environnement est appelée *politique* et notée $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$. Puisqu'il s'agit d'une probabilité, pour n actions possibles depuis l'état s, on a :

$$\sum_{i}^{n} \pi(\mathrm{A_t} = \mathrm{a}_i \mid \mathrm{S_t} = \mathrm{s}) = 1 \tag{2.4}$$

Une politique étant une probabilité de choisir chaque action selon un état, la probabilité de prendre chaque action depuis l'état s est notée $\pi(\mathrm{s})$, la probabilité de choisir une action a depuis l'état s est notée $\pi(\mathrm{a} \mid \mathrm{s})$. La prise de décision dans un MDP est représentée sous la forme d'arbre décisionnel présenté sur la Figure 2.2. Les MDP peuvent être finis ou infinis.

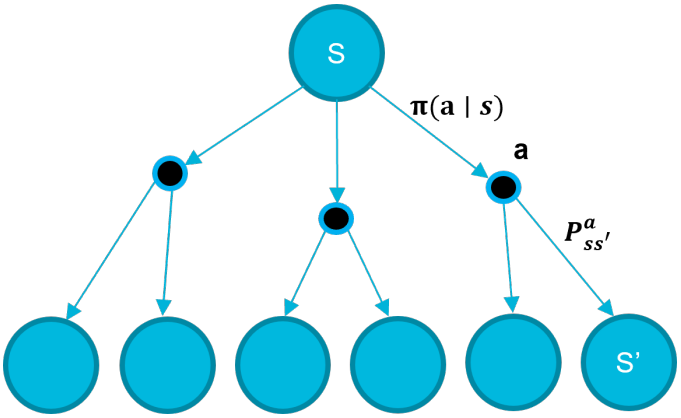

Figure 2.2 – Représentation d'une décision dans le MDP sous forme d'arbre

Si un MDP est fini, alors il existe au moins un état dans lequel l'interaction agent-environnement s'arrête. Cet état est appelé état terminal et l'instant auquel il est atteint est noté $t = T$. Si un MDP est fini, les interactions agent-environnement entre l'état initial et l'état terminal forment un **épisode**. L'enchaînement d'états et d'actions échantillonnés lors d'un épisode forment une **trajectoire** τ de l'état initial à l'état final. La politique utilisée pour générer une trajectoire $\pi(\tau)$ et les récompenses perçues sur la trajectoire s'expriment respectivement selon les Équations 2.5 et 2.6.

$$\pi(\tau) = \prod_{t=0}^{T} \pi(\mathrm{A}_t \mid \mathrm{S}_t) \tag{2.5}$$

$$R_\tau = \sum_{t=0}^{T} R_t \tag{2.6}$$

La politique d'une trajectoire est la probabilité de choisir chaque action état par état. C'est donc le produit des politiques de chaque état. Le but de l'apprentissage par renforcement est de trouver une politique π capable de maximiser les récompenses perçues par l'agent que le MDP soit fini ou non. Dans les sections suivantes on ne s'intéresse qu'aux MDP finis, car les problèmes liés à cette thèse sont résolus avec un horizon de temps fixé. L'objectif J à maximiser est donc une espérance s'écrivant selon les Équations (2.7, 2.8, 2.9).

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \mathrm{R}_t \mid \pi(\mathrm{S}_t, \mathrm{A}_t)\right] \tag{2.7}$$

$$= \sum_{\tau} \mathbb{P}\left(\tau \mid \pi\right) \mathrm{R}_\tau \tag{2.8}$$

$$= \sum_{\tau} \pi(\tau) r(\tau) \tag{2.9}$$

L'Équation 2.7 montre que c'est l'espérance des récompenses obtenues sur un épisode qui doit être maximisée. Or elle dépend de la politique à chaque état observé. L'Équation 2.8 apparait lorsque cette espérance s'exprime comme la somme des produits entre la valeur de la variable aléatoire R_τ et sa probabilité. Enfin, l'utilisation de l'Équation 2.3 permet l'obtention de l'Équation 2.9, car la somme des récompenses espérées sur les trajectoires échantillonnées dépend aussi de la politique π sur une trajectoire.

2.1.2 Principes de l'apprentissage par renforcement

Valeur d'état V

Les principes de l'apprentissage par renforcement (RL) reposent sur les interactions entre agent et environnement. Comme pour les MDP, la fonction à maximiser est l'objectif J. L'agent doit apprendre à choisir les meilleures actions dans chaque état dans le but de maximiser les récompenses perçues tout au long de ces interactions. De la même manière qu'en apprentissage supervisé une fonction coût doit être minimisée au fur et à mesure des interactions. Le terme d'entraînement désigne l'apprentissage de l'agent. L'entraînement doit permettre la convergence vers une séquence d'états pour lesquels les valeurs des récompenses sont les plus grandes possibles. Pour cela, l'agent utilise les récompenses perçues pour calculer des valeurs associées aux états de l'environnement. Ces valeurs permettent de cartographier les bonnes et mauvaises trajectoires pour maximiser les récompenses perçues. Des valeurs élevées correspondent aux états qui ont mené l'agent à percevoir des récompenses élevées. L'agent doit donc prendre des décisions qui maximisent les probabilités d'observer les états avec des valeurs maximales à chaque instant. Les valeurs d'états sont représentées par $\mathrm{V} : \mathcal{S} \to \mathbb{R}$ à chaque pas de temps t. Puisque la valeur dépend de la récompense perçue après décision de l'agent lorsque $\mathrm{S_t}$ est observé, $\mathrm{V}(\mathrm{S}_t)$ dépend du choix d'action de l'agent. $\mathrm{V}(\mathrm{S_t})$ dépend de la politique $\pi(\mathrm{S_t})$ de l'agent depuis l'état $\mathrm{S_t}$. On note alors $\mathrm{V}^{\pi}(\mathrm{s})$ la valeur d'un état s. La valeur d'un état ne peut qu'être estimée empiriquement, car elle dépend des récompenses perçues lors de l'interaction entre agent et environnement. On note $\hat{\mathrm{V}}$ l'estimation d'une valeur d'état. Lorsque l'agent a perçu une récompense $\mathrm{R_t}$ suite à une décision $\mathrm{A_t}$ prise depuis l'état $\mathrm{S_t}$, la mise à jour de l'estimation $\hat{\mathrm{V}}^{\pi}(\mathrm{S_t})$ se fait selon l'Assignation 2.10.

$$\hat{\mathrm{V}}^{\pi}(\mathrm{S_t}) \leftarrow \hat{\mathrm{V}}^{\pi}(\mathrm{S_t}) + \alpha \left[\mathrm{R_t} - \hat{\mathrm{V}}^{\pi}(\mathrm{S_t})\right] \tag{2.10}$$

Le coefficient $\alpha \in [0, 1]$ est le taux d'apprentissage (ou *learning rate*), c'est un hyper-paramètre qui nivelle la force de la mise à jour. À chaque nouvelle récompense reçue par l'agent pour un état s, la valeur de l'état est recalculée. Ainsi, plus un état sera observé, plus l'estimation de sa valeur sera précise. La récompense R_t dépend de la politique $\pi(\mathrm{S}_t)$ de l'agent. Le calcul de l'estimation de la valeur $\hat{\mathrm{V}}^{\pi}(\mathrm{s})$ peut s'écrire comme sur l'Assignation 2.11.

$$\hat{\mathrm{V}}^{\pi}(s) \leftarrow \hat{\mathrm{V}}^{\pi}(s) + \alpha \left[r(s, \pi(s)) - \hat{\mathrm{V}}^{\pi}(s)\right] \tag{2.11}$$

Dans la mesure où le choix de l'agent est fondé sur la maximisation d'un enchaînement de plusieurs valeurs d'état, la valeur d'un état doit prendre en compte les états suivants. Si la mise à jour de l'estimation $\hat{\mathrm{V}}^{\pi}$ se fait uniquement selon l'équation 2.10, l'agent ne considère pas les états futurs, mais uniquement la récompense immédiate. Un facteur d'actualisation (ou *discount factor*) $\gamma \in [0, 1]$ est un hyper-paramètre utilisé pour prendre en compte les récompenses futures. Pour $\gamma = 1$, toutes les récompenses futures ont le même impact que la récompense immédiate. Si $\gamma = 0$, seule la récompense immédiate est considérée dans le calcul de $\hat{\mathrm{V}}^{\pi}(\mathrm{s})$. L'ajustement de la valeur d'un état en fonction des suivants s'effectue de manière récursive et empirique. Lorsqu'un agent observe un état $\mathrm{S_t}$, il utilise l'estimation de la valeur de l'état suivant $\mathrm{S_{t+1}}$ pour mettre à jour $\hat{\mathrm{V}}(\mathrm{S_t})$. Si on sait que l'agent arrivera à un état s' depuis l'état s en appliquant sa politique $\pi(\mathrm{s})$, alors la mise à jour de $\hat{\mathrm{V}}^{\pi}(\mathrm{s})$ s'écrit selon l'Assignation 2.12.

$$\hat{\mathrm{V}}^{\pi}(s) \leftarrow \hat{\mathrm{V}}^{\pi}(s) + \alpha \left[r(s, \pi(s)) + \gamma \hat{\mathrm{V}}^{\pi}(s') - \hat{\mathrm{V}}^{\pi}(s)\right] \tag{2.12}$$

L'équation de mise à jour de l'estimation de la valeur d'un état est toujours de la forme $\mathrm{V} \leftarrow \mathrm{V} + \alpha(\mathrm{B} - \mathrm{V})$. Le terme multiplié par α est la fonction de coût (ou *loss function*). Il correspond à l'erreur de l'estimation et doit être minimisé. Le terme B est la valeur cible (ou *target*). C'est la valeur vers laquelle l'algorithme cherche à faire converger l'estimation à

chaque pas de temps. On définit donc la valeur d'un état, cible de l'estimation de l'agent à travers l'Équation 2.13.

$$\mathrm{V}^{\pi}(\mathrm{S_t}) = r(s, \pi(s)) + \gamma \mathrm{V}^{\pi}(s') \tag{2.13}$$

En réalité, l'application de la politique π(s) depuis l'état s ne mène pas forcément à l'état s′ si l'environnement est stochastique. L'état de l'environnement suivant la prise de décision n'est pas connu à l'avance. V^{π} dépend de la probabilité de transition $P_{ss'}^{\pi(s)}$. En explicitant cette dépendance, l'équation 2.13 devient 2.14.

$$\mathrm{V}^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s'} P_{ss'}^{\pi(s)} \mathrm{V}^{\pi}(s') \tag{2.14}$$

$$= r(s, \pi(s)) + \gamma \sum_{s'} \mathbb{P}\left(s' \mid s, \pi\left(s\right)\right) \mathrm{V}^{\pi}(s') \tag{2.15}$$

Enfin, la dépendance des valeurs d'état à la politique π s'explicite dans l'Équation 2.16.

$$\mathrm{V}^{\pi}(s) = \sum_{a} \pi\left(a \mid s\right) \left[r(s, \pi(s)) + \gamma \sum_{s'} P_{ss'}^{\pi(s)} \mathrm{V}^{\pi}(s') \right] \tag{2.16}$$

V se définit aussi sous la forme d'une espérance (voir Équation 2.17) :

$$\mathrm{V}^{\pi}(s) = \mathbb{E}\left[\mathrm{G_t} \mid \mathrm{S_t} = s; \pi\right] \tag{2.17}$$

avec $G_t = \sum_{\mathrm{k}=t}^{T} \gamma^{\mathrm{k\text{-}t}} \mathrm{R_k}$ appelé retour amorti à partir de l'instant t.
Chercher à maximiser $\mathrm{G_t}$ revient à maximiser les récompenses ultérieures à chaque prise de décision. La trajectoire maximisant le retour amorti sera identique à la trajectoire maximisant les récompenses.

Valeur de couple action-état

Une alternative courante à l'estimation des valeurs d'état est l'estimation de la valeur des couples action-état ou Q-valeur $\mathrm{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Cette valeur dépend aussi de la politique π.

$$\mathrm{Q}^{\pi}(\mathrm{s}, \mathrm{a}) = \mathbb{E}\left[\mathrm{G_t} \mid \mathrm{S_t} = \mathrm{s}, \mathrm{A_t} = \mathrm{a}; \pi\right] \tag{2.18}$$

La Q-valeur représente l'espérance du retour amorti suite à la prise de la décision a depuis l'état s à l'instant t. En développant l'Équation 2.18, la Q-valeur peut s'exprimer d'une autre manière :

$$\mathrm{Q}^{\pi}(\mathrm{s}, \mathrm{a}) = \mathbb{E}\left[\sum_{\mathrm{k}=t}^{T} \gamma^{\mathrm{t\text{-}k}} \mathrm{R_k} \mid \mathrm{S_t} = \mathrm{s}, \mathrm{A_t} = \mathrm{a}; \pi\right] \tag{2.19}$$

$$= \gamma^0 \sum_{\mathrm{s}'} \sum_{\mathrm{r}} \mathbb{P}\left(\mathrm{S_{t+1}} = \mathrm{s}', \mathrm{R_t} = \mathrm{r} \mid \mathrm{S_t} = \mathrm{s}, \mathrm{A_t} = \mathrm{a}\right) \left[\mathrm{r} + \gamma^1 \mathbb{E}\left[\sum_{\mathrm{k}=\mathrm{t}+1}^{T} \gamma^{\mathrm{k\text{-}t+1}} \mathrm{R_k} \mid \mathrm{S_{t+1}} = \mathrm{s}'; \pi \right] \right] \tag{2.20}$$

$$= r(\mathrm{s}, \mathrm{a}) + \gamma \sum_{s'} P_{ss'}^{a} \mathbb{E}\left[G_{\mathrm{t}+1} \mid \mathrm{S_{t+1}} = s'; \pi\right] \tag{2.21}$$

$$= r(\mathrm{s}, \mathrm{a}) + \gamma \sum_{\mathrm{s}'} P_{ss'}^{a} \mathrm{V}^{\pi}(\mathrm{s}') \tag{2.22}$$

La relation entre valeur d'état et Q-valeur (Équation 2.23) est obtenue à partir des Équations 2.22 et 2.16.

$$V^{\pi}(s) = \sum_{a} Q^{\pi}(\mathrm{s}, \mathrm{a}) \pi\left(\mathrm{a} \mid \mathrm{s}\right) \tag{2.23}$$

L'estimation empirique $\hat{\mathrm{Q}}$ de la Q-valeur selon la politique π se fait aussi itérativement et de manière récursive à mesure des interactions entre l'agent et l'environnement selon l'Assignation 2.24.

$$\hat{\mathrm{Q}}^{\pi}(\mathrm{s},\mathrm{a}) \leftarrow \hat{\mathrm{Q}}^{\pi}(\mathrm{s},\mathrm{a}) + \alpha\left[r(\mathrm{s},\mathrm{a}) + \gamma\hat{\mathrm{Q}}^{\pi}(\mathrm{s}',\mathrm{a}') - \hat{\mathrm{Q}}^{\pi}(\mathrm{s},\mathrm{a})\right] \tag{2.24}$$

Des valeurs peuvent être associées à des états ou à des couples action-état de manière à maximiser le retour amorti et donc le cumul des récompenses. Lorsque les valeurs de couples action-états sont bien approximées, la politique générant les trajectoires qui maximisent les récompenses est identifiable. Elle s'obtient en choisissant l'action pour laquelle la valeur Q est la plus élevée à chaque état.

Politique optimale

Afin de définir une politique optimale, il est nécessaire de comparer les différentes politiques. Une politique π est supérieure à une politique π' si toutes les valeurs d'état $\mathrm{V}^{\pi}(\mathrm{s})$ sont supérieures à $\mathrm{V}^{\pi'}(\mathrm{s})$, quel que soit s (Voir Équivalence 2.25).

$$\pi \geq \pi' \iff \forall \mathrm{s} \in \mathcal{S},\ \mathrm{V}^{\pi}(\mathrm{s}) \geq \mathrm{V}^{\pi'}(\mathrm{s}) \tag{2.25}$$

Une politique π^* est donc optimale si la fonction valeur d'état associée V^{π^*} est toujours supérieure aux autres, quels que soient les états parcourus (voir Équation 2.26).

$$\forall \mathrm{s} \in \mathcal{S},\ \mathrm{V}^{\pi^*}(\mathrm{s}) = \max_{\pi} \mathrm{V}^{\pi}(\mathrm{s}) \tag{2.26}$$

Les fonctions valeurs d'état ont une équation d'optimalité pour chaque choix d'action (Voir Équation 2.27). Maximiser une politique revient à choisir l'action qui maximise le retour espéré depuis chaque état.

$$\max_{\pi}\left(\mathrm{V}^{\pi}(\mathrm{s})\right) = \max_{\mathrm{a}}\left(r(\mathrm{s},\mathrm{a}) + \gamma\sum_{\mathrm{s}'} P^{a}_{ss'}\mathrm{V}^{\pi^*}(\mathrm{s}')\right) \tag{2.27}$$

Ceci est valable pour la fonction associée aux Q-valeurs (voir Équation 2.28).

$$\max_{\pi}\left(\mathrm{Q}^{\pi}(s,a)\right) = r(s,a) + \gamma\max_{a'}\left(\sum_{s'} P^{a}_{ss'}\mathrm{Q}^{\pi^*}(s',a')\right) \tag{2.28}$$

La politique optimale pour un état s et une action a dépend donc du choix de l'action ultérieure a' pour la fonction associée aux Q-valeurs. Il en est de même pour la fonction des valeurs d'état puisqu'elle dépend de l'action maximisant le retour amorti de l'état suivant s'. Une fois la cartographie des fonctions associées aux valeurs effective, l'action optimale a^* peut être déterminée à chaque instant selon l'Équation 2.29

$$a^* = \arg\max_{a}\left(\mathrm{Q}^{\pi^*}(s,a)\right) \tag{2.29}$$

En revanche, obtenir une politique optimale peut s'avérer très coûteux en temps de calcul selon la taille du MDP, à cause de l'exploration que nécessite la cartographie des fonctions V et Q.

2.1.3 Programmation dynamique

Lorsque les probabilités de transitions $P^{a}_{ss'}$ et la politique de l'agent $\pi(\mathrm{s},\mathrm{a})$ sont connues, l'environnement est parfaitement défini. Dans ce cas, on parle d'apprentissage par renforcement

avec modèle. Si ce n'est pas le cas, l'apprentissage par renforcement est sans modèle. La Programmation Dynamique ou *Dynamic Programming* (DP) est une méthode de résolution avec modèle. Elle est peu utilisable en pratique, car les environnements parfaitement connus sont rares. La DP est néanmoins présentée dans le but d'appréhender les méthodes de résolution de problèmes sans modèle. La DP est une succession de deux processus : l'évaluation de la politique et l'amélioration de la politique. Selon la manière dont ils se succèdent, on parle d'itération sur la politique ou d'itération sur la valeur.

Évaluation de la politique

Il a été notifié dans la sous-section 2.1.2 que les valeurs d'états et Q-valeurs sont calculées itérativement. En appelant $\mathrm{V}_k(\mathrm{s})$ la valeur de $\mathrm{V}^\pi(\mathrm{s})$ à la k-ième itération, l'Équation 2.16 peut être utilisée pour affecter itérativement des valeurs à $\mathrm{V}_k(\mathrm{s})$:

$$\forall \mathrm{s} \in \mathcal{S},\ \mathrm{V}_{k+1}(\mathrm{s}) \leftarrow \sum_{\mathrm{a}} \pi\left(\mathrm{a} \mid \mathrm{s}\right) \left[r(\mathrm{s}, \pi(\mathrm{s})) + \gamma \sum_{\mathrm{s}'} P_{\mathrm{ss}'}^{\pi(\mathrm{s})} \mathrm{V}_k(\mathrm{s}') \right] \tag{2.30}$$

Ainsi, en répétant l'Assignation 2.30, la politique d'un agent peut être évaluée.

Algorithme 1 Évaluation de la politique

Entrées : Politique π devant être évaluée, $\theta \in \mathbb{R}^+$, $\gamma \in [0, 1]$
Initialisation : $\mathrm{V}_0(s) = 0,\ \forall s \in \mathcal{S}$, $\Delta = 0$
Répéter :
  $\forall s \in \mathcal{S}$:
    $v \leftarrow \mathrm{V}(s)$
    $\mathrm{V}(s) \leftarrow \sum_a \pi\left(a \mid s\right) \left[r(s, \pi(s)) + \gamma \sum_{s'} P_{ss'}^{\pi(s)} \mathrm{V}(s') \right]$
    $\Delta \leftarrow \max\left(\Delta, |v - \mathrm{V}(s)|\right)$
Jusqu'à $\Delta < \theta$
Sortie : $\mathrm{V} \approx \mathrm{V}^\pi$

L'Algorithme 1 calcule itérativement la fonction valeur d'état V^π de chaque état s selon la politique π. Ainsi, il est possible d'évaluer une politique lorsque le MDP est basé sur un modèle.

Amélioration de la politique

Afin d'obtenir une politique optimale, il faut améliorer la politique au sens de l'Équation 2.25. L'amélioration d'une politique implique la modification de la probabilité de choisir certaines actions afin d'augmenter le retour amorti. Par exemple, si pour un état $\mathrm{s} \in \mathcal{S}$ il existe une action $\mathrm{a} \in \mathcal{A}$ telle que $\mathrm{Q}^\pi(\mathrm{s}, \mathrm{a}) > \mathrm{V}^\pi(\mathrm{s})$ alors une politique π' qui accroit la probabilité de choisir l'action a depuis l'état s est meilleure que π.

$$\mathrm{Q}^\pi(\mathrm{s}, \mathrm{a}) = \mathrm{Q}^\pi(\mathrm{s}, \pi'(\mathrm{s})) \geq \mathrm{V}^\pi(\mathrm{s}) \Rightarrow \mathrm{V}^{\pi'}(\mathrm{s}) \geq \mathrm{V}^\pi(\mathrm{s}) \tag{2.31}$$

Dans l'Implication 2.31, π' est supérieure à π seulement si elle lui est identique excepté pour l'état s. Si $\pi'(\mathrm{s}) = \arg\max_{\mathrm{a}} \left(Q^\pi(\mathrm{s}, \mathrm{a})\right)$, alors $\pi'(\mathrm{s}) \geq \pi(\mathrm{s})$ et π' est dite **gloutonne** selon a depuis l'état s. Dans le cas inverse, la politique est **douce**. En général, une politique est gloutonne si elle choisit uniquement l'action qui maximise la fonction valeur depuis chaque état.

L'Algorithme 2 montre l'étape d'amélioration d'une politique. La seule modification de la politique $\pi(s)$ d'un état s change la politique globale de l'agent. Ainsi, elle doit donc être réévaluée pour connaître la valeur de chacun de ses états.

Algorithme 2 Amélioration de la politique

Entrées : Politique π devant être améliorée, $\gamma \in [0,1]$
Initialisation :
Pour tout $s \in \mathcal{S}$:
$\pi(s) \leftarrow \arg\max_a \left(r(s,a) + \gamma \sum_{s'} P^a_{ss'} \mathrm{V}(s')\right)$
Sortie : π

Itération sur la politique

La succession d'évaluations et d'améliorations de politique est appelée itération sur la politique (Elena et al., 1996). Lorsqu'une politique est évaluée, elle peut être améliorée puis réévaluée jusqu'à convergence vers la politique optimale. L'Algorithme 3 calcule l'itération

Algorithme 3 Itération sur la politique

Entrées : $\theta \in \mathbb{R}^+$, $\gamma \in [0,1]$
Initialisation : $\forall s \in \mathcal{S} : \pi(s) \in \mathcal{A}(s),\ \mathrm{V}(s) \in \mathbb{R}$ choisis arbitrairement
Évaluation de la politique :
$\Delta \leftarrow 0$
Répéter :
$\forall s \in \mathcal{S}$:
$v \leftarrow \mathrm{V}(s)$
$\mathrm{V}(s) \leftarrow \sum_a \pi(a \mid s) \left[r(s,\pi(s)) + \gamma \sum_{s'} P^{\pi(s)}_{ss'} \mathrm{V}(s')\right]$
$\Delta \leftarrow \max(\Delta, |v - \mathrm{V}(s)|)$
Jusqu'à $\Delta < \theta$
Amélioration de la politique :
stable $\leftarrow 1$
$\forall s \in \mathcal{S}$:
$a \leftarrow \pi(s)$
$\pi(s) \leftarrow \arg\max_a \left(r(s,\pi(s)) + \gamma \sum_{s'} P^{\pi(s)}_{ss'} \mathrm{V}(s')\right)$
Si $a \neq \pi(s)$ **Alors** stable $\leftarrow 0$
Si stable $= 0$
Retourner à l'étape **Évaluation de la politique**
Sortie : V, π

sur la politique. Chaque itération induit une évaluation entière de la politique. Cela peut demander beaucoup de ressources de calcul. Il est possible de tronquer l'évaluation de la politique de plusieurs manières. La plus répandue est l'itération sur la valeur (Elena et al., 1996).

Itération sur la valeur

L'itération sur la valeur consiste à arrêter l'évaluation de la politique après une seule itération sur la politique. Ainsi les itérations d'évaluation et d'amélioration de la politique se font en une seule étape. À chaque évaluation, le calcul de valeur d'un état ne dépend plus que de l'action gloutonne.

$$\mathrm{V}^{\pi}(\mathrm{s}) \leftarrow \max_{\mathrm{a}} \left(r(\mathrm{s},\mathrm{a}) + \gamma \sum_{s'} P^a_{ss'} \mathrm{V}^{\pi}(\mathrm{s}')\right) \tag{2.32}$$

La seule différence avec le calcul d'une valeur d'état pour l'évaluation de politique réside dans la présence de l'opérateur $\max_{\mathrm{a}}$ dans l'Affectation 2.32 qui permet d'améliorer la politique en la rendant gloutonne et donc déterministe. L'Algorithme 4 représente l'itération sur la valeur.

Algorithme 4 Itération sur la valeur

Entrées : $\gamma \in [0, 1]$
Initialisation : $\mathrm{V}_0 = 0,\ \forall s \in \mathcal{S},\ \Delta = 0$
Répéter :
  $\forall s \in \mathcal{S}\ :$
    $v \leftarrow \mathrm{V}(s)$
    $\mathrm{V}(s) \leftarrow \max_a \left(r(s, \pi(s)) + \gamma \sum_{s'} P_{ss'}^{\pi(s)} \mathrm{V}(s') \right)$
    $\Delta \leftarrow \max(\Delta, |v - \mathrm{V}(s)|)$
Jusqu'à $\Delta < \theta$
Sortie : $\mathrm{V} \approx \mathrm{V}^*$

Ainsi, il existe différentes manières de calculer une politique optimale en DP. Le problème principal de ces méthodes réside dans le postulat que les probabilités de transition et le système de récompense soient parfaitement connus.

2.1.4 Algorithmes usuels d'apprentissage par renforcement

Les algorithmes de RL peuvent s'appliquer sans modèle. Les probabilités de transition et les récompenses s'infèrent itérativement à mesure des interactions entre agent et environnement. Parmi les modèles les plus classiques, la méthode de Monte-Carlo et la méthode de la différence temporelle (TD) seront présentées dans cette section.

Monte-Carlo

L'algorithme de Monte-Carlo (Sutton et al., 1995) simule un épisode entier avec une politique, puis utilise la moyenne des retours amortis obtenus pour mettre à jour les valeurs d'état. Comme dans la sous-section 2.1.3, l'algorithme évalue la politique et l'améliore. L'évaluation de la politique consiste à générer des épisodes d'interactions avec une politique fixée, enregistrer la valeur des retours amortis et calculer la valeur de chaque état grâce à la moyenne des retours amortis observés.

Algorithme 5 Évaluation de la politique de Monte-Carlo

Entrées : Politique π devant être évaluée, $\gamma \in [0, 1]$
Initialisation : $\forall s, a \in \mathcal{S} \times \mathcal{A}\ :\ G(s, a) = 0, N(s, a) = 0$
Répéter toujours :
  Générer un épisode de $t = 0$ à $t = T$ en utilisant π
  Pour chaque couple action-état (s, a) apparaissant dans l'épisode généré,
  $t = T - 1, T - 2, ..., 0$:
    $G(s, a) \leftarrow G(s, a) + \sum_{k=t}^{T} \gamma^{k-t} R_k$
    $N(s, a) \leftarrow N(s, a) + 1$
    $\mathrm{Q}(s, a) = \frac{1}{N(s,a)} G(s, a)$
Sortie : $\mathrm{Q} \approx \mathrm{Q}^\pi$

L'Algorithme 5 montre la manière d'évaluer une politique avec la méthode de Monte-Carlo. Un nombre suffisant de simulations permet d'obtenir une bonne approximation de V^π sans connaître $P_{ss'}^a$ ni $r(s, a)$. Cette méthode permet d'évaluer les valeurs d'état V de la même manière qu'il évalue Q dans l'Algorithme 5. En revanche, si la politique est gloutonne, certains états ne seront jamais observés. Dans ce cas, la politique sera évaluée en considérant uniquement les quelques états observés. Une solution consiste à améliorer la politique de manière douce en retournant une probabilité non-nulle de sélectionner chaque action depuis chaque état. La phase d'amélioration de politique est appelée contrôle en RL.

Algorithme 6 Contrôle Monte-Carlo *On-policy*

Entrées : Politique π arbitraire, $\gamma \in [0,1]$, $\varepsilon \in [0,1]$
Initialisation : $\forall s, a \in \mathcal{S} \times \mathcal{A}$: $G(\mathrm{S}_t, \mathrm{A}_t) = 0,\ N(s,a) = 0$
Répéter toujours :
Générer un épisode de $t = 0$ à $t = T$ en utilisant π
Pour chaque tuple $(\mathrm{S}_t, \mathrm{A}_t, \mathrm{R}_t)$ apparaissant dans l'épisode généré, $t = T-1, T-2, ..., 0$:
$G(\mathrm{S}_t, \mathrm{A}_t) \leftarrow G(\mathrm{S}_t, \mathrm{A}_t) + \sum_{k=t}^{T} \gamma^{k-t} R_k$
$N(\mathrm{S}_t, \mathrm{A}_t) \leftarrow N(\mathrm{S}_t, \mathrm{A}_t) + 1$
$\mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t) = \frac{1}{N(\mathrm{S}_t, \mathrm{A}_t)} G(\mathrm{S}_t, \mathrm{A}_t)$
Changer la politique en utilisant l'opérateur $\arg\max_a$:

$$\pi(a \mid \mathrm{S}_t) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(\mathrm{S}_t)|} & \text{si } a = \arg\max_a (Q(\mathrm{S}_t, a)) \\ \frac{\varepsilon}{|\mathcal{A}(\mathrm{S}_t)|} & \text{pour chacune des autres actions} \end{cases}$$

Dans l'Algorithme 6, un réel ε est utilisé pour déterminer une probabilité non-nulle de choisir chaque action. Si ε décroit au fur et à mesure des itérations, alors l'algorithme convergera vers une politique gloutonne. Ce compromis entre l'emploi d'une politique douce pour observer plus d'états ou de couples action-état et d'une politique gloutonne pour maximiser les retours amortis est appelé compromis exploration/exploitation. L'**exploration** consiste à prendre des décisions pour observer des états dans le but d'affiner l'évaluation. Tandis que l'**exploitation** consiste à prendre des décisions qui maximisent les retours amortis pour améliorer la politique. Avec l'Algorithme 6, la politique évaluée par l'agent est la politique qu'il déploie lors de l'échantillonnage d'épisodes. C'est un algorithme *On-policy*. Ces termes sont employés, car l'agent met à jour la politique à mesure qu'il l'utilise.

Algorithme 7 Contrôle Monte-Carlo *Off-policy*

Entrées : Politique π gloutonne arbitraire, politique b douce arbitraire, $\gamma \in [0,1]$
Initialisation : $\forall s, a \in \mathcal{S} \times \mathcal{A}$: $C(s,a) = 0, N(s,a) = 0$
Répéter toujours :
Générer un épisode de $t = 0$ à $t = T$ en utilisant b
$G \leftarrow 0$
$W \leftarrow 1$
Répéter pour chaque couple action-état $(\mathrm{S}_t, \mathrm{A}_t)$ apparaissant dans l'épisode généré, $t = T-1, T-2, ..., 0$:
$G \leftarrow \gamma G + R_{t+1}$
$C(\mathrm{S}_t, \mathrm{A}_t) \leftarrow C(\mathrm{S}_t, \mathrm{A}_t) + W$
$\mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t) = \frac{W}{C(\mathrm{S}_t, \mathrm{A}_t)} [G - Q(\mathrm{S}_t, \mathrm{A}_t)]$
Changer la politique en utilisant l'opérateur $\arg\max_a (Q(\mathrm{S}_t), a)$:
$\pi(\mathrm{S}_t) \leftarrow \arg\max_a (Q(\mathrm{S}_t, \mathrm{a}))$
Si $\mathrm{A}_t \neq \pi(\mathrm{S}_t)$:
Quitter la boucle de l'épisode généré
$W \leftarrow W \frac{1}{b(\mathrm{A}_t, \mathrm{S}_t)}$

Une autre méthode consiste à séparer la politique apprise $\pi(\mathrm{s}, \mathrm{a})$ de la politique déployée $b(\mathrm{s}, \mathrm{a})$ par l'agent lors de l'échantillonnage des épisodes. La politique utilisée pour échantillonner un épisode est appelée politique comportementale. Un tel algorithme est *Off-policy*. Un algorithme *Off-policy* peut apprendre une politique gloutonne tout en explorant pour mettre à jour la valeur des états ou les Q-valeurs. La politique apprise n'est pas la politique

déployée, car l'agent explore alors qu'il apprend une politique gloutonne. L'avantage d'une telle méthode est la souplesse de l'exploration, l'agent ne tend pas forcément vers une politique gloutonne dans son comportement et peut échantillonner des trajectoires variées.
L'Algorithme 7 est un exemple de Monte-Carlo utilisé en *Off-policy* pour une politique comportementale quelconque. L'algorithme de Monte-Carlo a l'avantage d'être utilisable sans modèle de l'environnement. C'est souhaitable pour la majorité des problèmes de contrôle, car les modèles sont rarement explicites. En revanche, les calculs des valeurs d'états sont indépendants. Pour que la méthode soit efficace, toutes les paires action-état doivent donc être explorées.

Différence temporelle (TD)

En apprentissage par différence temporelle (TD-*learning*), la valeur des états ou les Q-valeurs des états suivants sont calculées récursivement comme en DP, mais l'agent apprend sans modèle de l'environnement. L'apprentissage est réalisé à partir de l'échantillonnage de ses décisions comme la méthode de Monte-Carlo, mais il calcule la valeur de ses états et Q-valeurs en fonction des Q-valeurs des couples action-état ou de la valeur des états suivants. L'évaluation de la politique par TD ressemble à l'évaluation par Monte-Carlo à la différence qu'elle inclut la Q-valeur des couples action-état suivants.

La manière d'explorer la plus classique consiste à choisir un taux d'exploration $\varepsilon \in [0, 1]$. L'action qui maximise le retour amorti est prise avec une probabilité $1 - \varepsilon$, sinon une action aléatoire est choisie avec une probabilité ε. Une telle politique est qualifiée de ε-gloutonne. Comme la méthode de Monte-Carlo, l'apprentissage TD peut s'effectuer en *On-policy* ou en *Off-policy*. L'algorithme TD en *On-policy* est **SARSA** (Rummery, G.A et al., 1994) tandis que l'algorithme en *Off-policy* est le **Q-learning** (Watkins, Christopher J. C. H. et al., 1992). Dans les deux cas, la différence avec l'algorithme de Monte-Carlo est que les Q-valeurs sont calculées à chaque interaction entre l'agent et l'environnement et non à la fin d'un épisode d'interaction. L'algorithme SARSA met à jour $Q(\mathrm{S}_t, \mathrm{A}_t)$ en fonction de $Q(\mathrm{S}_{t+1}, \mathrm{A}_{t+1})$ issue de la politique comportementale à l'instant $t + 1$.

Algorithme 8 SARSA, apprentissage TD en *On-policy*

Entrées : Politique π gloutonne arbitraire, $\varepsilon \in [0, 1]$, $\gamma \in [0, 1]$, $\alpha \in [0, 1]$
Initialisation : $\forall s, a \in \mathcal{S} \times \mathcal{A}$: $\mathrm{Q}(s, a)$ arbitraire
Répéter pour chaque épisode :
  Initialiser S_0
  Choisir A_0 depuis S_0 en utilisant une politique dérivée de π (ε-gourmande par exemple)

  Répéter pour chaque pas de l'épisode $t = 0, 1, ..., T - 1$:
    Observer $\mathrm{R}_t, \mathrm{S}_{t+1}$
    Choisir A_{t+1} depuis S_{t+1} en utilisant une politique dérivée de π (ε-gourmande par exemple)
    $\mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t) \leftarrow \mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t) + \alpha \left[\mathrm{R}_t + \gamma \mathrm{Q}(\mathrm{S}_{t+1}, \mathrm{A}_{t+1}) - \mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t)\right]$
    $\mathrm{S}_t \leftarrow \mathrm{S}_{t+1}$
    $\mathrm{A}_t \leftarrow \mathrm{A}_{t+1}$
    Jusqu'à atteindre un état terminal

L'Algorithme 8 montre la manière dont SARSA met à jour itérativement les Q-valeurs en fonction des Q-valeurs des couples action-état suivants.
Contrairement à SARSA, l'algorithme de Q-learning (Algorithme 9) met à jour les Q-valeurs $Q(\mathrm{S}_t, \mathrm{A}_t)$ indépendamment de la politique utilisée à l'instant $t + 1$, mais en fonction de la Q-valeur de l'action maximisant le retour amorti depuis l'état S_{t+1}.

Algorithme 9 Q-learning, apprentissage TD en *Off-policy*

Entrées : Politique π gloutonne arbitraire, $\varepsilon \in [0,1]$, $\gamma \in [0,1]$, $\alpha \in [0,1]$
Initialisation : $\forall s, a \in \mathcal{S} \times \mathcal{A}$: $\mathrm{Q}(s,a)$ arbitraire
Répéter pour chaque épisode :
  Initialiser S_0
  Répéter pour chaque pas t **de l'épisode** $t = 0, 1, ..., T-1$:
    Choisir A_t depuis S_t en utilisant une politique dérivée de π (ε-gourmande par exemple)
    Observer $\mathrm{R}_t, \mathrm{S}_{t+1}$
    $\mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t) \leftarrow \mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t) + \alpha \left[\mathrm{R}_t + \gamma \underset{a}{max} \mathrm{Q}(\mathrm{S}_{t+1}, a) - \mathrm{Q}(\mathrm{S}_t, \mathrm{A}_t)\right]$
    $\mathrm{S}_t \leftarrow \mathrm{S}_{t+1}$
    Jusqu'à atteindre un état terminal

La différence entre l'apprentissage TD et la méthode de Monte-Carlo est que l'apprentissage TD se fait au sein d'un épisode, avec la relation de récurrence entre la valeur d'un état et de l'état suivant. En revanche, les deux méthodes nécessitent d'explorer tous les états ou couples action-état. Ces méthodes d'apprentissage par renforcement ne sont pas appropriées lorsque les espaces sont trop grands. De plus, à de nombreux problèmes sont associés des espaces d'état et d'action indénombrables. C'est le cas lorsque les valeurs d'état ou d'action sont prises dans un espace de variables continues. Dans ce cas, il y a une infinité de valeurs à calculer et aucun de ces algorithmes ne converge. La solution est l'apprentissage par renforcement profond.

2.2 Apprentissage par renforcement profond

La méthode de Monte-Carlo et l'apprentissage TD fonctionnent pour des environnements aux espaces d'état et d'action discrets. Les états ou les actions continus peuvent être discrétisés, la précision du modèle de l'environnement diminuera en conséquence. Plus la discrétisation est fine, plus le nombre d'états et d'actions nécessaires pour représenter l'environnement est important. L'estimation empirique de chaque valeur prend plus de temps dans des espaces larges. Avec l'apprentissage par renforcement profond, des réseaux de neurones sont utilisés pour prédire les Q-valeurs (voir sous-section 2.2.1) ou pour mettre à jour directement la politique d'un agent (voir sous-section 2.2.2).

2.2.1 Deep Q-learning

Le *Deep Q-Learning* (DQN) (Mnih et al., 2013) est une méthode fondée sur le Q-learning. On parle d'apprentissage profond, car l'estimation des Q-valeurs de chaque action vient d'un réseau de neurones. Ce réseau de neurones est appelé Q-réseau $\mathrm{Q}_\theta(s,a)$, avec θ son paramétrage. Ainsi, les valeurs de chaque état sont les entrées du Q-réseau, les sorties sont les Q-valeurs associées à chaque action. En conséquence, les actions sont en nombre fini. Le calcul des Q-valeurs est le résultat d'une régression effectuée par le Q-réseau. Le DQN est régi par la même équation que le Q-learning. On note $\mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t)$ la valeur du couple action-état $(\mathrm{S}_t, \mathrm{A}_t)$ calculée par le Q-réseau.

$$\mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t) \leftarrow \mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t) + \alpha \left(\mathrm{R}_t + \gamma \max_a \left(\mathrm{Q}_\theta(\mathrm{S}_{t+1}, a)\right) - \mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t)\right) \tag{2.33}$$

L'Assignation 2.33 est l'actualisation itérative des valeurs de Q_θ. Comme en Q-learning, l'opérateur $\max_a$ est utilisé pour la prédiction de la Q-valeur de l'instant suivant. Le problème mis en évidence par l'Assignation 2.33 est la non-stationnarité de la cible de l'apprentissage.

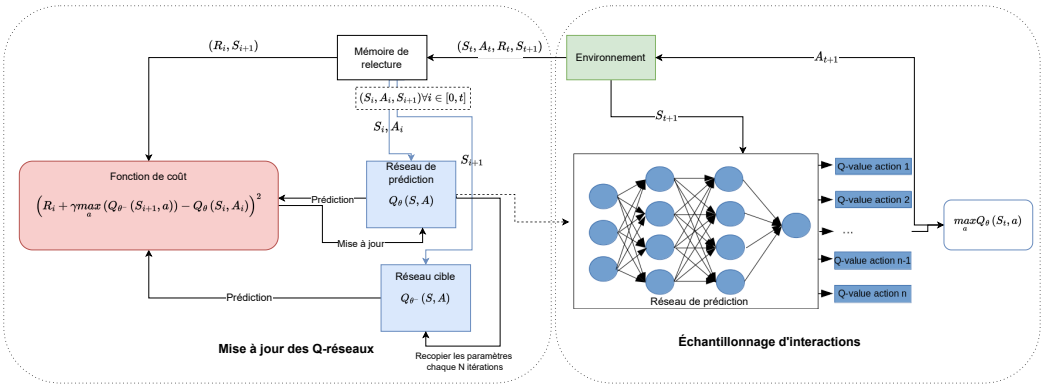

Figure 2.3 – Représentation schématique de l'algorithme de DQN

Comme vu à la Section 2.1, la mise à jour se fait selon une cible dont la prédiction de la Q-valeur doit se rapprocher. Ici, la cible dépend d'une prédiction faite par le Q-réseau avec le paramètre θ mis à jour. La valeur de la cible évolue donc au cours de l'apprentissage, la convergence n'est pas garantie. Une solution est de maintenir la valeur de θ pendant N_{cible} itérations. Il y a alors deux réseaux : un réseau de prédiction paramétré par θ est mis à jour à chaque itération d'entraînement, tandis qu'un réseau cible paramétré par θ^-, utilisé uniquement pour prédire la cible, est mis à jour toutes les N_{cible} itérations. L'Assignation 2.33 devient l'Assignation 2.34.

$$\mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t) \leftarrow \mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t) + \alpha \left(\mathrm{R}_t + \gamma \max_a \left(\mathrm{Q}_{\theta^-}(\mathrm{S}_{t+1}, a)\right) - \mathrm{Q}_\theta(\mathrm{S}_t, \mathrm{A}_t)\right) \tag{2.34}$$

La Figure 2.3 représente schématiquement le fonctionnement du DQN. Elle met en évidence l'utilisation de deux réseaux de neurones pour le calcul de la fonction coût L du DQN. Ce coût est l'écart quadratique entre la cible $\mathrm{R}_i + \gamma \max_a \left(\mathrm{Q}_{\theta^-}\left(S_{i+1}, a\right)\right)$ et la prédiction du Q-réseau $\mathrm{Q}_\theta\left(S_i, A_i\right)$ comme le montre l'Équation 2.35.

$$L\left(\theta\right) = \left(\mathrm{R}_i + \gamma \max_a \left(\mathrm{Q}_{\theta^-}\left(S_{i+1}, a\right)\right) - \mathrm{Q}_\theta\left(S_i, A_i\right)\right)^2 \tag{2.35}$$

Le Q-réseau peut aussi être vu comme un classificateur, car l'action associée à la plus grande Q-valeur prédite sera sélectionnée avec une probabilité plus grande dans le cas d'une politique comportementale ε-gloutonne. Sur la Figure 2.3, à droite, la manière d'échantillonner des tuples est représentée. Le Q-réseau est utilisé pour calculer les Q-valeurs et l'action est échantillonnée selon ε. La mise à jour des Q-réseaux est représentée à gauche de la figure. Les tuples échantillonnés sont stockés dans une mémoire de relecture, décrite dans le paragraphe suivant. Cette mémoire permet de mettre à jour le Q-réseau paramétré par θ selon l'Équation 2.35. Un réseau de neurones supplémentaire est utilisé pour calculer la cible de cette équation. C'est le Q-réseau mis à jour qui est utilisé pour échantillonner des interactions à droite de la figure.

La Q-valeur calculée influence les actions de l'agent et donc les prochains états en entrée du Q-réseau. Ces états consécutifs sont fortement corrélés, ce qui peut entraîner un apprentissage inefficace ou instable. En conséquence, une mémoire de relecture, visible sur la Figure 2.3, est utilisée. Cette mémoire stocke les tuples $(\mathrm{S}_t, \mathrm{A}_t, \mathrm{R}_t, \mathrm{S}_{t+1})$ échantillonnés par l'interaction de l'agent avec son environnement. Le Q-réseau peut prélever de manière aléatoire des tuples déjà échantillonnés par l'agent en tant qu'entrées pour son apprentissage. Il évite ainsi de se mettre à jour uniquement sur les choix et observations récents de l'agent. Ainsi, une méthodologie similaire au Q-learning peut s'appliquer sur des environnements aux états indénombrables grâce à l'utilisation d'un Q-réseau. En revanche, le DQN ne s'applique pas

aux espaces d'action continus puisque chaque sortie du Q-réseau est une valeur associée à chaque action. Le *Deep Q-learning* a évolué vers des algorithmes plus avancés comme le *Prioritized Experience Replay* (Schaul et al., 2016), le *Double DQN* (van Hasselt et al., 2015), ou le *dueling Deep Q-learning* (Z. Wang et al., 2015) qui sont des versions plus stables ou efficaces (en fonctionnant avec moins d'échantillons d'observation).

2.2.2 Méthodes fondées sur le gradient de politique

Une autre approche d'apprentissage par renforcement profond consiste à utiliser un ou plusieurs réseaux de neurones pour mettre à jour directement la politique d'un agent. Les sorties d'un tel réseau de neurones seraient les probabilités de choix de chaque action ou une distribution de probabilités si l'espace d'actions est continu. Soit $\tau \sim \pi_\theta$ ($\tau = (\mathrm{S}_0, \mathrm{A}_0, ..., \mathrm{S}_T)$) une trajectoire échantillonnée suivant la politique π_θ issue d'un réseau de neurones paramétré par θ, la fonction objectif à maximiser en lien avec la fonction récompense peut alors s'écrire selon l'Équation 2.36.

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[R_\tau\right] \tag{2.36}$$
$$= \int_\tau \mathbb{P}\left(\tau \mid \pi_\theta\right) R_\tau d\tau \tag{2.37}$$
$$= \int_\tau \pi_\theta(\tau) r(\tau) d\tau \tag{2.38}$$

L'Équation 2.37 s'obtient en exprimant l'espérance sur la trajectoire τ. L'Équation 2.3 permet de parvenir à l'Équation 2.38. Cette dernière équation s'obtient en exprimant la probabilité d'une trajectoire selon la politique π_θ par l'Équation 2.5. Comme dans les sections précédentes, l'objectif reste de modifier la politique afin de trouver la trajectoire qui maximise les récompenses. Les paramètres à modifier itérativement pour maximiser J sont θ. Comme les espaces sont continus, la mise à jour de θ se fait selon l'Assignation 2.39.

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta) \tag{2.39}$$

L'estimation du gradient $\nabla_\theta J(\theta)$, appelé **gradient de politique**, est nécessaire. D'après les Équations 2.38 et 2.5 ce gradient s'écrit selon l'Équation 2.40.

$$\nabla_\theta J(\theta) = \int_\tau \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau \tag{2.40}$$

$$\pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) = \nabla_\theta \pi_\theta(\tau) \tag{2.41}$$

Grâce à l'Équation 2.41 (Schulman, 2016), le gradient de politique s'exprime comme l'Équation 2.42.

$$\nabla_\theta J(\theta) = \int_\tau \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau \tag{2.40}$$
$$= \int_\tau \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) r(\tau) d\tau \tag{2.42}$$
$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\tau) R_\tau\right] \tag{2.43}$$

Puisque le gradient de politique s'exprime comme une espérance d'après l'Équation 2.43, il est possible de l'estimer empiriquement. Le gradient de politique se calcule en utilisant des données échantillonnées sur une trajectoire entière. Similairement à la méthode de Monte Carlo, la mise à jour du réseau de neurones se fait donc une fois l'échantillonnage d'un épisode terminé. Les récompenses échantillonnées R_τ ne dépendent pas directement de θ que l'on cherche à mettre à jour, mais de la trajectoire τ générée selon π_θ. La différenciation

cherchée l'est selon θ, on peut donc écrire l'équation 2.44.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathrm{S}_t, \mathrm{A}_t\right) R_\tau\right] \tag{2.44}$$

Une forme plus adaptée au calcul itératif du gradient de politique pour la mise à jour de θ s'écrit selon l'Équation 2.45, avec N le nombre de trajectoires générées.

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{\tau=\tau_1}^{\tau_N}\left[\sum_{t=1}^{T} \nabla_\theta log\pi_\theta\left(\mathrm{S}_t, \mathrm{A}_t\right) R_\tau\right] \tag{2.45}$$

Comme avec les méthodes de RL plus classiques, un facteur d'actualisation γ peut pondérer l'influence des récompenses lointaines dans le calcul du gradient de politique. Le retour amorti G_t peut donc être utilisé à la place de la somme des récompenses le long d'une trajectoire R_τ. C'est le cas dans l'algorithme REINFORCE (Sutton et al., 1999) (Algorithme 10). Il est prouvé (Schulman, 2016) que pour toute fonction b dépendant uniquement de l'état, l'Équation 2.46 est vérifiée.

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log(\pi_\theta(\mathrm{S}_t, \mathrm{A}_t))b(\mathrm{S}_t)\right] = 0 \tag{2.46}$$

b est appelé base de référence. Puisque l'Équation 2.46 est valide pour n'importe quel $b(\mathrm{S}_t)$, dépendant de l'état, il est possible de soustraire n'importe quelle base de référence à la récompense perçue dans l'Équation 2.45. Cette opération est présentée dans l'Équation 2.47. Cela permet de réduire la variance due à l'échantillonnage sans changer l'espérance utilisée pour calculer le gradient de politique.

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{1}^{N}\left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathrm{S}_t, \mathrm{A}_t\right)\right)\left(\sum_{t=1}^{T} R(\mathrm{S}_t, \mathrm{A}_t) - b(\mathrm{S}_t)\right)\right] \tag{2.47}$$

Si le gradient de politique est écrit avec l'Équation 2.47, la force et la direction de la mise à jour se déterminent selon l'écart entre la base de référence et les récompenses obtenues. Il est prouvé que la base de référence peut aussi être la fonction d'estimation de Q-valeur Q^{π_θ} bien qu'elle dépende des actions.
Ainsi, plusieurs algorithmes d'apprentissage par renforcement profond fondés sur le gradient de politique se différencient par la fonction qui leur sert de base de référence. Quelques exemples de fonction b avec le nom de l'algorithme associé sont montrés dans les Équations (2.48, 2.49, 2.50) (Konda et al., 1999 ; Mnih et al., 2016 ; Sutton et al., 1999) :

$$\nabla_\theta J(\pi_\theta) = \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\left(\nabla_\theta \log \pi_\theta(a_t|s_t)\right)\sum_{t'=t}^{T} G_{t'}\right] \qquad \textbf{REINFORCE} \tag{2.48}$$

$$= \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) Q^w(s_t, a_t)\right)\right] \qquad \textbf{Q Acteur-Critique} \tag{2.49}$$

$$= \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\left(\nabla_\theta \log \pi_\theta(a_t|s_t) A^w(s_t, a_t)\right)\right] \qquad \textbf{Avantage Acteur-Critique} \tag{2.50}$$

Avec w les paramètres du réseau de neurones associé à l'estimation des Q-valeurs, et A la fonction avantage qui s'écrit selon l'Équation 2.51.

$$A^\pi\left(\mathrm{S}_t, \mathrm{A}_t\right) = Q^\pi\left(\mathrm{S}_t, \mathrm{A}_t\right) - V^\pi\left(\mathrm{S}_t\right) \tag{2.51}$$

L'utilisation de cette fonction permet de diminuer la variance du produit qui apparaît dans le gradient de politique (Équation 2.47) pour chaque choix d'action. Les algorithmes dont les gradients de politique sont calculés selon les Équations (2.49 et 2.50) sont appelés Acteur-Critiques, car deux réseaux de neurones sont utilisés.

— L'**Acteur** met à jour la politique π dans la direction suggérée par le critique (par l'intermédiaire du gradient de politique). Il s'agit du réseau de neurones paramétré par θ.

— Le **Critique** estime la fonction valeur (Q, V, A, etc). Ce réseau de neurones est celui paramétré par w.

L'acteur et le critique sont complémentaires : Le critique calcule des fonctions de valeur (d'état, de couple action-état ou d'avantage) et l'acteur redéfinit une politique en fonction du critique. Q^w n'est pas mis à jour selon $\underset{a}{\text{argmax}} Q(\mathrm{S}_{t+1}, a)$ comme en Q-learning. Son fonctionnement ressemble davantage à SARSA, car il dépend de la politique utilisée pour échantillonner les interactions agent-environnement. Q^w s'ajuste selon la politique choisie et donc selon l'acteur. L'acteur et le critique sont interdépendants. L'espace des actions $\mathcal{A}$ étant continu, le Q-réseau critique prend en entrée un vecteur d'état et détermine une distribution de valeurs (selon la base de référence choisie) associée à l'espace d'actions de l'environnement. De nombreux algorithmes (DDPG Lillicrap et al., 2019, PPO Schulman et al., 2017b, TRPO Schulman et al., 2017a, A3C Mnih et al., 2016, SAC Haarnoja et al., 2018, TD3 Fujimoto et al., 2018) fondés sur le gradient de politique ont été construits pour répondre à des problématiques propres aux premiers algorithmes développés. Les espaces d'état et d'action, les systèmes de récompense des environnements, et la longueur des épisodes peuvent être très différents. Ces spécificités peuvent susciter des problèmes dans la convergence d'un entraînement. Les agents peuvent sous-explorer et apprendre des politiques de contrôles sous-optimales, perdre les acquis d'un entraînement à cause de gradients de mise à jour trop élevés. Les algorithmes présentés ont été développés pour répondre à ces problématiques spécifiques.

Algorithme 10 REINFORCE

Entrées : θ, paramètres d'une politique π_θ, $\alpha \in [0, 1]$
Initialisation : θ arbitraire
Générer un épisode $\mathrm{S}_0, \mathrm{A}_0, \mathrm{R}_0 ..., \mathrm{R}_T$ en suivant la politique π_θ :
Répéter pour chaque pas de l'épisode $t = 0, 1, ..., T-1$:
$G \leftarrow \sum_{k=t}^{T} \gamma^{k-1} \mathrm{R}_k$
$\theta \leftarrow \theta + \alpha G \nabla_\theta \log\left(\pi_\theta(\mathrm{S}_t, \mathrm{A}_t)\right)$

2.3 Apprentissage par renforcement hors ligne

Les méthodes vues jusqu'ici font partie de l'apprentissage par renforcement en ligne : l'agent interagit directement avec son environnement pour apprendre la politique de contrôle optimale. Ce n'est pas toujours possible : soit une simulation de l'environnement n'est pas toujours valide, soit l'utilisation d'un agent pendant son apprentissage sur un système réel présente des risques pour le système. De plus, le système de récompense peut être difficile à représenter y compris en simulation. Lorsque les objectifs d'un problème de contrôle sont multiples et complexes, il est difficile de les exprimer explicitement. L'apprentissage par renforcement hors ligne est une solution à ces problèmes puisque l'agent n'interagit pas avec son environnement, mais apprend à partir des interactions passées. Un démonstrateur ou expert interagit avec l'environnement, l'agent observe cette interaction et apprend. Ce démonstrateur peut être un humain, un algorithme ou un automate par exemple. Dans la sous-section 2.3.1, l'agent apprend une politique en n'observant que les actions et états

échantillonnés par le démonstrateur. Dans la sous-section 2.3.2, l'agent observe également les récompenses.

2.3.1 Apprentissage par imitation

Pour l'apprentissage de l'agent, il est parfois difficile de définir un système complexe de récompense. Récompenser la conduite d'une voiture autonome, par exemple, s'avère difficile compte tenu des nombreux objectifs (éviter les obstacles, respecter la signalisation, limiter les émissions...). Un objectif ne prévaut pas forcément sur les autres en toute circonstance. L'apprentissage par imitation utilise les interactions entre un démonstrateur et son environnement pour qu'un agent puisse concevoir la politique la plus proche possible de celle observée, sans utiliser de système de récompense explicite. Ces interactions observées sont appelées démonstrations.

Clonage comportemental

La forme la plus simple de l'apprentissage par imitation est le **clonage comportemental** ou *behavioral cloning* (BC) (Bratko et al., 1995). Pour l'agent, la politique π^* du démonstrateur est considérée optimale. Les actions et états issus de la trajectoire τ^* du démonstrateur sont observés. Un modèle d'apprentissage supervisé (classification) est utilisé pour reproduire π^* à partir des trajectoires observées. La cible de l'apprentissage est l'action choisie par le démonstrateur selon l'état. La fonction coût de cet apprentissage est l'écart entre l'action a^* prise par le démonstrateur et la politique $\pi_\theta(s)$ de l'agent. Le clonage comportemental n'est applicable qu'à des problèmes de complexité simple. En apprentissage supervisé, les données (entrées et cible) doivent être distribuées aléatoirement. Le nombre de démonstrations étant limité, de nombreux états ne sont jamais observés par l'agent. C'est le cas des états faisant suite aux actions que le démonstrateur ne prend pas. Si l'agent doit prendre une décision dans un état non observé, elle ne pourra pas correspondre à celle qu'aurait pris le démonstrateur. Ainsi, une prise de décision de l'agent trop différente de celle du démonstrateur peut le mener vers un sous-espace d'états non observés, ce qui peut s'avérer catastrophique.

Agrégation du jeu de données

L'**agrégation du jeu de données** ou *Dataset aggregation* (DAgger) (Ross et al., 2010) permet de compenser les faiblesses de BC. Le principe reste celui de l'apprentissage supervisé sur les données du démonstrateur. En revanche, avec DAgger l'agent échantillonne des interactions selon sa politique puis le démonstrateur indique quelle action il aurait pris pour chaque état observé sur la trajectoire τ de l'agent. L'agent compare sa politique avec les actions prises par le démonstrateur comme avec la méthodologie BC. Ainsi, la probabilité d'observer un sous-espace d'états inconnu est réduite pendant et après l'entraînement. En revanche, les actions du démonstrateur sont considérées comme issues d'une politique optimale, ce qui n'est pas systématiquement vrai.

Apprentissage par renforcement inverse

BC et DAgger ne fonctionnent pas dans les espaces d'état et d'action de dimensions importantes. Ils nécessitent un très grand nombre de démonstrations d'experts pour observer tous les états. L'apprentissage par renforcement inverse ou *Inverse RL* (IRL) (Andrew Y. Ng et al., 2000) est une classe d'algorithmes qui cherche à inférer un système de récompense pouvant justifier les décisions du démonstrateur. Ainsi, lorsqu'un système de récompense peut décrire le comportement de l'expert, l'agent peut se passer de démonstration et apprendre avec ce système de récompense. La forme la plus simple d'IRL est l'inférence d'une récompense linéaire selon un vecteur de combinaison des états, appelé vecteur des caractéristiques des états. On suppose qu'il existe une fonction $x : \mathcal{S} \to \mathbb{R}^n$ qui crée un vecteur des caractéristiques

des états à partir des états de l'environnement. L'hypothèse d'une récompense linéaire selon $x(s)$ pour chaque état $s \in \mathcal{S}$ est faite. L'inférence du système de récompense se fait en trouvant des coefficients de pondération pour chaque élément de ce vecteur.

$$R(s) = w^k x(s) \ (k \in \mathbb{N}) \tag{2.52}$$

Dans l'Équation 2.52, $w \in \mathbb{R}^n$ est le vecteur de poids associé aux caractéristiques des états pour un jeu de démonstrations donné. Le problème principal de ces méthodes est que plusieurs politiques sont optimales selon le système de récompense inféré, car la trajectoire du démonstrateur n'est pas prise en compte. Plus de détails sur l'IRL sont donnés dans l'Annexe A.1.

2.3.2 Apprentissage par renforcement par batch

L'apprentissage par renforcement par batch (batch RL) combine des éléments d'apprentissage par imitation et de RL classique. Le système de récompense est connu, mais l'agent apprend uniquement en observant des démonstrations. Il est ainsi possible d'apprendre une politique avec les avantages du RL, l'environnement peut avoir un comportement dynamique dépendant d'actions de contrôle et dépendre de données aléatoires, sans interaction agent-environnement. Dans cette sous-section deux types d'algorithmes de batch RL sont présentés. Les principes régissant le fonctionnement de l'algorithme d'itération ajustée sur Q puis l'algorithme de Q-learning contraint par batch sont expliqués.

Algorithme d'itération ajustée sur Q

Selon (Ernst et al., 2005), l'apprentissage en ligne, c'est-à-dire avec des interactions directes entre l'agent et l'environnement, atteint ses limites sans réseau de neurones, si les espaces représentant l'environnement sont très grands. C'est pourquoi l'algorithme d'itération ajustée sur Q (Ernst et al., 2005) a été développé. L'idée est d'utiliser des interactions pour échantillonner des tuples $(\mathrm{S_t}, \mathrm{A_t}, \mathrm{R_t}, \mathrm{S_{t+1}})$ en observant un démonstrateur. Une fois les tuples reçus, les Q-valeurs sont approximées par régression. L'approximation se fait en plusieurs étapes. Un premier algorithme de régression Q_0 prédit les récompenses $\mathrm{R_t}$ selon ses entrées $(\mathrm{S_t}, \mathrm{A_t})$. Les régressions suivantes $\hat{Q}_N, N \in \mathbb{R}$ prédisent la Q-valeur d'un tuple à l'instant t en estimant les sorties de la régression précédente à l'instant $t+1$ et la récompense immédiate. L'Équation 2.53 montre la prédiction réalisée pour un tuple $(\mathrm{S_t}, \mathrm{A_t})$.

$$\hat{Q}_N(\mathrm{S_t}, \mathrm{A_t}) = R_t + \gamma \max_{\mathrm{a}} \hat{Q}_{N-1}(\mathrm{S_{t+1}}, \mathrm{a}) \tag{2.53}$$

En itérant, l'influence de la Q-valeur des états suivants se propage dans l'estimation des Q-valeurs.

Algorithme 11 Itération ajustée sur Q

Entrées : $\gamma \in [0,1]$, plusieurs échantillons de 4 tuples $\{(\mathrm{S}_l, \mathrm{A}_l, \mathrm{R}_l, \mathrm{S}_{l'})\}, l \in [0, ..., \mid B \mid]$, K horizon d'optimisation
Initialisation : $\hat{Q}_0 = 0$ pour chaque paire action-état
Générer un épisode $\mathrm{S}_0, \mathrm{A}_0, \mathrm{R}_0 ..., \mathrm{R}_T$ en suivant la politique π_θ :
Pour n=1,..., N faire :
  Pour i = 1,..., l faire :
    $Q_{n,l} \leftarrow R_l + \gamma \max_a \hat{Q}_{n-1}$
  Entraîner l'algorithme de régression à estimer $\hat{Q}_n$ à partir de $\{((\mathrm{S}_l, \mathrm{A}_l), Q_{k,l}), l \in [0, ..., \mid B \mid]\}$
Sortie : $\hat{Q}_N$

L'Algorithme 11 montre le fonctionnement de l'itération ajustée sur Q. Plus tard, l'algorithme de l'itération ajustée sur Q par réseau de neurones (Riedmiller, 2005) a été développé. Les réseaux de neurones sont utilisés pour prédire les Q-valeurs par apprentissage supervisé. Bien que hors ligne, il est le précurseur du DQN puisque l'utilisation des échantillons collectés antérieurement ressemble au principe de la mémoire de relecture.

Algorithme de Q-learning contraint par batch

La mémoire de relecture est aussi utilisée de manière hors ligne dans d'autres algorithmes. En comparaison avec le DQN, il n'est pas nécessaire de mettre à jour le Q-réseau pendant l'interaction agent-environnement. Au lieu de cela, un démonstrateur remplit la mémoire de relecture de l'agent avec des tuples d'interaction et l'agent peut commencer l'apprentissage avec une mémoire complète. Ces algorithmes utilisent de manière plus efficace les données à disposition. En revanche, des défauts sont liés à l'apprentissage hors ligne. Le problème principal est la surestimation des Q-valeurs des paires action-état qui n'ont pas été observées dans les données à disposition. Les actions non-prises par le démonstrateur ont une plus grande valeur que les autres. C'est l'opposé de l'IRL qui considère que les états les plus observés sont optimaux. Cela est dû à une erreur d'extrapolation : les agents de RL hors ligne ne peuvent pas explorer et certaines paires d'actions-état peuvent ne pas être observées. Par conséquent, l'estimation des paires d'états et d'actions non contenues dans les données d'apprentissage peut être biaisée. Dans le RL en ligne, l'agent peut interagir avec l'environnement et visiter les états et les paires d'actions ayant des valeurs estimées élevées pour éventuellement les réajuster. L'algorithme de Q-learning contraint par batch ou *Batch Constrained Q-learning* (BCQ) (Fujimoto et al., 2019) est un algorithme conçu pour régler ce problème en utilisant des principes de RL profond et d'apprentissage par imitation. La politique apprise est "contrainte" d'éviter les actions qui ont peu de chances d'être prises par le démonstrateur. Une politique doit induire la même fréquence de visite état-action que le batch qui a permis son apprentissage tout en maximisant les récompenses. Les politiques apprises doivent sélectionner des actions en fonction de 3 objectifs :

1 Minimiser la distance entre les actions choisies et celles choisies par le démonstrateur.

2 Mener à des états où des observations familières peuvent être faites.

3 Maximiser la fonction Q-valeur.

Pour que l'objectif [1] soit atteint, un modèle génératif $G_\omega(\mathrm{s})$ est utilisé. Ses sorties sont des actions avec une forte probabilité d'être prises par le démonstrateur selon le jeu de données. $G_\omega(\mathrm{s})$ est un algorithme d'apprentissage par imitation, plus précisément de BC. BCQ fonctionne de deux manières différentes selon que l'espace d'actions est discret ou continu. Puisque dans cette thèse, l'espace d'actions est discrétisé, seul l'utilisation de BCQ sur un espace d'actions discontinu est introduit ici. Dans cette configuration, $G_\omega(\mathrm{s})$ prend un état en entrée et fournit la probabilité de prendre chaque action. L'idée est d'éliminer le biais de surestimation de la valeur Q pour les paires état-action les moins fréquemment observées. Ensuite, un algorithme de double DQN (van Hasselt et al., 2015) est utilisé pour prédire la paire d'états-actions ayant la valeur Q la plus élevée. En DQN, toutes les actions possibles sont considérées comme des choix potentiels et parmi elles, l'action ayant la Q-valeur la plus élevée est choisie. Avec BCQ le modèle génératif contraint le choix de l'action, des actions ne sont pas susceptibles d'être choisies par l'agent malgré une potentielle Q-valeur élevée. Ainsi, un seuil τ_{BCQ} est utilisé pour éliminer les actions peu susceptibles d'être observées dans la mémoire de relecture.

$$\pi(\mathrm{s}) = \underset{a \mid \frac{G_\omega(a|s)}{\max_{\tilde{a}} G_\omega(\tilde{a}|s)} > \tau_{\mathrm{BCQ}}}{\arg\max} Q_\theta(\mathrm{s}, \mathrm{a}) \tag{2.54}$$

L'Équation 2.54 montre le choix de la politique de BCQ pour chaque état s. Le seuil de BCQ est un compromis entre DQN et BC. S'il vaut 1, l'algorithme se comporte comme BC. S'il vaut 0, c'est un algorithme de double DQN. Cet algorithme parvient à obtenir des scores plus

élevés que les démonstrateurs qui ont généré sa mémoire de relecture sur des environnements classiques (Fujimoto et al., 2019).

2.4 Conclusion

Tout comme la programmation dynamique, l'apprentissage par renforcement est un outil de prises de décision séquentielles. Par rapport aux autres approches appliquées aux problèmes de Markov (programmation dynamique, commande prédictive, optimisation), le RL se distingue par plusieurs avantages. Contrairement à la programmation dynamique, cette méthodologie est compatible avec un environnement stochastique, et ne requiert aucune connaissance sur les probabilités de transition. Le RL s'applique au contrôle en temps réel sans nécessité de disposer de modèles prédictifs des variables aléatoires, ce qui n'est pas le cas de la commande prédictive. Enfin, les décisions sont prises en temps réel malgré les données aléatoires, une réponse non-linéaire de l'environnement aux actions de l'agent et l'absence de prédiction.

Si le système de récompense d'un agent est difficile à expliciter, l'apprentissage par imitation permet d'apprendre une politique de contrôle uniquement en observant des interactions avec l'environnement. Cet apprentissage est effectué en inférant les objectifs d'un démonstrateur uniquement en considérant sa politique optimale. Cette possibilité d'apprentissage hors ligne de l'agent RL permet d'éviter des interactions potentiellement dangereuses entre l'environnement et l'agent lors de son entraînement. En reprenant les principes du RL et de l'apprentissage par imitation, le RL hors ligne permet d'apprendre en observant du démonstrateur sans hypothèse d'optimalité sur les données observées. Ainsi, il est efficace sur un jeu de données limité et non-optimal. Les micro-réseaux électriques constituent un environnement liant à la fois modèles dynamiques représentables par des équations et données aléatoires. Une politique de contrôle (prises de décisions séquentielles) dans un tel environnement avec un horizon de temps fini peut s'établir avec un algorithme d'apprentissage par renforcement. Le choix de l'algorithme utilisé se fait selon la taille de l'environnement défini et selon le système de récompense de l'agent.

3

L'apprentissage par renforcement appliqué au contrôle des micro-réseaux

Les micro-réseaux permettent d'augmenter l'autoconsommation électrique des utilisateurs en rapprochant les points de demande, la production et le stockage électrique. Le réseau central de distribution est ainsi moins sollicité et les consommateurs gagnent en autonomie. L'autonomie obtenue par l'intermédiaire des micro-réseaux offre plusieurs avantages clés. Premièrement, elle confère une indépendance énergétique aux consommateurs, leur permettant d'être moins dépendants des fluctuations des approvisionnements et des prix de l'énergie. De plus, elle renforce la résilience en cas de panne du réseau central, assurant ainsi une continuité de service essentielle. Enfin, l'autonomie favorise une meilleure gestion de l'énergie, adaptée aux besoins spécifiques de l'utilisateur, tout en contribuant à la réduction des émissions de gaz à effet de serre par l'intégration d'énergies renouvelables. En revanche, la demande et la production renouvelables d'électricité sont des phénomènes stochastiques. La production d'énergie renouvelable n'est pas toujours en phase avec la demande d'énergie. Par exemple, la production d'énergie solaire est maximale pendant la journée, surtout en milieu de journée lorsque le soleil est à son zénith, alors que la demande d'énergie atteint souvent son pic en fin de journée et en soirée pour un profil de consommation de type résidentiel. Ce décalage peut entraîner des problèmes d'équilibre entre l'offre et la demande sur le réseau, et nécessite des solutions de stockage d'énergie efficaces et économiques pour stocker l'énergie produite pendant les périodes de faible demande et la mettre à disposition pendant les périodes de forte

demande. Le micro-réseau doit être contrôlé en maîtrisant le comportement dynamique des unités qui le composent et en anticipant ces phénomènes aléatoires. Le contrôle haut-niveau en temps réel consiste à prendre des décisions pour chaque unité contrôlable du micro-réseau avec un pas de temps relativement long (de plusieurs minutes à une heure).
Le RL est un outil efficace pour les problèmes de prises de décisions séquentielles avec phénomènes stochastiques et déterministes. Ce chapitre met en évidence la contribution de divers travaux appliquant des techniques d'apprentissage par renforcement au contrôle haut-niveau de micro-réseaux. La section 3.1 présente les différents problèmes de contrôle haut-niveau de micro-réseaux résolvables avec le RL. Les travaux montrant l'utilisation du RL pour résoudre ces problèmes sont mis en lumière dans les sections suivantes. La section 3.2 traite du problème de planification des systèmes de stockage énergétique. La section 3.3 présente des résolutions de problèmes de gestion de la demande. Enfin la section 3.4 montre l'application du RL pour l'administration du partage d'énergie entre plusieurs micro-réseaux.

3.1 Formulation des différents problèmes de contrôle haut niveau

Comme présenté au chapitre 1, le contrôle des micro-réseaux peut être catégorisé en trois niveaux selon la résolution temporelle des décisions de contrôle :

— Niveau primaire : Régulation réactive de la puissance entre différents onduleurs pour le contrôle de la tension et de la fréquence à une échelle de temps inférieure à la seconde.
— Niveau secondaire : Contrôle en régime permanent pour corriger les écarts de tension et de fréquence en programmant de manière optimale l'énergie dans la production et le stockage.
— Niveau tertiaire : Planification à long terme et routine de réparation pour un fonctionnement optimal du micro-réseau en terme de coûts.

Le type de contrôle étudié dans cette thèse est le contrôle haut-niveau (niveau tertiaire). Il est défini ici comme la prise de décisions sur les flux d'énergie dans un micro-réseau en fonctionnement normal, qu'il soit complètement isolé ou connecté au réseau central de distribution, avec une résolution temporelle minimale de 15 minutes. Par conséquent, les décisions qui doivent être prises à haute fréquence sont exclues. Le contrôle haut-niveau des micro-réseaux peut être séparé en trois catégories :

— Planification des systèmes de stockage et engagement d'unités
— Gestion de la demande
— Partage d'énergie entre plusieurs micro-réseaux

Le contrôle haut niveau en temps réel de micro-réseaux peut être réalisé par un algorithme fondé sur un système de règles, comme des méthodes déterministes ou de logique floue, sur des méthodes d'optimisation ou sur des méthodes d'apprentissage automatique. Néanmoins, la stochasticité de l'environnement rend difficile un contrôle optimal basé sur des règles de priorité lorsque la production d'énergie est d'origine renouvelable. La demande, le prix et la génération renouvelable du micro-réseau sont des phénomènes aléatoires auxquels le contrôle doit s'adapter. L'approche fondée sur un modèle physique complet du micro-réseau n'est donc pas adaptée aux incertitudes causées par ces phénomènes. Avec l'utilisation croissante de capteurs, les modèles d'apprentissage automatique sont des outils adaptés au contrôle de micro-réseaux. L'utilisation des données permet de construire des modèles permettant d'anticiper l'évolution des variables aléatoires. Le RL est un outil approprié pour le contrôle de haut niveau des micro-réseaux, car il permet de prendre des décisions de manière séquentielle et adaptative.

3.2 Planification des systèmes de stockage et engagement des unités

Une bonne utilisation de générateurs de secours et un stockage intelligent de l'énergie octroient plus d'autonomie aux micro-réseaux. Ils permettent aussi d'accroître leur rentabilité au regard des échanges avec le réseau extérieur. L'ajustement adéquat du contrôle de ces unités peut contribuer à la réduction des émissions de gaz à effet de serre associées à l'utilisation d'unités de production énergétique contrôlables. Une politique de contrôle efficace optimise ces objectifs, tels que l'autonomie, la rentabilité et la réduction des émissions, en considérant à la fois les contraintes liées aux unités et les variables stochastiques du problème.

Cette section met en évidence que la planification des systèmes de stockage et l'engagement des unités sont des thématiques de contrôle pour lesquelles le RL s'est avéré efficace dans de nombreux travaux. Une large majorité des études sur le RL appliquée à ce type de contrôle intègre une batterie électrochimique à l'architecture du micro-réseau. En revanche, les autres technologies intégrées à la structure des micro-réseaux varient, puisque chaque micro-réseau répond à une problématique unique s'inscrivant dans un marché ou une zone géographique spécifiques (Giraldez Miner et al., 2018). Les productions scientifiques revues dans cette section seront donc discriminées selon les technologies impliquées dans le micro-réseau étudié. Les véhicules électriques, ou *electric vehicles* (EV), sont pris en compte dans certaines études, et leur rôle au sein du micro-réseau, souvent décrit comme « stationnaire », varie selon les hypothèses émises. Ceci contraste avec les véhicules électriques eux-mêmes, qui peuvent être considérés comme des micro-réseaux mobiles, mais sont ici intégrés comme des unités au sein du micro-réseau stationnaire. S'ils peuvent uniquement être chargés, ils sont associés à une demande électrique. S'ils peuvent être chargés et déchargés, ils se comportent comme un stockage électrique avec des contraintes spécifiques.

Les systèmes de récompenses utilisés lors de l'entraînement de l'agent de RL permettent également de catégoriser les publications puisqu'ils traduisent de manière concise l'objectif des auteurs en matière de contrôle. Les récompenses ont des valeurs négatives et sont donc des pénalités. Le coût économique d'opération est l'objectif le plus fréquent. Il est associé à une pénalité lorsque le micro-réseau achète de l'énergie à l'extérieur et prend des valeurs positives lorsqu'il lui est possible de vendre de l'énergie. Lorsque le micro-réseau fonctionne en mode îloté, une pénalité est perçue lorsque l'approvisionnement de la demande en énergie n'est pas réalisé. Elle est similaire à la pénalité d'achat d'énergie à un réseau extérieur. D'autres pénalités courantes sont liées à l'émission de gaz à effet de serre, à la perte d'énergie produite et non-valorisée, et au non-respect de la contrainte de puissance au PCC (voir chapitre 1). L'autonomie, dans le contexte d'un micro-réseau, se réfère à l'indépendance vis-à-vis du réseau extérieur lorsqu'il fonctionne en mode connecté, et à la capacité du système à opérer sans perte lorsqu'il est en mode îloté. Puisque la programmation des batterie se fait selon son état de charge (SOC) pour *State of charge*, cette variable peut être associée à une pénalité de mauvaise gestion de la batterie. Enfin, des agents sont pénalisés en fonction de la dégradation des unités du micro-réseau.

Un autre critère permettant la catégorisation est la fonctionnalité du micro-réseau. Certains micro-réseaux sont conçus pour contrôler des flux énergétiques de différentes natures, tels que l'électricité, le gaz et la chaleur, un concept parfois appelé « planification énergétique multi-vectorielle » dans la littérature anglophone (Onen et al., 2022). Certaines études se distinguent par des caractéristiques spécifiques telles que l'exploitation du micro-réseau en mode connecté ou îloté, la mise en place de systèmes de tarification de l'énergie au sein du micro-réseau entre ses différentes unités, et l'utilisation simultanée de diverses technologies de stockage électrique.

Puisque la planification des unités de stockage est une partie essentielle au contrôle de micro-réseaux, elle est explorée dans la majorité des travaux sur lesquels cette section est

construite. Les différentes manières d'appréhender le contrôle des unités de stockage seront donc présentées tout au long de cette section selon le contexte des problèmes de contrôle exposés. Cette section s'articule en deux parties. L'engagement des unités pour le pilotage d'un micro-réseau est traité dans un premier temps. Puis, des problématiques de contrôle plus spécifiques sont examinées.

3.2.1 Engagement des unités

Le problème d'engagement des unités consiste à contrôler quand et à quelle puissance les unités de génération distribuées et contrôlables doivent fonctionner. Elles sont souvent alimentées par du fioul et ont donc un certain coût d'utilisation (économique et écologique), contrairement aux unités de production électrique renouvelable. Ji et al., 2019, Shuai et al., 2021 et Yoldas et al., 2020 contrôlent le point de fonctionnement à la fois des unités de génération et de stockage. Les coûts économiques liés à l'utilisation des générateurs contrôlables et à l'échange d'énergie avec le réseau central sont considérés dans les trois travaux. La pénalité reçue par l'agent de Shuai et al., 2021 se distingue par l'intégration d'une composante liée à la production d'énergie renouvelable non valorisée. Yoldas et al., 2020 pénalise les émissions dues aux générateurs contrôlables.
L'ajout du contrôle d'unités en plus de la planification de la batterie augmente la taille de l'espace d'actions d'un agent de RL. Avec un grand espace d'actions, les actions peuvent avoir un impact très différent sur l'environnement, ce qui peut entraîner une variation brusque des gradients lors de l'optimisation. Cette variation est due à la sensibilité de la fonction de Q-valeur, où de petites variations dans les actions peuvent provoquer des changements significatifs, rendant ainsi les gradients plus instables. Le risque d'instabilité de l'entraînement augmente avec la taille des espaces d'état et d'action. Pour cette raison, Guo et al., 2022 utilise de l'algorithme PPO lors de l'apprentissage de la politique de contrôle d'une batterie et d'une microturbine. Un mécanisme conservateur de mise à jour de politique est employé par PPO pour limiter les changements apportés à une politique existante. En contraignant les nouvelles politiques à rester proches des précédentes, ce mécanisme diminue le risque d'instabilité de l'apprentissage. B. V. Mbuwir et al., 2020 résout ce problème en distribuant les actions vers plusieurs agents, avec un agent par unité contrôlable. Les unités sont une batterie électrochimique et une pompe à chaleur. Le choix de l'apprentissage par renforcement hors ligne a été fait et il ne permet pas aux agents d'apprendre en considérant les actions de chacun puisqu'ils n'interagissent pas avec leur environnement. Il s'agit d'une configuration multi-agents non coordonnés. Afin de remédier à ce problème, l'optimisation du contrôle s'effectue en deux étapes. L'agent pilotant la pompe à chaleur optimise d'abord sa politique de contrôle et communique ses choix d'action à l'agent pilotant à la batterie. Ce dernier ajoute la consommation de la pompe à chaleur à la consommation totale au sein de son espace d'états.

Approche multi-agents. Une tendance croissante d'utilisation d'algorithmes de contrôle multi-agents est observée. Contrairement aux systèmes distribués à contrôle centralisé, chaque agent contrôle une unité en autonomie dans une approche multi-agents. La coopération n'est pas dirigée, on parle de coordination. Les objectifs des agents peuvent être différents et ils peuvent adopter des comportements coopératifs, compétitifs ou mixtes. Par exemple, dans un système d'enchère d'énergie, les agents liés à des unités de production sont vendeurs, ceux liés à des demandes sont acheteurs et les systèmes de stockage occupent les deux fonctions (Foruzan et al., 2018). L'objectif de chaque agent est de maximiser son profit tout en satisfaisant les besoins des consommateurs. Les récompenses liées au profit et à la satisfaction de la contrainte d'approvisionnement des consommations en énergie sont communes dans les travaux présentés par F.-D. Li et al., 2012, en plus d'une récompense propre à chaque agent. Ces travaux sont fondés sur les recherches de Dimeas et al., 2010 qui mettent en place un système multi-agents afin d'engager les unités après un blackout énergétique dans un micro-réseau en mode îloté. Le système de récompense défini prend

des valeurs négatives lorsque la dépendance au stockage virtuel (la puissance des unités de génération contrôlable) augmente.

Conversion d'énergie. Les micro-réseaux multi-vectoriels (aussi appelés *hubs énergétiques*) doivent assurer le contrôle de plusieurs flux tels que l'eau, le gaz, la chaleur et l'électricité. Leur pilotage nécessite le contrôle d'unités de conversion d'énergie. Par exemple, Sun et al., 2017 contrôle un hub énergétique transformant l'électricité, la chaleur urbaine et le gaz naturel en une forme d'énergie désirée selon la consommation avec un algorithme de Q-learning. La demande énergétique doit être approvisionnée sous la forme d'énergie appropriée, ce qui nécessite une bonne gestion des flux énergétiques selon les unités de production, de conversion et de stockage énergétiques disponibles sur le micro-réseau. Dreher et al., 2022 planifie la puissance d'opération d'un électrolyseur pour approvisionner une turbine à gaz dans un micro-réseau connecté à un réseau d'électricité et à un réseau de gaz naturel. L'algorithme PPO est utilisé et ses performances sont comparées à un algorithme basé sur des règles et à un algorithme de programmation dynamique.
Afin de contrôler l'état des systèmes de stockage électrique et thermique, et les flux d'énergie entre les différentes unités de production, de stockage et de consommation, Y. Ye et al., 2020 se sert de deep RL. L'algorithme DDPG est utilisé afin de prendre de nombreuses décisions à chaque pas de temps sans devoir discrétiser l'espace des actions. Cette planification est effectuée dans le but de réduire les coûts d'achat d'électricité et de gaz à des réseaux centraux et d'augmenter les revenus liés à la vente du surplus énergétique du micro-réseau. D. Qiu et al., 2022 a développé un algorithme DDPG guidé par des prédictions d'un réseau de neurones LTSM afin de contrôler des systèmes de stockage thermiques et électriques, d'un chauffe-eau, d'une pompe à chaleur et d'une unité de production combinée de chaleur et d'électricité fonctionnant au gaz naturel. Cet algorithme permet de réduire les violations de contraintes par l'agent tout en s'adaptant aux multiples incertitudes de l'environnement. L'utilisation de deep RL est devenue bien plus fréquente que celle des algorithmes de RL classiques pour ce type de contrôle qui requiert un choix multiple d'actions à chaque instant dans un environnement complexe.
Plus récemment, des techniques de contrôle multi-agents coordonnés sont développées dans le but de résoudre ces problèmes. En particulier, l'entraînement centralisé avec exécution décentralisée ou *Centralised training with decentralised execution* (CTDE) permet d'améliorer l'entraînement sans partager d'information entre les agents. L'algorithme multi-agents DDPG (MADDPG) est ainsi employé par G. Zhang et al., 2022 pour contrôler les flux d'électricité, de gaz et d'eau dans un hub énergétique. L'algorithme SAC est utilisé par Zhu et al., 2022 pour obtenir une politique « douce » pour le contrôle en temps réel de flux de puissance électrique, de gaz et de chaleur. Par rapport à l'algorithme classique, des modifications ont été apportées. Celles-ci incluent l'ajout de multiplicateurs de Lagrange pour empêcher l'agent de sortir des contraintes des systèmes de stockage. Un « mécanisme d'attention » a également été introduit pour guider l'exploration en se concentrant sur les informations pertinentes. Puisque plusieurs unités doivent être contrôlées simultanément, l'approche multi-agents est courante dans une optique d'engagement des unités.

Micro-réseaux en mode îloté. La gestion des générateurs contrôlables peut s'avérer cruciale pour la stabilité d'un micro-réseau en mode îloté. Un micro-réseau îloté est autonome et ne peut pas échanger d'énergie avec l'extérieur. Ces générateurs servent d'unités de secours lorsque les autres unités ne suffisent pas à fournir la demande énergétique. Les travaux de Kozlov et al., 2020 mettent en lumière une comparaison de plusieurs algorithmes voués à contrôler un système de stockage dans un micro-réseau comprenant des générateurs distribués. La capacité d'approvisionnement de la demande compte tenu de la décision sur la charge ou la décharge de la batterie est comblée par l'utilisation des générateurs de secours si elle n'est pas suffisante. S. Zhang et al., 2021 compare aussi la performance d'algorithmes de deep RL sur l'engagement des unités dans un micro-réseau en mode îloté comportant un générateur distribué et des panneaux PV pour la production d'électricité. DQN et TD3 obtiennent les meilleurs résultats. Il est conclu qu'il est préférable de vider la batterie avant

les épisodes d'ensoleillement afin de valoriser au maximum la production d'électricité d'origine photovoltaïque. Des coûts d'utilisation du générateur distribué sont ainsi évités.
L'autonomie d'un système îloté est primordiale. La planification de l'engagement des unités est couramment organisée avec un horizon temporel plus court que pour les autres systèmes afin de s'assurer que les contraintes d'équilibre soient atteintes à chaque instant. Deux horizons temporels de planification sont considérés dans les travaux de Lei et al., 2021. Une comparaison de l'incertitude sur l'utilisation d'un algorithme DDPG et DDPG récurrent est effectuée sur un horizon temporel de planification de la puissance d'opération d'un générateur diesel sur deux heures consécutives et deux jours consécutifs. Levent et al., 2021 traite le contrôle de générateurs distribués sous l'occurrence d'évènements rares dans les observations de l'agent. La longueur définie d'un épisode est alors d'un jour. Le but est de rendre la politique de contrôle applicable à un maximum de situations possibles. Un algorithme de machine learning supervisé est entraîné avec la table générée par le Q-learning afin d'utiliser les actions les plus susceptibles de mener à des récompenses plus élevées lorsque des événements rares se produisent. Des évènements à long terme doivent aussi être considérés. Totaro et al., 2021 tient compte des évènements tels que des changements dans la consommation d'électricité, la dégradation des panneaux PV et de la batterie, ainsi que les pannes de la batterie sur un pilotage en temps réel. L'algorithme utilisé est une nouvelle fois PPO pour garantir une monotonie dans l'évolution de la politique de contrôle.
Enfin, les micro-réseaux fonctionnant en mode îlotés peuvent aussi être des hubs énergétiques. Kofinas et al., 2018a met au point un algorithme multi-agents afin de contrôler l'opération d'une batterie, d'une PAC, d'un électrolyseur, d'un générateur distribué et d'une unité de désalinisation de l'eau.

La gestion des unités de production contrôlables s'avère cruciale lorsque le micro-réseau est îloté ou sujet à des flux énergétiques de différentes natures. Un espace d'actions continu et plus grand que pour la simple planification des systèmes de stockage est une difficulté inhérente à l'engagement des unités. Lorsqu'un micro-réseau fonctionne en mode îloté, des solutions de contrôle sont mises en œuvre pour maintenir son autonomie à court terme. Toutefois, la nature changeante de l'environnement du micro-réseau au fil de sa durée de vie requiert une attention particulière. Ainsi, un second niveau temporel de planification est de plus en plus pris en compte. Cela permet d'éviter que le système ne s'appuie que sur un apprentissage basé sur un environnement susceptible d'évoluer, ce qui pourrait compromettre l'efficacité du contrôle à long terme. Ces problématiques peuvent se combiner selon la structure et les objectifs associés au micro-réseau considéré.

3.2.2 Récompenses dans le contrôle des unités de production et de stockage

L'ingénierie de la récompense (Dewey, 2014) est cruciale pour l'apprentissage de l'agent. Le rôle de la récompense est d'influencer la politique de contrôle pour la mener vers un but désiré. Plusieurs formulations de la récompense sont possibles pour un même objectif et peuvent s'avérer plus ou moins efficaces en fonction de l'environnement (durée d'un épisode, pas de temps entre chaque prise de décision, contraintes...) et des paramètres propres à l'agent (facteur d'actualisation, taux d'apprentissage, mécanisme de mise à jour...). De plus, les objectifs peuvent être de différentes natures. La manière d'intégrer plusieurs récompenses de nature différente est un choix qui s'impose pour l'ingénierie de la récompense.

Coût économique d'opération. L'objectif le plus courant dans l'opération des systèmes de stockage et l'engagement des unités est l'optimisation du coût économique d'opération. Généralement, il inclut le coût de fonctionnement des unités contrôlables et les coûts et revenus liés à l'échange avec le réseau central. Cependant, ce coût est propre à l'environnement du micro-réseau et sa formulation dépend de sa constitution et de facteurs extérieurs. Par exemple, Kolodziejczyk et al., 2021 alloue la somme des coûts économiques à la fin d'une

journée comme récompense, alors que l'agent perçoit déjà un à un ces coûts distribués à chaque pas de temps. Les échanges avec un réseau extérieur peuvent être unilatéraux (possibilité de soutirage) ou bi-latéraux (soutirage et injection). L'établissement d'un contexte d'échange d'énergie diffère aussi selon les cas d'études. Ainsi, Cao et al., 2019 a pour objectif de minimiser le coût de l'énergie soutirée et de maximiser les revenus engendrés par la vente d'électricité à des EVs. L'évolution du prix de l'électricité est pris en compte dans les travaux de Berlink et al., 2015. La récompense perçue à chaque instant par l'agent est la différence entre les revenus d'injection d'électricité perçus par le micro-réseau avec le revenu qu'il aurait perçu en injectant la même énergie à un prix moyen. Les micro-réseaux fonctionnant en mode îloté ne peuvent pas percevoir de récompense liée aux échanges d'énergie (Levent et al., 2021, Hasanvand et al., 2020, Domínguez-Barbero et al., 2020). En revanche, la pénalisation du non-respect de la contrainte d'équilibre énergétique est similaire à celle du soutirage d'énergie à la différence qu'il n'y a pas de prix associé à la situation où l'énergie produite est insuffisante pour répondre à la demande.

Valorisation de la production locale d'électricité. Les micro-réseaux îlotés ne peuvent pas non plus injecter l'électricité générée en excès. Cette énergie non-valorisée est perdue lorsqu'elle n'est pas stockée ou directement consommée. La quantité d'énergie perdue est appelée énergie excédentaire et peut intervenir dans le système de récompense d'un algorithme de RL : Kozlov et al., 2020 contrôle une batterie électrochimique dans un micro-réseau en mode îloté en pénalisant l'électricité produite non-valorisée. Cette pénalité est vue comme un manque d'autonomie du micro-réseau qui ne valorise pas une partie de l'énergie produite localement. En opérant en mode connecté, les micro-réseaux ne peuvent pas toujours vendre d'électricité sans limites de puissance au réseau extérieur. Un seuil appelé puissance au point de couplage commun (PCC) est la limite de puissance injectable. Ainsi, toute configuration de micro-réseau est exposée à un excédent non-valorisé d'énergie. Shuai et al., 2021 pénalise la différence entre la puissance électrique productible maximale à celle effectivement produite par des générateurs éoliens et PV, la puissance d'injection à chaque pas de temps étant contrainte par la puissance de PCC.
Dans les travaux de Y. Liu et al., 2020 une récompense positive est attribuée à la puissance PV employée pour la charge de la batterie et une récompense négative à la puissance solaire au réseau central. Cette pénalité est assimilable à un manque d'autonomie du micro-réseau puisque cette énergie serait perdue sans l'injection au réseau central. Une récompense positive similaire est perçue par l'agent dans la publication de Kuznetsova et al., 2013, qui dépend de la proportion d'électricité d'origine éolienne dans la charge de la batterie. D'autres travaux pénalisent l'énergie excédentaire alors que le micro-réseau a la possibilité d'injecter l'électricité en surplus. X. Huang et al., 2021 et Leo et al., 2014 attribuent une pénalité au productible PV que la batterie ne peut pas utiliser pour se charger.

Dégradation des unités Le coût économique d'opération est le système de récompense le plus répandu en planification des batteries et engagement des unités. Puisque la batterie est une unité nécessitant un fort coût d'investissement et que son vieillissement dépend de la politique de contrôle, il est possible de prendre en compte sa dégradation dans le système de récompense d'un agent de RL. L'agent est récompensé de manière séquentielle. Après ses prises de décision, la récompense liée au vieillissement de la batterie doit donc être distribuée à chaque instant. Plusieurs méthodes de modélisation séquentielle de la dégradation de la batterie sont possibles. Chaque technologie de batterie électrochimique se dégrade de manière différente et le diagnostic réel de l'état de santé d'une batterie ne peut s'établir sans interrompre son opération au sein du micro-réseau. La récompense n'est qu'une prise en compte de l'influence des actions de l'agent sur l'état de santé de la batterie. Elle peut être plus ou moins réaliste dans la représentation de la dégradation de la batterie induite par les actions de l'agent. Venayagamoorthy et al., 2016 et Y. Liu et al., 2020 fixent une pénalité en fonction du nombre d'actions prises sur la charge ou la décharge de la batterie.
L'espérance de vie d'une batterie peut se modéliser selon le nombre de cycles d'utilisation (un cycle étant une charge complète suivie d'une décharge complète ou l'inverse) et la profondeur

de décharge (ou *Depth of discharge*) (DoD). La température est aussi un paramètre jouant sur le vieillissement d'une batterie. La majorité des modélisations de la dégradation d'une batterie sont linéaires. Par exemple, Shuai et al., 2021, Fang et al., 2019, Levent et al., 2019 et Yu et al., 2020 attribuent une pénalité linéaire fonction de la puissance de charge et de décharge de la batterie. Au contraire, W. Liu et al., 2017 se réfère à des paramètres de dégradation établis empiriquement et pénalise la dégradation de manière non-linéaire selon ces paramètres lorsque la batterie se décharge. La récompense liée au vieillissement de la batterie est construite à partir du DoD dans les travaux de Yoldas et al., 2020 et de Chen et al., 2018. Le vieillissement en fonction de DoD est exponentiel dans les travaux de Yoldas et al., 2020 et linéaire dans les travaux de Chen et al., 2018. Shang et al., 2020 établit une pénalité de vieillissement non-linéaire selon le nombre de cycles d'utilisation de la batterie et du SOC instantané. Récemment, une modélisation thermo-électrique de la dégradation de la batterie est employée dans le système de récompense de W. Lee et al., 2022. La composante négative liée à la dégradation de la batterie dans la récompense perçue par l'agent est souvent délibérément simpliste. Toutefois, cette simplicité est suffisante pour que la récompense guide l'apprentissage de l'agent vers une politique dégradant plus faiblement la batterie.
Les contributions les plus récentes intègrent des modèles plus complexes et réalistes du vieillissement de la batterie (Fallahifar et al., 2023, Seger et al., 2023). Cela pourrait être attribué aux progrès récents sur les algorithmes de RL.

Récompense et contraintes. La singularité de chaque micro-réseau implique une grande variété de systèmes de récompense. Parfois, des récompenses négatives sont attribuées lorsque les contraintes du micro-réseau ne sont pas respectées plutôt que de restreindre les actions de l'agent. Les contraintes dépendent des solutions technologies intégrées ou des préférences d'utilisation. Une liste exhaustive de ces récompenses n'a pas d'intérêt. Les plus fréquentes sont les pénalités liées au dépassement de certaines limites de SOC fixées selon les préférences d'utilisation des batteries (Samadi et al., 2020, Fang et al., 2020, Bian et al., 2020, Kofinas et al., 2018b et Venayagamoorthy et al., 2016). Une pénalité de violation de la contrainte de puissance au PCC est couramment appliquée aux micro-réseaux connectés (Ji et al., 2021, Q. Zhang et al., 2020b, Fang et al., 2019, B. V. Mbuwir et al., 2019 et Shang et al., 2020).

3.2.3 Problèmes de planification spécifique

Bien que de nombreuses études sont entreprises avec l'intention de développer des méthodologies généralisables, le pilotage d'un micro-réseau présente ses propres particularités. La géolocalisation, les modèles de consommation, les contraintes ou les résultats espérés sont des paramètres importants qui diffèrent d'une étude à l'autre. Cette sous-section met en lumière les problèmes particuliers de planification des systèmes de stockage et d'engagement des unités qui n'ont pas encore été traités. En particulier, la gestion de la charge d'EVs et la planification d'un ensemble de systèmes de stockage comportant plusieurs technologies sont examinées.

Gestion de la charge de véhicules électriques dans un micro-réseau stationnaire. Une station de charge d'EVs est un composant dont le flux de puissance unidirectionnel ou bidirectionnel peut être contrôlable. Elle peut être vue comme une unité de consommation avec des incertitudes sur ses besoins énergétiques et leurs occurrences. C'est le cas dans les travaux de Cao et al., 2019, dont le but est de contrôler un micro-réseau dans l'optique d'augmenter les revenus liés à la vente d'électricité aux EVs. L'agent entraîné a besoin d'observer le prix de l'électricité et la demande des utilisateurs d'EV pour ajuster la planification de la batterie. Y. Liu et al., 2020 et Kim et al., 2016 considère également la charge des EVs comme une demande incertaine dans l'organisation de charges et décharges des batteries (en plus du contrôle de certaines autres demandes et de générateurs distribués). Le nombre d'arrivées d'EVs à la station ainsi que la puissance requise pour leur charge sont considérées comme des états observables par l'agent dans l'étude de Hussain et al., 2022. L'étude va jusqu'à faire une distinction entre les véhicules privés et commerciaux, ainsi que selon les périodes,

qu'il s'agisse de périodes de vacances ou de jours de travail. Une étude comparative de SAC et DDPG comme algorithme de contrôle conclut à une meilleure vitesse et stabilité d'apprentissage de SAC pour ce problème.
Lorsque le flux d'énergie alloué aux EVs est bidirectionnel, la station de charge peut être perçue comme un stockage électrique supplémentaire. Toutefois, une caractéristique distinctive de cette approche est que le véhicule doit être chargé à un certain seuil avant un moment précis, assurant ainsi qu'il est prêt pour une utilisation ultérieure ou pour répondre à d'autres contraintes du système. Fang et al., 2019 établit un système multi-agents incluant un agent responsable de la charge des EVs. Il veille à ce que les EVs quittent la station avec assez d'électricité pour atteindre leur destination. La récompense négative attribuée à cet agent inclut le coût d'opération, le coût de dégradation de la batterie stationnaire et l'anxiété de l'utilisateur c'est-à-dire comme la crainte d'épuiser son énergie sur la route. L'anxiété augmente lorsque le SOC de l'EV diminue.
Le RL multi-agent est également utilisé par Sheikhi et al., 2016. Le micro-réseau considéré ne comporte pas d'unité de production ; la décharge des EVs et l'achat d'électricité au réseau central sont les seules solutions pour alimenter les demandes. L'algorithme de contrôle doit déterminer la puissance de charge et de décharge afin d'alimenter les demandes (incluant les EVs) au prix le plus bas et en respectant les délais. Une méthodologie CTDE est employée par Z. Ye et al., 2022. La décision sur la puissance de charge est prise de manière distribuée entre les différents chargeurs. Chaque action est déterminée avec un Q-réseau entraîné sur une mémoire de relecture partagée entre les agents.

Planification des batteries comportant plusieurs technologies. Dans le cadre de leur planification, les systèmes de stockage sont souvent représentés par une capacité énergétique, une plage de puissance et un rendement. Ces paramètres définissent les contraintes de l'environnement ou des actions de l'agent en charge de leur opération. Lorsque plusieurs unités de stockage de même technologie sont intégrées au micro-réseau, la capacité augmente et leur opération est distribuée. En revanche, les unités sont soumises aux mêmes contraintes imposées par la technologie. Si au contraire plusieurs types de stockage sont présents, le pilotage pourrait gagner en flexibilité grâce à la variété des contraintes. Ainsi, l'efficacité du contrôle du micro-réseau augmente selon les objectifs considérés si le contrôle des unités est géré de manière coordonnée. Une combinaison commune est l'usage d'une batterie électrochimique comme stockage court terme et d'un système de stockage d'hydrogène (avec conversion énergétique par électrolyseur et PAC) comme stockage long terme. François-Lavet et al., 2016 contrôle ainsi la planification d'un réservoir de dihydrogène en choisissant l'utilisation d'une PAC ou d'un électrolyseur à chaque heure. Une batterie, des générateurs distribués et un stockage hydrogène sont contrôlés en temps réel par un algorithme de deep RL dans les travaux de Sidorov et al., 2020. Un système multi-agents est mis en œuvre afin de distribuer les actions de contrôle au sein d'un environnement similaire dans la contribution de Kofinas et al., 2018a. Les agents sont indépendants et apprennent avec un algorithme de Q-learning. La responsabilité du pilotage d'une unité est assignée à chaque agent. Deux agents sont nécessaires pour contrôler le stockage long terme puisqu'il est constitué par deux unités. L'approche de Hasanvand et al., 2020 adopte une perspective orientée vers les flux. L'agent pilote les flux énergétiques entre les stockages court terme et long terme, ainsi qu'entre les unités de production. Deux batteries électrochimiques différentes sont pilotées dans la contribution de X. Qiu et al., 2016. L'une est une batterie redox vanadium et l'autre une batterie au plomb. L'apprentissage de l'agent se fait avec des pénalités liées à la perte de charge des batteries utilisées, la pénalité est plus grande si une batterie non-disponible est sollicitée. Tel qu'exposé à la sous-section 3.2.2, un système de stockage est modélisé simplement dans le cadre de son contrôle séquentiel par RL. Cela peut rendre l'utilisation de plusieurs technologies superflue. Des publications récentes utilisent des modèles dynamiques précis des batteries pour leur contrôle par RL. Cela concerne Harrold et al., 2021 qui compare la performance de plusieurs algorithmes de deep RL sur le contrôle d'une batterie dans une simulation réaliste. Les puissances de charge et de décharge de la batterie sont déterminées

selon une fonction non-linéaire du SOC. L'algorithme DDPG développe une politique de contrôle meilleure que celle issue d'un apprentissage avec le DQN. L'auteur a en revanche développé un algorithme de DQN distributif qui surpasse les performances de DDPG dans l'environnement réaliste. Plus la représentation du comportement dynamique des différentes batteries sera précise, plus les différences entre les technologies seront prononcées, cela offrira plus de possibilités pour répondre à des besoins spécifiques.

3.3 Gestion de la demande

La gestion de la demande consiste à incorporer des degrés de liberté supplémentaires dans le contrôle en supervisant la demande énergétique. La consommation de certains appareils peut être déplacée vers des périodes où le système a une capacité d'approvisionnement plus importante. Les bâtiments sont équipés d'installations consommatrices d'électricité, comme le chauffage, la ventilation ou la climatisation dont la puissance d'utilisation peut être contrôlée en continu. Le développement des EVs et l'inclusion de bornes de charges dans les micro-réseaux offrent de nouvelles opportunités en terme de gestion de l'énergie. Les objectifs principaux de la gestion de micro-réseaux orientée sur la demande sont la réduction du coût énergétique, le respect de l'équilibre et le confort de l'utilisateur. Pour y parvenir, plusieurs modes d'actions sont possibles ; cette section s'articule autour des différentes pratiques de pilotage de la demande.

3.3.1 Délestage de la charge

Le délestage est la réduction de la charge interrompant la consommation d'utilisateurs ou d'appareils. C'est une mesure à adopter en cas d'urgence ou d'incapacité à rétablir l'équilibre entre la production et la consommation électrique. C'est une mesure qui concerne plutôt les micro-réseaux fonctionnant en mode îloté. Toutes les méthodologies de contrôle revues dans cette sous-section sont conçues pour cette configuration. J. Yang et al., 2022 entraîne un agent à délester des machines en cas de besoin sur une chaîne de production industrielle. La production étant effectuée par batch, l'impact du délestage d'une machine peut s'étendre, car les machines sont disposées en séquence. L'agent affine sa politique de délestage en recevant une récompense pour chaque produit fini. C. Wang et al., 2020 a développée un agent de Q-learning pour délester des consommations lorsqu'un micro-réseau a des difficultés d'équilibre. L'agent choisit la consommation à délester parmi une liste. Les consommations sont classées par ordre de priorité et la pénalité reçue par l'agent pour le délestage d'une consommation est nivelée par un facteur de priorité. Cette pénalité est plus grande si l'équilibre n'est pas rétabli en dépit de l'action de l'agent.
Nie et al., 2020 présente un système multi-agents contrôlant à la fois la génération et la consommation d'électricité. Ainsi, des récompenses liées au coût économique d'opération et des demandes non-approvisionnées sont attribuées aux agents de délestage. Les agents de production doivent choisir le point de fonctionnement des générateurs électriques afin de satisfaire les demandes avec le surplus de production le plus faible possible. Par conséquent, la récompense des agents de production est la puissance de chaque générateur multipliée par un facteur de résilience et une pénalité qui représente la violation de la contrainte. Le délestage peut être effectué avec des méthodes similaires aux enchères dans de tels systèmes multi-agents. Les agents sont associés à une unité et émettent une offre à chaque pas de temps. Les agents rattachés aux unités de génération vendent l'électricité, les agents de consommation l'achètent et les agents représentant les unités de stockage choisissent s'ils sont vendeurs ou acheteurs à chaque pas de temps. Un système multi-agents d'un micro-réseau constitué d'un centre hospitalier et de résidences est développé par Hu et al., 2020. Les agents sont liés aux différents points de consommation. Chaque agent décide de la part de la consommation totale d'électricité qui lui sera attribuée. La technique de résolution du jeu instauré entre les agents est différente selon qu'un agent est lié à une demande indispensable

(l'hôpital par exemple) ou non. Le délestage est une mesure très restrictive qui doit être adoptée en cas de nécessité dans un micro-réseau isolé. Des stratégies moins restrictives pour les utilisateurs peuvent être adoptées.

3.3.2 Déplacement de la charge

Déplacement de la puissance de charge. La possibilité de déplacer ou non la puissance de charge dépend de la nature de la demande. La décision prise ajuste l'énergie allouée à une demande sur un pas de temps ; l'appareil à l'origine de la demande doit donc être en mesure de fonctionner à une plus faible puissance. Ainsi, lorsque l'équilibre énergétique d'un micro-réseau est incertain ou si le prix de l'électricité est élevé, la puissance de consommation des appareils concernés peut être baissée. Un agent de DDPG contrôle le système de chauffage, ventilation et climatisation (CVC) d'une maison dans les travaux de Yu et al., 2020. Les informations sur le prix de l'électricité, les températures intérieure et extérieure et l'instant temporel sont présentes dans l'espace d'états de l'agent qui ajuste l'énergie allouée à la charge de la batterie et au CVC. L'objectif de ce contrôle est la réduction des coûts économiques tout en maintenant le confort des résidents. Un cas d'étude similaire est mené par Ji et al., 2021. La réduction de puissance des appareils est pénalisée par une fonction quadratique intégrant un coefficient qui reflète la sensibilité de l'utilisateur à la réduction de charge pour chaque appareil.
La confidentialité des données est un frein dans le contrôle de la consommation des appareils. Qin et al., 2021 traite ce problème dans une méthodologie de contrôle d'un générateur distribué, d'un climatiseur et d'une borne de charge d'EVs. Certaines informations ne sont intentionnellement pas observables par l'agent. Une méthode de décomposition de la récompense par action est utilisée pour palier ce problème. L'auteur aborde le problème en créant une récompense spécifique conçue pour refléter les conséquences individuelles de ses composants sur la récompense globale. Par exemple, le coût d'opération du générateur contrôlable est directement lié à ses consignes de fonctionnement, une composante de la récompense perçue est liée à ce coût. La conclusion suggère l'utilisation de systèmes multi-agents pour la résolution des problèmes de confidentialité. Une étude comparative de systèmes multi-agents et mono-agent incluant le déplacement de la puissance de charge est réalisée par Perera et al., 2021. La politique obtenue par le contrôle centralisé par un agent semble meilleure que celle du système multi-agents pour un même nombre d'itérations d'entraînement. Avec ce système multi-agents, chaque agent ajuste sa propre politique de contrôle au cours de l'apprentissage, perçue par les autres comme une modification de l'environnement. Sans mesures spécifiques, comme le partage d'informations ou la coordination assistée entre les agents, cela peut engendrer de l'instabilité dans le processus d'entraînement. Hurtado et al., 2018 propose une comparaison entre la performance d'un système multi-agents coopératif décentralisé et celle des approches de contrôle centralisé pour le contrôle de la puissance allouée à des bâtiments avec des demandes flexibles.

Déplacement temporel de charge. Le déplacement temporel d'une charge consiste à reprogrammer la plage d'une consommation de durée fixe (machine à laver ou lave-vaisselle par exemple). Cela apporte de la flexibilité au micro-réseau, mais peut nuire à la satisfaction des utilisateurs. Des travaux de recherche définissent la flexibilité d'une demande comme une fenêtre temporelle propre à chaque appareil durant laquelle la consommation doit être effectuée. Les demandes sont dites flexibles quand elles permettent d'ajuster la consommation d'énergie, facilitant ainsi sa gestion. À l'inverse, les demandes rigides requièrent une satisfaction immédiate et sans compromis. Le Q-learning est utilisé dans les travaux de Cicirelli et al., 2020 afin de contrôler les demandes flexibles. Chaque demande est caractérisée par une date limite et un temps restant d'opération. L'agent choisit le moment de démarrage de l'opération de chaque appareil. L'opération d'un appareil ne peut pas être arrêtée avant la fin de sa durée. Sheikhi et al., 2016 implémente également le Q-learning pour décaler les consommations, en plus de planifier la charge d'EVs.

Une méthodologie novatrice est établie par Mathew et al., 2020. Le contexte dans lequel l'agent de DQN évolue est un jeu de Tétris. La largeur de la fenêtre de jeu est assimilée au temps dans une journée et la hauteur à l'énergie consommée. L'agent peut donc déplacer l'énergie allouée aux demandes à gauche (plus tôt) ou à droite (plus tard). On observe que l'agent a tendance à créer des pics de consommation en heures creuses. L'auteur ajuste ce comportement en ajoutant une récompense négative affectée aux pics de consommation.
Les consignes de contrôle peuvent simultanément inclure le déplacement de la charge en terme de temps et de puissance. L'agent de Q-learning de X. Huang et al., 2021 contrôle la génération, le stockage et la consommation électrique des appareils d'une maison dans laquelle les utilisateurs spécifient le délai maximal acceptable pour chaque consommation électrique. Y. Liu et al., 2020, S. Lee et al., 2019 et F.-D. Li et al., 2012 optimisent le coût d'un micro-réseau en jouant sur les puissances et les temps de consommation. En plus de ces actions, l'agent de C. Huang et al., 2022 planifie l'opération d'une batterie. Un mixte entre DDPG et DQN a été développé dans cette étude, car l'espace d'actions prend des valeurs continues (puissances) et discrètes (instants de démarrage de l'opération des appareils). Ainsi, les progrès dans le domaine du deep RL ont permis l'utilisation d'un seul agent pour gérer efficacement la demande, même lorsqu'elle implique de nombreuses variables de nature différente.

3.3.3 Signaux de prix

La variation du prix de l'électricité a un impact sur l'opération des unités des micro-réseaux (voir section 3.2). Elle influence aussi les demandes électriques flexibles. Afin de déplacer les demandes, R. Lu et al., 2019 a construit et entraîné un réseau de neurones prédisant le prix de l'électricité. C. Zhang et al., 2019 entraîne un réseau de neurones LSTM pour prédire le prix de l'électricité et la puissance de production PV pour modifier les charges déplaçables du micro-réseau. Dans de nombreux travaux, c'est l'agent de contrôle qui adapte le prix de l'énergie pour que les demandes flexibles s'ajustent. Les travaux de Xu et al., 2020 portent sur la réduction des pics de demande grâce à des signaux de prix choisis par un agent de Q-learning. La méthodologie développée prédit la production PV à chaque pas de temps et estime le comportement des consommateurs face au prix communiqué. Kim et al., 2016 a construit un système multi-agents pour à la fois choisir les prix de l'électricité vus par le micro-réseau et adapter les demandes flexibles en fonction. L'agent de Q-learning chargé de la tarification considère la consommation totale à chaque décision. Cette consommation totale se compose de la demande non-satisfaite au pas de temps précédent, de la demande à satisfaire pour les appareils dont le fonctionnement a été activé et des nouvelles demandes.
Des signaux de prix peuvent également être transmis entre les différentes unités du micro-réseau pour un système multi-agents. De cette manière, un marché électrique interne au micro-réseau existe et permet d'atteindre différents objectifs. Les contributions Fang et al., 2020 et Fang et al., 2019 développent ainsi un système multi-agents de Q-learning. À chaque agent est associée une unité de consommation (incluant une borne de charge d'EVs) de production ou de stockage électrique. Les producteurs d'énergie choisissent un prix et une quantité d'électricité à disposition, les agents demandeurs ajustent leur demande (selon les méthodes évoquées aux sous-sections précédentes) et un agent trouve un prix d'équilibre du marché interne. Il calcule quelle quantité d'électricité chaque agent achète et vend et à quel prix. À chaque instant l'électricité en surplus ou en déficit est vendue ou achetée au réseau central à un prix défini par le problème. Le système de récompense des agents est axé sur les coûts et revenus ainsi que sur la satisfaction des demandes. Shojaeighadikolaei et al., 2020 a développé un système multi-agents basé sur le DQN pour parvenir à la réduction du coût de l'électricité d'une manière analogue. Un agent émet un signal de prix selon la demande prévue, la production et le prix de l'électricité. Les agents reçoivent le prix et contrôlent en conséquence le système de stockage qui leur est associé. Les travaux de Nakabi et al., 2021 font la distinction entre des demandes contrôlables et non contrôlables dans le système de gestion de l'énergie. La consommation thermique du bâtiment est régulée par un agent,

tandis que la demande électrique résidentielle est influencée par le prix généré. Les demandes non contrôlables sont modélisées en tant que tendance, dont une partie est sensible aux variations du prix de l'électricité. L'agent associé à l'EMS prend 4 décisions à chaque instant : il contrôle les demandes thermiques, le prix de l'électricité virtuel au sein du micro-réseau, et établit un ordre de priorité sur les unités à mettre en service en cas d'excès ou de déficit énergétique.
La réactivité d'un micro-réseau face aux signaux de prix peut être finement orchestrée par la planification stratégique de ses systèmes de stockage, générateurs et demandes ajustables. Cependant, il est souvent plus courant d'opter pour influencer la demande en jouant sur les signaux de prix plutôt que pour contrôler directement la consommation d'énergie. Cette approche est en accord avec les préoccupations relatives à la confidentialité et à l'autonomie des utilisateurs.

3.4 Échange d'énergie entre plusieurs micro-réseaux

L'échange d'énergie entre plusieurs micro-réseaux permet de réduire leur dépendance au réseau central de distribution. Si un micro-réseau est en situation de déficit énergétique, les autres micro-réseaux servent de réserve. Si un micro-réseau est en surproduction énergétique, il peut tirer profit de cette surproduction en fournissant l'énergie excédentaire à ses voisins. Des méthodes algorithmiques comme le RL facilitent l'anticipation de la situation des autres micro-réseaux pour allouer l'électricité en cas de besoin. La coordination du contrôle des micro-réseaux est déterminante pour l'autonomie de l'ensemble.

3.4.1 Commerce d'énergie de pair à pair

Le commerce d'énergie de pair à pair (P2P) est un moyen de favoriser l'échange local entre les différents acteurs. L'utilisation d'un seul agent de RL pour trouver des solutions aux problèmes d'équilibre agrégés ne peut pas être envisagée pour plusieurs raisons. La centralisation de l'information n'est pas réaliste pour des raisons de confidentialité. De plus, la réalisation de cette approche est entravée par la nécessité de prendre plusieurs décisions avec un espace d'états vaste incluant les observations de chaque micro-réseau. Pour ces raisons, les échanges P2P entre micro-réseaux avec l'utilisation du RL sont réalisés en systèmes multi-agents.

Apprenants indépendants. Les apprenants indépendants sont des agents de RL capables d'apprendre sans coordination artificiellement imposée. Dans cette configuration, les agents sont associés à un micro-réseau ou à un composant de micro-réseau et n'observent que leurs propres actions et états. Les actions des autres agents sont perçues comme une réponse dynamique de l'environnement. Chaque micro-réseau fait une offre d'échange à travers un agent commercial qui cherche à maximiser une récompense individuelle et non centralisée. Dans les travaux de Chen et al., 2019, Xiao et al., 2017, Bharadwaj et al., 2018 et X. Lu et al., 2019, la quantité d'énergie échangée est une action individuelle de chaque micro-réseau. Les agents formulent une demande ou une offre d'électricité. Lorsque les propositions concordent (si un micro-réseau compte vendre de l'électricité et qu'un micro-réseau a l'intention d'en acheter), la quantité d'énergie échangée est le minimum des deux propositions. Les déficits et excès d'électricité restants sont comblés par le réseau central. Les systèmes d'échange P2P peuvent avoir une architecture centralisée selon la typologie du regroupement de micro-réseaux. Selon Zhou et al., 2019, une production PV d'électricité sert de réserve partagée entre diverses maisons, avec ou sans stockage. Seules les maisons avec batteries utilisent un EMS intelligent basé sur le RL pour ajuster la charge en fonction des prix du marché et du

réseau central. Les échanges énergétiques sont faits soit avec la réserve, soit avec le réseau central. Des systèmes d'enchères sont également étudiés pour implémenter les interactions de P2P. Les micro-réseaux peuvent être acheteurs ou vendeurs d'électricité et émettent des offres de prix et quantités à chaque pas de temps. La manière de déterminer le prix d'échange et la quantité d'électricité échangée entre les micro-réseaux est montrée sur la figure 3.1. La quantité d'électricité échangée au sein du groupe de micro-réseaux correspond à la quantité offerte et demandée à un prix inférieur au prix d'équilibre. Les excès et manques d'électricité restants sont compensés par le réseau central de distribution. La contribution de N. Wang et al., 2019 montre l'élaboration d'une architecture modifiée de l'algorithme de Q-learning appelée Q-Cube pour trouver une politique d'enchères satisfaisante selon les objectifs.

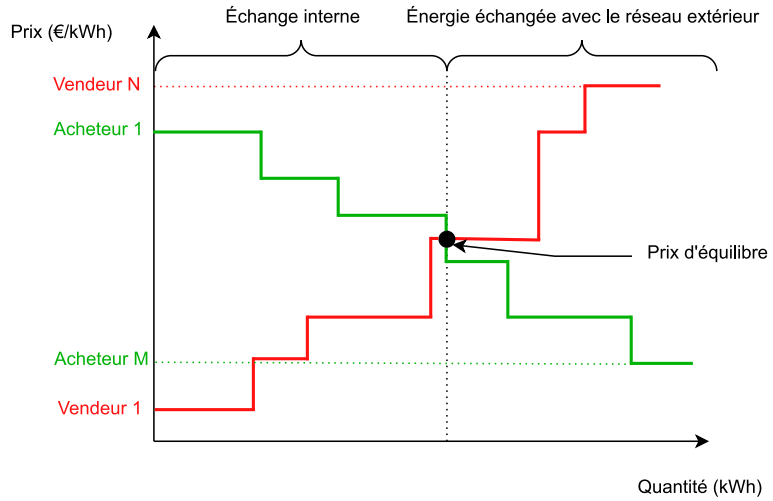

Figure 3.1 – Système d'enchères internes avec N micro-réseaux vendeurs (en rouge) et M acheteurs (en vert). Le prix et la quantité d'équilibre se lisent sur les axes à la jonction entre les courbes.

Toutefois, les apprenants indépendants sont exposés à des problèmes de convergence. Ebell et al., 2019 a comparé la dépendance au réseau central de deux micro-réseaux entre des EMS ayant appris de manière indépendante et deux EMS qui partagent des informations. L'apprentissage des EMS indépendants converge vers une politique menant à un score comparable aux EMS partageant des informations au prix d'un temps d'entraînement cinq fois supérieur. La confidentialité implique l'opacité des informations observables par les agents ce qui peut conduire à un entraînement moins stable à mesure que la politique de contrôle des autres apprenants évolue.

Systèmes multi-agents coopératifs Le partage d'information permet d'éviter des problèmes de stabilité et de convergence dans l'entraînement de plusieurs agents servant d'EMS dans un système de partage d'énergie entre plusieurs micro-réseaux. Lorsque c'est possible, l'intégration des états et actions des autres agents oriente l'apprentissage d'un agent. Y. Yang et al., 2019 intègre une variable abstraite à l'espace d'états des agents. Elle offre une vue de la dynamique des décisions collectives, et l'auteur a inclus un élément dans les récompenses pour encourager la diversité des contributions et éviter des actions uniformes.
Un algorithme d'acteur-critique hors-ligne est développé par W. Liu et al., 2017 avec prise en considération des états des agents voisins dans la mise à jour du critique. Toutes les informations de transitions sont stockées après avoir été échantillonnées puis les poids des réseaux de neurones sont ajustés de manière synchrone. La coordination des agents gagne de l'intérêt dans le commerce local de l'énergie. CTDE est utilisé par Fang et al., 2020 pour gérer un problème d'échange d'énergie dans lequel les micro-réseaux émettent des offres et

demandes ciblées les uns envers les autres. L'algorithme de MADDPG est employé dans les travaux de G. Gao et al., 2020 pour optimiser des échanges d'énergie régis par un système d'enchères. Chen et al., 2022 construit un algorithme de MATD3 (proche de MADDPG dans son fonctionnement) afin de superviser les flux électriques, thermiques et d'eau d'un hub énergétique constitué de plusieurs micro-réseaux. Les échanges se font en P2P entre des micro-réseaux identifiés comme commerciaux, industriels et résidentiels. T. Zhang et al., 2022 se sert de MADDPG la planification des batteries, de certaines demandes et de l'énergie échangée. Les buts de l'opération sont la minimisation des échanges avec le réseau central, de la dégradation des batteries, de l'inconfort des utilisateurs et des coûts liés aux transactions avec le marché local. Même si l'information n'est requise que pour l'entraînement (le critique n'est pas utilisé en exécution), le partage d'informations est nécessaire et l'algorithme ne permet donc pas une confidentialité totale entre utilisateurs.

3.4.2 Échanges d'énergie avec configuration hiérarchique

Les agents utilisés pour orchestrer l'échange d'énergie entre les micro-réseaux ne sont pas toujours au même niveau. La participation d'acteurs comme le gestionnaire du réseau de distribution (GRD) peut être envisagée dans le processus. Il peut exploiter des signaux de prix de l'électricité pour influencer le comportement d'un micro-réseau afin de préserver sa stabilité. La coopération entre les acteurs locaux du groupement de micro-réseaux peut aussi être limitée ou inexistante.

Lorsque les groupements de micro-réseaux fonctionnent de manière hiérarchique, un acteur central prend des décisions à un niveau plus haut que l'EMS de chaque micro-réseau. Ses objectifs peuvent être différents de ceux des autres acteurs (agents associés aux unités ou aux micro-réseaux). Dans l'étude menée par Qazi et al., 2020, les EMS de chaque micro-réseau communiquent avec un agent central du groupement dans un ensemble de micro-réseaux non-reliés au réseau central de distribution. Un agent de DQN coordonne la coopération des micro-réseaux, planifiant générateurs, systèmes de stockage et échanges entre micro-réseaux après avoir reçu leurs informations, visant à rendre le groupement autonome tout en minimisant les coûts économiques. Le GRD peut être exposé à des problèmes de congestion si la puissance injectée par les micro-réseaux est trop importante au PCC. B. V. Mbuwir et al., 2019 s'intéresse à la résolution des problèmes de congestion en influençant le comportement des micro-réseaux avec un système de prix. La décomposition duale (Boyd et al., 2008) est utilisée pour distribuer le prix virtuel de l'énergie sans connaître la politique de contrôle des EMS et évitant ainsi tout problème de confidentialité.
Une manière d'encourager la consommation locale d'électricité et ainsi d'éviter des pics de demande pour le GRD est d'établir une interaction compétitive entre le GRD et le groupe de micro-réseaux. Le GRD peut limiter les pics en émettant un prix local. Ce prix de l'électricité sert de prix d'échange interne au groupe de micro-réseaux, et doit être plus intéressant que le prix d'injection ou de soutirage proposé par le réseau central de distribution. Le GRD peut apprendre à anticiper la réponse du groupe de micro-réseaux aux variations des signaux de prix. Les objectifs du groupe de micro-réseaux et du GRD diffèrent, les interactions sont dites compétitives. Q. Zhang et al., 2020a a développé un algorithme bi-niveaux avec lequel le GRD apprend à fixer un prix de l'énergie sans recevoir d'informations supplémentaires des micro-réseaux, à part la puissance au PCC. Un algorithme d'optimisation sous contrainte est alors utilisé pour résoudre localement la gestion des flux énergétiques selon le prix fixé par le GRD. La méthodologie développée par Du et al., 2020 permet au GRD d'apprendre la réponse des micro-réseaux avec un algorithme RL de Monte Carlo et d'adapter le prix local de l'électricité de manière à diminuer les pics. Cette diminution et le respect de la contrainte de puissance au PCC sont garanties par un algorithme génétique. Il est développé pour fixer les prix d'échange de l'électricité. Les EMS sont entraînés avec DDPG pour intégrer ce prix dans leur politique de contrôle. La problématique de confidentialité a incité Xiong et al., 2022 à instaurer une couche supplémentaire entre le GRD et l'ensemble de micro-réseaux. Cette

couche prédit la puissance au PCC pour chaque micro-réseau dans un intervalle de prédiction de la production et de demande ainsi que du prix au détail envoyé par le GRD. L'agent du GRD est entraîné par RL avec l'algorithme de TRPO. L'objectif est la maximisation des gains de vente d'électricité aux micro-réseaux et la récompense est calculée à partir des puissances prédites aux PCCs.

3.5 Conclusion

Le RL évite le besoin de modèles prédictifs pour les données aléatoires et s'adapte à des environnements plus complexes. Sur un micro-réseau isolé du réseau central, l'autonomie est un enjeu majeur qui requiert une gestion attentive pour éviter les déséquilibres entre la demande et la production. Le pilotage des systèmes de stockage et des générateurs contrôlables est une stratégie clé pour maintenir l'équilibre et ainsi assurer l'autonomie du micro-réseau. Les environnements de contrôle deviennent plus complexes, avec des espaces d'action et d'état de plus grande dimension. L'évolution des algorithmes a conduit au développement de systèmes coordonnés multi-agents, qui se sont révélés particulièrement efficaces. L'utilisation d'algorithmes de RL profond et plus précisément d'algorithmes incorporant un gradient de politique croit de manière significative. Les modèles du stockage électrique employés sont également de plus en plus sophistiqués.
Dans l'objectif de compenser un déficit énergétique ou d'optimiser les coûts d'opération d'un micro-réseau, la demande peut être contrôlée par l'EMS. Le RL permet d'adapter les décisions de contrôle aux habitudes de consommation énergétique variables. La demande peut être délestée ou déplacée, dans le temps ou en puissance. La gestion de la demande est souvent couplée au contrôle des systèmes de stockage et de production. La coordination des différentes unités est renforcée avec un système de tarification virtuelle créant un marché énergétique local au sein du micro-réseau. Le compromis entre des objectifs antinomiques tels que le confort des utilisateurs et les coûts du micro-réseau démocratisent l'utilisation de systèmes multi-objectifs dans la gestion de la demande. L'élaboration de systèmes de récompenses ingénieux ou la mise en œuvre de systèmes multi-agents permettent de contrôler les demandes selon des objectifs contradictoires. Le contrôle centralisé reste une méthodologie largement appliquée.
Les échanges d'électricité entre les micro-réseaux contribuent à réduire la dépendance des systèmes énergétiques décentralisés au réseau central et aux générateurs contrôlables. Le contrôle de l'énergie entre plusieurs micro-réseaux peut être centralisé ou décentralisé. Chaque micro-réseau doit parvenir à optimiser ses flux énergétiques internes, en prenant en considération les actions des micro-réseaux voisins. Une stratégie bi-niveaux respecte la confidentialité des utilisateurs et facilite la coordination entre les entités. De nombreux travaux se concentrent sur la gestion ou la diminution de la complexité des espaces d'état et d'action. L'importance du temps de calcul indique un intérêt croissant pour l'application en temps réel du RL dans ce domaine.

Synthèse de la partie I

Le développement de micro-réseaux électriques vient en partie de la nécessité de décentraliser la production d'énergies renouvelables, tout en assurant la stabilité des réseaux de distribution. Les micro-réseaux doivent être autonomes car ils peuvent être amenés à fonctionner en mode îloté. Leur dimensionnement optimal est nécessaire pour garantir leur autonomie sans engendrer des coûts excessifs. À cause des fluctuations dans la production d'énergie renouvelable et des incertitudes sur la demande électrique, la qualité du contrôle exercé par un outil de gestion de l'énergie (EMS) est cruciale. Ce contrôle s'effectue avec un algorithme basé sur des règles, de l'optimisation hors-ligne, de la commande prédictive ou avec une approche basée sur un agent. L'apprentissage par renforcement est une méthode stochastique adaptative, idéale pour gérer les fluctuations de la production d'énergie renouvelable et la demande incertaine.
Un algorithme d'apprentissage par renforcement est basé sur un agent et est adapté aux prises de décisions séquentielles. La politique développée est apprise par essais et erreurs grâce à des systèmes de récompenses qui permettent d'associer des valeurs à des états et des prises de décisions dans un environnement. Aucune connaissance préalable de probabilité de transition d'un état à un autre n'est requise. Ceci fait de l'apprentissage par renforcement une classe d'algorithmes particulièrement adaptée aux données aléatoires. L'intégration de réseaux de neurones profonds aux algorithmes d'apprentissage par renforcement a rendu la prise de multiples décisions possible dans des environnements complexes. Le fonctionnement d'un micro-réseau combine modèles dynamiques des unités et éléments aléatoires. L'apprentissage par renforcement est donc adapté à la résolution des problèmes de contrôle de ces systèmes.
L'application de l'apprentissage par renforcement au pilotage des micro-réseaux est récente et pourtant très répandue. Le domaine du contrôle, qu'il s'agisse d'unités pilotables, de gestion de la demande ou d'échanges entre micro-réseaux, dépend des possibilités offertes par la typologie du micro-réseau. Les algorithmes permettent de déplacer la demande d'énergie en fonction des habitudes de consommation et des coûts, tout en étant capable de coordonner diverses unités au sein du micro-réseau et de planifier les échanges énergétiques entre plusieurs micro-réseaux. Les consignes de fonctionnement sont centralisées ou distribuées avec un système multi-agents selon la nature du problème à résoudre. Le but est de parvenir à un équilibre énergétique, tout en prenant en compte des facteurs tels que le confort des utilisateurs et la réduction des coûts économiques.
Après avoir examiné les catégories de contrôle, les objectifs de dimensionnement et exploré les bases théoriques de l'apprentissage par renforcement, nous avons mené une analyse de son application au contrôle des micro-réseaux. Dans le cadre de cette thèse, un outil méthodologique a été développé pour dimensionner et piloter un micro-réseau. Comment peut-on appliquer l'apprentissage par renforcement au pilotage du micro-réseau en vue de son dimensionnement optimal ? Lors de la phase de dimensionnement, les préférences en temps réel des utilisateurs pour la demande ne sont pas connues. Le dimensionnement et le contrôle

ne portent que sur un unique micro-réseau sans prendre en considération la possibilité d'être raccordé à d'autres systèmes autonomes. C'est pour cette raison que l'outil se concentre principalement sur la planification du stockage d'énergie du micro-réseau. L'approche retenue pour le contrôle est centralisée, car seule la gestion du stockage électrique est concernée par les consignes de l'EMS. La formulation du problème de contrôle et de dimensionnement, ainsi que les hypothèses et la modélisation adoptées pour le micro-réseau sont présentées au Chapitre 4.

II

Contributions scientifiques

4

Méthodologie et modélisation

Un outil méthodologique dont le but est de dimensionner et contrôler un micro-réseau est développé tout au long de ce travail de thèse. Dans ce contexte, le dimensionnement d'un micro-réseau consiste à choisir les caractéristiques nominales des unités qui le composent dans le but de minimiser les coûts sur toute sa durée de vie. La minimisation des coûts du micro-réseau concerne à la fois l'optimisation des coûts d'investissements et d'opération. Les coûts d'opération dépendent de la stratégie de contrôle en temps réel adoptée : une stratégie efficace minimise l'utilisation de sources d'énergie coûteuses et optimise celle du stockage pour équilibrer l'offre et la demande. Certaines contraintes d'opération du micro-réseau sont déterminées par son dimensionnement. La phase de fonctionnement du micro-réseau est donc à prendre en compte dans la phase de dimensionnement.

La simulation des opérations d'un micro-réseau permet de calculer son coût pour un dimensionnement donné. La modélisation des unités doit être modulable dans cette simulation afin que la méthodologie développée soit adaptable à diverses typologies de micro-réseaux. Le dimensionnement revient à choisir la valeur des caractéristiques des modules qui constituent le micro-réseau : les unités sont classées par fonction (production renouvelable, stockage court terme ou long terme, consommation...) formant ainsi des modules. Leurs caractéristiques sont en partie conditionnées par les décisions de dimensionnement et sont utilisées pour fixer les contraintes au problème de contrôle. Les modules dépendent aussi de facteurs non-dimensionnables tels que les données aléatoires et des contraintes techniques spécifiques au contexte d'application. L'approche modulaire permet à un utilisateur d'intégrer les spécificités de son installation.

Ce chapitre présente la méthodologie élaborée afin de coupler le contrôle (section 4.2) et le dimensionnement (section 4.3) du micro-réseau, ainsi que la modélisation retenue pour simuler le comportement dynamique du micro-réseau (section 4.1).

4.1 Modélisation du micro-réseau

La modélisation des unités dans les problèmes de contrôle et de dimensionnement du micro-réseau est présentée à la sous-section 4.1.1. La sous-section 4.1.2 établi les contraintes propres à chacune des unités. Enfin, la sous-section 4.1.3 présente les modèles des unités du micro-réseau en phase d'opération.

4.1.1 Organisation modulaire du problème

L'outil méthodologique présenté à travers cette thèse est un outil de dimensionnement et de contrôle de micro-réseaux dont la production d'électricité est d'origine renouvelable. Le dimensionnement d'un micro-réseau et le choix de la stratégie de contrôle répondent à des objectifs préalablement définis.

Le micro-réseau type étudié dans cette thèse se compose d'une unité de production solaire PV, d'un électrolyseur haute pression, d'une PAC ainsi qu'un réservoir de dihydrogène (H_2). Ces unités sont présentées sur la Figure 4.1. Les modules paramétrables sont le stockage électrique long terme, court terme et la production d'électricité renouvelable. L'EMS constitue aussi un module puisque selon les critères de contrôles choisis, le micro-réseau ne sera pas piloté de la même manière. La consommation est un module puisque les données dépendent de l'utilisateur, en revanche, elle n'est pas paramétrable lors du dimensionnement.

Le stockage long terme se compose de deux modules, l'électrolyseur et la PAC dont les puissances (respectivement $P^{\text{elec}}_{\text{max}}$ et $P^{\text{PAC}}_{\text{max}}$) peuvent-être choisies. L'électrolyseur utilisé est de type à membrane échangeuse de proton (PEM). Un compresseur est nécessaire au stockage à haute pression. Cela a pour effet l'augmentation de la densité énergétique du stockage. Aucune caractéristique du compresseur n'est une variable de dimensionnement puisque ce changement de pression est nécessaire au stockage stationnaire d'hydrogène quelle que soit la puissance de l'électrolyseur. La PAC est équipée d'une technologie PEM. La taille du réservoir H_2 est fixe pour des raisons d'espace, il occupe un volume de $0{,}5\,\text{m}^3$. Son rendement est pris en compte dans le rendement de l'électrolyseur. Pour une utilisation stationnaire, le gaz est stocké à une pression de 200 bar (Hancke et al., 2022) à une température de 15 °C.

Le stockage électrique court terme est la batterie électrochimique dont la capacité peut être choisie. La batterie est équipée d'une technologie lithium fer-phosphate ($LiFePO_4$). Elle fonctionne à $1C$ en puissance maximale, c'est-à-dire qu'elle peut se charger ou se décharger entièrement en une heure.

Le module de génération d'électricité d'origine renouvelable correspond aux panneaux solaires PV dont la puissance crête est une variable de dimensionnement. Les panneaux PV sont installés à Albi (France) et génèrent de l'énergie en fonction de données de productible PV.

Les données de consommation, d'ensoleillement ou le volume du réservoir d'hydrogène sont invariables dans le problème présenté. La consommation considérée est de type résidentielle. Les caractéristiques invariables des modules constituent la base du problème et sont à fixer avant la résolution par la méthodologie de dimensionnement proposée. Les modules ont donc des caractéristiques variables et invariables qui déterminent les conditions de la simulation de l'opération du micro-réseau. Les espaces des caractéristiques variables et invariables de l'ensemble des modules dans un problème de dimensionnement sont respectivement notés $\mathcal{D}$ et $\mathcal{N}$. Les caractéristiques variables et invariables des différents modules sont présentées sur la Figure 4.1.

Les variables en rouge sur la Figure 4.1 sont celles qui seront systématiquement optimisées lors du dimensionnement du micro-réseau. Certains modules comme la consommation des utilisateurs et l'EMS n'ont aucune caractéristique ajustable en phase de dimensionnement. La consommation des utilisateurs est fixe, car elle dépend d'un historique de données. L'EMS

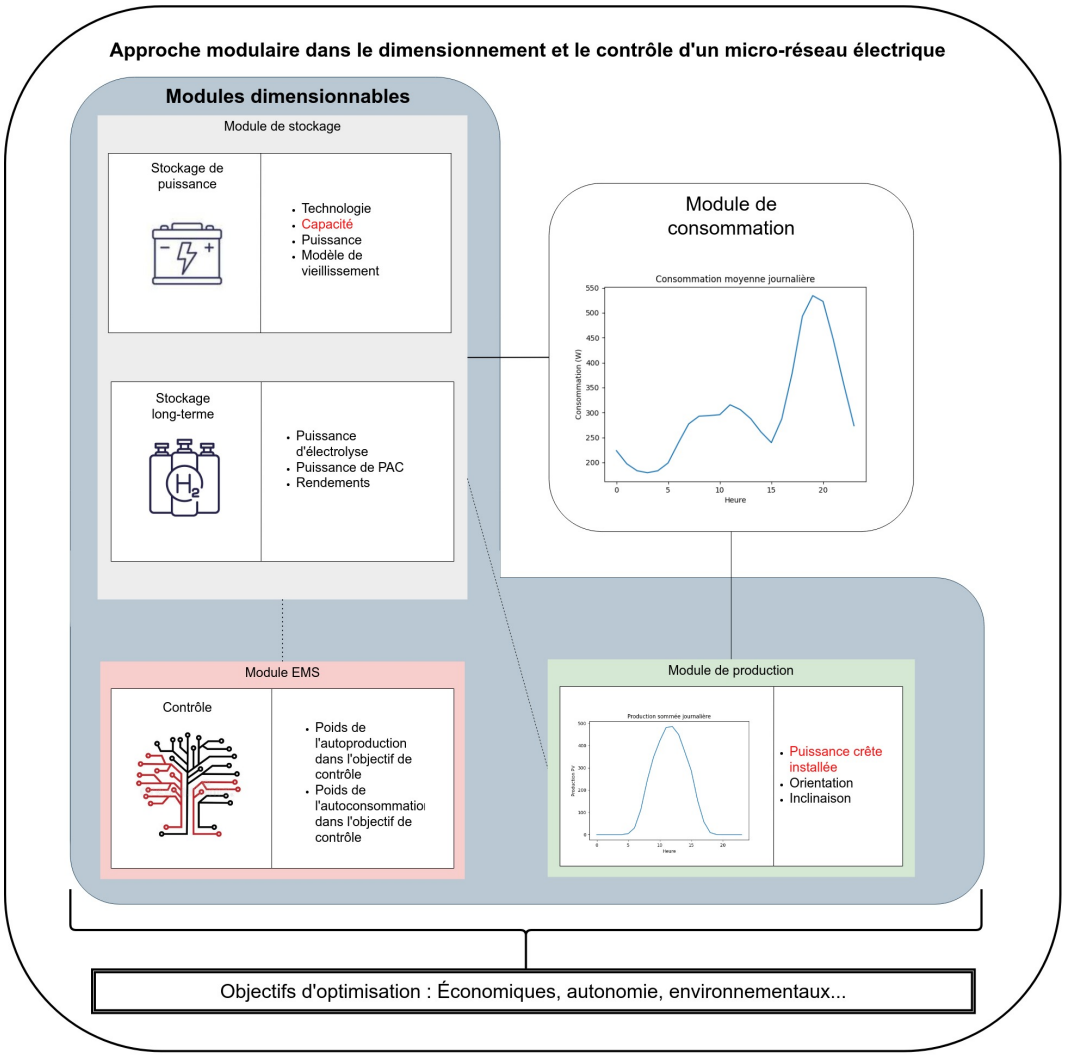

Figure 4.1 – Représentation de l'approche modulaire du dimensionnement d'un micro-réseau, les variables de dimensionnement retenues pour l'étude sont en rouge

fixe les objectifs de contrôle du micro-réseau. Ils peuvent être envisagés de diverses manières. Un micro-réseau peut répondre à des objectifs techniques (garantir l'approvisionnement énergétique des utilisateurs, bien répartir l'énergie entre usagers), environnementaux (limiter les émissions de gaz à effet de serre) et économiques (rentabiliser les échanges avec le réseau principal). La définition de ces objectifs déterminera la politique de contrôle en phase d'opération du micro-réseau. Les objectifs optimisés en phase de fonctionnement sont indépendants des objectifs de dimensionnement. La stratégie de contrôle est optimisée en simulation, avec un horizon temporel de l'ordre de l'année et un pas de temps d'une heure entre chaque décision de contrôle.

La modélisation des unités dans la simulation de la phase d'opération du micro-réseau intègre les contraintes de chaque module. La sous-section 4.1.2 définit les contraintes résultantes pour le micro-réseau type étudié. La sous-section 4.1.3 présente le comportement dynamique des composants du micro-réseau lors de la simulation.

4.1.2 Fonctionnement du micro-réseau

La dynamique de réponse du micro-réseau est soumise à des contraintes. Ces contraintes sont propres à chaque unité et peuvent dépendre des variables de dimensionnement (ou autres) d'une unité.

Contraintes sur les panneaux solaires PV

La puissance produite par les panneaux PV dépend de données météorologiques. Ces données sont conditionnées par la localisation des panneaux, leur inclinaison et leur puissance nominale choisie en phase de dimensionnement. En particulier, l'Équation 4.1 montre que la puissance produite ne peut pas dépasser la puissance nominale.

$$\forall t \in [\![0; T]\!],\ 0 \leq P^{PV}(t) \leq P^{PV}_{nom} \tag{4.1}$$

avec P^{PV} la puissance des panneaux PV et P^{PV}_{nom} la puissance nominale des panneaux PV (en W)

Contraintes sur la batterie

Les contraintes de la batterie portent principalement sur l'énergie stockée et la puissance à laquelle elle se charge et décharge. L'Équation 4.2 montre que le SOC est compris entre des valeurs maximales et minimales à chaque instant.

$$\forall t \in [\![0; T]\!],\ SOC_{min} \leq SOC(t) \leq SOC_{max} \tag{4.2}$$

avec $SOC_{min} \geq 0$ et $SOC_{max} \leq 1$. Il augmente lorsque la batterie se recharge et diminue si elle se décharge. L'Équation 4.3 montre que la charge et la décharge de la batterie ne peuvent pas s'effectuer en même temps :

$$\forall t \in [\![0; T]\!],\ 0 \leq \delta^{batt}_{charg}(t) + \delta^{batt}_{dech}(t) \leq 1 \tag{4.3}$$

avec δ^{batt}_{charg} un booléen valant 1 si la batterie se charge et 0 sinon et δ^{batt}_{dech} un booléen valant 1 si la batterie est en décharge et 0 sinon. Les puissances de charge et de décharge sont bornées puisque la batterie se charge et décharge à 1C au maximum (voir sous-section 4.1.1). La valeur de ces puissances se calcule selon l'Équation 4.4 :

$$P^{batt}_{max} = \frac{(SOC_{max} - SOC_{min})\, E^{batt}_{nom}}{\Delta t} \tag{4.4}$$

avec P_{max}^{batt} la puissance maximale de charge ou de décharge selon le mode d'opération de la batterie et Δt le pas de temps. Ainsi, la puissance de la batterie est contrainte à chaque instant selon l'Équation 4.5 :

$$\forall t \in [\![0; T]\!],\ 0 \leq P_j^{batt}(t) \leq P_{max}^{batt}\delta_j(t) \tag{4.5}$$

avec j le mode de fonctionnement de la batterie (j= charg ou dech).

Contraintes sur le stockage d'hydrogène

Le volume du stockage hydrogène est fixé à 0,5 m^3. La pression de stockage est de 200 bar. L'hypothèse d'une température de stockage de 15 °C est formulée. En utilisant la loi des gaz parfaits, la masse d'hydrogène stockable vaut $m_{max}^{H_2} = 8{,}42\,kg$. L'utilisation de la loi des gaz parfaits est une approximation, elle n'est en pratique pas utilisable à 200 bars et la compressibilité de H_2 devrait être considérée. L'électrolyseur a un rendement de conversion d'électricité en masse d'hydrogène de $\eta^{elec} = 61{,}6\,kWh/kg^{H_2}$ (Hancke et al., 2022), soit $1030\,kWh/m^{3H_2}$. La PAC a un rendement de conversion d'électricité en masse d'hydrogène de $\eta^{PAC} = 16{,}66\,kWh/kg^{H_2}$ (Dufo-López et al., 2007), soit $280\,kWh/m^{3H_2}$. Le pourcentage d'hydrogène stocké par rapport à la capacité de stockage est appelé niveau d'hydrogène ou *Level of Hydrogen* (LOH). Des contraintes s'appliquent sur cette variable à chaque pas de temps comme le montre l'Équation 4.6.

$$\forall t \in [\![0; T]\!],\ 0 \leq LOH(t) \leq 1 \tag{4.6}$$

La PAC et l'électrolyseur sont toujours utilisés à puissance nominale sauf si la quantité d'énergie stockée ne le permet pas. Les contraintes liées à l'utilisation de l'électrolyseur sont explicitées dans les Équations 4.7 et 4.8 :

$$\forall t \in [\![0; T]\!],\ P_{max}^{elec} = \min\left(P_{nom}^{elec}, \frac{\left(m_{max}^{H_2} - m^{H_2}(t)\right)\eta^{elec}}{\Delta t}\right) \tag{4.7}$$

$$\forall t \in [\![0; T]\!],\ P^{elec}(t) = \left\{0, P_{max}^{elec}\right\} \tag{4.8}$$

avec $m^{H_2}(t)$ la masse d'hydrogène stockée à l'instant t. Si le réservoir d'hydrogène ne dispose pas du volume libre nécessaire pour accumuler l'hydrogène formé par électrolyse, la puissance d'électrolyse correspond à la puissance nécessaire pour le remplir totalement. De même, la PAC ne peut pas être utilisée à puissance maximale si trop peu d'hydrogène est stocké. Les Équations 4.9 et 4.10 montrent la contrainte encadrant la puissance de la PAC.

$$\forall t \in [\![0; T]\!],\ P_{max}^{PAC} = \min\left(P_{nom}^{PAC}, \frac{\left(m^{H_2}(t) - m_{min}^{H_2}\right)\eta^{PAC}}{\Delta t}\right) \tag{4.9}$$

$$\forall t \in [\![0; T]\!],\ P^{PAC}(t) = \left\{0, P_{max}^{PAC}\right\} \tag{4.10}$$

L'électrolyseur et la PAC ne peuvent pas être utilisées en même temps comme le montre l'Équation 4.11.

$$\forall t \in [\![0; T]\!],\ 0 \leq \delta_{elec}(t) + \delta_{PAC}(t) \leq 1 \tag{4.11}$$

avec δ_{elec} et δ_{PAC} des booléens associés respectivement à l'électrolyseur et à la PAC, dont la valeur est de 1 si l'unité est en cours de fonctionnement et 0 sinon. La puissance sortant de l'ensemble du stockage hydrogène P^{H_2} s'écrit selon l'Équation 4.12.

$$P^{H_2}(t) = P^{PAC}(t)\delta_{PAC}(t) - P^{elec}(t)\delta_{elec}(t) \tag{4.12}$$

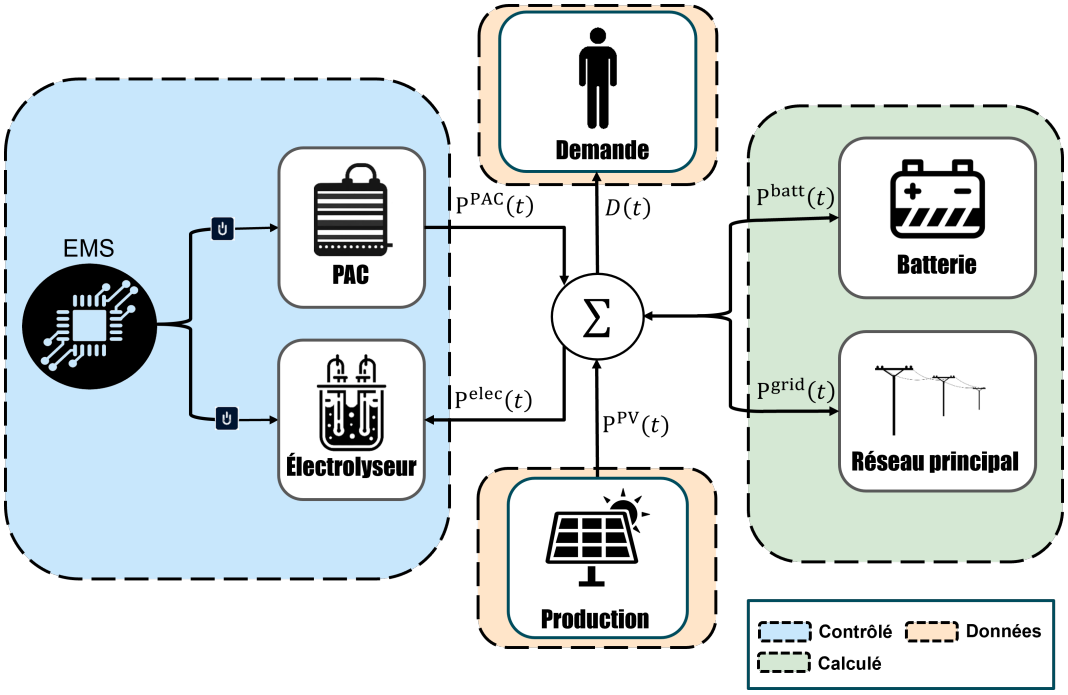

Figure 4.2 – Schéma de principe des différentes puissances à déterminer à chaque instant dans le micro-réseau considéré selon les unités.

4.1.3 Modèle dynamique

Dans le micro-réseau modélisé, l'EMS contrôle uniquement l'électrolyseur et la PAC. Il calcule des signaux de commande de manière séquentielle qui prennent la forme d'interrupteurs (*on*/*off*). La simulation de l'environnement intègre ces commandes ainsi que les données utilisées. Elle détermine les variables et permet d'observer divers indicateurs sur la situation du micro-réseau à chaque pas de temps. Une représentation en schéma de principe des flux de puissance en fonction des unités du micro-réseau considéré est présentée sur la Figure 4.2. La puissance de fonctionnement de l'électrolyseur $P^{elec}(t)$ et de la PAC $P^{PAC}(t)$ dépendent directement de la consigne de l'EMS. Ce sont des flux contrôlés. Les puissances en sortie des panneaux PV $P^{PV}(t)$ et les demandes $D(t)$ sont issues des données historiques à disposition. Enfin, les valeurs des puissances de charge ou de décharge de la batterie $P^{batt}(t)$ ainsi que des puissances injectées ou soutirées du réseau central $P^{grid}(t)$ sont issues de calculs.
Selon les commandes choisies, un déséquilibre énergétique dans le micro-réseau doit être comblé afin que la production et la demande soient égales. Pour cela, un ordre de priorité est établi entre les unités pour compenser le déséquilibre. La puissance du stockage long terme est calculée par l'EMS. La batterie est passive, elle n'est pas contrôlée par l'EMS et se charge ou se décharge selon la nature du déséquilibre. La batterie stocke l'électricité restant en excès ou fournit l'électricité manquante après la prise en compte de la charge ou de la décharge du stockage hydrogène. Enfin, si la batterie ne peut pas assurer l'équilibre énergétique du micro-réseau, l'électricité est injectée ou soutirée au réseau central de distribution. La quantité d'énergie injectée ou soutirée se détermine grâce à l'Équation 4.13 :

$$E^{grid} = \left| D(t) - P^{H_2}(t)\Delta t - P^{PV}(t)\Delta t + P^{batt}_{charg}\Delta t - P^{batt}_{dech}\Delta t \right| \tag{4.13}$$

avec E^{grid} l'énergie injectée ou soutirée au réseau central de distribution, $D(t)$ la demande à chaque instant, P^{batt}_{charg} et P^{batt}_{dech} les puissance de charge et de décharge de la batterie respectivement. Si E^{grid} est positif, la demande électrique n'est pas totalement approvisionnée,

l'énergie correspondante est soutirée du réseau central. Au contraire, si elle est négative, l'électricité en surplus est injectée au réseau central.

Électrolyseur et PAC

Le signal envoyé par l'EMS indique la puissance de l'électrolyseur et de la pile à combustible entre l'instant t et t+Δt. Il est reçu par l'environnement comme un signal d'interrupteur indiquant l'utilisation ou non de l'électrolyseur ou de la PAC. Une seule des deux unités peut être utilisée à chaque pas de temps, comme l'illustre l'Équation 4.11. La valeur de la puissance d'opération de l'électrolyseur ou de la PAC est déterminée grâce aux Équations 4.7 à 4.10. Une fois ces puissances déterminées, le LOH est calculé pour l'instant t+Δt avec l'Équation 4.14 :

$$\mathrm{LOH}\,(\mathrm{t}+\Delta\mathrm{t}) = \mathrm{LOH}(\mathrm{t}) + \frac{\dot{\mathrm{m}}^{\mathrm{elec}}(\mathrm{t})\Delta\mathrm{t} - \dot{\mathrm{m}}^{\mathrm{PAC}}(\mathrm{t})\Delta\mathrm{t}}{\mathrm{m}^{\mathrm{H_2}}_{\mathrm{max}}} \tag{4.14}$$

avec $\dot{\mathrm{m}}^{\mathrm{elec}}(\mathrm{t})$ et $\dot{\mathrm{m}}^{\mathrm{PAC}}(\mathrm{t})$ respectivement le débit massique d'hydrogène entrant et sortant du réservoir.

Demande nette et batterie

Le rôle du stockage long terme est déterminé à chaque instant par la commande de l'EMS. Si elle consiste à utiliser la PAC, c'est une unité productrice d'électricité. Si l'électrolyseur est utilisé, son rôle est celui d'un consommateur d'électricité. L'énergie fournie ou demandée par le stockage long terme, les données de productible PV et la demande électrique permettent de déterminer la demande nette $\mathrm{D}^{\mathrm{net}}$ à chaque instant selon l'Équation 4.15 :

$$\mathrm{D}^{\mathrm{net}}(\mathrm{t}) = \mathrm{D}(\mathrm{t}) - \mathrm{P}^{\mathrm{H_2}}(\mathrm{t})\Delta\mathrm{t} - \mathrm{P}^{\mathrm{PV}}(\mathrm{t})\Delta\mathrm{t} \tag{4.15}$$

avec Δt le pas de temps considéré. Ici, il vaut 1 h. Le micro-réseau est en surproduction ou surconsommation selon le signe de $\mathrm{D}^{\mathrm{net}}$. Si $\mathrm{D}^{\mathrm{net}} > 0$, alors la demande est plus importante que la production et inversement si $\mathrm{D}^{\mathrm{net}} < 0$. Pour plus de clarté dans les paragraphes suivants, la notation de capacité instantanée $\mathrm{E}^{\mathrm{batt}}(\mathrm{t})$, capacité maximale $\mathrm{E}^{\mathrm{batt}}_{\mathrm{max}}$ et capacité minimale $\mathrm{E}^{\mathrm{batt}}_{\mathrm{min}}$ de la batterie est adoptée. Ces valeurs s'expriment respectivement selon les Équations 4.16, 4.17 et 4.18.

$$\mathrm{E}^{\mathrm{batt}}(\mathrm{t}) = \mathrm{SOC}(\mathrm{t})\mathrm{E}^{\mathrm{batt}}_{\mathrm{nom}} \tag{4.16}$$

$$\mathrm{E}^{\mathrm{batt}}_{\mathrm{max}} = \mathrm{SOC}_{\mathrm{max}}\mathrm{E}^{\mathrm{batt}}_{\mathrm{nom}} \tag{4.17}$$

$$\mathrm{E}^{\mathrm{batt}}_{\mathrm{min}} = \mathrm{SOC}_{\mathrm{min}}\mathrm{E}^{\mathrm{batt}}_{\mathrm{nom}} \tag{4.18}$$

Décharge de la batterie

Si l'énergie stockée dans la batterie électrochimique est suffisante, la batterie peut délivrer l'énergie nécessaire pour l'équilibre du système. Cette condition se vérifie en considérant le rendement de la batterie d'après l'Équation 4.19 :

$$\mathrm{P}^{\mathrm{batt}}_{\mathrm{dech}}(\mathrm{t})\Delta\mathrm{t} = \mathrm{D}^{\mathrm{net}}(\mathrm{t}) \text{ si } \mathrm{D}^{\mathrm{net}}(\mathrm{t}) \leq \left(\mathrm{E}^{\mathrm{batt}}(\mathrm{t}) - \mathrm{E}^{\mathrm{batt}}_{\mathrm{min}}\right)\mu^{batt} \tag{4.19}$$

avec μ^{batt} le rendement de la batterie. Dans ce cas, le SOC est déterminé en considérant la décharge effectuée et le rendement de la batterie selon l'Équation 4.20.

$$\mathrm{SOC}\,(\mathrm{t}+\Delta\mathrm{t}) = \frac{\left(\mathrm{E}^{\mathrm{batt}}(\mathrm{t}) - \frac{\mathrm{D}^{\mathrm{net}}(\mathrm{t})}{\mu^{\mathrm{batt}}}\right)}{\mathrm{E}^{\mathrm{batt}}_{\mathrm{nom}}} \tag{4.20}$$

L'énergie reçue par le micro-réseau est toujours inférieure à l'énergie déchargée. Si la demande nette est compensée par la décharge, alors l'énergie perdue par la batterie est $\frac{\mathrm{D^{net}(t)}}{\mu^{\mathrm{batt}}}$.

Si l'énergie stockée dans la batterie n'est pas suffisante, la batterie se décharge à puissance maximale pendant l'intervalle de temps et le manque d'électricité est soutiré au réseau central. Dans ce cas, la décharge de la batterie est limitée par le SOC à l'instant t. La valeur de l'énergie déchargée par la batterie s'exprime alors selon l'Équation 4.21 :

$$\mathrm{P^{batt}_{dech}(t)\Delta t} = \left(\mathrm{E^{batt}(t) - E^{batt}_{min}}\right)\mu^{\mathrm{batt}} \text{ si } \mathrm{D^{net}(t)} > \left(\mathrm{E^{batt}(t) - E^{batt}_{min}}\right)\mu^{batt} \tag{4.21}$$

Puisque la batterie atteint sa valeur minimale d'énergie stockée, la détermination du SOC après sa décharge est immédiate d'après l'Équation 4.22.

$$\mathrm{SOC\,(t + \Delta t) = SOC(t) - \frac{E^{batt}(t) - E^{batt}_{min}}{E^{batt}_{nom}} = SOC_{min}} \tag{4.22}$$

L'électricité soutirée au réseau central de distribution est déterminée à l'aide de l'Équation 4.23 :

$$\mathrm{E^{grid}_{sout}(t) = P^{batt}_{dech}(t)\Delta t - D^{net}(t)} \tag{4.23}$$

En résumé, l'Équation 4.24 montre le fonctionnement de la batterie en décharge selon sa capacité et la demande nette.

$$\mathrm{P^{batt}_{dech}(t)\Delta t} = \begin{cases} \mathrm{D^{net}} & \text{si } \mathrm{D^{net}(t)} \leq \left(\mathrm{E^{batt}(t) - E^{batt}_{min}}\right)\mu^{batt} \\ \left(\mathrm{E^{batt}(t) - E^{batt}_{min}}\right)\mu^{\mathrm{batt}} & \text{sinon} \end{cases} \tag{4.24}$$

L'Équation 4.25 décrit le calcul du SOC à l'instant suivant la décharge selon les cas :

$$\mathrm{SOC\,(t + \Delta t)} = max\left(\mathrm{SOC_{min}}, \frac{\mathrm{E^{batt}(t) - D^{net}(t)}}{\mathrm{E^{batt}_{nom}}}\right) \tag{4.25}$$

Charge de la batterie

Si $\mathrm{D^{net}} < 0$, le surplus d'électricité doit être absorbé par la charge de la batterie. Puisque la batterie est contrainte par un niveau de charge maximal, elle ne peut pas toujours absorber ce surplus d'énergie. Dans ce cas, l'énergie restante est injectée au réseau central ou perdue.

Si la capacité libre de la batterie est suffisante pour absorber le surplus énergétique, la demande nette est supérieure à l'écart entre l'énergie stockée et l'énergie maximale stockable à l'instant t. L'énergie utilisée pour la charger s'exprime selon l'Équation 4.26 :

$$\mathrm{P^{batt}_{charg}(t)\Delta t} = -\mathrm{D^{net}(t)} \text{ si } \mathrm{D^{net}(t)} \geq \frac{\left(\mathrm{E^{batt}(t) - E^{batt}_{max}}\right)}{\mu^{batt}} \tag{4.26}$$

La condition de capacité libre inclut le rendement, car il limite l'énergie effectivement stockée par rapport à l'électricité absorbée. Conformément à l'Équation 4.27, le SOC se calcule de manière linéaire en multipliant l'énergie chargée par l'efficacité μ^{batt}.

$$\mathrm{SOC\,(t + \Delta t) = SOC(t) - \frac{D^{net}(t)\mu^{batt}}{E^{batt}_{nom}}} \tag{4.27}$$

Une partie de l'électricité absorbée est perdue puisque le rendement de charge est pris en compte dans le calcul du SOC.

Si la capacité libre de la batterie est insuffisante, la batterie se recharge jusqu'à atteindre sa valeur maximale d'énergie stockée. Les conditions de cette charge sont déterminées à l'aide de l'Équation 4.28.

$$\mathrm{P}^{\mathrm{batt}}_{\mathrm{charg}}(\mathrm{t})\Delta\mathrm{t} = \frac{\mathrm{E}^{\mathrm{batt}}_{\mathrm{max}} - \mathrm{E}^{\mathrm{batt}}(\mathrm{t})}{\mu^{\mathrm{batt}}} \text{ si } \mathrm{D}^{\mathrm{net}}(\mathrm{t}) < \frac{\left(\mathrm{E}^{\mathrm{batt}}(\mathrm{t}) - \mathrm{E}^{\mathrm{batt}}_{\mathrm{max}}\right)}{\mu^{batt}} \tag{4.28}$$

Le rendement est de nouveau situé au dénominateur de l'expression de manière à atteindre une énergie stockée de $\mathrm{E}^{\mathrm{batt}}_{\mathrm{max}}$. Le SOC s'exprime donc selon l'Équation 4.29.

$$\mathrm{SOC}\,(\mathrm{t}+\Delta\mathrm{t}) = \mathrm{SOC}(\mathrm{t}) + \frac{\left(\mathrm{E}^{\mathrm{batt}}_{\mathrm{max}} - \mathrm{E}^{\mathrm{batt}}(\mathrm{t})\right)}{\mathrm{E}^{\mathrm{batt}}_{\mathrm{nom}}} = \mathrm{SOC}_{\mathrm{max}} \tag{4.29}$$

L'énergie injectée au réseau central de distribution peut être déterminée à l'aide de l'Équation 4.30.

$$\mathrm{E}^{\mathrm{grid}}_{\mathrm{inje}}(\mathrm{t}) = \mathrm{P}^{\mathrm{batt}}_{\mathrm{charg}}(\mathrm{t})\Delta\mathrm{t} + \mathrm{D}^{\mathrm{net}}(\mathrm{t}) \tag{4.30}$$

L'Équation 4.31 synthétise la dynamique de charge de la batterie électrochimique selon les conditions sur le stockage libre.

$$\mathrm{P}^{\mathrm{batt}}_{\mathrm{charg}}(\mathrm{t})\Delta\mathrm{t} = \begin{cases} -\mathrm{D}^{\mathrm{net}} & \text{si } \mathrm{D}^{\mathrm{net}}(\mathrm{t}) \geq \frac{\left(\mathrm{E}^{\mathrm{batt}}(\mathrm{t}) - \mathrm{E}^{\mathrm{batt}}_{\mathrm{max}}\right)}{\mu^{batt}} \\ \frac{\mathrm{E}^{\mathrm{batt}}_{\mathrm{max}} - \mathrm{E}^{\mathrm{batt}}(\mathrm{t})}{\mu^{\mathrm{batt}}} & \text{sinon} \end{cases} \tag{4.31}$$

Conformément à l'Équation 4.32, le SOC peut se calculer :

$$\mathrm{SOC}\,(\mathrm{t}+\Delta\mathrm{t}) = min\left(\mathrm{SOC}_{\mathrm{max}}, \mathrm{SOC}(\mathrm{t}) - \frac{\mathrm{D}^{\mathrm{net}}(t)\mu^{\mathrm{batt}}}{\mathrm{E}^{batt}_{\mathrm{nom}}}\right) \tag{4.32}$$

4.2 Méthodologie de contrôle

Le contrôle du micro-réseau consiste en l'élaboration d'une stratégie pour optimiser les objectifs associés à l'EMS. Cette stratégie est développée à partir de simulations du micro-réseau. Elle est ensuite testée sur une simulation à horizon temporel fixe et les coûts d'opérations sont calculés. Les objectifs du contrôle et dépendent de chaque micro-réseau, voire de chaque utilisateur. L'apprentissage par renforcement sera utilisé pour obtenir une politique de contrôle adaptée aux variables aléatoires. L'application de l'apprentissage par renforcement comme outil de résolution d'un problème séquentiel nécessite d'interagir avec un environnement réel ou d'utiliser une simulation réaliste de cet environnement. La sous-section 4.2.1 détaille les moyens d'implémentation de la méthodologie utilisée pour cette interaction. La manière d'appliquer le RL à cette problématique de contrôle de micro-réseau est présentée dans la sous-section 4.2.2. Enfin, la sous-section 4.2.3 établit les objectifs mis en place et le dimensionnement choisi.

4.2.1 Implémentation du problème de contrôle

Le développement de l'algorithme de contrôle ainsi que la simulation du micro-réseau sont réalisés avec Python. La librairie utilisée pour construire la simulation de l'environnement est gym (Brockman et al., 2016), développée par OpenAI. Gym permet de construire ou d'utiliser des environnements d'apprentissage par renforcement pour entraîner des agents. Leur utilisation a l'avantage de servir de point de référence pour comparer différents algorithmes

de RL. Gym est utilisée dans le but de construire la simulation dynamique du micro-réseau dans un formalisme propre à l'utilisation du RL. L'environnement gym est une classe qui interagit avec un agent défini par l'utilisateur. La continuité et les dimensions des espaces d'état et d'action sont définies lors de l'appel de la classe. Un environnement gym contient plusieurs fonctions :

— *reset* permet de réinitialiser tous les états observables et non observables par l'agent. Elle est appelée dans le but de commencer un nouvel épisode. Elle renvoie en sortie les premiers états observables par l'agent pour chaque épisode.

— *step* reçoit en entrée les actions de l'agent à un pas de temps. Les changements de l'environnement dus aux actions et aux autres facteurs (données de consommation, de productible renouvelable) sont déterminés. Dans la fonction *step*, les caractéristiques des unités, les données et les actions de l'agent sont intégrées dans le processus algorithmique. Des nouveaux états et une récompense sont ainsi déterminés et transmis à l'agent. De même l'information sur l'éventuelle fin de l'épisode est déterminée et transmise à l'agent.

— *render* est une fonction destinée à l'utilisateur qui permet de visualiser différents indicateurs de l'environnement pour suivre la simulation.

Deux objectifs peuvent être à l'origine des interactions entre l'agent et l'environnement.

— L'agent peut être en phase d'entraînement et doit apprendre des récompenses et états pour adapter sa politique de contrôle dans un but de maximisation des récompenses.

— En phase de validation, l'interaction a lieu pour mesurer la qualité de la politique apprise par l'agent selon les récompenses perçues.

Les deux phases s'enchaînent au cours de l'apprentissage de l'agent. Puisqu'il explore pendant son entraînement, les indicateurs de l'environnement ne sont pas représentatifs de la politique "réelle" de l'agent : certaines actions sont choisies aléatoirement. Au bout d'un certain nombre t_{eval} d'itérations pendant lesquelles l'agent apprend, sa politique doit être évaluée. Les agents ont été développés grâce à la librairie PyTorch. Une fois un épisode initialisé, l'agent envoie une action à l'environnement, et reçoit un tuple composé des états de l'environnement à l'instant suivant, d'un booléen indiquant si l'épisode est terminé ou en cours et de la récompense. Avec ces informations, l'agent choisit l'action suivante selon sa politique de contrôle. Cette politique est mise à jour au fur et à mesure des interactions si l'agent est en phase d'entraînement (voir section 2.2). Si l'agent est en phase de validation, sa politique apprise doit être testée. Il l'utilise donc sur un épisode entier sans explorer ni apprendre. Le processus itératif de l'interaction entre l'agent et l'environnement est décrit sur la Figure 4.3. Cette représentation n'inclut que l'entraînement de l'agent pour des raisons de visibilité. Le processus démarre à t = 0 (en haut à gauche). À chaque appel de *reset*, le numéro N^{ep} de l'épisode généré augmente, le processus se termine lorsque le nombre maximal d'épisode N^{ep}_{max} est atteint. À gauche de la figure, l'environnement est représenté dans l'encadré en pointillé. Il est développé selon le formalisme gym. À droite de la figure se trouve le script de l'agent. Son entraînement apparaît aussi encadré en pointillé et s'effectue avec une classe utilisant la librairie PyTorch.

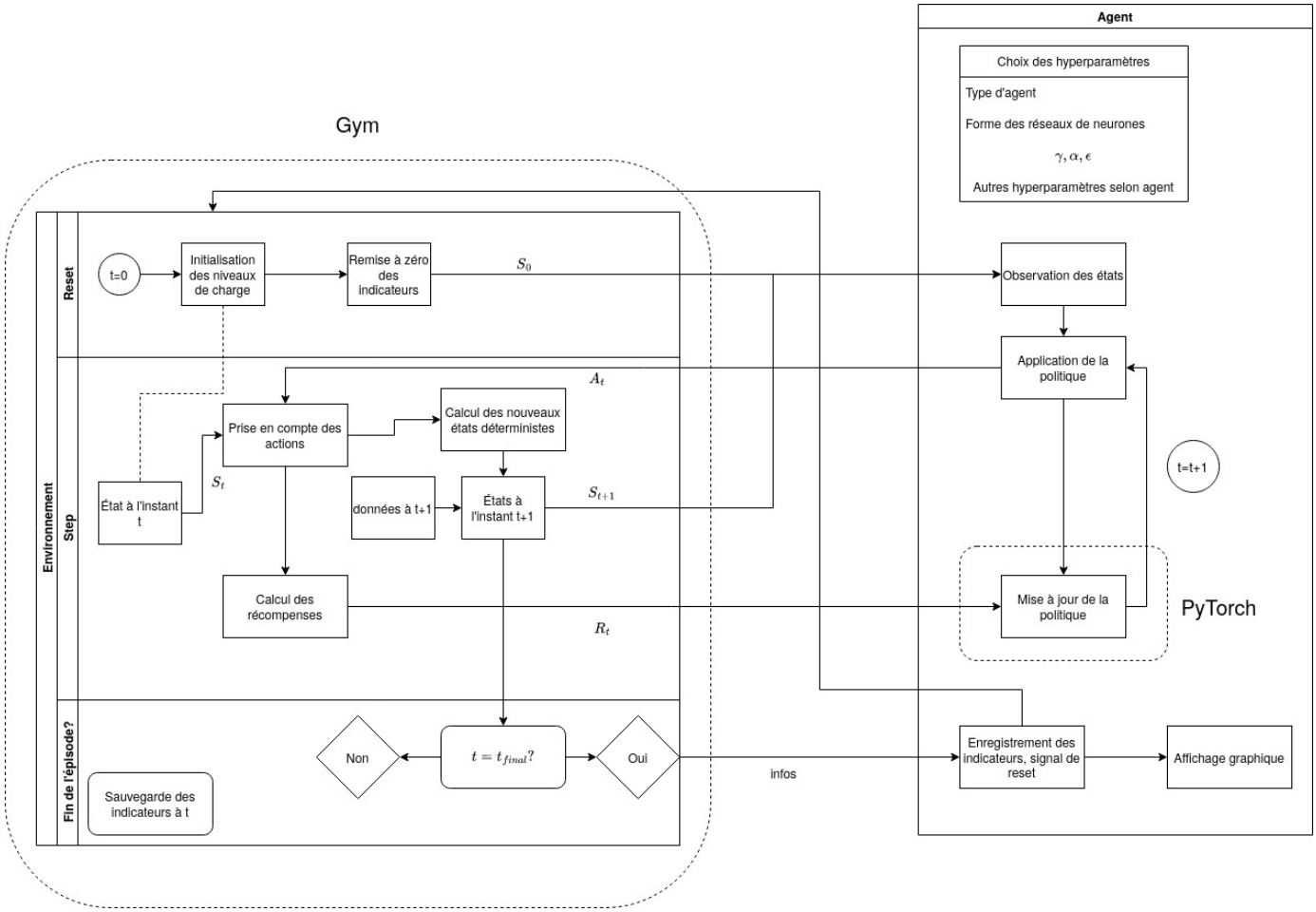

Figure 4.3 – Représentation de l'interaction entre agent et environnement.

Le score de validation d'un agent est la somme cumulée des récompenses obtenues lors d'un épisode de validation. Lorsqu'un score de validation est meilleur que les précédents, la politique qui a permis d'obtenir ce score est enregistrée dans un fichier. Seule la meilleure politique obtenue est conservée. Le nombre d'épisodes maximum n'est pas la seule condition d'arrêt des interactions. Lorsque l'agent est en phase d'entraînement, un critère d'arrêt utilisé dépend de la patience N_{patience} : l'entraînement s'arrête si la politique n'est pas améliorée après $N_{patience}$ épisodes consécutifs. L'Algorithme 12 montre la méthodologie adoptée pour l'entraînement de l'agent.

Algorithme 12 Méthodologie d'entraînement et de validation

Entrées : Agent (EMS), environnement (env), $N^{\text{ep}}_{\text{max}}$, N_{patience}, t_{eval}
Initialisation : Espace d'états $\mathcal{S}$, espace d'actions $\mathcal{A}$, $N^{\text{ep}} = 0$, patience $= 0$, score $= -\infty$, t $= 0$
Tant que $N^{\text{ep}} < N^{\text{ep}}_{\text{max}}$ **ET patience** $< N_{\text{patience}}$ **répéter :**
  $N^{\text{ep}} \leftarrow N^{\text{ep}} + 1$
  s $\leftarrow$ env.*reset*
  R $\leftarrow 0$
  terminal $\leftarrow$ Faux
  Tant que terminal=Faux :
    t $\leftarrow$ t $+ 1$
    a $\leftarrow$ EMS(s)
    s, R, terminal $\leftarrow$ env.*step*(a)
    Si $\text{t}/\text{t}_{\text{eval}} \in \mathbb{N}$ **:**
      $\text{score}_N \leftarrow 0$
      s $\leftarrow$ env.*reset*
      R $\leftarrow 0$
      terminal $\leftarrow$ Faux
      Tant que terminal=Faux :
        a $\leftarrow$ EMS(s)
        s, R, terminal $\leftarrow$ env.*step*(a)
        $\text{score}_N \leftarrow \text{score}_N + \text{R}$
      Si $\text{score}_N >$ **score :**
        Enregistrer l'agent
        score $\leftarrow \text{score}_N$
      Sinon :
        patience $\leftarrow$ patience $+ 1$

4.2.2 Espace d'états et espace d'actions

Espace d'états

L'espace d'états $\mathcal{S}$ permet de générer un vecteur constitué de 4 éléments dans ce cas d'étude. Ils sont observables par l'agent à chaque pas de temps. L'espace d'états est constitué du SOC, de la consommation, de la production PV et de la distance au solstice d'été. Les données de consommation ont été obtenues via AutoCalSol, outil développé par l'Institut National de l'Énergie Solaire (INES). La somme annuelle des demandes équivaut à la consommation électrique moyenne d'un Français en 2022 selon l'INSEE (2,21 MWh). Les données de la production PV ont été récupérées sur la base de données SARAH-2 de PVGIS, mise à disposition par la Commission européenne. Enfin, la distance au solstice d'été est le jour le plus long de l'année dans l'hémisphère nord auquel appartient le lieu de l'étude (21 juin). Cette varible a été utilisée comme état par François-Lavet et al., 2016 pour le contrôle d'un micro-réseau. Elle s'exprime en jours et permet à l'agent de se situer dans l'année. Elle est donc intégrée à l'espace d'états. L'Équation 4.33 montre les notations adoptées pour l'ensemble des états :

$$\forall \text{S} \in \mathcal{S},\ \text{S}_\text{t} = \left\{ \text{SOC(t)}, \text{D(t)}, \text{P}^{\text{PV}}\text{(t)}, \text{dist}^{\text{sols}}\text{(t)} \right\} \tag{4.33}$$

avec $\text{dist}^{\text{sols}}$(t) la distance temporelle au solstice d'été au temps t. Elle est intégrée à l'espace d'état pour prendre en compte la variation annuelle de la production solaire PV. À l'exception de la distance au solstice d'été, les valeurs attribuées des états se situent dans des espaces continus. L'espace d'états est défini comme un espace continu. Toutes les valeurs sont

normées :

$$\forall \mathrm{S} \in \mathcal{S},\ \mathrm{S} \in [0,1]^{|\mathcal{S}|} \tag{4.34}$$

L'adéquation de ces états avec le problème de contrôle a été vérifiée à travers différents tests effectués avec l'apprentissage d'un agent de Q-learning. Dans les tests effectués, l'agent apprend toujours mieux avec tous ces états que lorsque l'un d'entre eux n'est pas observable. Ces états sont couramment utilisés dans des travaux portant sur le contrôle haut-niveau des micro-réseaux (François-Lavet et al., 2016, Ali et al., 2021, B. Mbuwir et al., 2017).

Espace d'actions

Les décisions de contrôle de l'EMS concernent le stockage long terme. Consommer l'hydrogène revient à utiliser la PAC, tandis que sa charge se fait par électrolyse de l'eau. Les actions consistent donc à utiliser l'électrolyseur à puissance maximale, utiliser la PAC à puissance maximale ou ne rien faire. L'espace d'actions se représente ainsi :

$$\forall \mathrm{A} \in \mathcal{A},\ \mathrm{A_t} = \left\{ \mathrm{P^{H_2}} \right\} \tag{4.35}$$

L'espace d'actions est donc un espace d'une dimension comportant uniquement la puissance issue du stockage hydrogène. Cette puissance peut donc prendre trois valeurs selon l'Équation 4.36 :

$$\mathrm{P^{H_2}} = \left\{ \mathrm{P}_{\mathrm{max}}^{PAC}, 0, -\mathrm{P}_{\mathrm{max}}^{\mathrm{elec}} \right\} \tag{4.36}$$

avec $\mathrm{P^{H_2}}$ la puissance sortante du stockage hydrogène. Si cette puissance est positive, le stockage hydrogène se décharge. Si elle est négative, il se recharge par électrolyse de l'eau et cette consommation s'additionne aux demandes d'électricité du micro-réseau. Trois valeurs d'action étant possibles, l'espace d'actions est donc discret.

Pas de temps et horizon de simulation

Les données de production PV et de demande d'électricité collectées sont organisées avec un pas de temps horaire, ce qui signifie que chaque enregistrement de données est effectué à intervalles réguliers d'une heure.

La base de données SARAH-2 a permis d'extraire 11 ans de données (du 1er janvier 2010 au 31 décembre 2020). L'horizon temporel du contrôle du micro-réseau peut donc aller jusqu'à 11 ans. L'horizon temporel de contrôle pour le calcul du coût d'opération du micro-réseau dans une optique de dimensionnement optimal sera de 10 ans. Cependant, des horizons temporels de 1 à 10 ans seront utilisés dans les chapitres 5 et 6.

L'horizon temporel permet d'établir la longueur d'un épisode. Au minimum, un épisode représentera 1 an d'interactions entre l'agent et l'environnement avec un pas de temps horaire. Dans la simulation du micro-réseau, cela représente 8760 calculs d'état du micro-réseau et 8759 consignes d'action communiquées par l'EMS.

4.2.3 Cadre des études sur le contrôle du micro-réseau

Le cadre de référence des chapitres portant sur le contrôle du micro-réseau est défini dans cette sous-section. Des paramètres tels que la valeur des variables de dimensionnement et l'objectif de contrôle de l'EMS sont établis.

Algorithme de référence

Afin de pouvoir comparer les politiques de contrôle des agents entraînés, un algorithme de référence est introduit. Le fonctionnement de l'algorithme de référence ne varie pas selon

les cas d'études qui seront abordés. Cet algorithme est fondé sur un système de règles et n'apprend pas. Les opérations de l'électrolyseur et de la PAC sont déterminées en fonction de l'écart entre la demande et la production à chaque instant. Si $D - P^{PV} > 0$ alors la PAC est utilisée. Si $D - P^{PV} < 0$ alors l'électrolyseur est utilisé. Leur puissance d'utilisation est la puissance maximale, déterminée de la même manière qu'à la sous-section 4.1.2. L'éventuel déséquilibre énergétique restant est comblé par la batterie si son état le permet. Le mode opératoire de charge et décharge a été présenté à la sous-section 4.1.3.

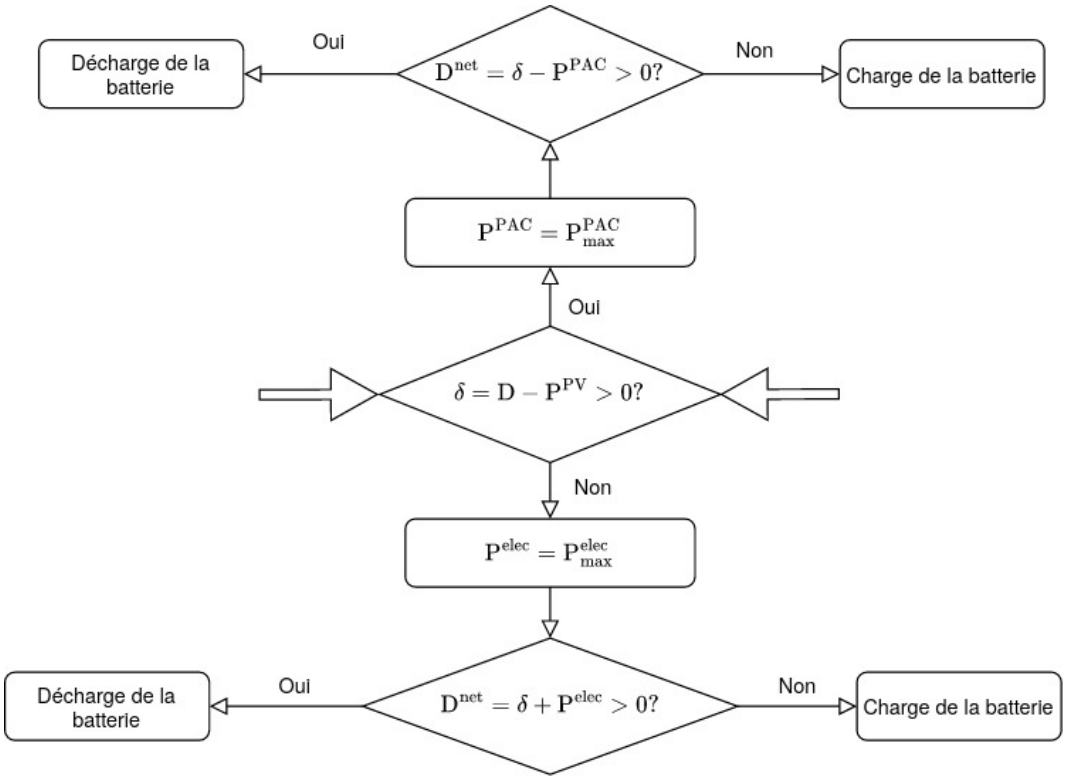

Figure 4.4 – Description de l'algorithme de référence de contrôle

Une représentation schématique de l'algorithme décrit est visible sur la Figure 4.4.

Objectifs de contrôle

Les objectifs de contrôle sont déterminés pour orienter la politique de l'EMS. Ils sont calculés de manière séquentielle afin de constituer le système de récompense de l'agent pendant son apprentissage. Ils peuvent être économiques, environnementaux ou techniques. Puisque la production locale d'électricité est photovoltaïque, la réduction des gaz à effet de serre n'est pas envisagée comme objectif de contrôle du micro-réseau. La structure du micro-réseau ne permet que d'optimiser des objectifs liés aux échanges d'électricité avec le réseau central de distribution.

La minimisation des coûts économiques consiste à minimiser le coût de soutirage de l'électricité et à maximiser les revenus liés à l'injection vers le réseau central. L'objectif d'autonomie relative au réseau central revient à minimiser à la fois l'électricité soutirée et injectée au réseau. Selon les horizons temporels de simulation considérés, la quantité d'électricité injectée ou soutirée varie considérablement. Le taux d'autoconsommation τ_{autocons} et le taux

d'autoproduction τ_{autoprod} sont les indicateurs à maximiser par l'EMS. Ils sont définis par les Équations 4.37 et 4.38.

$$\tau_{\text{autocons}} = 1 - \sum_{t=0}^{T} \frac{E_{\text{inje}}(t)}{P^{PV}(t)} \tag{4.37}$$

$$\tau_{\text{autoprod}} = 1 - \sum_{t=0}^{T} \frac{E_{\text{sout}}(t)}{D(t)} \tag{4.38}$$

La majorité des études sur le contrôle ne minimisent que le coût de soutirage du réseau central dans la littérature (voir section 3.2). Les études pour lesquelles le prix de l'électricité est constant ne considèrent que le taux d'autoproduction comme objectif à minimiser par l'EMS. Les autres choix d'objectif de l'EMS seront considérés pendant le dimensionnement sous contrôle optimal afin de déterminer leur impact sur les objectifs économiques de dimensionnement.

Dimensionnement retenu pour les études de contrôle

Il est nécessaire de définir un espace borné de l'espace vectoriel $\mathcal{D}$ dans lequel les variables de dimensionnement prendront leur valeur. Ainsi, l'optimisation ne s'effectue pas en cherchant des valeurs dans un espace infini. Le dimensionnement du micro-réseau utilisé lors d'études sur le contrôle, notamment au chapitre 5, sera choisi dans cet intervalle.

Une simulation sans stockage hydrogène a été effectuée avec différentes valeurs pour la puissance des panneaux PV et la capacité de la batterie installées. La batterie opère en production ou en consommation d'électricité en fonction du signe de la différence entre la demande et la puissance du productible PV à chaque instant pour combler le déséquilibre du micro-réseau. L'horizon temporel de cette simulation est d'un an. Le taux d'autoproduction a été relevé pour chaque couple de variables afin de déterminer un intervalle de valeurs pertinent. Les résultats de ces simulations sont montrés sur la Figure 4.5. Le taux d'autoproduction est tracé en fonction du rapport de la puissance nominale PV sur la capacité nominale de la batterie. Afin de pouvoir identifier la valeur des couples $\left[P^{PV}_{nom}, E^{batt}_{nom}\right]$, la puissance crête des panneaux PV est représentée par un jeu de couleurs. Les valeurs de la puissance crête des panneaux PV et de la capacité de la batterie ont respectivement été prises entières dans les intervalles $[\![1\,\text{kWc}, 11\,\text{kWc}]\!]$ et $[\![1\,\text{kWh}, 11\,\text{kWh}]\!]$. La zone dense dans la partie haute de la figure montre qu'un taux d'autoproduction proche de 1 peut être approché sans utilisation d'un stockage hydrogène. Inclure des valeurs plus élevées de chaque variable de dimensionnement n'a donc pas de pertinence dans la recherche d'un dimensionnement optimal avec stockage long terme inclus. D'après la Figure 4.5, les domaines de valeur des variables de dimensionnement sont $[\![3\,\text{kWc}, 11\,\text{kWc}]\!]$ et $[\![2\,\text{kWh}, 11\,\text{kWh}]\!]$.

Il a été choisi une installation PV de puissance nominale de 5 kWc et une capacité de batterie nominale de 5 kWh. De cette manière, l'impact de l'incorporation d'un stockage hydrogène contrôlable peut être évalué, car le taux d'autoproduction n'est pas proche de 1. De plus, le rôle de la batterie est celui d'un stockage de puissance (en stockant l'électricité le jour pour la nuit par exemple) tandis que le stockage hydrogène doit se comporter comme un stockage à long terme en stockant l'électricité plusieurs mois. La puissance d'électrolyse et d'opération de la PAC a été fixée à 1 kW. Ainsi, leurs puissances sont plus faibles que celle de la batterie (si elle est suffisamment chargée). La capacité du stockage hydrogène reste supérieure à celle de la batterie électrochimique.

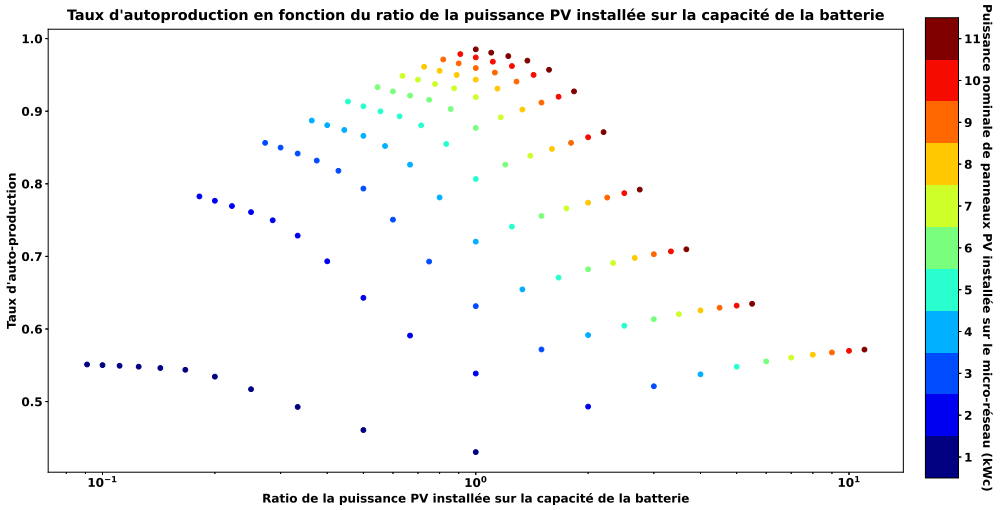

Figure 4.5 – Taux d'autoproduction obtenu par simulation de micro-réseau électrique sans stockage long terme, en fonction de la puissance PV installée et du rapport entre la puissance PV et la capacité de batterie installée

4.3 Dimensionnement du micro-réseau

Le dimensionnement du micro-réseau consiste à minimiser la totalité les coûts économiques liés au micro-réseau en choisissant la valeur des variables de dimensionnement. L'approche modulaire retenue permet d'itérer sur certaines variables liées aux composants du micro-réseau. Étant donné que les coûts d'opération dépendent du contrôle, le coût d'un dimensionnement est déterminé en fonction de la politique de contrôle adoptée pour ce dimensionnement. La sous-section 4.3.1 présente les coûts à minimiser. Les variables de dimensionnement et la méthodologie adoptée sont détaillées à la sous-section 4.3.2.

4.3.1 CAPEX et OPEX

Les objectifs de dimensionnement considérés sont principalement économiques. Ces coûts économiques peuvent se décomposer en deux parties :

— Les dépenses d'exploitation (OPEX) sont les charges courantes : le budget alloué à la maintenance C_{maint} et les coûts et revenus liés aux objectifs d'opération C_{ope}.

— Les dépenses de capital (CAPEX) sont les dépenses d'investissement C_{inv} et le coût de remplacement des unités C_{remp} : ils sont considérés conjoitement en répartissant le coût d'investissement de chaque unité sur sa durée de vie.

L'OPEX intègre seulement les coûts de maintenance C_{maint} qui s'élèvent à 2 % du coût d'investissement par an pour la plupart des unités (Diaf et al., 2008, Nagapurkar et al., 2019) et le coût de remplacement des batteries C^{batt}_{rempl} qui est le coût d'une batterie lissé sur le nombre d'années nécessaire à une dégradation de 30 % de sa capacité.

Les coûts des différentes unités sont listés dans la Table 4.1. Le coût d'investissement lié aux différentes unités est décomposé de la manière suivante :

— Les panneaux PV coûtent $C^{PV}_{inv} = 1100\,€/kWc$ à l'achat (*Coûts et rentabilités du grand photovoltaïque en métropole continentale* 2019).

Panneaux PV	Batterie	Électrolyseur	PAC	Réservoir H_2	Compresseur
1100 €/kWc	151 €/kWh	1200 €/kW	4000 €/kW	50 €/50L à 200 bar	3800 €

Table 4.1 – Coût d'installation de différentes unités du micro-réseau

— Le coût d'investissement de la batterie électrochimique en fonction de sa capacité nominale est de $C_{inv}^{batt} = 151\,€/kWh$ pour une utilisation stationnaire.
— L'électrolyseur PEM coûte $C_{inv}^{elect} = 1200\,€/kW$ (Salehmin et al., 2022). Le compresseur coûte $C_{inv}^{comp} = 3800\,€$ pour une compression de 15 bar à 200 bar (Hancke et al., 2022).
— Le prix d'achat de la PAC vaut $C_{inv}^{PAC} = 4000\,€/kW$ (B. Li et al., 2017, Dufo-López et al., 2007).
— Le réservoir H_2 occupe un volume de $0{,}5\,m^3$ et stocke l'hydrogène à une pression de 200 bar. À cette pression, un tel volume coûte $C_{inv}^{réservoir} = 500\,€$.

Le coût d'investissement global C_{inv} se décompose selon l'Équation 4.39 :

$$C_{inv} = C^{réservoir} + P_{nom}^{elec} \times C_{nom}^{elec} + C^{comp} + P_{nom}^{PAC} \times C_{inv}^{PAC} + E_{nom}^{batt} \times C_{inv}^{batt} + P_{nom}^{PV} \times C_{inv}^{PV} \quad (4.39)$$

avec P_{nom}^{elec}, P_{nom}^{PAC} et P_{nom}^{PV} les puissances nominales de l'électrolyseur, de la PAC et des panneaux solaire PV respectivement. E_{nom}^{batt} est la capacité nominale de la batterie.

Le coût d'opération C_{ope} se détermine en fonction des coûts de l'énergie soutirée au réseau central et aux revenus générés par l'injection d'électricité au réseau central. Il s'écrit selon l'Équation 4.40.

$$C_{ope} = \sum_{t=1}^{T} \left(E_{sout} C_{sout}(t) - E_{inje} C_{inje}(t) \right) \quad (4.40)$$

avec E_{sout} et E_{inje} respectivement l'énergie soutirée et injectée au réseau central de distribution (en kWh), C_{sout} et C_{inje} le revenu de soutirage et le coût d'injection (en €/kWh) et T est l'horizon temporel. Le dimensionnent préalable de chaque module impose des contraintes pour la simulation, plus de détails sur ces contraintes ont été donnés à la Section 4.2. Selon les caractéristiques des modules choisis, le comportement dynamique des unités en phase d'opération sera différent et la quantité d'électricité achetée ou vendue au réseau central aussi. Le dimensionnement répond à un objectif d'optimisation en minimisant le CAPEX selon l'Équation 4.41 :

$$\min_{d \in \mathcal{D}} C_{inv}(d,n) + C_{ope}(d,n) + C_{maint}(d,n) + C_{remp}(d,n), \;\; n \in \mathcal{N} \quad (4.41)$$

avec d et n des valeurs de variables dimensionnables et non-dimensionnables respectivement.

4.3.2 Méthodologie de dimensionnement

Le dimensionnement des modules consiste à déterminer des valeurs des variables de l'espace $\mathcal{D}$ pour optimiser un critère économique. Les variables de dimensionnement sont entières. Dans le Chapitre 5, seules P_{nom}^{PV} et E_{nom}^{batt} sont considérées comme variables. Au Chapitre 6, P_{nom}^{elec} et P_{nom}^{PAC} sont ajoutées aux variables de dimensionnement.

La Figure 4.6 présente l'algorithme d'optimisation. À chaque itération, les variables de dimensionnement sont choisies et le coût est calculé. Puisque le coût d'opération et de maintenance sont à déterminer, l'élaboration d'une politique de contrôle pour chaque itération de dimensionnement est nécessaire : un agent est entraîné par apprentissage par renforcement à chaque itération. Chaque agent est entraîné sur une simulation de l'environnement

dont le dimensionnement correspond aux valeurs des variables d'optimisation. Lorsqu'une politique est apprise, elle est appliquée pour déterminer le coût d'opération du micro-réseau. Cette application de la politique s'effectue dans une simulation du micro-réseau dont les données aléatoires sont issues de jeux de données différents des données employées pour l'apprentissage afin de ne pas biaiser les résultats. L'optimisation du dimensionnement est qualifiée d'optimisation bi-niveaux puisqu'à chaque itération une politique de contrôle est apprise par apprentissage par renforcement. Le niveau le plus haut est appelé la boucle externe et le niveau le plus bas est la boucle interne. La boucle externe requiert des résultats issus d'une optimisation complète de la boucle interne à chacune de ses itérations. Dans cette configuration, la boucle externe est l'optimisation globale du dimensionnement et la boucle interne le RL appliqué au contrôle du micro-réseau. Chaque niveau d'optimisation est représenté par une couleur sur la Figure 4.6.

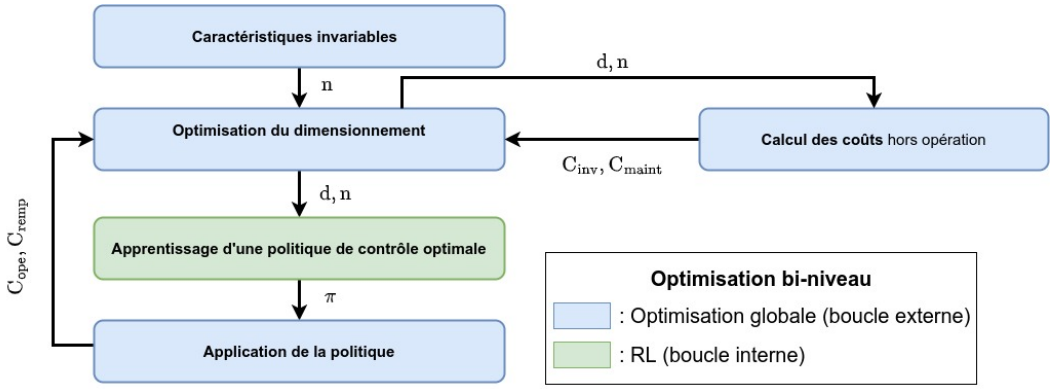

Figure 4.6 – Diagramme représentant la méthode de dimensionnement par optimisation bi-niveaux

Ainsi, les coûts liés au contrôle du micro-réseau sont optimisés à chaque itération du dimensionnement. L'optimisation du dimensionnement est non-linéaire en raison de l'intégration du contrôle dans le calcul des coûts. L'algorithme d'optimisation choisi est le recuit simulé (Kirkpatrick et al., 1983). Plus de détails sur l'algorithme d'optimisation et son choix seront donnés au chapitre 6.

4.4 Conclusion

L'application du RL au problème de contrôle permet d'adapter les performances de fonctionnement selon les dimensionnements du micro-réseau et sans modèle prédictif. Après avoir examiné les bases théoriques du RL et son application dans le pilotage des micro-réseaux, une méthodologie pour l'intégrer dans le processus de dimensionnement d'un micro-réseau sous contrôle optimal a été développée.
Le pilotage et le dimensionnement sont abordés de manière conjointe à travers un algorithme d'optimisation. Cet algorithme itère sur différents dimensionnements d'unités et calcule les coûts sur un horizon de temps défini. Les coûts d'opération sont déterminés à partir d'une phase de contrôle guidée par une politique apprise via un algorithme de RL.
Le dimensionnement du micro-réseau s'effectue avec un objectif de minimisation des coûts. Le micro-réseau est modélisé de façon modulaire afin pouvoir appliquer la méthodologie développée à d'autres architectures de micro-réseaux. Les modules sont caractérisés par des variables, certaines ne sont pas itérables durant le processus de dimensionnement et d'autres le sont. Ces modules permettent de fixer les contraintes de pilotage du micro-réseau et les

objectifs de l'EMS.
Un dimensionnement a été établi pour les études portant sur le contrôle d'un micro-réseau. Ces études incluent l'apprentissage d'une politique de contrôle cohérente avec les objectifs définis dans ce chapitre dans différents niveaux de complexité de simulation. Elles sont présentées dans le Chapitre 5. La réduction du temps de calcul pour la phase de contrôle est cruciale, étant donné son intégration dans un processus d'optimisation plus large. L'implémentation de l'optimisation du dimensionnement avec une méthode pour réduire le temps d'apprentissage ainsi que l'analyse des résultats sont détaillées dans le Chapitre 6.

5

Contrôle à long terme du système de stockage d'un micro-réseau

Le contrôle des unités de stockage d'un micro-réseau avec production électrique d'origine renouvelable est essentiel pour maintenir son autonomie et augmenter sa rentabilité. L'apprentissage par renforcement (RL) (Chapitre 2) permet de développer une politique de contrôle malgré les incertitudes liées aux phénomènes aléatoires et le comportement dynamique complexe du système (Chapitre 3). La modélisation du micro-réseau et la méthodologie de contrôle ont été présentées au Chapitre 4. Le développement d'une politique de contrôle est l'aboutissement de l'entraînement d'un agent en interaction avec une simulation de l'environnement. La stabilité de l'entraînement et la garantie de convergence vers une politique efficace sont recherchées. Pour cela, le choix des paramètres d'apprentissage et l'analyse des politiques de contrôle vont être réalisés sur une simulation simple et avec un horizon temporel court. Ces démarches sont détaillées à la section 5.1. Une fois les paramètres d'apprentissage établis, la modélisation du micro-réseau évolue progressivement pour devenir plus réaliste. Les paramètres d'apprentissage sont réajustés en fonction des modifications apportées à la simulation. En particulier, un modèle de vieillissement non-linéaire de la batterie est incorporé à la simulation. Des modifications sont également effectuées pour que les agents apprennent des politiques de contrôle qui demeurent efficaces pour des simulations avec un horizon temporel plus long que lors de l'apprentissage. Cette étude est présentée à la section 5.2.
Enfin, les politiques développées sont analysées à travers diverses visualisations. Des indicateurs liés à différents objectifs d'opération permettent de conclure sur l'intérêt du stockage hydrogène dans la configuration étudiée. Cette analyse est discutée dans la section 5.3.

5.1 Apprentissage d'une stratégie de contrôle du stockage hydrogène sur un horizon d'un an

La politique de contrôle du système de gestion de l'énergie (EMS) est obtenue à l'issue d'une phase d'apprentissage d'un agent de RL. Un dimensionnement des unités du micro-réseau a été établi au chapitre 4 pour l'élaboration de cette politique. Les valeurs de dimensionnement des unités sont données sur la Table 5.1.

Paramètre	P_{nom}^{PV} (kWc)	E_{nom}^{batt} (kWh)	P_{nom}^{elec} (kW)	P_{nom}^{PAC} (kW)
Valeur	5	5	1	1

Table 5.1 – Dimensionnement des unités du micro-réseau pour son contrôle sur un an.

Les panneaux PV sont la source de génération électrique locale du micro-réseau. La batterie a un rôle de stockage à court terme tandis que le stockage hydrogène est envisagé comme une solution de stockage à long terme. L'EMS contrôle le stockage hydrogène, en choisissant le mode d'opération de l'électrolyseur et de la pile à combustible (PAC) à chaque instant. Le comportement dynamique du micro-réseau est simulé sur un an avec des données de productible PV et de demande électrique. L'influence des hyper-paramètres sur la qualité de l'apprentissage est évaluée dans la sous-section 5.1.1. Le RL étant un algorithme d'apprentissage automatique, une étude de la politique apprise par un agent entraîné est nécessaire afin d'assurer la qualité de la stratégie de contrôle développée. Cette étude est conduite à la sous-section 5.1.2.

5.1.1 Implémentation de l'algorithme et choix des hyper-paramètres

Les différentes classes algorithmiques du RL ont été présentées dans le chapitre 2. L'électrolyseur et la PAC sont les unités pilotables et fonctionnent à puissance nominale lorsque les contraintes le permettent. L'espace d'actions est donc discret : l'agent choisit le mode d'opération de la PAC et de l'électrolyseur. L'espace d'états est constitué de quatre variables continues : l'état de charge de la batterie (SOC), la demande, le productible PV et la distance en jours au solstice d'été (voir section 4.2). L'algorithme retenu pour le contrôle de ces unités est le DQN pour sa compatibilité avec l'environnement construit.

Bien que l'algorithme de DQN soit couramment employé, il existe différentes approches d'implémentations qui lui sont associées. Certaines présentent des fonctionnalités supplémentaires, comme le *Prioritized Experience Replay* (Schaul et al., 2016), le *Double DQN* (van Hasselt et al., 2015), le *dueling Deep Q-learning* (Z. Wang et al., 2015). Chaque configuration du DQN implémenté peut se distinguer par le nombre et la valeur de ses paramètres. Les configurations varient car elles dépendent de l'environnement, des espaces d'état et d'action associés et du système de récompense de l'agent. Les hyper-paramètres les plus répandus du RL (α, ε, γ) et un hyper-paramètre propre au DQN (N_{cible}) ont été introduits au chapitre 2. Les paramètres de l'algorithme de DQN employé et leur fonction sont présentés ici. Enfin, une analyse de l'impact des différents paramètres est menée.

Implémentation du DQN

Le développement de la politique de contrôle d'un agent requiert trois modes d'interaction avec son environnement : apprenttissage, validation et test.

1 - Apprentissage ou entraînement. Durant la phase d'apprentissage, l'agent interagit avec son environnement selon une politique ε-gourmande. Cela lui permet de prendre des actions aléatoires avec une probabilité ε pour diversifier les observations faites de l'environnement. Il met à jour sa politique de contrôle en associant des valeurs aux couples état-actions observés.

Exploration. La probabilité de prendre des actions aléatoires est le taux d'exploration ε. Sa valeur diminue de manière linéaire pendant la phase d'apprentissage. Le nombre d'itérations pendant lequel ε décroit se note ε_{dec}. La pente de décroissance τ_ε du taux d'exploration peut se déterminer grâce à l'Équation 5.1.

$$\tau_\varepsilon = \frac{\varepsilon_{\text{ini}} - \varepsilon_{\text{fin}}}{\varepsilon_{\text{dec}}} \tag{5.1}$$

Les tuples d'état, d'action et de récompense observés par l'agent grâce à l'exploration et à l'exploitation (utilisation de la politique développée par l'agent) sont appelés tuples d'expérience $\{S_i, A_i, R_i, S_{i+1}\}$. Ils sont stockés dans une mémoire de relecture. Cette mémoire est indispensable dans la mise à jour la politique de l'agent.
Rappel : S_i, A_i, R_i et S_{i+1} sont respectivement les état observés, l'action choisie, la récompense reçue et les nouveaux états observés à l'instant i.

Mise à jour. L'algorithme de DQN fonctionne avec deux réseaux de neurones : un Q-réseau principal Q_θ qui prend les décisions, et un Q-réseau cible Q_{θ^-} qui lui permet de stabiliser son apprentissage (voir section 2.2). Le réseau principal est mis à jour toutes les t_{train} interactions avec son environnement. La valeur la plus courante de t_{train} est 1 pour qu'il soit mis à jour à chaque interaction. Sa mise à jour s'effectue en associant des valeurs à des couples état-action selectionnés aléatoirement dans la mémoire de relecture en fonction des récompenses reçues. La méthode du gradient s'applique sur l'écart entre la prédiction du réseau principal et celle du réseau cible lors de cette mise à jour. Le Q-réseau cible est mis à jour moins fréquemment. Il recopie les paramètres du Q-réseau principal toutes les N_{cible} mises à jour du Q-réseau principal, soit toutes les $t_{\text{train}} \times N_{\text{cible}}$ itérations.

Mémoire de relecture. La mémoire de relecture est essentielle pour l'entraînement de l'agent. La mise à jour du Q-réseau principal Q_θ s'effectue selon l'Équation 2.33. Un Q-réseau est constitué d'une multitude de poids et biais (paramètres) interconnectés utilisés pour estimer l'ensemble des Q-valeurs. Une mise à jour selon un seul tuple d'observation suffit à changer toutes les Q-valeurs en sortie de Q_θ. Des lots ou *batchs* de tuples sont utilisés. La taille n_{batch} de ces batchs est paramétrable et fera l'objet d'une étude de sensibilité. Les tuples constituant les batchs doivent être selectionnés de manière aléatoire pour éviter au Q-réseau d'apprendre sur des séquences de prises de décisions consécutives. La capacité de la mémoire de relecture est paramétrable. Lorsque le nombre d'observations échantillonnées dépasse cette capacité, chaque nouvelle observation remplace la plus ancienne.

2 - Validation Des phases de validation ont lieu périodiquement pendant l'apprentissage. La meilleure politique est identifiée grâce à une phase de validation. Il s'agit d'un épisode simulé pendant lequel l'agent ne prend pas de décision aléatoire. Les récompenses cumulées obtenues en validation sont sauvegardées pour visualiser l'évolution de l'apprentissage de l'agent. L'agent ayant obtenu le meilleur score en validation est enregistré. Le nombre d'itérations t_{eval} entre chaque validation est un paramètre réglable. La validation sert de critère d'arrêt de l'entraînement : l'entraînement s'arrête lorsque le score en validation n'a pas augmenté depuis un nombre de validations appelé patience. Un autre critère d'arrêt consiste à définir un nombre d'épisodes d'entraînement maximal $N^{\text{ep}}_{\text{max}}$.

3 - Test Une fois l'entraînement terminé, l'agent ayant obtenu le meilleur score de validation, doit être testé. Ces tests sont réalisés avec des jeux de données différents de ceux de l'entraînement et de la validation. Cela permet de vérifier que l'agent n'a pas été sujet au sur-apprentissage et peut s'adapter à de nouvelles conditions de simulation. Les tests permettent de comparer les scores de plusieurs agents indépendamment des données utilisées lors de leur entraînement. C'est utile pour choisir un jeu de paramètres d'apprentissage sans le biais du surapprentissage.

Le processus itératif de l'entraînement de l'agent est montré avec l'Algorithme 13. La phase de test n'est pas représentée car elle intervient lorsque l'entraînement est terminé. Les paramètres décrits balisent le fonctionnement de l'algorithme au cours de l'apprentissage. Leur impact sur la politique apprise doit être maintenant explicité.

Algorithme 13 Apprentissage d'une politique de contrôle par un agent

Entrées : Agent (EMS), environnement (env), N^{ep}_{max}, $N_{patience}$, t_{eval}, t_{train}, n_{batch}, N_{cible}, ε_{ini}, ε_{fin}, ε_{dec}
Initialisation : terminal = Faux, Espace d'états $\mathcal{S}$, espace d'actions $\mathcal{A}$, $N^{ep} = 0$, patience = 0, score = $-\infty$, t = 0
Tant que terminal=Faux :
  Échantillonner aléatoirement a $\in \mathcal{A}$
  a $\leftarrow$ EMS(s)
  s', R, terminal $\leftarrow$ env.*step*(a)
  Ajouter {s,a,R,s'} à la mémoire de relecture
  s $\leftarrow$ s'
Tant que $N^{ep} < N^{ep}_{max}$ ET patience $< N_{patience}$ répéter :
  $N^{ep} \leftarrow N^{ep} + 1$
  s $\leftarrow$ env.*reset*
  R $\leftarrow$ 0
  terminal $\leftarrow$ Faux
  Tant que terminal=Faux :
    $t \leftarrow t + 1$
    Échantillonner $\epsilon \in [0, 1]$
    Si $\epsilon < \varepsilon$ **:**
      Échantillonner aléatoirement a $\in \mathcal{A}$
    Sinon :
      a $\leftarrow$ EMS(s)
    s', R, terminal $\leftarrow$ env.*step*(a)
    Ajouter {s,a,R,s'} à la mémoire de relecture
    s $\leftarrow$ s'
    Si $t/t_{train} \in \mathbb{N}$:
      $N_{train} \leftarrow N_{train} + 1$
      Échantillonner n_{batch} tuples dans la mémoire de relecture
      Mettre à jour Q_θ selon α, γ, et les tuples échantillonnés
      $\varepsilon \leftarrow \varepsilon - \varepsilon_{dec}$
      Si $N_{train}/N_{cible} \in \mathbb{N}$:
        $Q_{\theta^-} \leftarrow Q_\theta$
    Si $t/t_{eval} \in \mathbb{N}$:
      Valider la politique apprise par l'agent (voir Algorithme 12)
      Enregistrer une politique dans le cas ou elle est meilleure que la politique précédemment enregistrée et réinitialiser patience (voir Algorithme 12)
      patience $\leftarrow$ patience + 1

Influence et choix des différents hyper-paramètres

La valeur des hyper-paramètres influence l'apprentissage de différentes manières (Kiran et al., 2022). Ils ont une influence sur la vitesse à laquelle une politique de contrôle satisfaisante est trouvée, sur la qualité de la politique apprise selon les objectifs définis et sur la reproductibilité des résultats obtenus (Eimer et al., 2023).

Méthodologies de l'étude paramétrique. Une valeur spécifique d'un hyper-paramètre peut être optimale pour un ensemble particulier de valeurs des autres hyper-paramètres, mais cela peut varier si la valeur des autres hyper-paramètres change. En conséquence, ces hyper-paramètres ne peuvent pas être sélectionnés indépendamment les uns des autres. De nombreuses méthodes pour choisir leurs valeurs existent. La recherche manuelle (Hinton, 2012) est un ajustement de la valeur des hyper-paramètres par essais et erreurs jusqu'à l'obtention d'un ensemble satisfaisant. Ce processus peut être chronophage et entraîner des biais lorsque certains sous-ensembles de valeurs sont négligés. La recherche par grille (Liashchynskyi et al., 2019) consiste à choisir un ensemble de valeurs possibles par hyper-paramètre, puis d'évaluer le score en phase de test de la politique apprise pour chaque combinaison de valeurs. La recherche aléatoire (Bergstra et al., 2012) explore également des combinaisons de valeurs, mais sans recourir à des ensembles prédéfinis. Cette approche s'avère utile et plus rapide que la recherche par grille, notamment lorsque les hyper-paramètres sont nombreux et qu'ils prennent leurs valeurs dans de vastes espaces.
Dans cette étude, le nombre d'hyper-paramètres est restreint mais le nombre de valeurs combinées est important. Les hyper-paramètres de l'algorithme de DQN ont été déterminés à l'aide de la méthode de recherche par grille.

Jeux de données. Afin d'éviter le sur-apprentissage de l'agent sur les données d'entraînement, deux jeux de données ont été utilisés. Les données de productible PV proviennent de la base de données SARAH-2 de PVGIS (Šúri et al., 2005). Les données de productible PV d'entraînement sont celles de la ville d'Albi en 2013 et les données de test sont celles de la ville de Toulouse en 2018.

Paramètres immuables. Certains paramètres doivent être déterminés pour définir le cadre de l'étude de sensibilité des autres paramètres. Les objectifs sont formulés de manière à attribuer une récompense négative à l'agent lorsque de l'électricité est soutirée au réseau central de distribution. La minimisation de cette pénalité revient à maximiser le taux d'autoproduction τ_{autoprod} (voir Équation 4.38). Le système de récompense de l'agent influence la politique apprise par l'agent. Les politiques développées par le biais de systèmes de récompense différents feront l'objet d'une étude comparative dans cette section.
L'agent est entraîné à chaque itération, c'est-à-dire à chaque fois qu'il interagit avec l'environnement ($t_{\text{train}} = 1$). La validation de la politique apprise par l'agent a lieu toutes les 4 itérations ($t_{\text{eval}} = 4$). Le seuil de patience des agents validations vaut 1000.

Les effets de chaque hyper-paramètre seront analysés séparément. Chaque hyper-paramètre sera évalué selon ses valeurs possibles dans le cadre de la recherche par grille. Il convient de noter que pour chaque hyper-paramètre évalué, les valeurs utilisées pour les autres hyper-paramètres sont celles finalement sélectionnées à l'issue de l'étude.

Mémoire de relecture et exploration

La taille de la mémoire de relecture a été choisie de manière empirique. Elle est suffisamment grande pour permettre le stockage de chaque tuple observé jusqu'à la fin de l'entraînement. Une grande mémoire de relecture réduit la corrélation entre les tuples échantillonnés, garantissant une meilleure hétérogénéité des données d'apprentissage.

Une stratégie de décroissance linéaire a été adoptée pour le taux d'exploration ε, initié à une valeur $\varepsilon_{\text{ini}} = 1$ et progressivement réduit pour atteindre une valeur minimale de $\varepsilon_{\text{fin}} = 0.001$. La taille de la mémoire de relecture est d'un million de tuples. Le nombre d'itérations d'entraînement est donc $\varepsilon_{\text{dec}} = 1\,\text{e}^{6}$ puisque $\text{t}_{\text{train}} = 1$. D'après l'Équation 5.1, τ_{ε} vaut $0.99\,\text{e}^{-6}$. Ainsi, l'exploration est maximale lorsque le nombre d'interactions échantillonnées dans la mémoire est minimale. L'exploration diminue progressivement à mesure que la mémoire de relecture se remplit. Ce mécanisme a permis d'atteindre un équilibre efficace entre exploration et exploitation tout au long du processus d'apprentissage.

Taille de batch

La taille des batchs a été itérée sur les valeurs 32, 64, 128 et 256. La récompense cumulée obtenue en test est similaire pour chaque taille de batch. En revanche, la Figure 5.1 montre que l'amplitude des mises à jour du Q-réseaux Q_{θ} (aussi appelée *fonction perte*) est sujette à des oscillations importantes pour des tailles de batchs plus faibles. Une taille de batch trop petite peut entraîner des fluctuations perceptibles dans l'évolution de la fonction perte, créant un certain niveau de bruit et d'oscillations durant l'entraînement (Smith et al., 2018). La fonction perte est plus lisse quand la taille du batch augmente, ce qui stabilise l'apprentissage. Des batchs plus grands ont été considérés mais les résultats de convergence et de politique développée sont similaires au batch de taille 256 pour des besoins de mémoire et puissance de calcul plus élevés. La courbe correspondant à une taille de batchs de 256 présente les plus faibles oscillations et est retenue pour les étapes ultérieures de l'étude.

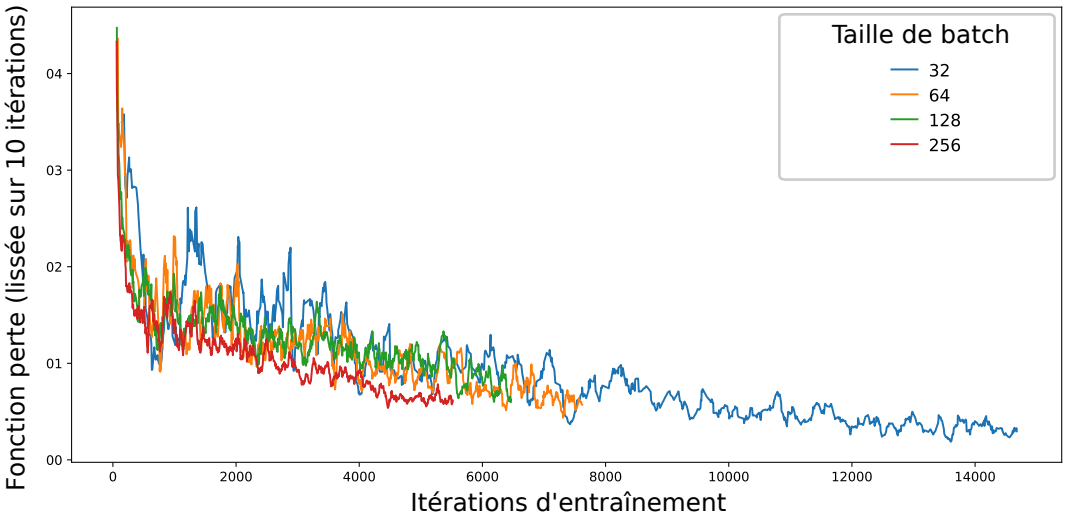

Figure 5.1 – Évolution de la fonction perte durant l'entraînement pour des tailles de batchs différentes.

Taux d'apprentissage.

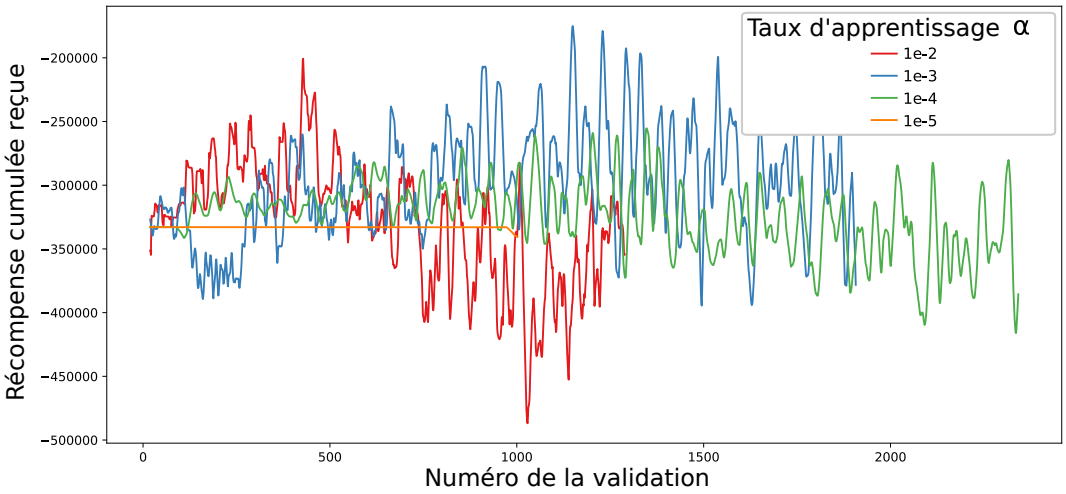

Figure 5.2 – Évolution des récompenses reçues au long des validations pour différents taux d'apprentissage.

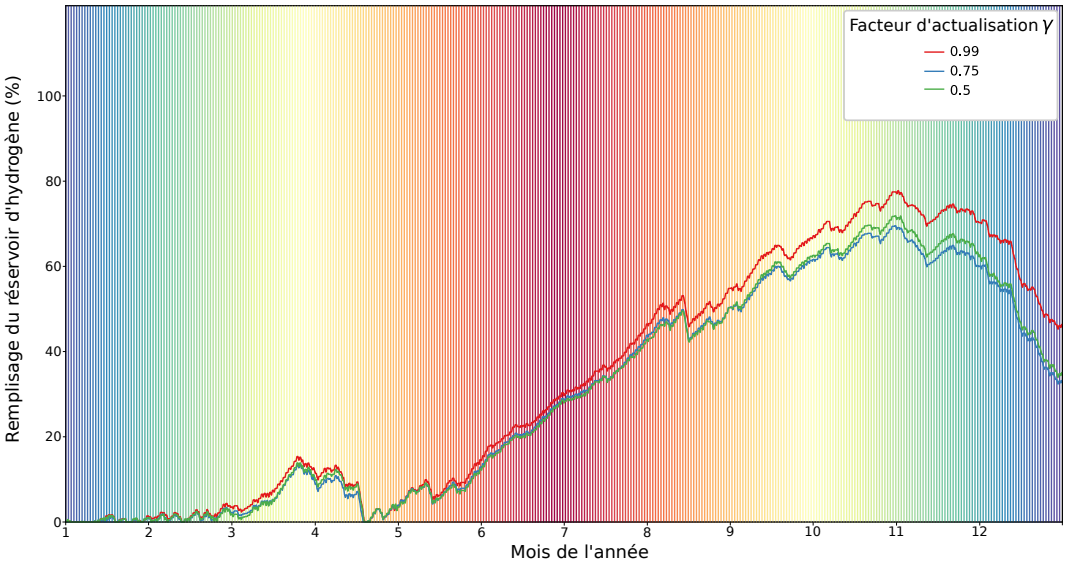

Figure 5.3 – LOH durant un an de test pour des agents entraînés avec des facteurs d'actualisation différents.

La sélection du taux d'apprentissage α est un compromis entre vitesse d'apprentissage et stabilité du modèle (voir Assignation 2.33). La Figure 5.2 présente les courbes d'apprentissage pour des taux d'apprentissage de $1\,\mathrm{e}^{-2}$, $1\,\mathrm{e}^{-3}$, $1\,\mathrm{e}^{-4}$ et $1\,\mathrm{e}^{-5}$. Si α est élevé, l'amplitude des mises à jour est plus grande, l'agent apprend avec des gradients plus élevés à chaque échantillon et l'apprentissage peut diverger. Dans le cadre du DQN, un taux d'apprentissage insuffisant peut causer des problèmes de convergence. En effet, le Q-réseau cible stabilise l'entraînement en permettant une convergence par palier du modèle (voir sous-section 2.2.1). L'évolution trop lente du modèle par rapport à celle du Q-réseau cible limite cet effet et nuit à la stabilité de l'apprentissage. Un taux d'apprentissage de $1\,\mathrm{e}^{-3}$, bien que générant des oscillations plus importantes que le taux de $1\,\mathrm{e}^{-4}$, donne des performances globalement

meilleures. $1\,e^{-3}$ est un choix approprié qui équilibre vitesse d'apprentissage et stabilité de convergence.

Facteur d'actualisation

Le facteur d'actualisation γ introduit au chapitre 2 permet de pondérer l'importance des récompenses futures par rapport aux récompenses immédiates. Trois valeurs de γ ont été testées : 0.99, 0.75 et 0.5. Les récompenses cumulées reçues au terme de l'entraînement sont similaires dans les trois cas. Une visualisation du stockage hydrogène en fonction de γ est représentée sur la Figure 5.3. Les couleurs de fond permettent de représenter les saisons : les couleurs sont plus chaudes l'été et froides l'hiver. Chaque simulation commence en janvier, période de l'année où le stockage d'énergie est limité en raison de conditions d'ensoleillement défavorables et d'un niveau d'hydrogène initial à zéro. Il apparaît que la quantité d'hydrogène stocké pour $\gamma = 0.99$ est supérieure. La quantité d'hydrogène présente dans le réservoir en fin de simulation est plus importante. Le facteur d'actualisation $\gamma = 0.99$ est choisi.

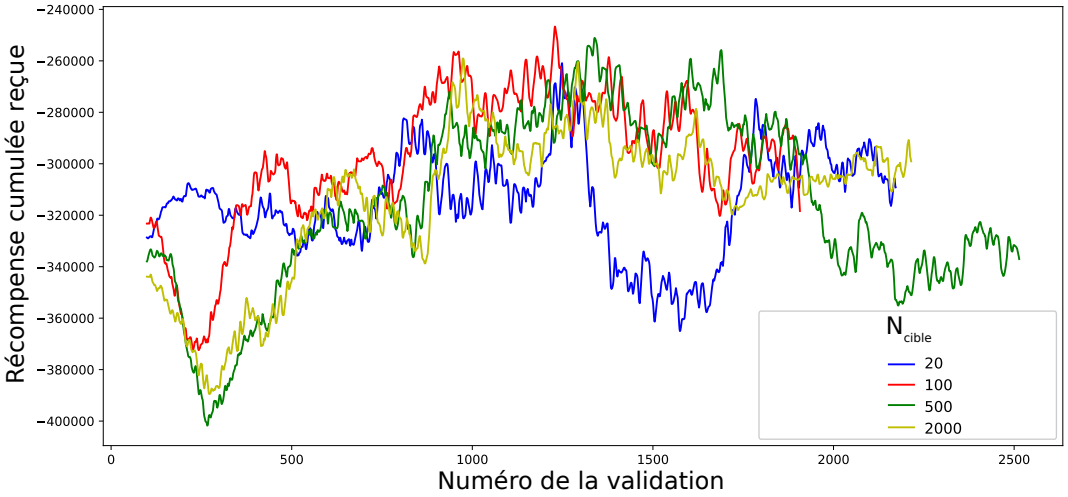

Figure 5.4 – Évolution des récompenses cumulées au long des validations pour différents intervalles de mise à jour du Q-réseau cible.

Nombre d'itérations avant la mise à jour du Q-réseau cible.

Le paramètre du nombre d'itérations N_{cible} avant la mise à jour du Q-réseau cible dans le DQN joue un rôle important dans la stabilisation de l'apprentissage (voir section 2.2). La valeur recherchée pour ce paramètre doit permettre au Q-réseau principal d'effectuer ses prédictions avec une cible stable pendant un certain nombre d'itérations d'entraînement. Dans le cadre de cette étude, il a été constaté que le choix d'un nombre d'itérations égal à 100 avant la mise à jour du réseau cible conduisait à des performances légèrement supérieures en terme de récompense cumulée, comme le montre la Figure 5.4. Pour rappel, le seuil patience étant égal à 1000, l'entraînement s'arrête 1000 itérations après l'atteinte du score maximal. Les courbes sont lissées pour plus de visibilité et le maximum réel des récompenses reçues n'apparaît donc pas sur les courbes.

5.1.2 Analyse de la stratégie apprise par l'agent

Les paramètres retenus sont listés sur la Table 5.2

Paramètre	t_{eval}	t_{train}	$N_{patience}$	ε_{ini}	ε_{fin}	ε_{dec}	n_{batch}	N_{cible}	α	γ
Valeur	4	1	1000	1	0.001	1 e 6	256	100	0.001	0.99

Table 5.2 – Valeur des paramètres du DQN retenus.

Une fois les paramètres d'apprentissage déterminés, l'analyse détaillée de la stratégie ou politique apprise par les agents va être présentée dans cette section. L'évolution de la politique apprise au fur et à mesure de l'apprentissage d'un agent sera considérée dans un premier temps. Une comparaison de performances entre un agent de DQN et un algorithme basé sur des règles présenté au chapitre 4 sera menée. Enfin, une étude de sensibilité de l'algorithme sera conduite selon différentes conditions d'initialisation et horizons temporels.

Politiques développées au cours de l'apprentissage

Évolution de la stratégie de stockage au cours de l'apprentissage. L'étude de l'évolution de la politique apprise par l'agent au cours de son entraînement permet de comprendre la manière dont elle s'est développée. Les actions de l'agent portent sur la gestion du stockage à long terme du micro-réseau en utilisant l'électrolyseur ou la PAC, c'est pourquoi la visualisation choisie pour comprendre la politique porte sur le tracé de l'équivalent énergétique de la quantité de H_2 stocké au cours du temps (voir Figure 5.5).

La politique initiale de l'agent n'implique jamais l'usage de l'électrolyseur. À mesure de l'entraînement (itération 44), sa fréquence d'utilisation augmente, son emploi est souvent suivi par l'usage de la PAC, ce qui suggère une utilisation à court terme de ce stockage. Aux itérations suivantes, la politique évolue progressivement en stockant une quantité significative d'énergie au printemps, en été et en automne. La PAC est utilisée assez régulièrement, car la courbe d'énergie stockée a fréquemment une pente décroissante. Elle est en revanche bien plus mise à contribution en l'hiver. L'aire sous les courbes augmente avec le nombre d'itérations. La forme des courbes est globalement conservée jusqu'à la fin du processus d'apprentissage. Le stockage est accru au cours de l'apprentissage jusqu'à convergence vers la politique finale (obtenue à l'itération 3616).

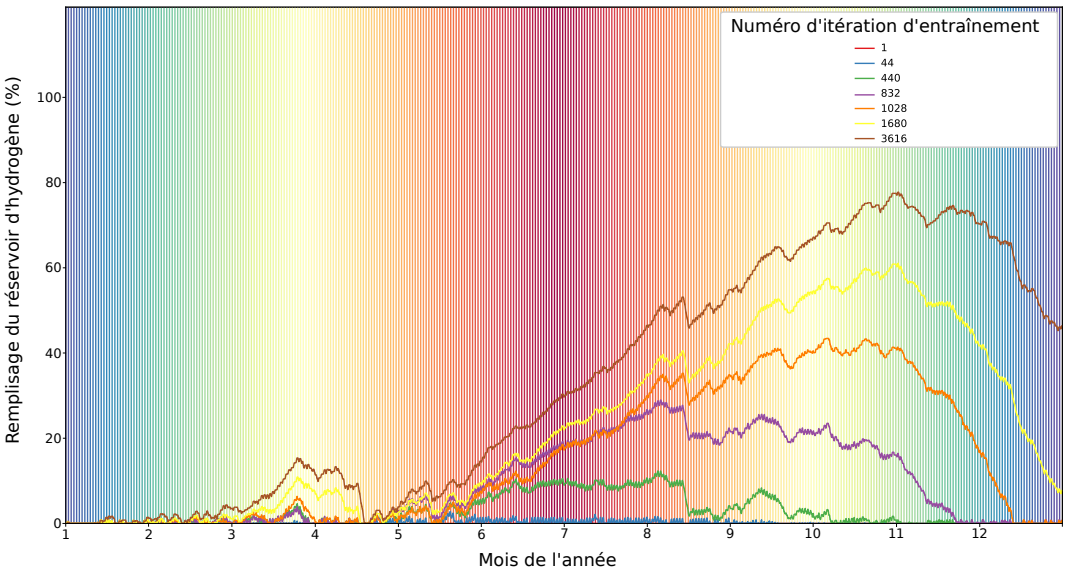

Figure 5.5 – LOH sur un an pour à différentes étapes de l'apprentissage de l'agent.

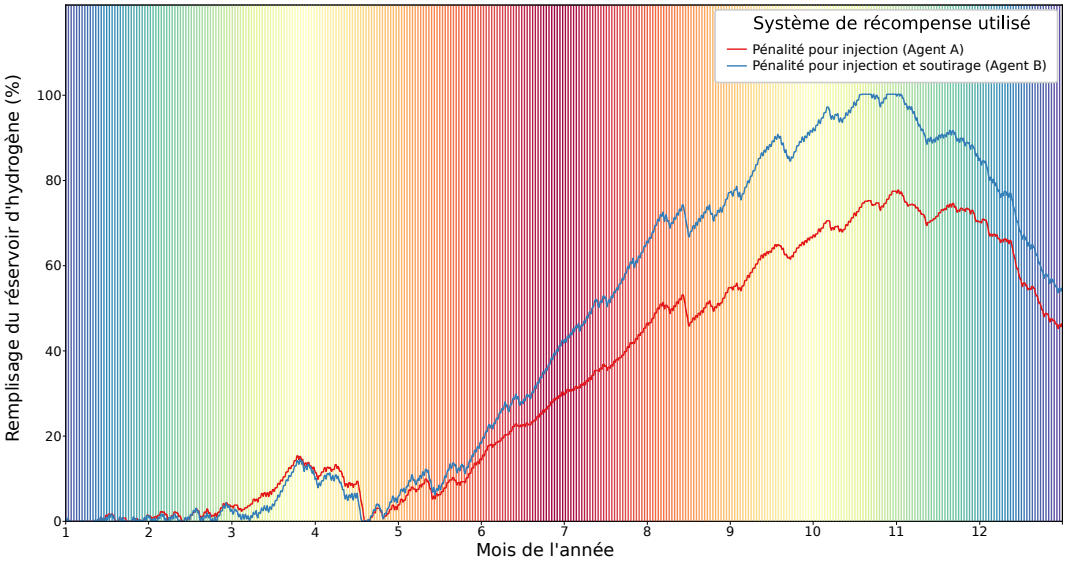

Figure 5.6 – LOH sur un an selon le système de récompense défini lors de l'entraînement de l'agent.

Effet des différentes récompense. Puisque les objectifs de l'EMS peuvent varier selon les besoins de l'utilisateur, il est essentiel d'examiner leur influence sur la stratégie adoptée par l'agent. L'apprentissage de différents agents a été mené en utilisant plusieurs systèmes de récompense (Père et al., 2022). La stratégie développée par un agent formé pour maximiser le taux d'auto-production τ_{autoprod} en pénalisant le soutirage au réseau central (agent A) est comparée à celle élaborée par un autre agent qui, en plus de maximiser τ_{autoprod}, cherche également à maximiser le taux d'autoconsommation τ_{autocons} (agent B). L'agent B est pénalisé à la fois pour l'énergie soutirée et injectée au réseau central de distribution. Les taux d'autoconsommation et d'autoproduction ont respectivement été définis par les Équations 4.37 et 4.38. La quantité de H_2 stocké selon le système de récompense est présentée sur la Figure 5.6. L'agent B a stocké le plus d'énergie, avec un maximum atteint en octobre. Ce système de récompense incite explicitement à minimiser les échanges avec le réseau central. Les excès d'énergie étant pénalisés autant que les déficits, l'énergie produite est stockée lorsqu'elle dépasse la consommation. L'agent A a stocké une quantité d'hydrogène légèrement inférieure. Il a moins stocké l'énergie en excès. Cela peut s'expliquer par le faible rendement de ce type de stockage et par la contrainte qu'a l'agent à l'utiliser à puissance fixe.

Comportement de l'agent selon différentes configurations initiales

L'efficacité d'un algorithme de RL ne se limite pas seulement à sa capacité à apprendre et à optimiser des politiques dans des environnements spécifiques. Une caractéristique tout aussi importante est son adaptabilité à des conditions initiales variables. En effet, l'EMS peut être amené à piloter le stockage H_2 sans qu'il soit initialement vide, ou que sa mise en service soit effectuée en début d'année.

Effet de la charge initiale du stockage hydrogène. Un agent a été soumis à un entraînement avec un réservoir d'hydrogène initialement rempli à 66% de sa capacité. Sa politique de contrôle du stockage hydrogène est comparée à celle de l'agent ayant appris avec un réservoir vide en début de simulation. Pour réaliser cette comparaison, ces agents sont testés sur les deux environnements : Les Figures 5.7 et 5.8 montrent respectivement l'évolution de l'hydrogène stocké dans une simulation commençant avec un stockage à 66% de sa capacité et un stockage vide pour les deux agents.

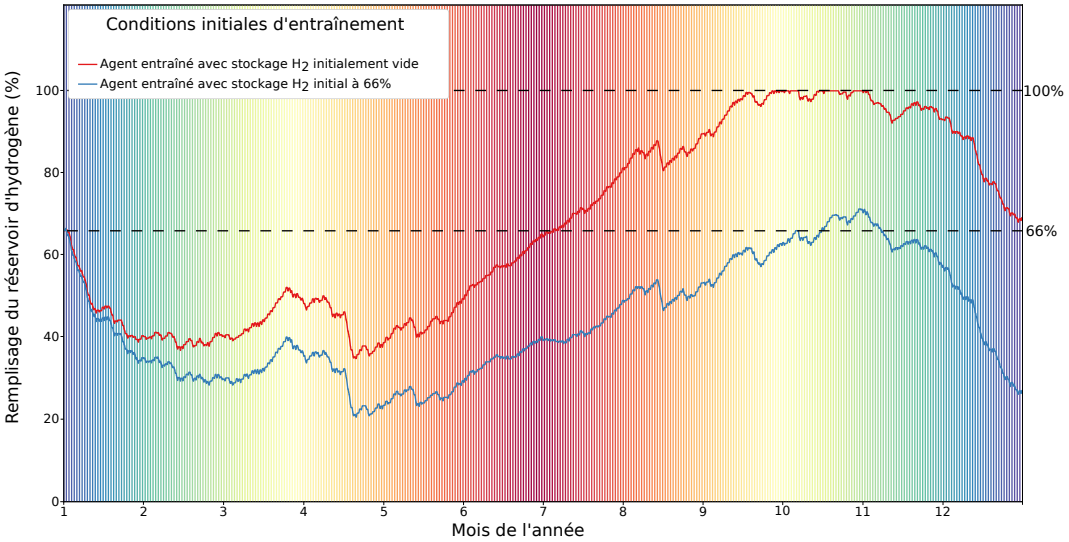

Figure 5.7 – LOH sur un an pour des agents ayant appris avec des initialisations différentes pour le stockage d'énergie. Cas d'une simulation avec stockage initial à 66% de sa capacité.

La politique résultante semble significativement différente pour les deux agents. Notamment, l'agent entraîné avec un environnement initialisé avec un stockage vide atteint la capacité maximale de stockage à plusieurs reprises sur l'environnement avec initialisation à 66% (Figure 5.7).

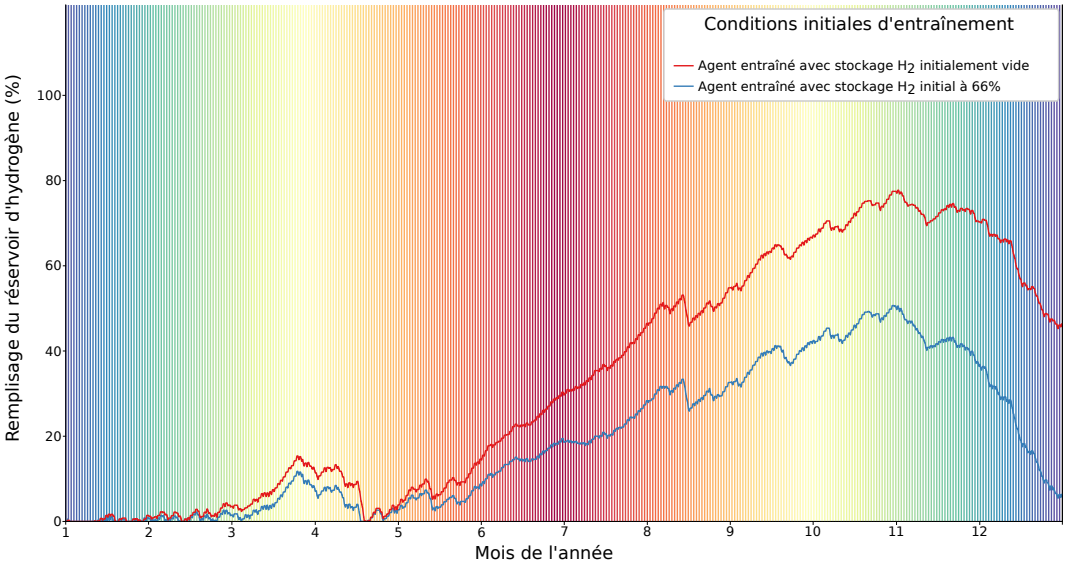

Figure 5.8 – LOH sur un an pour des agents ayant appris avec des initialisations différentes pour le stockage d'énergie. Cas d'une simulation avec stockage initial vide.

Les politiques apprises sont aussi différentes sur l'environnement pour lequel la réserve de H_2 initiale est vide (Figure 5.8) : la courbe rouge est toujours au dessus de la courbe bleue. L'agent qui a appris à contrôler un stockage hydrogène initialement vide a tendance à stocker plus d'énergie.

Les différents taux d'autoproduction obtenus sur cet environnement par les agents entraînés et par l'algorithme déterministe introduit au chapitre 4 sont comparés sur la Table 5.3. Ils sont calculés uniquement sur cet environnement car la charge initiale du stockage long terme biaise les résultats en augmentant le taux d'autoproduction sur l'autre environnement. L'algorithme déterministe révèle un taux d'autoproduction inférieur par rapport aux deux

Algorithme	$\tau_{\mathbf{autoprod}}$
Déterministe	0.78
DQN (LOH initial à 66 %)	0.94
DQN (LOH initial à 0 %)	0.97

Table 5.3 – Comparaison des taux d'autoproduction selon différents algorithmes et conditions initiales d'entraînement sur l'environnement sans charge initiale du stockage H_2.

agents de DQN. Il y a un écart de score de 3% entre les deux algorithmes de DQN. Il n'est pas significatif malgré les différences observées dans l'évolution du stockage sur les deux environnements.

Effet de la date initiale de simulation. Deux agents sont entraînés sur un environnement dont les épisodes commencent à une date différente. Par défaut, la date initiale d'un épisode est le 1er janvier. Le deuxième agent est entraîné sur des épisodes commençant au solstice d'été, soit le 21 juin. L'énergie stockée sous forme de H_2 pour les deux agents est présentée sur la Figure 5.9 dans le cas d'une simulation commançant au solstice d'été.

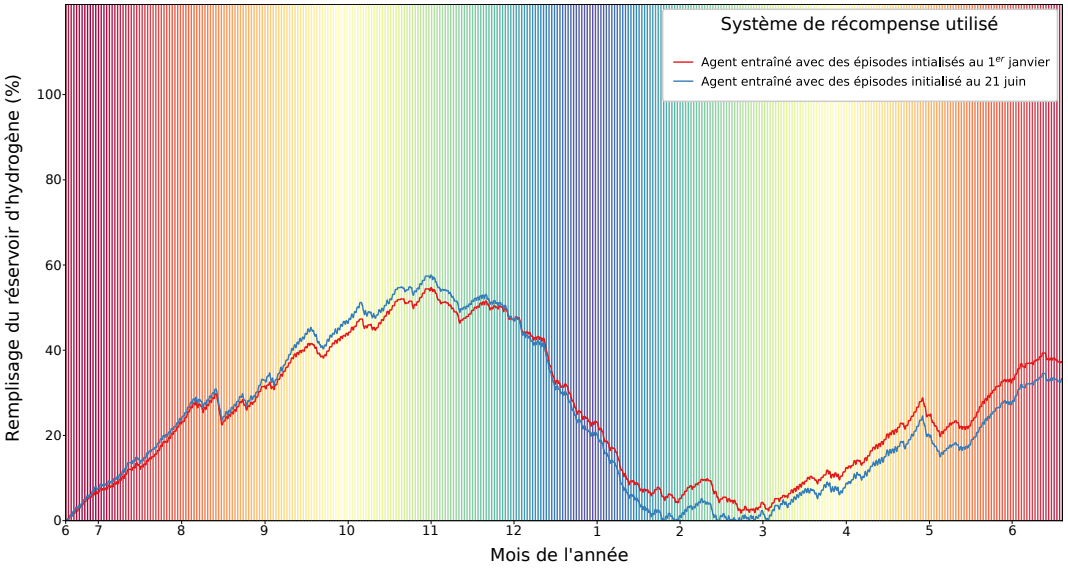

Figure 5.9 – LOH sur un épisode de test d'un an simulé avec épisodes commençant le 21 juin. La stratégie de deux agents est comparée, l'un entraîné sur cet environnement, l'autre non.

Les taux d'autoproduction obtenus par les deux agents sont très proches et sont respectivement de 0.9651 et 0.9690. La stratégie développée semble s'adapter à la date initiale des épisodes.

5.2 Contrôle à long terme avec dégradation de la batterie

L'intérêt de l'utilisation du RL dans le contrôle de micro-réseaux électriques est sa capacité à apprendre des stratégies dans des environnements non-linéaires et stochastiques. Des hypothèses simplificatrices ou des modèles de prédiction ne sont pas nécessaires pour ces algorithmes qui apprennent directement des interactions avec l'environnement. La politique apprise n'est pas forcément optimale, mais elle est adaptative, capable de fonctionner avec des données aléatoires et disparates.
Dans cette section, un niveau de complexité supplémentaire est introduit dans la simulation dynamique des unités pour vérifier l'adaptabilité du RL à un environnement plus réaliste. L'objectif de l'apprentissage de l'agent est ici la recherche d'un compromis entre la dégradation des unités et le coût d'opération du micro-réseau. L'allongement de l'horizon temporel permet d'évaluer plus précisément les coûts d'opération du système, au prix de temps d'apprentissage globalement plus longs. Pour diminuer le temps d'apprentissage, des agents sont entraînés en interaction sur des environnements avec des horizons temporels inférieurs ou égaux à l'horizon temporel de test. Des agents entraînés dans de telles conditions sont capables de s'adapter si leurs scores en phase de test sont aussi bons que ceux des agents entraînés avec de plus longs horizons temporels. Dans ce cas, le temps d'apprentissage pourrait être réduit. La sous-section 5.2.1 présente des modèles de vieillissement de la batterie électrochimique pour les intégrer à la simulation dynamique du micro-réseau. Un modèle linéaire en fonction du nombre de cycles d'utilisation de la batterie, et un autre non-linéaire et fonction de la profondeur de décharge (DOD) sont envisagés. Une étude de l'influence du mode de dégradation sur la politique apprise par l'agent est menée à la sous-section 5.2.2. Des comparaisons entre les stratégies développées par les EMS sont effectuées pour déterminer si la dégradation de la batterie les influence. Inversement, l'impact de la politique de contrôle de l'EMS sur l'usure de la batterie est étudié. Enfin, bien que l'horizon temporel de simulation soit allongé pour observer la dégradation progressive de la batterie, l'apprentissage est accéléré en considérant des épisodes plus courts. La sous-section 5.2.3 présente le développement d'un agent qui, bien qu'entraîné sur un horizon temporel réduit, parvient à maintenir un niveau de performance comparable à celui des agents entraînés sur un horizon correspondant à la phase de test.

5.2.1 Modèles de dégradation de la batterie

Plusieurs approches ont été envisagées pour modéliser la dégradation d'une batterie électrochimique avec plus ou moins de précisions : L'absence de dégradation, la dégradation linéaire en fonction du nombre de cycles et la dégradation non-linéaire en lien avec la profondeur de décharge.

Objectifs et hypothèses. Contrairement aux sections précédentes, où les contraintes sur le SOC de la batterie sont appliquées explicitement pour la maintenir hors de situations de charge et de décharge excessives, l'approche adoptée ici est différente. Le SOC n'est plus contraint mais l'agent reçoit une pénalité supplémentaire pour l'usure de la batterie engendrée par ses charges et décharges excessives. Cette pénalité dépend de la capacité dégradée et elle est indexée sur le coût de remplacement de la batterie.
Les actions de l'agent n'influencent pas directement le SOC, mais la charge ou la décharge du stockage hydrogène conditionnent le flux énergétique entrant ou sortant de la batterie à chaque instant. En apprenant le lien entre ses actions et le SOC de la batterie, l'agent pourrait minimiser les pénalités liées à la dégradation de la batterie. Ceci pourrait permettre de tirer un meilleur parti de la flexibilité du stockage d'hydrogène, et de réduire l'impact sur la dégradation de la batterie, tout en minimisant les coûts. Cette hypothèse sera étudiée

dans cette section. L'impact de la température sur la dégradation de la batterie ainsi que sa dégradation naturelle sont négligés.

Limites des modélisations existantes. De nombreux modèles de l'usure d'une batterie lithium-ion sont utilisés dans la littérature. Fallahifar et al., 2023 calcule le SOC moyen d'une batterie sur un intervalle de temps pour déterminer la dégradation associée. le calcul du SOC moyen sur un intervalle de temps rend les décisions séquentielles de l'agent dépendantes les unes des autres, ce qui n'est pas adapté avec l'approche par RL considérée. Maheshwari et al., 2020 utilise la puissance de décharge de la batterie pour déterminer sa dégradation. La dégradation en fonction de la puissance de décharge est propre à chaque batterie et ne peut être connue qu'après des tests en laboratoire. Une batterie se décharge à 1 C si elle peut se décharger complètement en une heure. Le terme 'C' fait référence à la capacité nominale de la batterie en Ah ou mAh, et indique que le courant de décharge est équivalent à cette capacité par heure. Deux points de mesure étant disponibles (l'un pour une décharge à 1 C, l'autre à 2 C), la courbe de dégradation en fonction de la puissance de décharge est supposée linéaire entre 0 C et 1 C et entre 1 C et 2 C. La puissance de charge et de décharge évolue selon la capacité de la batterie, elle est donc spécifique au dimensionnement de la batterie utilisée dans ces travaux. Pour cette raison, cette approche n'est pas utilisable pour le contrôle du micro-réseau. La dégradation en fonction du DOD, mesurée en laboratoire pour des batteries Lithium Nickel Manganèse Cobalt, est présentée par Maheshwari et al., 2018. Cette approche donne une valeur de dégradation absolue (en mAh) pour le passage d'un SOC à un autre, ce qui dépend encore de la capacité nominale de la batterie.
Les principales problématiques liées à la modélisation du vieillissement de la batterie sont dues à la nature empirique des données. Les résultats présentés en terme de capacité dégradée ne peuvent être généralisés qu'aux batteries spécifiquement testées en laboratoire. La valeur de dégradation du modèle construit est en Ah et dépend de la capacité nominale de la batterie. Un modèle généralisable doit exprimer la dégradation en pourcentage de la capacité nominale de la batterie pour une technologie et une configuration donnée.

Approche considérée. La modélisation du micro-réseau va intégrer la dégradation de la batterie à travers un modèle d'usure relatif à la capacité nominale $\mathrm{E}_{\mathrm{nom}}^{\mathrm{batt}}$ de la batterie. L'usure s'exprime alors en pourcentage et dépend de l'usage de la batterie. Les modèles de dégradation utilisés permettent de calculer la capacité restante $\mathrm{E}_{\mathrm{max}}^{\mathrm{batt}}$ de la batterie selon la forme générale donnée par l'Équation 5.2.

$$\mathrm{E}_{\mathrm{max}}^{\mathrm{batt}}(X) = \mathrm{E}_{\mathrm{nom}}^{\mathrm{batt}}(1 - \delta(X)) \tag{5.2}$$

X est une variable propre à l'usage de la batterie. Elle peut être le nombre de cycles d'utilisation, le SOC ou le DOD. Des fonctions de dégradation $\delta(X)$ seront employées pour chaque modèle considéré.

Modèle de dégradation linéaire. Le modèle employé pour représenter la dégradation linéaire de la batterie est basé sur le nombre de cycles (Le et al., 2023). Un cycle correspond à une charge complète de la batterie suivie d'une décharge complète. La dégradation δ_{lin} peut s'écrire selon l'Équation 5.3.

$$\delta_{\mathrm{lin}}\left(\mathrm{n}^{\mathrm{cycle}}\right) = A_{\mathrm{lin}}^{\mathrm{deg}} \times \mathrm{n}^{\mathrm{cycle}} \tag{5.3}$$

Avec $\mathrm{n}^{\mathrm{cycle}}$ le nombre de cycles à un instant t et $A_{\mathrm{lin}}^{\mathrm{deg}} = \frac{30\%}{\mathrm{n}_{\mathrm{max}}^{\mathrm{cycle}}}$, $\mathrm{n}_{\mathrm{max}}^{\mathrm{cycle}}$ étant le nombre de cycles avant d'atteindre l'état de dégradation maximale de la batterie. L'état de dégradation maximale d'une batterie lithium ion stationnaire est de 30% (voir Chapitre 4). Le taux de dégradation par cycle employé dans cette étude est calculé à partir de l'étude de Johnen et al., 2021, qui évalue une dégradation de 20% pour 3000 cycles. $\mathrm{n}_{\mathrm{max}}^{\mathrm{cycle}}$ est donc pris égal à 4500. Puisque la relation est linéaire, la dégradation peut se déterminer pour toute portion

de cycle. Par exemple, une charge de 20% de la capacité de la batterie suivie d'une décharge de 20% correspond à 0.2 cycle. Une charge seule de 30% équivaut à effectuer 0.15 cycle. Les dégradations associées se calculent avec l'Équation 5.3.

Modèle de dégradation non-linéaire. Le modèle de dégradation non-linéaire retenu dépend du DOD. Les formes les plus couramment adoptées pour représenter la capacité dégradée par rapport à la capacité nominale de la batterie sont exponentielle, polynomiale et sigmoïdale (Johnen et al., 2021). Les coefficients des fonctions de dégradation relatives sont déterminés par des tests en laboratoire. Comme des données adaptées sur ces coefficients sont publiés par Fallahifar et al., 2023 pour la batterie $LiFePO_4$, ce modèle est choisi pour cette étude. La fonction de dégradation de la batterie s'exprime en fonction du DOD selon l'Équation 5.4.

$$\deg\left(\mathrm{DOD}\right) = \left(A_{\mathrm{nl}}^{\mathrm{deg}}\mathrm{DOD}^2 + B_{\mathrm{nl}}^{\mathrm{deg}}\mathrm{DOD}\right)\mathrm{E}_{\mathrm{nom}}^{\mathrm{batt}} \tag{5.4}$$

$A_{\mathrm{nl}}^{\mathrm{deg}}$ et $B_{\mathrm{nl}}^{\mathrm{deg}}$ valent respectivement $4.83\,\mathrm{e}^{-4}$ et $2.38\,\mathrm{e}^{-5}$ (Fallahifar et al., 2023). La dégradation de la batterie δ_{nl} se détermine en décharge, donc lorsque le DOD augmente avec le temps, par l'Équation 5.5.

$$\delta_{\mathrm{nl}}\left(\mathrm{DOD(t+1)}\right) = \delta_{\mathrm{nl}}\left(\mathrm{DOD(t)}\right) + \int_{\mathrm{DOD(t)}}^{\mathrm{DOD(t+1)}} \deg(x)dx \tag{5.5}$$

Comparaison des modèles de dégradation. La dégradation de la capacité de la batterie est calculée selon les deux modèles grâce à une simulation du contrôle du micro-réseau. L'EMS utilisé pour cette simulation provient de la section 5.1. La simulation se déroule sans tenir compte de la dégradation de la batterie. Après simulation, les modèles calculent la dégradation potentielle à chaque instant, en se basant sur le nombre de cycles et le DOD pour les modèles linéaire et non-linéaire respectivement. Les capacités dégradées sont comparées. 298.5 cycles ont été observés avec la simulation. Le modèle de dégradation linéaire a conduit à une dégradation finale, δ_{lin}, de 1.99%. Selon cette estimation, la batterie atteindrait une dégradation de 30%, et serait ainsi remplacée, après une durée de 15 ans. Le modèle non-linéaire a déterminé une dégradation significativement plus élevée, de $\delta_{\mathrm{nl}} = 5\%$ en un an, suggérant un besoin de remplacement de la batterie après seulement 6 ans.

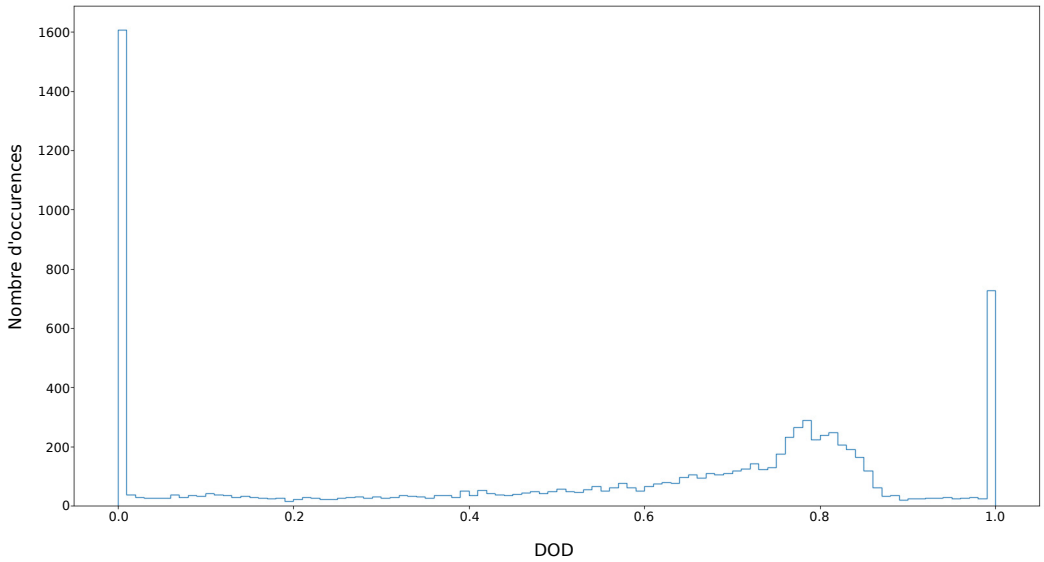

Figure 5.10 – Distribution des DOD calculés pendant la simulation sur un épisode d'un an.

Ajustement du coefficient de dégradation linéaire. Afin de comprendre la différence entre les valeurs de dégradations calculées par les deux modèles, la distribution des valeurs de DOD lors de la simulation est présentée sur la Figure 5.10. La distribution montre que la batterie est régulièrement chargée ou déchargée au maximum. Lorsque ce n'est pas le cas, les DOD sont souvent supérieurs à 0.6. Les phénomènes de charge et de décharge sont d'une intensité significative, ce qui ne contribue pas à préserver la durabilité de la batterie selon le modèle non-linéaire de dégradation.
Le coefficient linéaire $A_{\text{lin}}^{\text{deg}}$ du modèle de dégradation de la batterie semble sous-estimé. Cette situation peut provenir de l'hypothèse selon laquelle le modèle linéaire a été estimé en considérant que la batterie est utilisée de manière à préserver sa durée de vie. Afin d'examiner cette hypothèse, un calcul de dégradation est effectué avec l'Équation 5.4 pour une valeur faible de DOD (0.35). En utilisant ce résultat fixe, deg(0.35), dans l'Équation 5.5, la dégradation est linéaire selon le nombre de cycles. Une dégradation de la batterie de 2.1% sur un an est obtenue. Ce résultat est proche de la dégradation déterminée par le modèle linéaire sur la simulation. Par conséquent, le coefficient $A_{\text{lin}}^{\text{deg}}$ est trop faible par rapport à l'utilisation de la batterie sans contrainte de SOC. Ce coefficient a été réajusté pour correspondre à l'utilisation moins préservante de la batterie par l'agent.
Le nouveau taux se rapproche de celui utilisé dans l'étude de Le et al., 2023, dont la batterie se dégrade de 2.9% en 300 cycles. Un taux arrondi à 1% pour 100 cycles est finalement retenu, soit $\text{n}_{\text{max}}^{\text{cycle}} = 3000$.

5.2.2 Influence du vieillissement de la batterie sur la politique apprise

L'effet de la modélisation de la dégradation de la batterie sur le comportement de l'agent est étudié dans cette sous-section. Un horizon temporel de 10 ans est choisi pour simuler un impact significatif de la dégradation de la batterie sur le comportement de l'agent. Un état observable supplémentaire relatif à la dégradation de la batterie est introduit, ainsi qu'une récompense négative additionnelle liée à cette dégradation. Ces éléments fournissent à l'agent plus d'informations sur la batterie pour mieux adapter sa politique. L'espace d'états $\mathcal{S}$ est alors constitués de 5 éléments :

$$\forall \text{S} \in \mathcal{S},\ \text{S}_\text{t} = \left\{ \text{SOC(t)}, \text{D(t)}, \text{P}^{\text{PV}}\text{(t)}, \text{dist}^{\text{sols}}\text{(t)}, \delta_s^{\text{batt}}\,\text{(t)} \right\} \tag{5.6}$$

Le terme δ_s^{batt} est l'état correspondant à la dégradation de la batterie. Il est calculé selon l'Équation 5.7 à chaque pas de temps.

$$\delta_s^{\text{batt}}\,\text{(t)} = \frac{\text{E}_{\text{nom}}^{\text{batt}} - \text{E}_{\text{max}}^{\text{batt}}\,\text{(t)}}{\text{E}_{\text{nom}}^{\text{batt}}} \tag{5.7}$$

La récompense négative δ_r^{batt} liée à la dégradation est déterminée selon l'Équation 5.8 à chaque pas de temps.

$$\delta_r^{\text{batt}}\,\text{(t)} = \frac{\text{E}_{\text{max}}^{\text{batt}}\,\text{(t)} - \text{E}_{\text{max}}^{\text{batt}}\,\text{(t-1)}}{\text{E}_{\text{nom}}^{\text{batt}}\,\text{(t)}} = -\left(\delta_s^{\text{batt}}\,\text{(t)} - \delta_s^{\text{batt}}\,\text{(t-1)}\right) \tag{5.8}$$

Cadre de l'étude. Afin de garantir la décorrélation des données d'entraînement et de test, la production PV de la région d'Albi entre 2010 et 2019 est utilisée pour l'entraînement, alors que pour le test, la production PV de Toulouse entre 2005 et 2020 est utilisée. Les données employées pour le test ne suivent pas une séquence chronologique continue. Elles incluent les années 2005 à 2009 et l'année 2020, qui ne figurent pas dans l'entraînement.
Dans cette étude, des agents sont entraînés dans des environnements où la dégradation de la

batterie est modélisée de manière linéaire (agent A) et non-linéaire (agent B). Les agents sont ensuite testés dans les deux environnements. Cette approche permet de comparer la politique apprise par les agents dans ces différents contextes afin de déterminer l'impact de la dégradation de la batterie et de sa modélisation sur le comportement de l'agent.

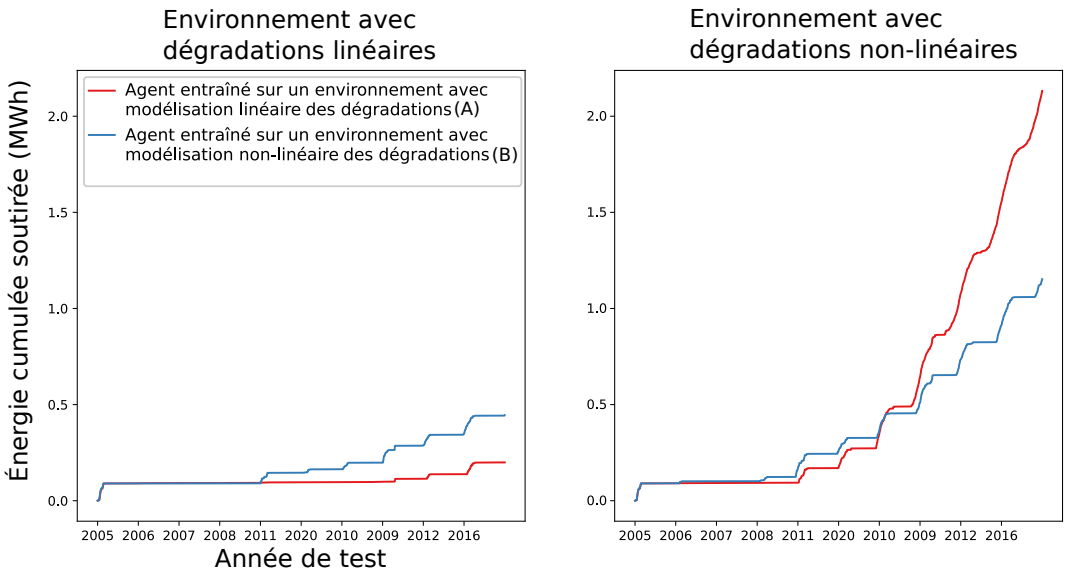

Figure 5.11 – Énergie soutirée au réseau central dans un environnement pour lequel la batterie se dégrade de manière linéaire (à gauche) et non-linéaire (à droite) en fonction de l'environnement d'entraînement des agents.

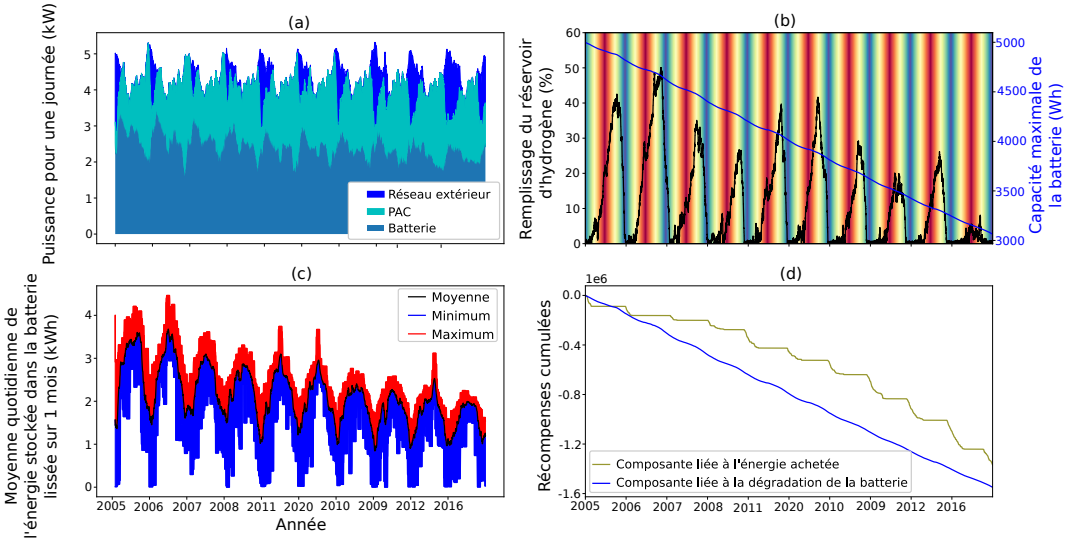

Figure 5.12 – Visualisation d'indicateurs pour le contrôle par un agent entraîné sur 10 ans. (a) Distribution quotidienne de l'origine de l'approvisionnement en réponse à une demande nette négative. (b) Quantité d'hydrogène stocké et capacité maximale de la batterie. (c) Indicateurs statistiques sur la moyenne quotidienne du SOC sur une période plus longue. (d) Décomposition des pénalités reçues par l'agent.

Énergie soutirée au réseau central. Les résultats sont illustrés à la Figure 5.11. Dans l'environnement avec dégradation linéaire de la batterie (à gauche), l'agent A, entraîné dans un environnement similaire, soutire moins d'énergie que l'agent B. La quantité d'électricité soutirée est plus importante pour l'agent B à partir de la cinquième année. Cet écart se creuse progressivement. Dans l'environnement incluant une modélisation non-linéaire de la dégradation de la batterie (à droite), l'agent B soutire moins d'électricité que l'agent A dès la troisième année. L'énergie soutirée est plus importante dans l'environnement avec une dégradation non-linéaire de la batterie.
En se basant sur ces observations, il est possible de conclure que la manière dont la dégradation de la batterie est modélisée a un impact sur le comportement appris par l'agent. De plus, l'agent entraîné dans un environnement avec une modélisation de la dégradation identique à celle de l'environnement de test obtient sytématiquement de meilleurs scores. La politique apprise s'adapte au modèle de dégradation de la batterie. Pour la suite de l'étude, la modélisation non-linéaire de la dégradation de la batterie est adoptée afin de vérifier la capacité de l'algorithme à garantir une politique de contrôle correcte dans un environnement plus complexe.

Analyse de la stratégie développée avec dégradation de la batterie. Pour comprendre la stratégie développée par l'agent et identifier les facteurs influents, une visualisation de l'évolution de plusieurs indicateurs est montrée sur la Figure 5.12. Les résultats affichés sont obtenus suite à une simulation sur 10 ans avec une dégradation non-linéaire de la batterie en fonction du DOD. La visualisation (a) affiche la provenance des flux énergétiques qui équilibrent la demande nette négative pour chaque jour de l'année. Elle montre que la contribution de la batterie à l'approvisionnement de la demande nette diminue avec le temps. Les pics observés montrent que sa contribution est plus forte pendant une période de l'année, mais leur amplitude diminue au cours de la simulation. Cette diminution est visiblement en partie comblée par le soutirage au réseau central, qui augmente avec le temps. La visualisation (b) représente la quantité de H_2 stocké et la capacité maximale de la batterie. La quantité de H_2 stocké diminue progressivement à mesure que la batterie se dégrade. Cette tendance sera analysée à la section 5.3. La visualisation (c) montre les valeurs maximales, minimales et moyennes de l'énergie quotidienne stockée dans la batterie lissée sur un mois. La batterie est régulièrement vide et l'énergie stockée au maximum diminue progressivement. Enfin, la visualisation (d) compare les deux composantes des récompenses reçues par l'agent. La pénalité liée à la dégradation de la batterie devient plus importante que la pénalité de soutirage d'électricité au réseau central de distribution dès la deuxième année de simulation. L'EMS n'adopte pas la même stratégie de contrôle si la batterie se dégrade ou non. Malgré l'ajustement du coefficient linéaire de dégradation de la batterie pour calibrer les résultats du modèle linéaire avec ceux du modèle non-linéaire, une simulation longue montre que le choix du modèle a un impact significatif sur la stratégie apprise. La quantité d'électricité soutirée au réseau central augmente significativement après quelques années de simulation lorsque le modèle de dégradation diffère de celui utilisé en entraînement comparé à la politique développée avec ce même modèle.

5.2.3 Choix d'un horizon temporel d'apprentissage

L'influence de l'horizon temporel des épisodes d'entraînement sur la politique de l'agent est ici étudiée. L'objectif principal est d'identifier si un horizon temporel d'entraînement plus court pourrait être utilisé pour élaborer une politique de contrôle efficace sur une période de 10 ans, avec l'ambition de réduire les temps de calcul de l'apprentissage.

Méthodologie. Pour atteindre cet objectif, plusieurs agents seront entraînés dans des environnements où la dégradation de la batterie est modélisée de manière non-linéaire. Chaque agent sera entraîné avec un horizon temporel différent. Les politiques apprises seront ensuite testées sur un environnement indépendant, dont l'horizon temporel s'étend sur 10 ans. Les

données de productible PV sont les mêmes que pour les environnements de la sous-section 5.2.2. La performance des politiques sera évaluée en se basant sur la quantité d'énergie soutirée du réseau central.

Premiers résultats. Les résultats observés selon l'horizon temporel d'entraînement sont montrés dans la Table 5.4.
L'énergie soutirée au réseau central est nettement inférieure pour un agent entraîné sur un horizon temporel de 10 ans. Dans cette configuration, il est préjudiciable de réduire l'horizon temporel en phase d'apprentissage tout en conservant un taux d'autoproduction acceptable.

Horizon temporel d'entraînement (en années)	5	6	7	8	9	10
Énergie soutirée (MWh)	3	2.1	1.5	1.8	1.5	1.2

Table 5.4 – Énergie soutirée sur un épisode de test sur 10 ans selon l'horizon temporel de l'entraînement des agents.

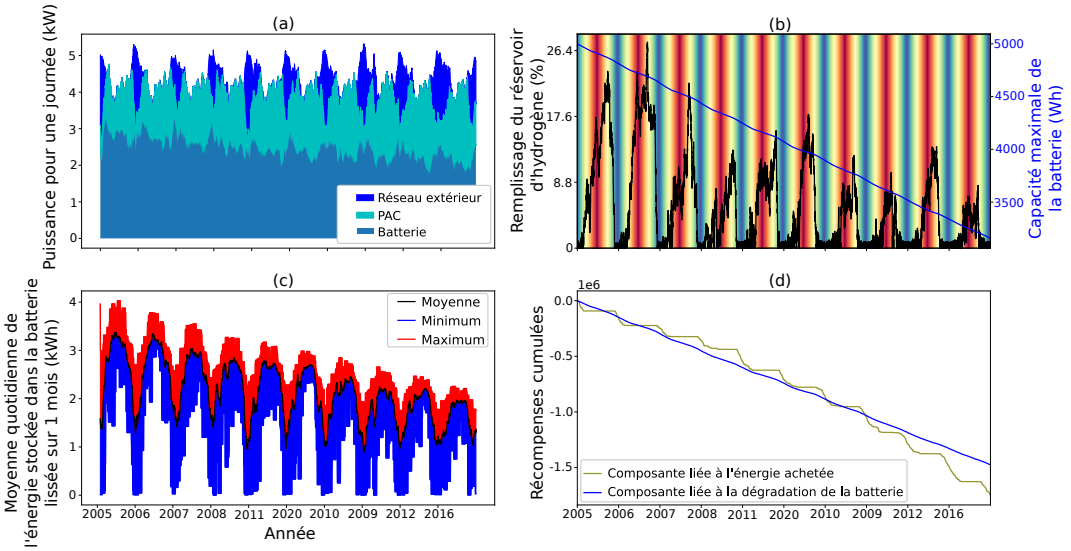

Figure 5.13 – Visualisation d'indicateurs pour le contrôle par un agent entraîné sur 10 ans avec un espace d'états à 6 dimensions. (a) Distribution quotidienne de l'origine de l'approvisionnement en réponse à une demande nette négative. (b) Quantité d'hydrogène stocké et capacité maximale de la batterie. (c) Indicateurs statistiques sur la moyenne quotidienne de l'énergie stockée dans la batterie sur une période plus longue. (d) Décomposition des pénalités reçues par l'agent.

Ajout d'un état supplémentaire. Pour remédier à ce problème, l'état de charge du réservoir d'hydrogène (H_2), également appelé *Level of Hydrogen* (LOH), est intégré à l'espace d'états de l'agent. L'hypothèse sous-jacente est que cela permettrait à l'agent de discerner des motifs annuels directement liés à ses actions pour les reproduire sur les années additionnelles de l'horizon temporel de test. L'introduction d'un nouvel état peut nuire à l'apprentissage d'un agent. Pour s'assurer que la qualité de la politique apprise ne s'est pas dégradée, des agents sont entraînés puis testés selon les mêmes configurations que précédemment. Leur espace d'états est décrit par l'Équation 5.9.

$$\forall \mathrm{S} \in \mathcal{S},\ \mathrm{S_t} = \left\{ \mathrm{SOC(t)}, \mathrm{D(t)}, \mathrm{P^{PV}(t)}, \mathrm{dist^{sols}(t)}, \delta_s^{\mathrm{batt}}\,(\mathrm{t})\,, \mathrm{LOH}\,(\mathrm{t}) \right\} \tag{5.9}$$

La Figure 5.13 présente les différentes visualisations d'indicateurs liés au contrôle du micro-réseau par un agent entraîné et testé sur un horizon temporel de 10 ans. On observe des différences avec la Figure 5.12 (agent avec 5 états observables). L'énergie soutirée au réseau central de distribution augmente et la dégradation de la batterie est plus faible (Figure 5.13 (a) et (d)). La récompense négative reçue pour l'énergie soutirée devient plus importante que celle liée à la dégradation de la batterie. Sur la Figure 5.13 (b), on observe que les courbes de la quantité de H_2 stocké sont moins lisses et les pics sont de plus faible amplitude. La présence de bruit dans les actions de charge et de décharge empêche une accumulation de l'énergie dans le réservoir. L'aspect bruité de l'énergie stockée résultant des actions de l'agent suggère une convergence vers une politique qui semble différente des stratégies de contrôle précédemment développées. L'ajout d'une dimension à l'espace d'états complexifie l'entraînement. L'agent doit prendre en compte une information supplémentaire à chaque itération. Les différences dans la forme du stockage H_2 pourraient provenir d'un entraînement instable.

Ajustement des paramètres d'apprentissage. Sous l'hypothèse que l'ajout d'un état supplémentaire nuit à la convergence de l'apprentissage de l'agent avec les paramètres utilisés, une nouvelle étude paramétrique a été menée. Les résultats en phase de validation sont meilleurs pour une fréquence d'entraînement de $t_{train} = 4$, un facteur d'actualisation de $\gamma = 0.99$, avec une taille de batch de 64. La Figure 5.14 montre les visualisations associées au test de l'agent sur 10 ans après la modification des paramètres.
Les pénalités, énergie soutirée et dégradation de la batterie, sont inférieures (Figure 5.14 (d)) à celles de l'agent entraîné avec 5 états observables. La batterie est légèrement plus dégradée qu'avant l'ajustement des paramètres d'apprentissage mais l'énergie soutirée au réseau central est bien plus faible. La Figure 5.14 (a) montre que le réseau central n'approvisionne pas le micro-réseau sur une année complète. Enfin, les actions de l'agent sur le stockage d'hydrogène semblent plus lisses et la quantité d'hyrogène stocké est trois fois plus élevée (Figure 5.14 (b)). On remarque toutefois que l'amplitude des pics de stockage d'énergie l'été diminue avec le temps de simulation.

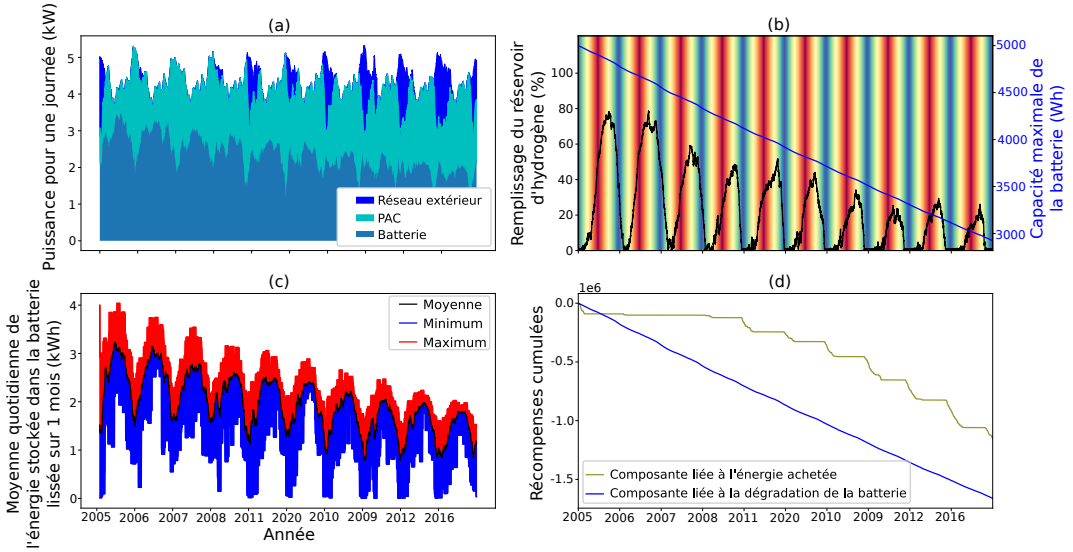

Figure 5.14 – Visualisation d'indicateurs pour le contrôle par un agent entraîné sur 10 ans avec un espace d'états à 6 dimensions après modification des paramètres. (a) Distribution quotidienne de l'origine de l'approvisionnement en réponse à une demande nette négative. (b) Quantité d'hydrogène stocké et capacité maximale de la batterie. (c) Indicateurs statistiques sur la moyenne quotidienne du SOC sur une période plus longue. (d) Décomposition des pénalités reçues par l'agent.

Étude de l'effet de l'horizon temporel après ajustement des paramètres pour 6 états observables. Les agents ainsi paramétrés sont entraînés sur des environnements avec divers horizons temporels de simulation. La dégradation des batteries et les quantités d'énergie soutirée au réseau central de distribution obtenues en test sont affichées sur la Figure 5.15. Les résultats pour le soutirage au réseau central s'avèrent similaires pour les agents entraînés sur des horizons temporels de 6 ans et plus. L'agent entraîné avec un horizon temporel de 3 ans présente des performances inférieures pour l'énergie soutirée au réseau central (à gauche de la Figure 5.15). L'agent entraîné sur 5 ans produit des résultats qui semblent similaires à ceux des agents formés sur 6 et 10 ans lors des premières années de simulation. Ces résultats se dégradent en fin de la simulation.
À droite de la figure 5.15, on observe que l'état final de la batterie est plus dégradé pour les agents entraînés sur 5 ans et 10 ans. Aucun lien clair entre l'horizon temporel de simulation d'entraînement et la politique de préservation de la batterie n'a pu être établi.
Afin de réduire le temps de calcul, dans les sections suivantes les agents seront entraînés avec un horizon temporel de simulation de 6 ans pour des tests sur 10 ans.

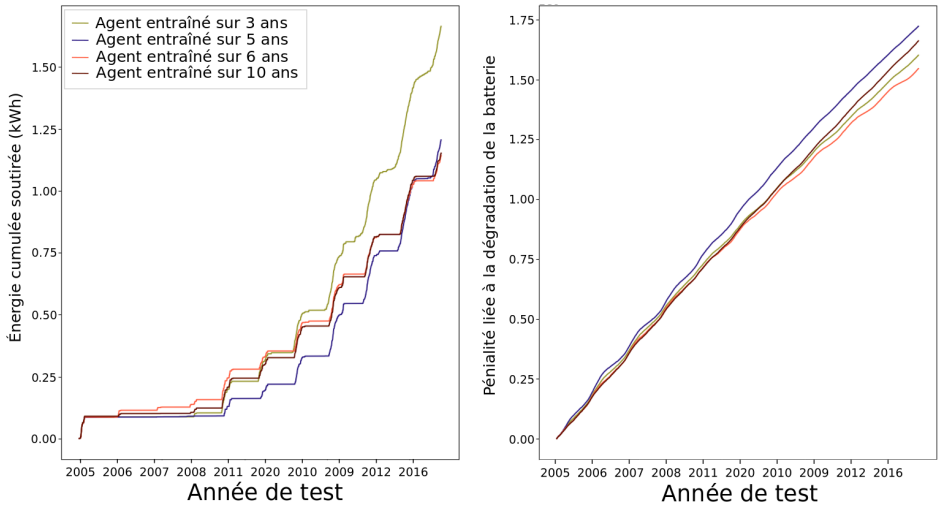

Figure 5.15 – Électricité soutirée au réseau central de distribution (à gauche) et capacité de la batterie (à droite) selon l'horizon temporel d'entraînement des agents.

5.3 Analyse de la stratégie de contrôle et évaluations avec des indicateurs clés

L'analyse de la politique de contrôle développée par l'agent et ses résultats sont étudiés dans cette section. Les horizons temporels d'entraînement et de test sont toujours de 6 et 10 ans respectivement. La configuration du micro-réseau évolue selon les besoins des analyses menées. L'étude de cas principale considérée est le contrôle d'un micro-réseau équipé de 5 kWc de panneaux PV et d'une batterie d'une capacité de 5 kWh se dégradant de manière non-linéaire selon le DOD. Les données de demande sont celles d'un français ayant souscrit à un abonnement d'une puissance inférieure à 6 kVA. La demande annuelle vaut 2220 kWh. Les résultats d'opération sont analysés de manière à comprendre la stratégie développée par l'agent sur la gestion du stockage long terme. Cette analyse qualitative s'appuie sur des visualisations de variables lors des épisodes de tests de l'agent. Des analyses quantitatives sont également effectuées grâce au suivi d'indicateurs pour évaluer la performance du contrôle sur plusieurs axes.

5.3.1 Étude des politiques de contrôle et analyse de leur performance

Le déploiement de l'algorithme de DQN a permis de développer une stratégie de contrôle du stockage hydrogène sur 10 ans de simulation. Puisqu'il s'agit d'un algorithme d'apprentissage automatique, l'interprétation directe des paramètres appris est impossible. Des visualisations générées à partir de simulations ont été construites pour faciliter l'analyse.
Ces représentations du comportement dynamique du micro-réseau permettent de formuler des hypothèses sur la politique adoptée par l'agent.

Évolution de la stratégie au cours de la simulation

L'utilisation du stockage hydrogène comme stockage long terme par l'agent a été confirmée à la sous-section 5.2.3. La quantité d'hydrogène stocké atteint des pics en se rechargeant pendant l'été pour satisfaire les demandes hivernales. En revanche, l'amplitude de ces pics diminue avec le temps. Deux hypothèses peuvent être formulées pour expliquer ce phénomène : soit la fréquence d'utilisation de l'électrolyseur diminue, soit celle de la pile à combustible augmente durant les périodes ensoleillées.

Fréquence d'utilisation de l'électrolyseur et de la PAC. Pour écarter ou confirmer ces hypothèses, l'usage de l'électrolyseur et de la PAC est analysé en début et en fin de simulation. La visualisation de leur nombre d'utilisations est présentée sur la Figure 5.16 pour deux années distinctes de test. Les deux années choisies sont 2006 et 2016 et correspondent respectivement à la seconde et dernière années de simulation. La première année, 2005, n'est pas représentative car la simulation démarre avec un stockage H_2 vide. La figure 5.16 montre que l'électrolyseur est autant utilisé en début qu'en fin de simulation. La PAC semble plus sollicitée en fin de simulation, y compris durant les journées ensoleillées.

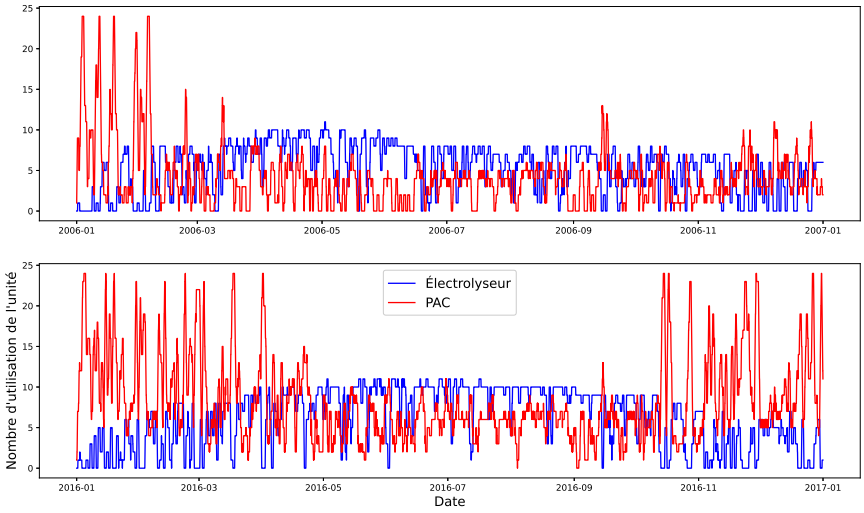

Figure 5.16 – Nombre d'utilisations quotidiennes de l'électrolyseur et de la PAC en 2006 (en haut) et en 2016 (en bas).

Analyse de l'usage du stockage H_2. Pour comprendre l'augmentation de l'utilisation de la PAC en période estivale, les flux énergétiques dans le micro-réseau sont examinés.

Description des figures. Les Figures 5.17 et 5.18 montrent la répartition des flux entrants dans le micro-réseau sur une semaine d'été et d'hiver respectivement au début et en fin de simulation. Ces semaines ont été choisies car la demande nette suit une évolution comparable sur les deux années étudiées. La courbe de demande nette (rouge) correspond à la soustraction de la production PV à la demande. L'influence de la production PV est nettement observable puisque la demande nette est négative en journée. Les flux négatifs indiquent que la puissance sort du micro-réseau, c'est le cas lorsque l'énergie est injectée au réseau central ou lorsque les systèmes de stockage se chargent.
L'usage de l'électrolyseur et de la PAC sont respectivement représentés par des traits verticaux vers le bas et vers le haut. La puissance demandée pour l'électrolyse est supérieure à la puissance produite par l'utilisation de la PAC. En effet, le rendement d'électrolyse influence la quantité d'énergie effectivement stockée mais pas le flux énergétique sortant du micro-réseau. Inversement, le rendement de la PAC diminue le flux énergétique reçu par le micro-réseau par rapport à la puissance sortante du stockage H_2.
L'intensité de couleur de la courbe de puissance de la batterie est inversement proportionnelle au SOC. Pour plus de visibilité, toutes les figures agrandies représentant ainsi les flux énergétiques hebdomadaires sont à disposition en Annexe A.2.

Été. Sur la Figure 5.17, l'électrolyseur semble être utilisé exclusivement lorsque la demande nette est négative et plus particulièrement lorsque la batterie est rechargée. Ce phénomène s'observe sur la zone A de la Figure 5.17. L'électrolyseur est employé lorsque la puissance d'utilisation de la batterie passe de valeurs négatives à zéro. Les conditions d'utilisation de l'électrolyseur semblent moins restrictives en fin de simulation puisqu'il est employé plusieurs fois alors que sa puissance dépasse la demande nette en valeur absolue. Cela provoque des décharges de la batterie pour rétablir l'équilibre (zone B). Le SOC se situe régulièrement à un niveau faible lorsque la demande nette est positive en fin de simulation, ce qui provoque l'utilisation de la PAC (zone C) et fait diminuer la quantité d'hydrogène stocké. L'affaiblissement du niveau du SOC en fin de simulation plutôt qu'au début pourrait être lié à la forte dégradation du système de stockage.

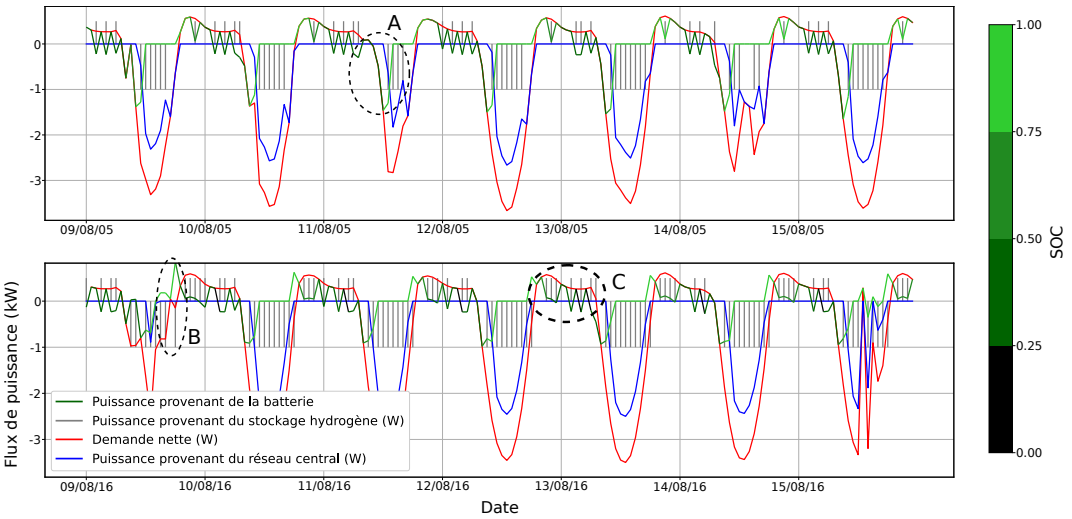

Figure 5.17 – Flux énergétiques du micro-réseau pour une semaine d'août en début (2005) et en fin (2016) de simulation.

Hiver. Les différents flux énergétiques en périodes moins ensoleillées (octobre 2005 et 2016) sont représentés sur la figure 5.18. La visualisation montre que la batterie atteint des faibles SOC pour les deux années (zones A). La PAC est utilisée en soutien de la batterie et intervient comme génération électrique supplémentaire lorsque le SOC de la batterie

est trop faible (zones A). Le SOC est plus fréquemment faible en fin de simulation (2016).

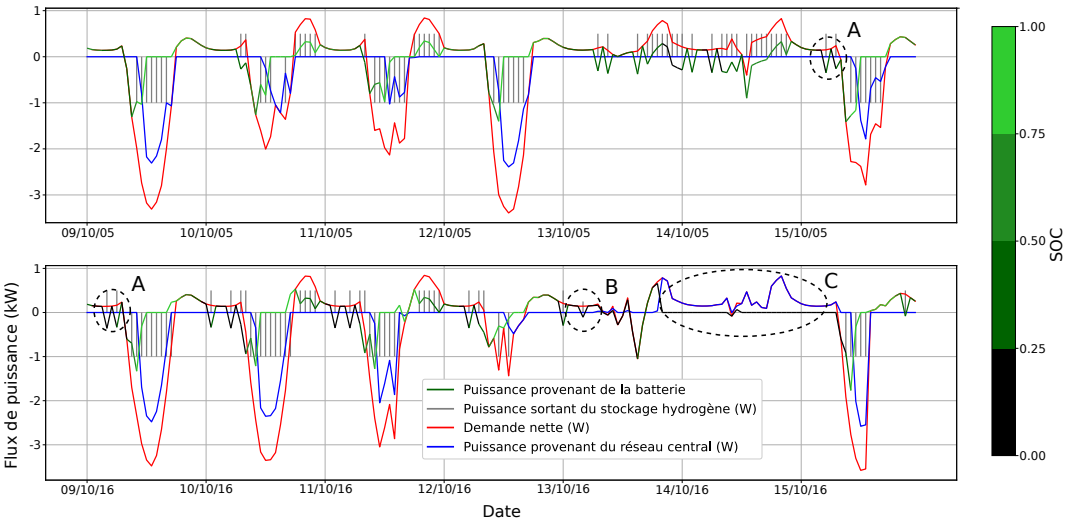

Figure 5.18 – Flux énergétiques du micro-réseau pour une semaine d'octobre en début (2005) et en fin (2016) de simulation.

La zone B de la Figure 5.18 montre une puissance de PAC inhabituellement faible le 13 octobre 2016. Cette situation ne se produit que lorsque le stockage hydrogène atteint une réserve nulle d'après les contraintes de l'utilisation de la PAC définies par les Équations 4.9 et 4.10. Puisque le stockage H_2 est vide, la PAC n'est plus utilisée jusqu'à l'électrolyse suivante (le 15 octobre 2016). C'est le réseau central de distribution qui satisfait la majorité de la demande jusqu'à cette date (zone C).

En fin de simulation, la PAC est employée pour remplacer la batterie dégradée afin de subvenir aux demandes à court terme. L'amplitude des pics d'énergie stockée est réduite à cause de cet usage répété de la PAC pendant des périodes ensoleillées (Figure 5.17). De ce fait, le stockage H_2 se vide rapidement lors des périodes peu ensoleillées alors que la batterie est dégradée. Les besoins de consommation sont donc en grande partie satisfaits par le réseau central (5.18) à la fin de la simulation.
La dégradation trop importante de la batterie empêche le stockage hydrogène de se remplir convenablement en été et intensifie sa fréquence de décharge en hiver. Cette hypothèse peut se vérifier en confrontant la stratégie de l'EMS à celle apprise dans un environnement où la batterie est remplacée lorsqu'elle est usée.

Influence du vieillissement de la batterie sur la stratégie de contrôle

Afin d'évaluer le rôle de la dégradation de la batterie sur la politique apprise par l'EMS, des environnements intégrant différentes caractéristiques de la batterie ont été introduits.
Un apprentissage de l'EMS a été effectué dans un environnement où la batterie est remplacée lorsque son usure dépasse 30%. La visualisation de différents suivis d'indicateurs suite à cet entraînement est présentée sur la Figure 5.19. La politique de contrôle du stockage hydrogène apprise est différente des autres vues précédemment, puisque le LOH n'est jamais nul d'après la Figure 5.19 (b) au contraire de la Figure 5.14 (b). La Figure 5.19 (d) montre que la pénalité reçue pour la dégradation de la batterie est bien plus importante que précédemment.
La batterie semble davantage sollicitée lorsque la simulation inclut son remplacement. Les conséquences secondaires de la dégradation de la batterie sur les pénalités assignées à l'agent sont atténuées car la quantité d'énergie soutirée a significativement diminué.

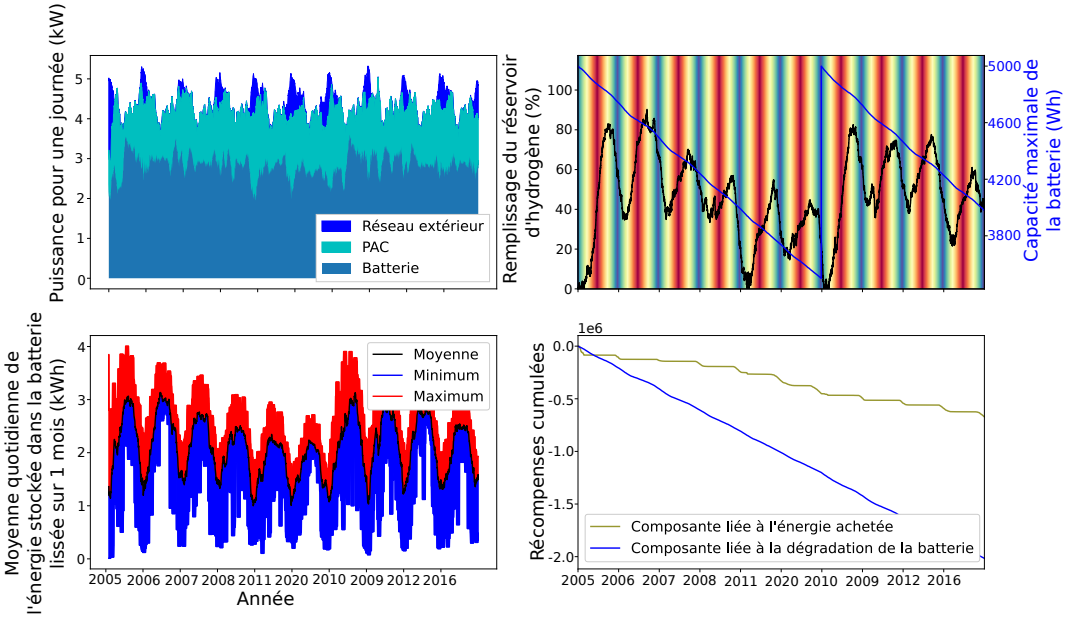

Figure 5.19 – Visualisation de la politique d'un agent entraîné dans un environnement dans lequel la batterie est remplacée lorsque trop usée. (a) Distribution quotidienne de l'origine de l'approvisionnement en réponse à une demande nette négative. (b) LOH et capacité maximale de la batterie. (c) Indicateurs statistiques sur la moyenne quotidienne de l'énergie stockée dans la batterie sur une période plus longue. (d) Décomposition des pénalités reçues par l'agent.

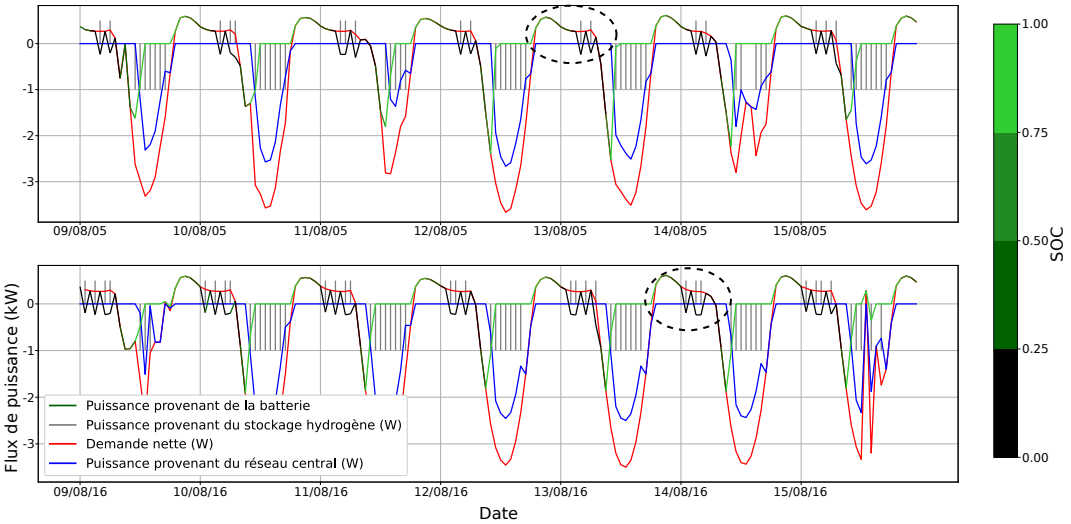

Figure 5.20 – Flux énergétiques du micro-réseau pour une semaine d'août en début (2005) et en fin (2016) de simulation. L'environnement inclut un remplacement des batteries usées.

Analyse des flux énergétiques avec remplacement de la batterie. Les flux énergétiques en simulation pour l'agent ayant appris sur cet environnement sont présentés sur les Figures 5.20 et 5.21. En période ensoleillée (août), la PAC est peu utilisée en début et en fin de simulation (voir zones en pointillés sur la Figure 5.20). Ce comportement ressemble à

la stratégie observée sur la Figure 5.17. La politique de stockage de l'hydrogène sur cette semaine semble dépendre de la capacité de la batterie.
Un comportement nouveau est observé sur la Figure 5.21 : l'agent semble prioriser l'utilisation de la batterie par rapport à celle de la PAC. Ce phénomène est particulièrement visible sur les zones en pointillés de la Figure 5.21. Le SOC de la batterie diminue considérablement lorsque la demande nette est positive, ce qui se traduit par des DOD élevés et donc une usure plus importante de la batterie. Cette moindre utilisation de la PAC explique la raison pour laquelle le LOH n'est presque jamais nul.

Contrainte de SOC. Pour observer la politique adoptée par l'agent sans usure drastique de la batterie, un environnement dont le SOC de la batterie est contraint a été établi. Sur cet environnement, le SOC est contraint entre 30% et 80%. Ainsi, la dégradation de la batterie est limitée (voir section 5.2). L'EMS doit compter sur une capacité et une puissance plus faible du stockage à court terme. Les résultats obtenus après l'apprentissage sont présentés sur la Figure 5.22. La Figure 5.22 (a) montre que la part de la batterie est fréquemment plus faible que celle du stockage hydrogène dans la satisfaction de la demande. L'électricité est toujours soutirée du réseau central de distribution pour équilibrer la production et la demande. La batterie semble sous-dimensionnée dans cette configuration du micro-réseau. La Figure 5.22 (b) illustre une utilisation à court terme du stockage hydrogène. Bien que la batterie soit moins dégradée (Figure 5.22 (c,d)), la viabilité et l'autonomie du micro-réseau sont faibles car une grande quantité de l'énergie consommée est importée. La capacité disponible de la batterie influence la politique de contrôle apprise par un agent. L'intérêt du stockage hydrogène comme stockage long terme semble amoindri à mesure que la puissance transmise par la batterie diminue. L'hydrogène tend alors à se comporter comme un substitut de la batterie et sert de stockage à court terme.

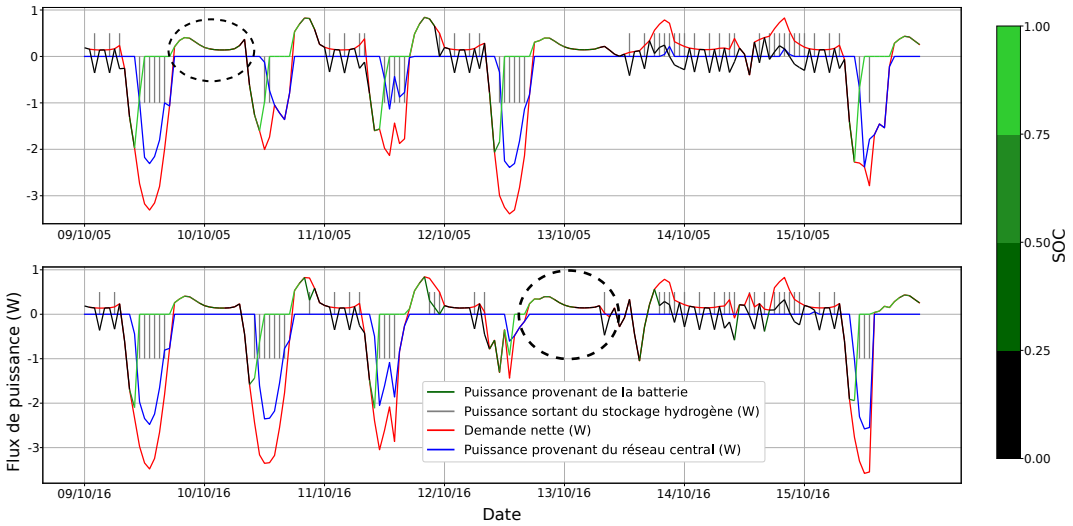

Figure 5.21 – Flux énergétiques du micro-réseau pour une semaine d'octobre en début (2005) et en fin (2016) de simulation. L'environnement inclut un remplacement des batteries usées.

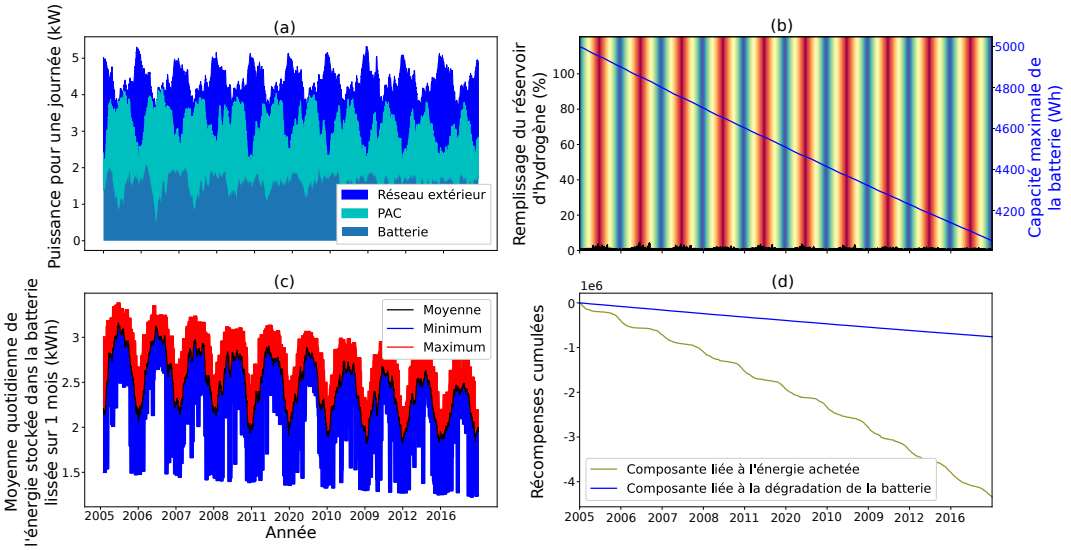

Figure 5.22 – Visualisation de la politique d'un agent entraîné dans un environnement dans lequel le SOC est contraint. (a) Distribution quotidienne de l'origine de l'approvisionnement en réponse à une demande nette négative. (b) LOH et capacité maximale de la batterie. (c) Indicateurs statistiques sur la moyenne quotidienne de l'énergie stockée dans la batterie sur une période plus longue. (d) Décomposition des pénalités reçues par l'agent.

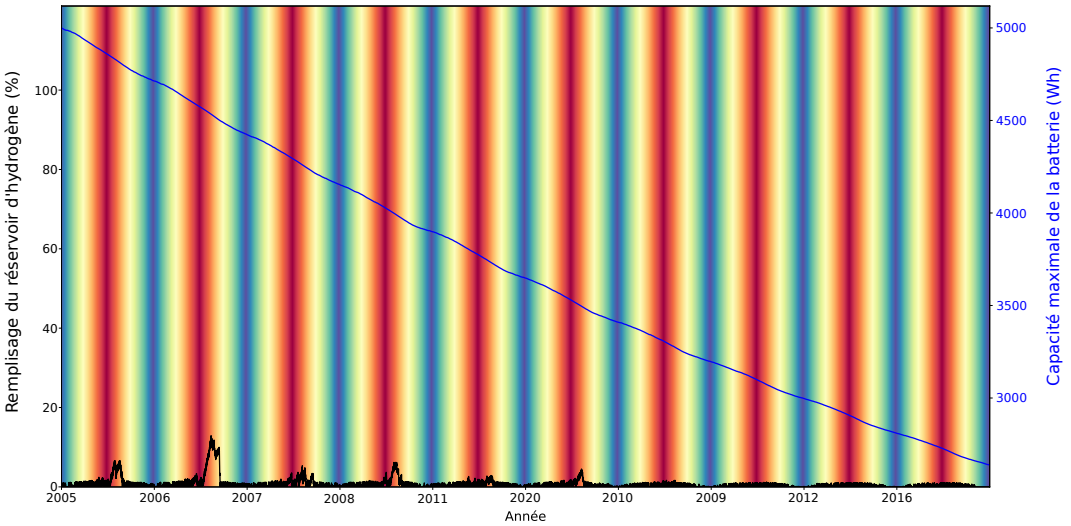

Figure 5.23 – LOH et capacité maximale de la batterie. La somme des données de consommation électriques utilisées est supérieure aux simulations précédentes.

Influence de la demande

On étudie ici l'influence du niveau de demande électrique sur la politique apprise par l'agent lors de son entraînement.
Un entraînement a été réalisé à partir de nouvelles données de consommation. Ces données, fournies par Enedis, correspondent à la consommation moyenne d'un foyer français sur un an, soit 4679 kWh/an. Cette demande est sensiblement supérieure aux consommations utilisées jusqu'à présent (2220 kWh/an). La capacité nominale de la batterie et la puissance

nominale des panneaux PV restent inchangées. Le SOC de la batterie n'est pas contraint, sa valeur est comprise entre 0 et 1. La politique de contrôle apprise par l'agent est présentée sur la Figure 5.23. On peut observer que les pics d'hydrogène stocké sont à nouveau de faible amplitude, avec une valeur maximale de 58,9 kWh. Une diminution de la puissance PV installée entraînerait un effet similaire dans la politique de contrôle. L'intérêt d'un stockage long terme d'énergie dans un micro-réseau dépend ainsi fortement du niveau de demande et/ou des capacités de production et de stockage. Une demande trop importante - ou des productions ou capacités de batteries trop faibles - entraînent donc l'EMS à piloter le stockage hydrogène à un usage court terme.

5.3.2 Indicateurs de la performance

La performance de la stratégie développée peut être évaluée dans plusieurs dimensions, au travers de plusieurs indicateurs pertinents. Les indicateurs proposés ont été introduits aux chapitres 1 et 4. Dans cette section, les résultats sont analysés selon des dimensions économiques, écologiques et techniques.

Indicateurs économiques

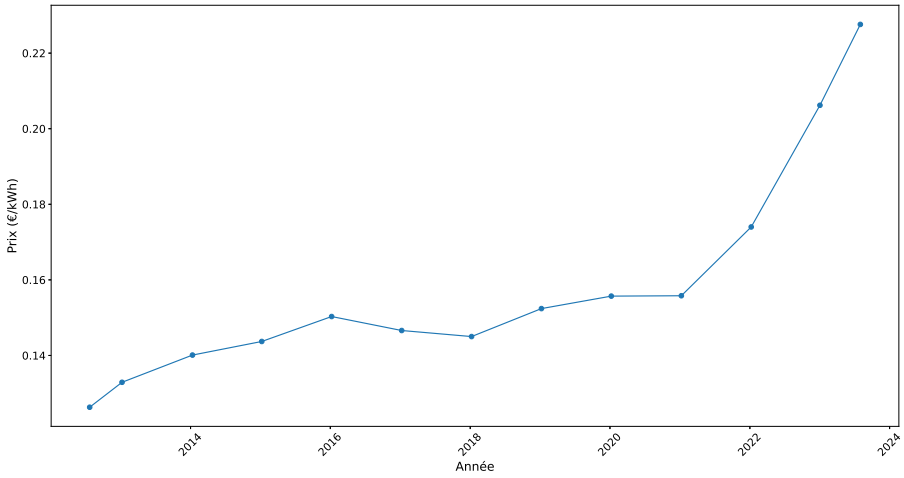

Figure 5.24 – Évolution du prix de l'électricité en France (source des données : EDF).

Le coût nivelé de l'énergie (LCE), introduit avec l'Équation 1.1, permet de déterminer le coût d'un kWh électrique en considérant à la fois le CAPEX et l'OPEX du micro-réseau. Pour l'étude de la viabilité économique du micro-réseau (relié au réseau central), le LCE doit être comparé à une valeur seuil comme par exemple le tarif d'électricité proposé par un fournisseur. Ce tarif influence le coût d'opération du micro-réseau puisqu'il permet de déterminer le coût de l'électricité achetée et vendue. Depuis août 2023, le tarif d'électricité est de 0,2276 €/kWh en France (voir l'évolution de ce tarif ces dix dernières années sur la Figure 5.24). Le tarif de revente d'électricité est de 0,1339 €/kWh pour un particulier en autoconsommation avec une installation PV de puissance inférieure à 9 kWc (source : EDF ENR). L'Équation 5.10 donne la formule de calcul du LCE. Seules la demande, l'injection et le soutirage d'électricité sont fonction du temps et vont dépendre du fonctionnement du

micro-réseau.

$$\mathrm{LCE} = \sum_{t=0}^{T} \frac{C^{comp} + P_{nom}^{elec} C_{inv}^{elec} + P_{nom}^{PAC} C_{inv}^{PAC} + P_{nom}^{PV} C_{inv}^{PV} + P_{nom}^{batt} C_{inv}^{batt} + E_{sout}(t) C_{sout} - E_{inj}(t) C_{inj}}{D(t)} \tag{5.10}$$

T est l'horizon temporel d'opération du micro-réseau.

Résultats. Le LCE moyen obtenu sur un horizon de 10 ans, avec dégradation non-linéaire de la batterie et sans remplacement est de 0,2758 €/kWh. Il est significativement meilleur que le LCE obtenu avec un algorithme basé sur des règles, pour lequel il est de 0,4123 €/kWh. Cependant, il reste supérieur au tarif du réseau centralisé. Le micro-réseau n'est donc pas attractif de ce point de vue. La répartition des coûts dans le calcul du LCE est représentée sur la Figure 5.25 pour les deux algorithmes.
On observe que la part de l'achat d'électricité au réseau central est très faible lorsque le DQN est déployé. L'opération du micro-réseau génère davantage de revenus grâce à l'injection d'électricité qu'elle n'engendre de coûts dus au soutirage. Lorsque le revenu d'injection et le coût de soutirage sont négligés, le LCE calculé vaut 0,2833 €/kWh.

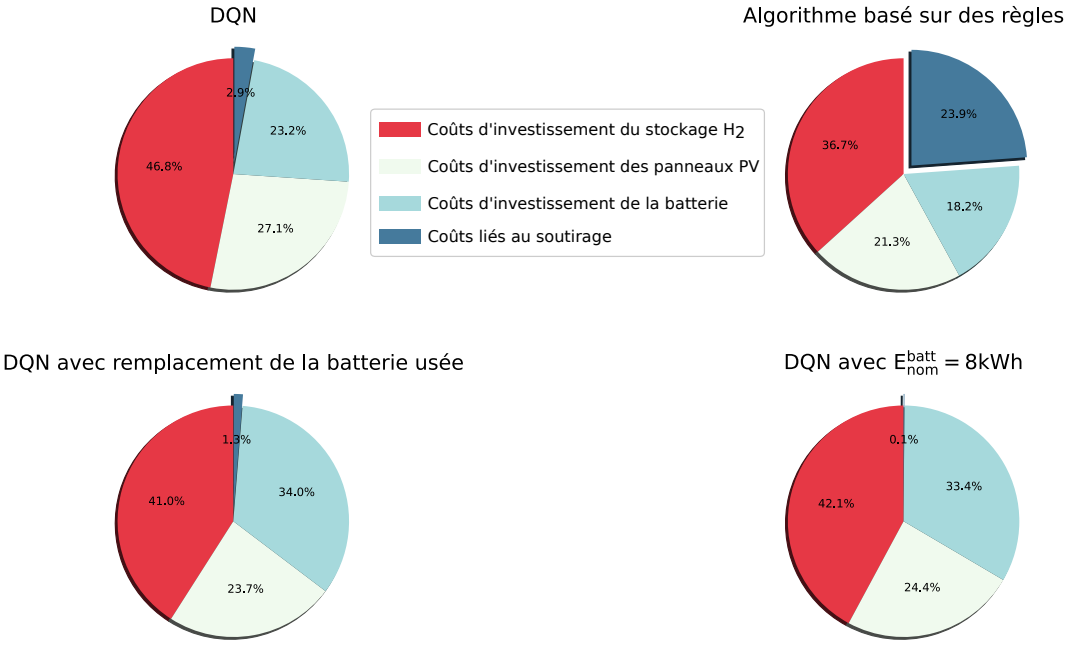

Figure 5.25 – Répartition des coûts dans le calcul du LCE.

Remplacement de la batterie. Dans le cas où la batterie est remplacée, la part de l'achat d'électricité au réseau central diminue dans le calcul du LCE, mais l'investissement lié à la batterie augmente. Le LCE moyen vaut 0,31780 €/kWh.

Accroissement de la capacité de la batterie. Lorsque la capacité de la batterie est plus élevée, la part de l'électricité soutirée devient négligeable dans les coût totaux mais l'investissement lié à l'achat de la batterie est plus conséquent. Le LCE moyen est de 0,2830 €/kWh.

Accroissement de la demande. Lorsque la consommation électrique augmente, le LCE diminue mais la part de l'énergie soutirée au réseau central augmente. Le LCE vaut 0,2198 €/kWh en moyenne. En revanche, le stockage hydrogène ne sert que de substitut à la batterie (voir sous-section 5.3.1). L'investissement majoritaire étant lié au stockage

hydrogène, il pourrait être remplacé par une batterie de capacité plus haute, pour un coût moindre.

Indicateurs écologiques et d'autonomie

Les taux d'autoproduction τ_{autoprod} et d'autoconsommation τ_{autocons}, dont les expressions sont données aux Équations 4.38 et 4.37 respectivement, sont à maximiser par les agents grâce aux systèmes de récompense. Ils sont représentés sur la Figure 5.26. Dans les périodes de fort ensoleillement, le taux d'autoproduction augmente, tandis que le taux d'autoconsommation tend à diminuer.

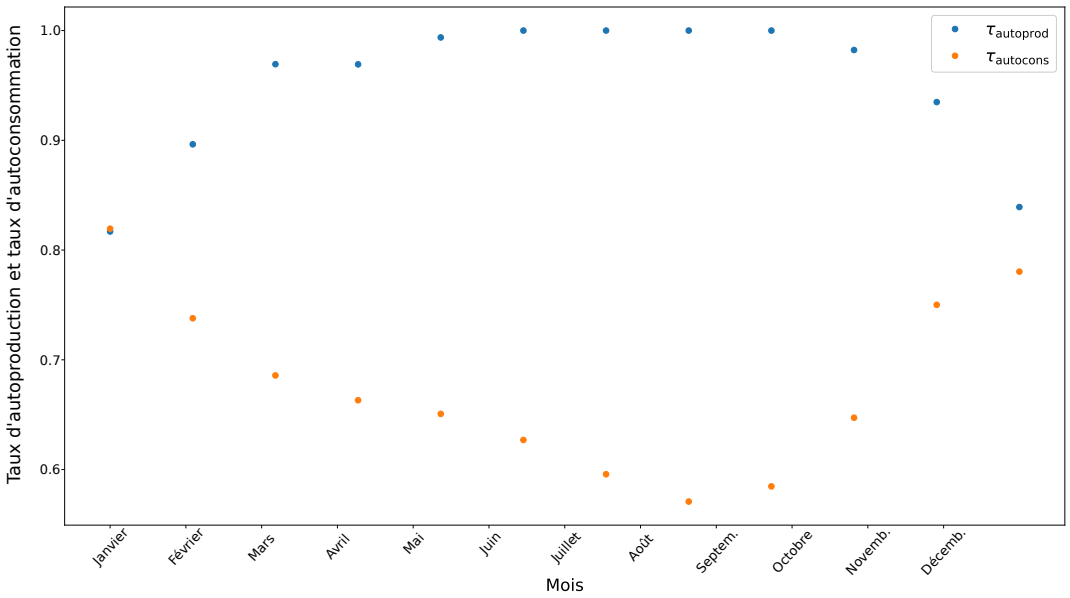

Figure 5.26 – Taux d'autoproduction et taux d'autoconsommation mensuels moyens sur 10 ans.

En moyenne sur l'année, les taux d'autoproduction et d'autoconsommation sont respectivement de 0.9496 et 0.6808.

Effet de la récompense. Pour des entraînements où l'agent doit minimiser les flux d'échange avec le réseau central, les taux varient peu (voir Table 5.5).

Accroissement de la demande. Avec une demande plus importante, le taux d'autoconsommation augmente ce qui est cohérent puisqu'il est en lien avec la proportion d'électricité consommée qui est produite localement (voir Table 5.6). Le taux d'autoproduction diminue puisqu'une part plus importante d'énergie est importée du réseau central.

Accroissement de la capacité de la batterie. Les taux d'autoproduction et d'autoconsommation sont donnés dans la Table 5.6 pour une capacité nominale de la batterie égale à 8 kWh, selon la nature de la récompense de l'agent. Ces évolutions sont cohérentes avec un gain en autonomie du micro-réseau. La probabilité de perte de charge (LPSP), introduite au chapitre 1, permet de visualiser le temps de coupure de l'approvisionnement énergétique pour le micro-réseau. La Figure 5.27 présente les variations mensuelles du LPSP pour l'entraînement effectué avec l'algorithme de DQN. Des pics apparaissent lors des mois les moins ensoleillés. La présence de ces pics prononcés indique que la politique de contrôle développée sur cet environnement ne permet pas d'assurer une très grande autonomie au système. Le LPSP moyen est de 6,46 %. Ainsi, si le micro-réseau fonctionne en mode îloté, il

a une probabilité moyenne de $6,46\%$ de ne pas pouvoir satisfaire les demandes à chaque pas de temps, ce qui n'est pas suffisant pour assurer une bonne autonomie (H. Yang et al., 2008).

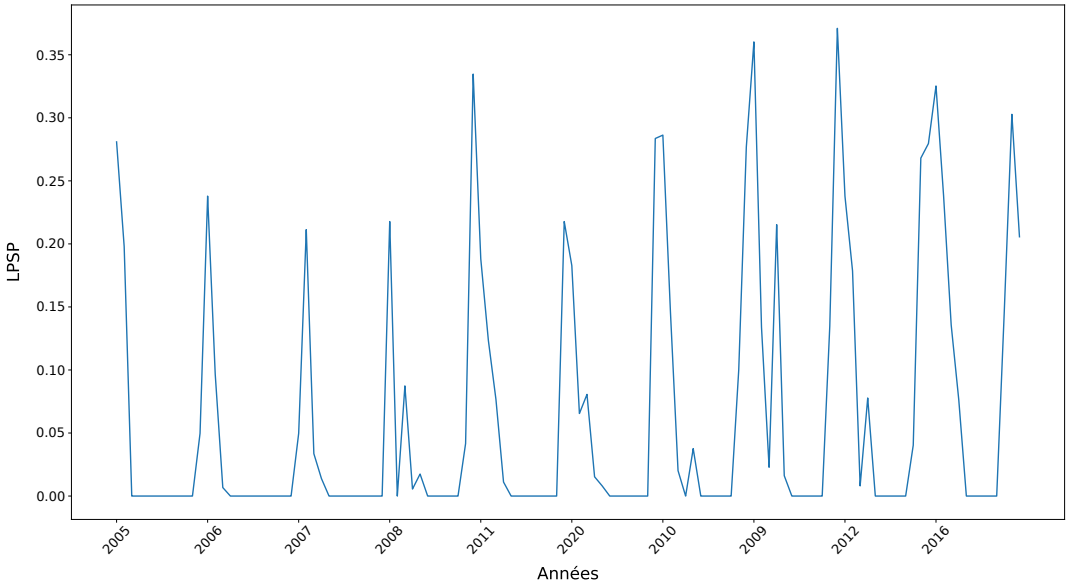

Figure 5.27 – Évolution du LPSP calculé mensuellement.

Enfin, le ratio d'excès d'énergie (REE) et la fraction d'énergie renouvelable (FER) (voir Équations 1.2 et 1.3) valent respectivement 33.6% et 98.0%. Cette valeur de FER montre que l'énergie consommée est en très grande majorité d'origine renouvelable. En revanche, le REE est trop élevée : l'énergie injectée au réseau central représente un tiers de l'électricité produite et importée. Les panneaux PV sont bien dimensionnés pour subvenir aux besoins de consommation électrique des utilisateurs, mais une plus grande capacité de batterie permettrait d'absorber l'énergie excédentaire pour minimiser le REE.

5.3.3 Analyse de la performance

L'algorithme de DQN permet à l'EMS de développer une stratégie de contrôle cohérente adaptée aux paramètres et aux caractéristiques du micro-réseau. Les résultats obtenus sont meilleurs que ceux d'un algorithme déterministe basé sur des règles pour tous les indicateurs considérés.
La dégradation de la batterie a pour conséquence de contraindre l'EMS à recourir au stockage hydrogène pour compenser les défaillances à court terme. Lorsque la batterie est usée, la PAC est plus sollicitée dans des périodes où la production d'électricité d'origine PV est forte, la quantité d'hydrogène stocké est donc limitée.
Des indicateurs ont été calculés sur des simulations de contrôle avec des données de test. L'horizon temporel est de 10 ans alors que les agents ont été entraînés sur des épisodes de 6 ans. La demande est issue des profils de consommation Enedis pour un français ayant souscrit à une puissance inférieure à 6 kVA. La Table 5.5 récapitule les différents indicateurs calculés. Le remplacement de la batterie permet au système de gagner en autonomie en valorisant plus l'énergie produite localement tout en diminuant sa dépendance au réseau central de distribution. En revanche, son remplacement fait augmenter le coût du micro-réseau. Le coût d'investissement lié au stockage hydrogène est très important et nuit à la viabilité économique du micro-réseau. Il n'est pas viable en mode connecté puisque le LCE dépasse le tarif du réseau central quelque soit l'algorithme de contrôle développé. Le système n'est pas non-plus viable en mode îloté car la probabilité de panne (donnée par le LPSP) est

importante. Le contrôle du micro-réseau équipé d'une batterie ayant une capacité supérieure $\left(\mathrm{E}_{\mathrm{nom}}^{\mathrm{batt}} = 8\mathrm{kWh}\right)$ montre de bons résultats pour l'autonomie du micro-réseau et pourrait être déployé pour fonctionner en mode îloté. En revanche, le taux d'autoconsommation et le REE suggèrent que le surplus d'électricité d'origine renouvelable produite pourrait être mieux valorisé.

Algorithme	**LCE**	τ_{autoprod}	τ_{autocons}	**LPSP**	**REE**	**FER**
DQN	0.2450	0.9496	0.6908	0.0646	0.3361	0.9795
DQN (pénalité pour énergie soutirée et injectée)	0.2524	0.9433	0.6965	0.0787	0.3171	0.9768
DQN (batterie remplacée)	0.2872	0.9747	0.6447	0.0521	0.3800	0.9895
DQN (batterie à 8kWh)	0.2164	0.9978	0.5869	0.002	0.4562	0.9991
Déterministe	0.3815	0.5018	0.6183	0.5006	0.3252	0.8281

Table 5.5 – Comparaison des indicateurs pour différents algorithmes et paramètres du micro-réseau.

La demande nette a un impact fort sur le rôle du stockage hydrogène au sein du micro-réseau. Plus la demande nette est élevée et plus le stockage hydrogène est utilisé comme substitut de la batterie. Le stockage hydrogène se comporte comme un stockage à long terme si la capacité de la batterie est augmentée avec la demande. La Table 5.6 résume les résultats obtenus avec des simulations incluant des demandes électriques plus importantes.

Algorithme	**LCE**	τ_{autoprod}	τ_{autocons}	**LPSP**	**REE**	**FER**
DQN	0.2023	0.7614	0.7777	0.3251	0.2068	0.8404
DQN (pénalité pour énergie soutirée et injectée)	0.2126	0.7418	0.8057	0.3480	0.1785	0.8305
DQN (batterie à 8kWh)	0.2031	0.8367	0.8064	0.2167	0.1852	0.8816
DQN (batterie à 8kWh, pénalité pour énergie soutirée et injectée)	0.2635	0.7937	0.8803	0.2704	0.1188	0.8573
Déterministe	0.2411	0.5230	0.6936	0.6529	0.2375	0.7358
Déterministe (batterie à 8kWh)	0.2487	0.5518	0.7122	0.6136	0.2266	0.7481

Table 5.6 – Comparaison des indicateurs pour différents algorithmes et paramètres du micro-réseau. La demande est celle d'un foyer résidentiel français.

Lorsque l'agent est entraîné à minimiser à la fois l'énergie injectée et soutirée au micro-réseau, le taux d'autoconsommation augmente et le REE diminue. Les revenus de la vente d'électricité au réseau central contribuent à la diminution du LCE. De plus, une dégradation des taux d'autoproduction, FER et LPSP est constatée.
Les LPSP observés montrent qu'aucune de ces configurations ne pourrait subvenir convenablement aux demandes de consommation si le micro-réseau fonctionne en mode îloté.

5.4 Conclusion

Dans ce chapitre, l'application du RL au contrôle du stockage à long terme dans un micro-réseau a été explorée. Initialement, l'analyse de l'apprentissage et de la politique adoptée par l'agent a été réalisée sur un horizon temporel d'une année. Pour optimiser l'efficacité de l'apprentissage, les hyper-paramètres de l'agent DQN ont été sélectionnés en utilisant

une méthode de recherche par grille. Il a été démontré que l'agent est capable d'apprendre efficacement avec plusieurs systèmes de récompense s'ils octroient des récompenses similaires en termes d'ordre de grandeur et de fréquence d'attribution. Une étude de sensibilité sur la capacité à généraliser la politique apprise a montré une adaptabilité à la date initiale d'un épisode, mais pas à la charge initiale de la réserve d'hydrogène H_2 au début d'un épisode.
Plusieurs modèles de dégradation de la batterie ont été établis : un modèle linéaire basé sur le nombre de cycles d'utilisation de la batterie et un modèle non linéaire dépendant du DOD instantané. Ces modèles ont été intégrés dans le module de la batterie de l'environnement et un horizon temporel de simulation étendu a été utilisé pour mesurer leur impact. Il a été observé que le choix du modèle de dégradation a une influence sur la politique apprise par l'agent. Cela souligne l'adaptabilité de la politique apprise en fonction des interactions de l'agent avec l'environnement.
Enfin, l'analyse de la stratégie développée par l'EMS montre que le rôle du stockage hydrogène varie selon les conditions de l'environnement. Bien qu'il soit autonome, le micro-réseau n'est pas rentable pour une demande faible dans la configuration choisie. Il l'est pour une demande plus conséquente mais en stockant une quantité très limitée d'énergie à long terme. Il n'est donc pas résilient pour une forte demande et les configurations, notamment le dimensionnement des unités, doivent être adaptées afin qu'il puisse fonctionner en mode îloté. L'étude du dimensionnement sous contrôle optimal du micro-réseau est donc essentielle pour assurer un système stable et rentable.

6

Dimensionnement sous contrôle optimal

6.1 Transfert de politique de contrôle par apprentissage par renforcement hors ligne

La conception de l'EMS d'un micro-réseau basé sur un agent entraîné avec un algorithme de DQN nécessite un temps de calcul conséquent. L'agent débute avec une politique de contrôle aléatoire et interagit avec l'environnement pour collecter des observations et faire évoluer sa politique de contrôle au cours de son entraînement. Or, le développement d'une politique quasi-optimale dépend du dimensionnement du micro-réseau (voir section 4.3). Pour chaque dimensionnement des équipements du micro-réseau, une stratégie de contrôle différente est développée, ce qui engendre des temps de calcul conséquents.
La méthode proposée pour résoudre ce problème stipule qu'une politique apprise pour un dimensionnement donné pourrait servir de base initiale pour entraîner un agent sur un environnement de dimensionnement différent, plutôt que de commencer avec des actions aléatoires. Un nouvel agent pourrait ainsi commencer son entraînement en apprenant d'interactions provenant d'une stratégie construite pour optimiser les mêmes objectifs dans un environnement relativement proche. La politique apprise serait transférée d'un environnement

à un autre. Cette section décrit le développement d'une méthode de transfert de politique afin d'écourter le temps nécessaire au développement d'une politique de contrôle satisfaisante par l'EMS. La sous-section 6.1.1 présente la méthodologie du transfert de politique par apprentissage par renforcement hors ligne. De nouveaux paramètres sont établis. La valeur de ces nouveaux paramètres est exposée à la sous-section 6.1.2. Enfin, une analyse des résultats d'optimisation et de la méthodologie, et une conclusion sont établies à la sous-section 6.1.3.

6.1.1 Formulation du problème de transfert de politique par apprentissage par renforcement hors ligne

Le dimensionnement du micro-réseau s'effectue sur deux niveaux. À haut niveau (en boucle principale ou externe), les caractéristiques des équipements du micro-réseau sont choisies selon un coût global économique. Les variables de dimensionnement sont la puissance crête des panneaux photovoltaïques (PV) $\mathrm{P}_{\mathrm{nom}}^{\mathrm{PV}}$ et la capacité nominale de la batterie $\mathrm{E}_{\mathrm{nom}}^{\mathrm{batt}}$ dans le travail présenté. Un algorithme d'optimisation est employé pour choisir les valeurs qui minimisent le coût économique global. Ces coûts globaux dépendent notamment du contrôle du micro-réseau, qui est effectué à plus bas niveau (en sous-boucle ou boucle interne) et des coûts d'investissement et de maintenance. La politique de contrôle développée dépend du dimensionnement du micro-réseau, et doit être apprise pour tout nouveau dimensionnement, c'est-à-dire pour chaque itération de la boucle principale.
Ainsi dans le cadre d'une conception optimale du micro-réseau par une approche bi-niveaux, il sera nécessaire qu'à chaque itération de l'optimiseur où de nouveaux dimensionnements sont choisis, un nouvel agent soit entraîné pour générer une meilleure politique de contrôle de l'EMS. Par conséquent, le dimensionnement bi-niveaux du micro-réseau implique la réalisation d'une multitude d'entraînements. Cette section établit une méthodologie visant à réduire le temps d'entraînement de l'agent. Contrairement à la section 5.1 du chapitre précédent, les études vont être menées pour plusieurs dimensionnements afin de vérifier l'intérêt de la méthodologie proposée. Comme dit précédemment, les variables de dimensionnement sont la puissance nominale des panneaux PV installés $\mathrm{P}_{\mathrm{nom}}^{\mathrm{PV}}$ et la capacité nominale de la batterie électrochimique $\mathrm{E}_{\mathrm{nom}}^{\mathrm{batt}}$. Un agent associé à l'EMS du micro-réseau dont le dimensionnement est $\left[\mathrm{P}_{\mathrm{nom}}^{\mathrm{PV},j}, \mathrm{E}_{\mathrm{nom}}^{\mathrm{batt},j}\right]$ est appelé agent j.
La méthodologie développée consiste à utiliser la politique de contrôle de l'EMS pour alimenter la mémoire de relecture d'un agent i au début de son apprentissage. L'objectif attendu est une réduction du temps d'entraînement sans dégrader la politique apprise. En effet, la suppression de la phase d'exploration initiale doit permettre de réduire le nombre d'itérations nécessaires. L'EMS dont la politique est réemployée pour accélérer l'entraînement d'un autre EMS est appelé *démonstrateur*. Le démonstrateur interagit avec l'environnement $\left[\mathrm{P}_{\mathrm{nom}}^{\mathrm{PV},i}, \mathrm{E}_{\mathrm{nom}}^{\mathrm{batt},i}\right]$ de l'agent i. Ses actions sont donc choisies à partir d'une politique apprise sur un environnement au dimensionnement différent.
Une fois la mémoire de relecture complétée par le démonstrateur, l'agent i apprend de manière hors ligne (sans interagir avec l'environnement). L'algorithme utilisé pour l'apprentissage hors ligne est le BCQ (voir section 2.3).
Pendant le processus d'optimisation bi-niveaux, un EMS peut soit apprendre à partir de sa propre interaction avec son environnement avec l'algorithme de DQN soit apprendre de manière hors ligne en observant les interactions d'un autre EMS déjà entraîné grâce à l'algorithme de BCQ. L'apprentissage par BCQ n'est possible que si d'autres agents ont déjà été entraînés lors d'itérations antérieures dans le processus d'optimisation. Un schéma de principe de l'utilisation de BCQ pour la méthodologie proposée d'optimisation bi-niveaux est présenté sur la Figure 6.1.

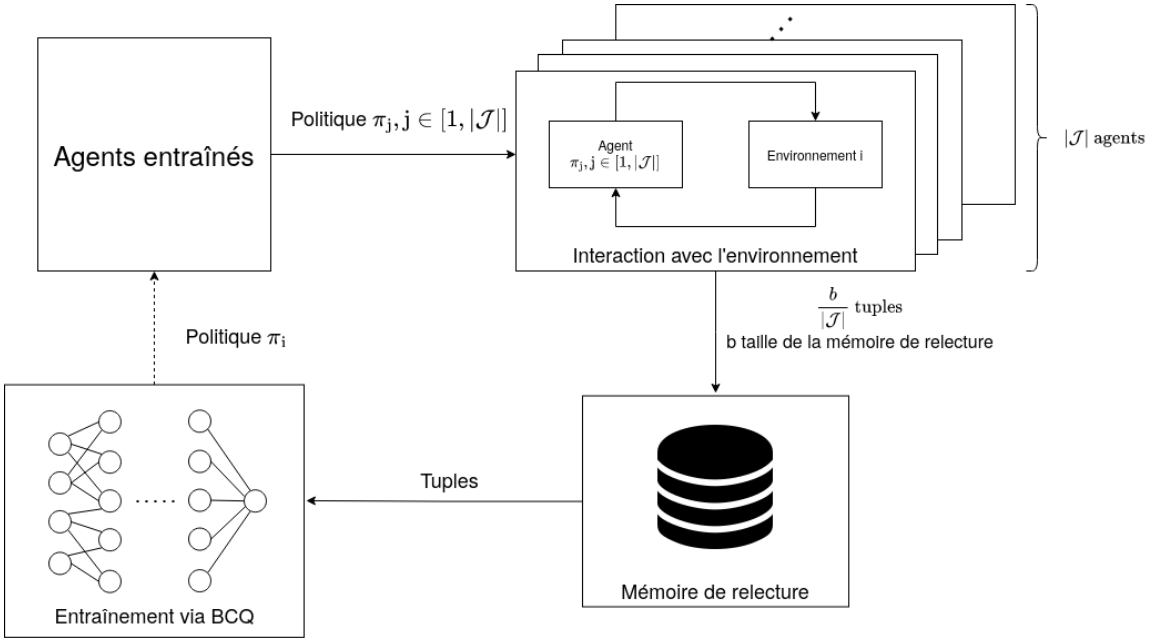

Figure 6.1 – Principe de l'entraînement de l'agent i. Une mémoire de relecture est complétée par des interactions entre $|j|$ agents déjà entraînés sur d'autres environnements et l'environnement i.

L'algorithme de BCQ a été présenté au chapitre 2, nous rappelons ici l'Équation 2.54 :

$$\pi(\mathrm{s}) = \underset{a|\frac{G_\omega(a|s)}{\max_{\hat{a}} G_\omega(\hat{a}|\mathrm{s})} > \tau_{\mathrm{BCQ}}}{\arg\max} Q_\theta(\mathrm{s}, \mathrm{a}) \tag{2.54}$$

Cet algorithme entraîne un agent i à choisir une action a depuis un état s selon sa probabilité d'être prise par le démonstrateur et son potentiel dans l'environnement de l'agent i. L'algorithme de BCQ nécessite qu'un démonstrateur existe et ne peut donc pas être utilisé dès la première itération de dimensionnement du micro-réseau. Pour la mise en œuvre de BCQ dans ce travail, deux hypothèses sont formulées :

1 La proximité entre la valeur des variables de dimensionnement d'un démonstrateur et de l'agent i a un impact sur la qualité de la politique apprise. Autrement dit, plus le dimensionnement du micro-réseau du démonstrateur est distant de celui de l'environnement de l'agent i, moins la politique apprise par l'agent i sera performante.

2 L'utilisation d'un grand nombre de démonstrateurs pour remplir la mémoire de l'agent i améliore l'apprentissage en mettant en jeu une plus grande variété de politiques.

Afin de vérifier ces hypothèses, des comparaisons entre l'apprentissage effectué selon la méthodologie proposée et des méthodes plus classiques ont été effectuées. En particulier, les temps d'entraînement et les taux d'autoproduction τ_{autoprod} sont systématiquement comparés à ceux obtenus avec l'algorithme de DQN et un algorithme basé sur des règles. Dans cette section, les performances de l'agent obtenu seront évaluées sur des données de productible PV de 2018 alors que les entraînements sont menés sur des données de 2020.

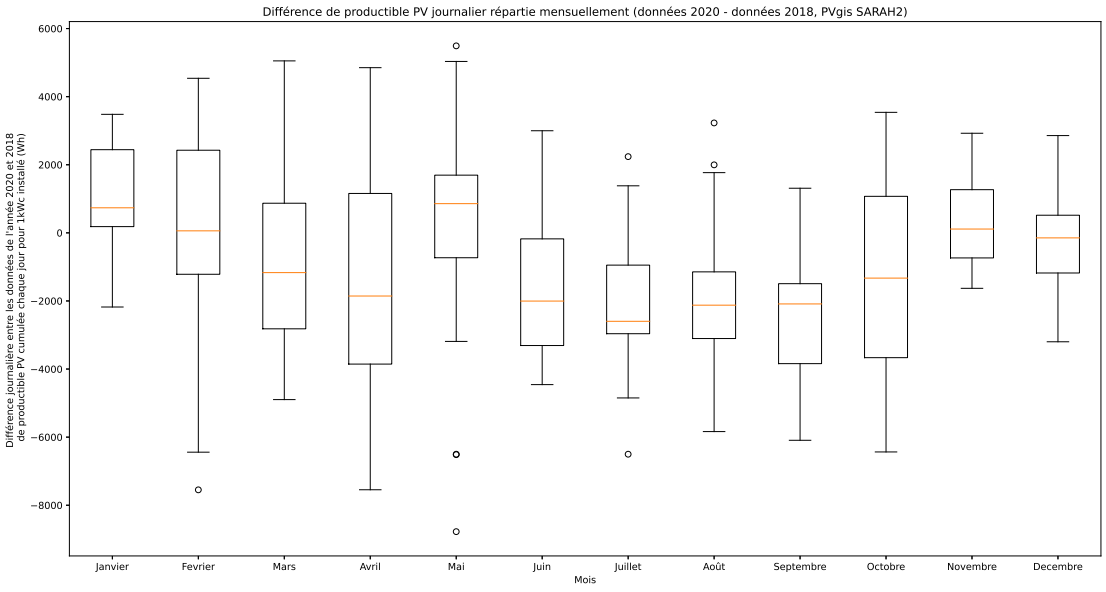

Figure 6.2 – Diagramme en boîte de la différence par mois du productible PV journalier entre le jeu d'entraînement et le jeu de test.

La Figure 6.2 montre les différences de productible PV journalier pour une installation de 1 kWc entre les données d'entraînement et les données de test. Le productible PV quotidien moyen est de 3540 Wh pour l'ensemble des données. Les disparités entre les données d'entraînement et de test sont significatives.

6.1.2 Application de l'apprentissage par renforcement hors ligne et choix des paramètres

Nous allons étudier ici l'influence du nombre de démonstrateurs (cf. hypothèse [2]) et de la proximité des variables de dimensionnement entre les environnements des démonstrateurs et de l'environnement i (cf. hypothèse [1]).

La vérification de l'hypothèse [1] nécessite de définir une distance pour évaluer la proximité entre les variables de dimensionnement de l'agent à entrainer et celles d'un démonstrateur. Cette distance est définie selon l'Équation 6.1 comme la distance euclidienne des valeurs des variables de dimensionnement des micro-réseaux.

$$\mathrm{d_j} = \sqrt{\left(\mathrm{P}_{\mathrm{nom}}^{\mathrm{PV},i} - \mathrm{P}_{\mathrm{nom}}^{\mathrm{PV},j}\right)^2 + \left(\mathrm{E}_{\mathrm{nom}}^{\mathrm{batt},i} - \mathrm{E}_{\mathrm{nom}}^{\mathrm{batt},j}\right)^2} \tag{6.1}$$

L'agent j déjà entraîné dont le but est de fournir la mémoire de relecture de l'agent i est appelé *voisin* de l'agent i. Deux critères sont définis pour identifier quels sont les agents voisins de l'agent i :

— La distance maximale $\mathrm{d_{max}}$ entre les valeurs des variables de dimensionnement.

— Le nombre minimum J de voisins nécessaires pour que l'agent i soit entraîné avec l'algorithme de BCQ plutôt qu'avec l'algorithme de DQN.

Hyper-paramètres	Valeur
τ_{BCQ}	0.3
Probabilité d'action aléatoire dans un épisode à bruit important	0.3
p	0.9
ϵ	0.05

Table 6.1 – Hyper-paramètres utilisés pour l'algorithme de BCQ

Si les distances d_j sont inférieures à d_{max} pour au moins J environnements, alors l'agent i sera entraîné de manière hors ligne avec un algorithme de BCQ. Si ce n'est pas le cas, l'agent i sera entraîné avec l'algorithme de DQN et interagira directement avec son environnement pendant l'apprentissage. Afin de déterminer d_{max} et J, le taux d'autoproduction obtenu en phase de test selon le nombre de voisins et leur distance à l'environnement i est considéré. Pour éviter la surreprésentation d'états similaires dans la mémoire de relecture pouvant compromettre la capacité de généralisation d'un agent entraîné par BCQ, des décisions aléatoires sont prises lors de chaque épisode avec une certaine probabilité :

— Chaque épisode échantillonné a une probabilité p d'être un épisode « à bruit important ». Dans ce cas, la probabilité de prendre des décisions aléatoires est plus élevée.

— Si un épisode échantillonné n'est pas « à bruit important », il y a une probabilité ε plus faible de choisir une action aléatoire à chaque instant.

Ces probabilités d'exploration sont des hyper-paramètres choisis par l'utilisateur. La version discrète de BCQ utilisée admet également τ_{BCQ} comme hyper-paramètre supplémentaire (voir section 2.3). La valeur des hyper-paramètres utilisés est précisée dans la Table 6.1.
L'échantillonnage et le stockage dans la mémoire de relecture pour entraîner un agent i de BCQ à partir des interactions avec un démonstrateur j sont présentée sur la Figure 6.3.
Sur cette figure, un démonstrateur unique est considéré et les tuples des séquences échantillonnées en interaction avec l'environnement i occupent la totalité de la mémoire de relecture. Lorsque plusieurs démonstrateurs sont voisins avec l'agent i, chacun interagit avec l'environnement i de manière à générer une proportion égale d'échantillons dans la mémoire de relecture. L'entraînement d'un agent via BCQ commence une fois sa mémoire de relecture complète. Comme indiqué dans les travaux de Fujimoto et al., 2019, l'algorithme de BCQ montre de bonnes performances pour une mémoire contenant au moins un million de tuples. Un épisode se termine après une prise de décision à chaque heure pendant un an. 8759 tuples de type {s,a,R,s'} sont donc échantillonnés pendant un épisode. Un million de tuples sont collectés par interaction avec l'environnement en 115 épisodes. Par conséquent, les interactions entre les démonstrateurs et l'environnement génèrent 115 épisodes.

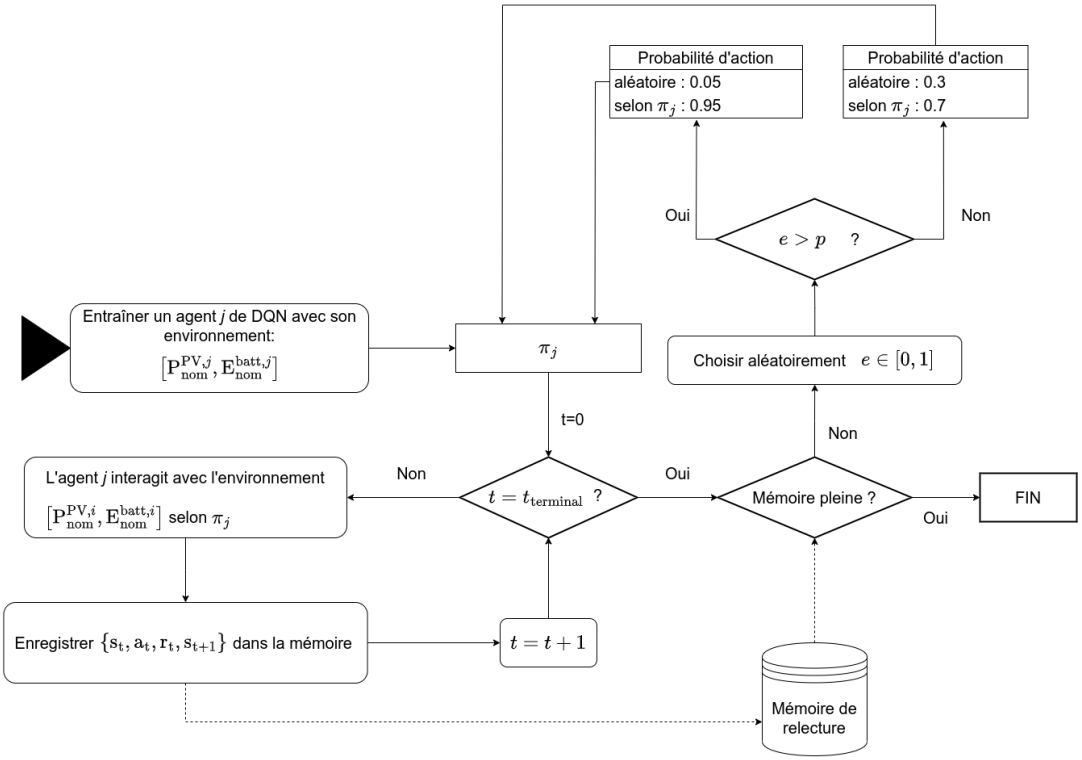

Figure 6.3 – Échantillonnage et stockage de la mémoire de relecture pour entraîner un agent i via BCQ à partir d'un agent démonstrateur j.

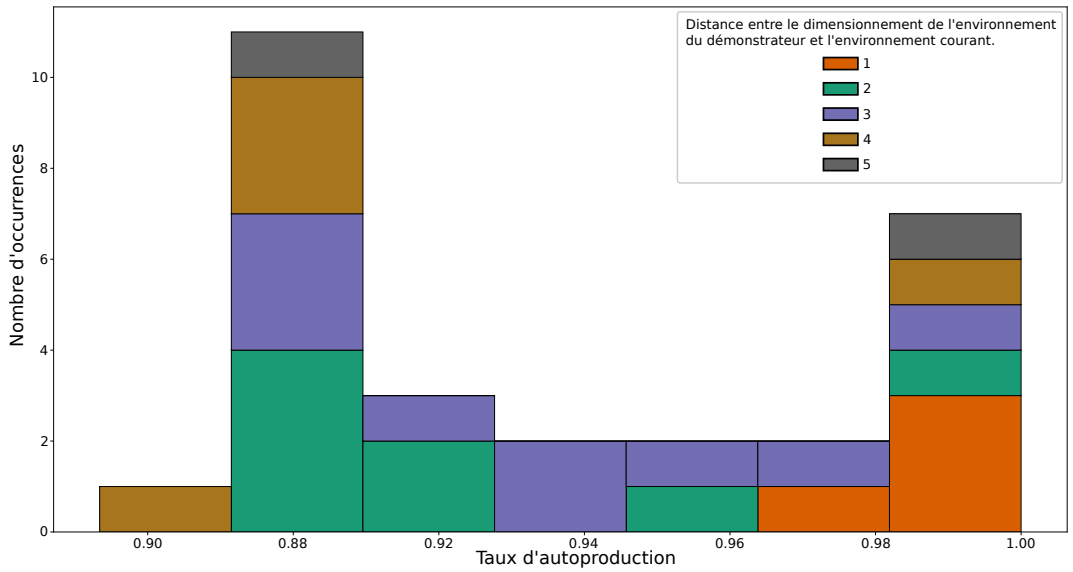

Figure 6.4 – Histogramme des taux d'autoproduction obtenus en fin d'entraînement de l'algorithme BCQ avec un unique démonstrateur. Les couleurs sont associées aux distances entre les environnements du démonstrateur et de l'agent apprenant.

L'étude paramétrique pour déterminer $\mathrm{d_{max}}$ et J s'effectue avec les variables de dimensionnement égales à $[5\,\mathrm{kWc}, 5\,\mathrm{kWh}]$. Dans un premier temps, l'agent de BCQ est entraîné avec

un unique démonstrateur dans le but d'évaluer la distance minimale entre les variables de dimensionnement des deux environnements. Les résultats obtenus sont représentés sous forme d'histogramme sur la Figure 6.4.

L'histogramme présente la répartition des scores suite à l'entraînement d'agents avec un seul voisin démonstrateur. Chaque couleur représente la distance d entre l'environnement et le démonstrateur. Il est observé que les performances sont globalement supérieures avec $d = 1$. Des écarts entre les différents taux d'autoproduction sont observés pour des agents entraînés avec un voisin situé à une distance supérieure. La valeur associée à d_{max} est donc 1 pour garantir une politique de contrôle performante selon le taux d'autoproduction. Ainsi, les démonstrateurs sont tous les agents entraînés pour lesquels la distance euclidienne entre le dimensionnement de leur environnement d'entraînement et de l'environnement i vaut 1.

Une seconde expérience est mise en œuvre en variant le nombre de démonstrateurs pour alimenter la mémoire de relecture de l'agent entraîné. Les résultats de l'expérience précédente sont pris en considération pour relever le nombre de démonstrateurs voisins de l'agent i parmi les démonstrateurs impliqués. Ceux-ci sont situés à une distance $d = 2$ au maximum. Les histogrammes montrant le nombre d'occurrences de scores par tranche de taux d'autoproduction après entraînement sont montrés en fonction du nombre de voisins impliqués dans la génération de la mémoire de relecture de l'agent entraîné sur la Figure 6.5. La figure est scindée en deux graphiques, l'un exposant les résultats pour une mémoire générée par deux démonstrateurs et l'autre pour une mémoire générée par trois démonstrateurs.

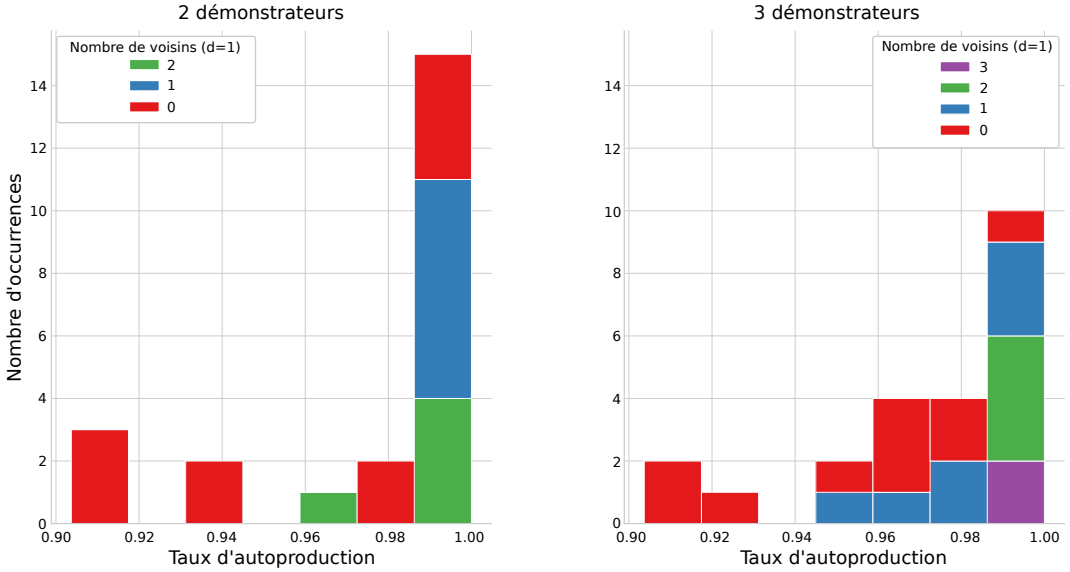

Figure 6.5 – Histogrammes montrant les taux d'autoproduction obtenus en fin d'entraînement de BCQ avec 2 démonstrateurs (à gauche) et avec 3 démonstrateurs (à droite) en fonction du nombre de voisins parmi les démonstrateurs.

En l'absence de démonstrateur voisin de l'agent i entraîné, les scores présentent une variation plus élevée dans les deux cas. La qualité (mesurée par le taux d'autoproduction) de la politique de contrôle de l'agent augmente globalement avec la proportion de voisins parmi les démonstrateurs. Les scores obtenus sont considérés comme acceptables lorsque pour n démonstrateurs, n-1 sont voisins de l'agent i. En revanche, puisque la proportion de voisins dans le nombre de démonstrateurs influe sur la qualité de la politique apprise, les démonstrateurs utilisés seront voisins de l'agent i pour son entraînement. Le nombre minimal de voisins requis pour entraîner un agent via la méthodologie de transfert de politique développée est de $J = 1$. Si plusieurs autres agents déjà entraînés sont voisins de l'agent i à

entraîner, ils seront utilisés à proportion égale pour échantillonner des tuples d'expérience. Une fois les valeurs des paramètres d_{max} et J déterminées, il convient maintenant de procéder à l'évaluation de l'efficacité et de la performance du transfert de politique.

6.1.3 Analyse et conclusion

Une comparaison entre le temps de calcul et le score obtenu par la politique apprise par la méthodologie de transfert hors ligne et les résultats de l'algorithme de DQN et d'un algorithme déterministe basé sur des règles (détaillé à la section 4.2) doit être effectuée. Cette étude est menée sur différents dimensionnements (P_{nom}^{PV} et E_{nom}^{batt}) de micro-réseaux. Rappelons que l'objectif pour le fonctionnement du micro-réseau considéré est l'utilisation appropriée du réservoir H_2 comme dispositif de stockage à long terme. Les valeurs des variables de dimensionnements admises ont été données à la section 4.2 du chapitre 4. Puisque l'algorithme basé sur des règles n'est pas soumis à un apprentissage, son temps de calcul est négligé. Le score employé pour cette comparaison est le taux d'autoproduction $\tau_{autoprod}$ obtenu sur un épisode en phase de test.

Afin de prévenir tout biais dans l'analyse de leur performance, les algorithmes de DQN et BCQ disposent des mêmes hyper-paramètres, lesquels ont été explicités à la section 5.1. La condition d'arrêt de l'apprentissage est une patience correspondant à 1000 itérations d'entraînement. Les paramètres propres à l'algorithme de BCQ sont présentés dans la Table 6.1. La valeur des paramètres associés au transfert de politique sont $d_{max} = 1$ et $J = 1$ conformément à la sous-section 6.1.2. Chaque algorithme de BCQ est entraîné avec des échantillons d'expériences provenant d'un unique démonstrateur. Chaque démonstrateur impliqué est entrainé avec l'algorithme de DQN. L'étude comparative est menée sur les environnements correspondant aux valeurs suivantes des variables de dimensionnement :

— $[8\,\text{kWc}, 8\,\text{kWh}]$ est un dimensionnement pour lequel la valeur $\tau_{autoprod} = 1$ est facilement obtenue.

— $[8\,\text{kWc}, 5\,\text{kWh}]$, $[5\,\text{kWc}, 8\,\text{kWh}]$ et $[5\,\text{kWc}, 5\,\text{kWh}]$ sont des dimensionnements pour lesquels la valeur $\tau_{autoprod} = 1$ peut être atteinte.

— $[3\,\text{kWc}, 2\,\text{kWh}]$ constitue un micro-réseau aux variables de dimensionnement minimales.

Les résultats sont présentés dans la Table 6.2.

L'agent entraîné par BCQ montre parfois de meilleurs résultats que l'agent entraîné avec l'algorithme de DQN. En revanche, des disparités dans le score apparaissent selon l'environnement sur lequel le démonstrateur a été entraîné. Le temps d'entraînement moyen d'un agent avec l'algorithme de DQN est de 32 minutes. Il vaut 9 minutes pour un entraînement effectué avec l'algorithme de BCQ. Les agents sont développés avec la librairie python *PyTorch* avec trois réseaux de 256 neurones entièrement connectés. La machine utilisée est un PC avec 16 Gio de RAM, équipé de *Intel i5 8th Generation*. L'apprentissage d'un agent hors ligne via BCQ est toujours plus rapide que l'apprentissage d'un agent de DQN dans le même environnement. L'utilisation de l'algorithme de BCQ dans une démarche de transfert de politique est donc un outil approprié pour le dimensionnement optimal d'un micro-réseau sous contrôle quasi-optimal.

6.2 Méthode d'optimisation globale par méta-heuristique pour le dimensionnement bi-niveaux du micro-réseau

La contextualisation du dimensionnement du micro-réseau considéré a été présentée à la section 4.3. La minimisation du coût économique global du micro-réseau est l'objectif principal

Dimensionnement ($\lfloor P_{\text{nom}}^{\text{PV}}; E_{\text{nom}}^{\text{batt}} \rfloor$)	Algorithme	Environnement du démonstrateur	τ_{autoprod}	Temps d'apprentissage
[8; 8]	DQN	-	1	36 min
	Déterministe	-	0.8775	-
	BCQ	[8; 9]	0.9999	8 min
		[9; 8]	0.9775	8 min
		[7; 8]	1	14 min
		[8; 7]	1	16 min
[5; 8]	DQN	-	0.9344	43 min
	Déterministe	-	0.8179	-
	BCQ	[5; 7]	1	17 min
		[5; 9]	0.9539	10 min
		[4; 8]	1	10 min
		[6; 8]	0.9355	12 min
[8; 5]	DQN	-	0.9501	26 min
	Déterministe	-	0.8437	-
	BCQ	[8; 6]	1	5 min
		[8; 4]	1	8 min
		[9; 5]	0.9501	6 min
		[7; 5]	1	6 min
[5; 5]	DQN	-	0.9691	31 min
	Déterministe	-	0.7766	-
	BCQ	[5; 4]	0.9999	8 min
		[5; 6]	1	8 min
		[4; 5]	0.9711	10 min
		[6; 5]	1	9 min
[3; 2]	DQN	-	0.7672	26 min
	Déterministe	-	0.5927	-
	BCQ	[4; 2]	0.7592	7 min
		[3; 3]	0.8025	6 min

Table 6.2 – Comparaison du taux d'autoproduction et du temps d'apprentissage obtenus avec des agents de DQN et de BCQ selon plusieurs configurations de dimensionnement du micro-réseau simulé.

recherché. Il est explicité avec l'Équation 4.41 qui est rappelée ici :

$$\min_{d \in \mathcal{D}} \mathrm{C_{inv}}(d,n) + \mathrm{C_{ope}}(d,n) + \mathrm{C_{maint}}(d,n) + \mathrm{C_{remp}}(d,n),\ n \in \mathcal{N} \tag{4.41}$$

avec $\mathrm{C_{inv}}$, $\mathrm{C_{ope}}$, $\mathrm{C_{maint}}$ et $\mathrm{C_{remp}}$ respectivement les coûts d'investissement, d'opération, de maintenance et de remplacement des unités. $\mathcal{D}$ et $\mathcal{N}$ sont respectivement les espaces de valeur des variables dimensionnables et non-dimensionnables. Le choix de l'algorithme d'optimisation est justifié à la sous-section 6.2.1. La configuration de l'algorithme pour le mettre en adéquation avec le problème d'optimisation est décrite dans la sous-section 6.2.2.

6.2.1 Caractérisation du problème d'optimisation et choix d'un algorithme

La résolution d'un problème d'optimisation consiste à trouver les valeurs d'un jeu de variables qui minimisent ou maximisent une fonction objectif avec ou sans contrainte.
Comme défini dans la section 4.1, les variables de dimensionnement sont $\mathrm{P_{nom}^{PV}}$ et $\mathrm{E_{nom}^{batt}}$, respectivement. Ce sont des variables entières dont le domaine de définition est $\{[\![3\,\mathrm{kWc}, 11\,\mathrm{kWc}]\!], [\![2\,\mathrm{kWh}, 11\,\mathrm{kWh}]\!]\}$. C'est un problème d'optimisation combinatoire comme les valeurs sont dénombrables. La fonction objectif combine le CAPEX et l'OPEX. Le CAPEX dépend

linéairement des deux variables de dimensionnement choisies. L'expression de l'OPEX n'est pas linéaire selon ces variables. L'OPEX dépend de la politique de contrôle délivrée par l'agent. La fonction objectif à minimiser est donc non-linéaire. Sa convexité n'est pas garantie. Les algorithmes d'optimisation déterministes utilisent des méthodes analytiques. Leur dépendance à la modélisation du problème, les incertitudes et la complexité de la modélisation les rendent peu adaptés au dimensionnement du micro-réseau.
Les méthodes de résolution stochastiques ont été conçues pour gérer les problèmes exposés à du bruit et des incertitudes sur les mesures (C. Li et al., 2021). Le données ne sont pas considérées comme fixes et exactes et des modèles probabilistes sont utilisés pour la résolution du problème. Parmi ces approches, les algorithmes heuristiques sont élaborés autour de règles empiriques mais ne sont pas justifiées par une théorie mathématique. Ils sont le plus souvent spécifiques au problème traité.

Méta-heuristiques

De nombreux critères de classification des algorithmes d'optimisation méta-heuristiques existent (Stegherr et al., 2022). Ce algorithmes peuvent être inspirés de phénomènes naturels, évolutifs ou non, fonctionner avec une population de candidats ou un seul point, garder en mémoire la meilleure solution trouvée ou non, etc... La Figure 6.6 inspirée de Harifi et al., 2021, classe les algorithmes méta-heuristiques selon la manière dont ils explorent l'espace des solutions.

Algorithmes évolutionnistes. Les algorithmes évolutionnistes s'inspirent des principes de l'évolution biologique. Ce sont des algorithmes à population, c'est-à-dire que plusieurs candidats sont générés et la valeur de la fonction objectif associée à chacun est calculée à chaque itération. Ces candidats sont appelés individus et sont générés par évolution des individus de l'itération précédente. Les individus avec les meilleures valeurs de la fonction objectif sont sélectionnés et peuvent se reproduire en croisant la valeur de leurs variables pour l'exploitation. Ils peuvent aussi muter et modifier légèrement la valeur d'un de leur composant pour l'exploration. Enfin, des individus disparaissent et sont remplacés par les individus issus de croisements ou générés aléatoirement.

Algorithmes inspirés de la nature. Les algorithmes inspirés de la nature sont issus d'observations de phénomènes naturels. Ils sont adaptés pour résoudre des problèmes d'optimisation. C'est généralement la communication entre les individus d'une population dans la nature qui est reproduite par les algorithmes pour minimiser un ou plusieurs objectifs. Ces algorithmes sont des algorithmes à population. Par exemple, les algorithmes par essaim particulaire reproduisent le comportement des oiseaux en essaim ou des poissons en banc. Chaque candidat, appelé particule dans ce contexte, se déplace dans l'espace de recherche vers une direction donnée par sa meilleure solution individuelle et par la meilleure solution du groupe.

Algorithmes basés sur une trajectoire. Les algorithmes basés sur une trajectoire ne sont généralement pas à population. Un seul candidat explore l'espace de recherche. Les algorithmes ont une manière différente de contrôler la trajectoire du candidat pour maximiser l'exploration et l'exploitation afin de s'approcher rapidement d'une solution optimale globale. La taille réduite de l'espace de recherche rend la recherche exhaustive envisageable. Cela consiste à calculer une solution pour chaque candidat possible dans l'espace de recherche.

Le coût en calcul de l'entraînement d'un agent de RL pour le contrôle d'un micro-réseau sur un horizon temporel long est important. De ce fait, une méta-heuristique est à privilégier pour parcourir efficacement l'espace des solutions.
Les méthodes basées sur une population pourraient être "lourdes" pour un espace de recherche

de taille réduite. Pour cette raison, une méthode basée sur une trajectoire est privilégiée. Les méta-heuristiques basées sur la trajectoire sont sensibles aux optima locaux puisqu'un seul point est considéré dans l'espace des candidats. Lorsqu'une région semble intéressante, l'exploitation suggère de choisir des candidats voisins au candidat actuel, pour les étapes suivantes de la trajectoire. C'est une recherche d'optimum local. L'exploration consiste alors à incorporer des mécanismes de déplacements aléatoires dans l'espace de recherche. Cela permet de faire évoluer la région de recherche afin d'empêcher la convergence éventuelle vers une solution sous-optimale.

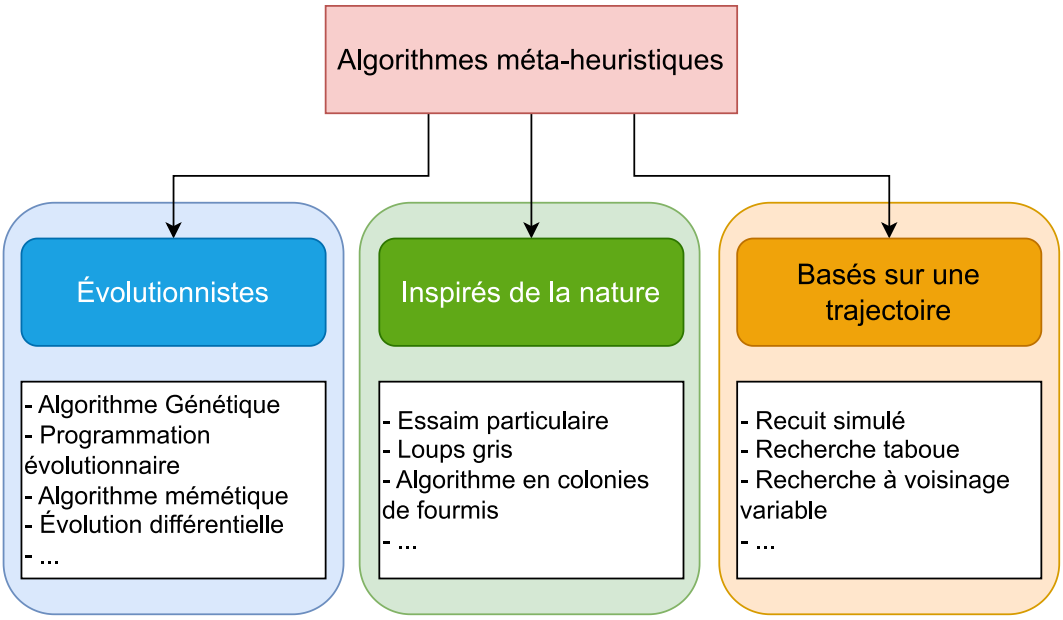

Figure 6.6 – Exemple de classification d'algorithmes d'optimisation méta-heuristiques. Inspiré de Harifi et al., 2021

Parmi les méta-heuristiques basées sur une trajectoire, quatres sont répandues et adaptables à de nombreux problèmes. Elles sont caractérisées dans la Table 6.3.

Descriptions des principaux algorithmes basés sur une trajectoire

Recuit simulé. Le recuit simulé (Kirkpatrick et al., 1983) est inspiré de la recuisson des métaux. L'objectif est de minimiser l'énergie libre des particules constituant un métal. Pour cela, le recuit consiste d'abord à augmenter la température avant de refroidir lentement pour atteindre un état plus stable. En optimisation, l'algorithme accepte des candidats dont la solution est moins bonne que la solution actuelle avec une probabilité élevée. À mesure des itérations, cette probabilité diminue et la trajectoire des candidats s'oriente vers les régions à plus haut potentiel. La probabilité d'accepter de moins bons résultats assure l'exploration et sa diminution participe à la capacité d'exploitation. Cette technique est simple à mettre en place et s'applique aussi bien aux variables continues que discrètes.

Recherche taboue. La recherche taboue (Glover, 1986) parcourt l'espace des candidats en gardant en mémoire les solutions déjà visitées dans une liste taboue. Ainsi, la trajectoire de l'algorithme ne peut plus inclure les candidats déjà vus ce qui contribue à sa capacité d'exploration. Une mémoire à long terme permet de conserver les meilleures solutions obtenues pour l'exploitation de l'algorithme. La recherche taboue fonctionne avec des variables

discrètes mais peut être adaptée à des variables continues (Siarry et al., 1997).

Recherche à voisinage variable. La recherche à voisinage variable (Mladenović et al., 1997) explore l'espace des solutions avec une définition variable de la notion de voisinage. Les candidats sont choisis dans un ensemble défini de solutions proches, appelé "voisinage", basé sur une certaine métrique ou règle de proximité. L'exploration consiste à varier la définition de ce voisinage autour d'un candidat pour élargir l'espace des solutions possibles. Si une meilleure solution est trouvée, l'algorithme se recentre sur sa région en ajustant cette définition. L'algorithme s'applique aux variables discrètes comme aux variables continues.

Méthode	**Exploration**	**Exploitation**	**Types de variables**	**Complexité**
Recuit Simulé	Accepte des solutions moins bonnes pour explorer.	Probabilité de choisir une solution moins bonne diminue avec les itérations.	Convient à des variables discrètes et continues.	Faible à modérée.
Recherche Taboue	Liste taboue empêche la redondance, favorisant l'exploration.	Favorise les meilleures solutions récentes en utilisant des mémoires à court et long termes.	Principalement pour des variables discrètes, adaptable aux variables continues.	Modérée.
Recherche à Voisinage Variable	Utilise différentes structures de voisinage pour une exploration variée.	Se concentre sur des zones prometteuses en changeant de structure de voisinage.	Convient à des variables discrètes et continues.	Modérée.
GRASP	Construction aléatoire mais gourmande d'une solution.	Améliore la solution via une recherche locale.	Principalement pour des variables discrètes, adaptable aux variables continues.	Modérée à élevée.

Table 6.3 – Comparaison des méthodes basées sur une trajectoire.

GRASP. Enfin, le GRASP (pour *Greedy Randomized Adaptive Search Procedure*, Feo et al., 1989) est un algorithme qui alterne deux étapes : une étape de construction et une étape d'amélioration. Comme le nom de l'algorithme l'indique, l'étape de construction est gourmande et aléatoire. Une liste restreinte de candidats est complétée de manière gourmande. L'amélioration consiste à explorer localement la région du candidat sélectionné pour atteindre une solution optimale. Généralement, ces étapes se succèdent et l'algorithme enregistre la meilleure solution observée. L'étape de construction peut être guidée par une autre méthode méta-heuristique : Une liste taboue permet par exemple d'éviter de considérer des candidats déjà observés lors de l'initialisation de la construction (Casado et al., 2022). L'algorithme est construit pour optimiser des variables discrètes mais peut s'appliquer à des variables

continues (M. J. Hirsch et al., 2007). La complexité de l'algorithme est modérée à élevée, en fonction des heuristiques utilisées lors de la construction ou de la recherche locale.
Le recuit simulé sera l'algorithme employé pour l'optimisation du dimensionnement dans la suite de ces travaux. Il requiert moins de puissance de calcul et est adaptable facilement à de nombreux problèmes. L'adaptabilité de l'algorithme permet d'ajouter progressivement des variables au problème d'optimisation en modifiant peu de paramètres. Ces paramètres sont présentés à la sous-section 6.2.2.

6.2.2 Recuit simulé pour le dimensionnement du micro-réseau

Principes du recuit simulé

L'algorithme de recuit simulé est initialisé avec un candidat choisi aléatoirement dans l'espace de recherche. La valeur de la fonction objectif associée à une solution est appelée *énergie*, elle est symbolisée par la lettre E. La probabilité P d'accepter un candidat est donnée par l'Équation 6.2.

$$\mathrm{P}\left(\Delta \mathrm{E}\right) = \begin{cases} 1, & \text{si } \Delta \mathrm{E} > 0 \\ e^{\frac{(-\Delta \mathrm{E})}{\mathrm{T}}}, & \text{sinon} \end{cases} \tag{6.2}$$

ΔE est l'écart entre l'énergie de l'itération en cours et l'itération précédente. Plus la température est haute et plus la probabilité de choisir un candidat plus mauvais est élevée. La température peut diminuer de différentes manières. Les plus courantes sont des diminutions linéaires, géométriques, logarithmiques ou exponentielles. Les paramètres de ces différentes fonctions appelés coefficients de refroidissement sont ajustables par l'utilisateur. La température peut ne pas diminuer à chaque itération, on parle alors d'attente d'équilibre thermodynamique. Le nombre d'itérations à température fixe avant équilibre thermodynamique est également ajustable.
La manière de générer des nouveaux candidats est réglable. Classiquement, un pas maximal $\mathrm{s_i^{max}}$ est défini à chaque itération pour chaque variable i et le candidat suivant sera généré avec un écart $\Delta \mathrm{s_i}$, $\Delta \mathrm{s_i} \in [-\mathrm{s_i^{max}}, \mathrm{s_i^{max}}]$. En résumé, les paramètres ajustables du recuit simulé sont la température initiale T_0, les coefficients de refroidissement, la manière d'initialiser la trajectoire, le nombre d'itérations avant équilibre thermodynamique, le critère d'arrêt (la température finale T_f ou un nombre d'itérations) et l'écart entre les candidats.

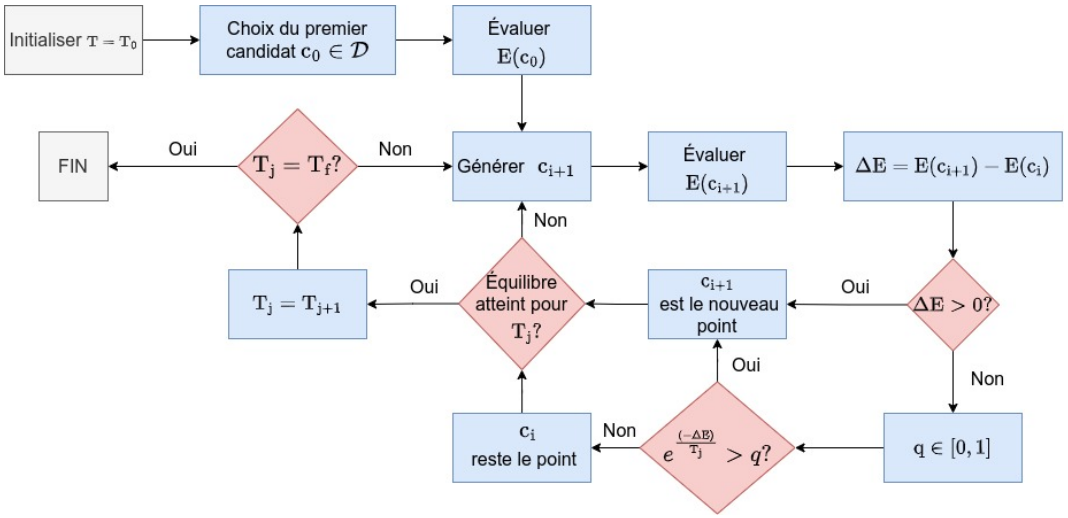

Figure 6.7 – Diagramme illustrant les principes itératifs du recuit simulé.

La Figure 6.7 permet d'illustrer le fonctionnement itératif du recuit simulé. Une fois la température T_0 initialisée, les itérations commencent par le choix aléatoire d'un candidat dans l'espace de recherche. Ce candidat sera le premier point de la trajectoire, qui servira à générer d'autres candidats. La fonction objectif est utilisée pour évaluer le candidat. Un autre candidat est ensuite généré en ajoutant le vecteur Δs_i à chaque variable du premier point. L'écart d'énergie entre les deux candidats permet de définir lequel servira de second point sur la trajectoire.
Si $\Delta E > 0$, alors le second candidat sera le second point et les prochains candidats seront générés à partir de ses coordonnées dans l'espace des solutions. Dans le cas contraire, la température T_0 permet de calculer la probabilité d'utiliser le second candidat dans la trajectoire avec l'Équation 6.2. Un nouveau candidat est généré et le processus est réitéré jusqu'à atteindre l'équilibre thermodynamique. Une fois atteint, la température diminue selon les règles de refroidissement et le processus recommence. Lorsqu'il y a un équilibre thermodynamique ($T = T_f$) ou lorsque le nombre d'itérations maximal est atteint, l'algorithme s'arrête et la solution est le point correspondant à la plus faible énergie observée.

Application au dimensionnement du micro-réseau

Comme défini précédemment, on cherche donc à déterminer les variables de dimensionnement de certaines unités. Les variables de dimensionnement fixées pour les autres unités, ainsi que les profils de demandes et les conditions météorologiques établissent le cadre du problème étudié. Les variables préfixées sont la capacité du stockage hydrogène, la somme des demandes annuelles, l'origine des données, la puissance maximale de fonctionnement de la batterie (voir section 4.1). L'objectif de l'apprentissage d'un agent de RL pour le contrôle du micro-réseau est aussi prédéfini.

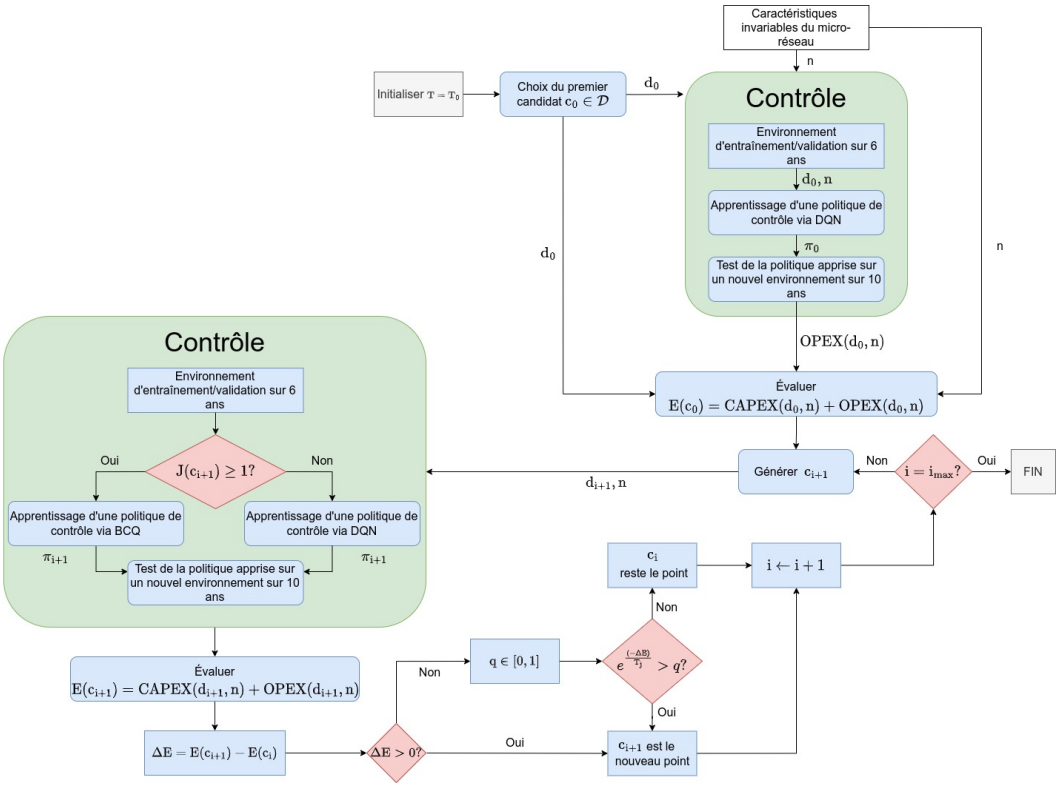

Figure 6.8 – Diagramme illustrant les principes de la méthodologie développée.

La Figure 6.8 illustre l'application du recuit simulé au dimensionnement sous contrôle optimal du micro-réseau électrique.

Initialisation et premier candidat Le premier candidat est sélectionné aléatoirement par l'algorithme de recuit simulé. L'algorithme de DQN permet d'apprendre une politique de contrôle sur l'environnement généré par le premier candidat. Les phases d'entraînement et de validation s'effectuent avec les mêmes données sur un horizon temporel de 6 ans. Une fois établie, la politique de contrôle est testée avec des données différentes de celles de l'entraînement et un horizon temporel de 10 ans. C'est lors de la phase de test que le critère lié au contrôle pour le dimensionnement du micro-réseau est calculé. Ce critère est l'OPEX puisque le coût global correspond à l'énergie à minimiser.
Le CAPEX est calculé directement à partir des coûts d'investissement et maintenance présentés à la section 4.3.

Énergie des nouveaux candidats Un nouveau candidat est désigné de manière aléatoire selon la longueur maximale de pas autorisé pour chaque variable depuis le point en cours. Afin de calculer son énergie, un nouvel apprentissage d'un agent est effectué avec le dimensionnement associé au candidat. Si la distance euclidienne entre le candidat et au moins un des anciens candidats vaut 1, le candidat a au moins un voisin et BCQ est employé pour l'apprentissage d'une politique de contrôle. Dans ce cas, les agents entraînés sur d'autres environnements remplissent la mémoire de relecture de l'agent à entraîner. Chaque démonstrateur créé une portion égale d'échantillons dans la mémoire de relecture. S'il n'y a qu'un démonstrateur, il en génère la totalité. Les démonstrateurs interagissent avec l'environnement défini par le candidat à évaluer.
Si le candidat n'a pas de voisin parmi les candidats précédents, DQN est employé. Quelque soit l'algorithme d'apprentissage utilisé, l'énergie du candidat est calculée de la même manière que le premier candidat.

Trajectoire Lorsque l'énergie d'un nouveau candidat a été évaluée, l'écart d'énergie entre ce candidat et le point de la trajectoire de l'algorithme détermine si le candidat évalué est le prochain point de la trajectoire. La méthodologie développée par le recuit simulé permet d'accepter un candidat à énergie plus élevée que le point actuel. Un nouveau candidat est systématiquement généré après le choix du point de la trajectoire et ces étapes sont répétées jusqu'à l'arrêt de l'algorithme. À chaque génération de candidat, un pas, représenté par un entier relatif, est choisi aléatoirement à partir de la solution courante. Ce pas détermine la direction et l'amplitude du changement pour obtenir le candidat suivant. Il est donc possible que l'algorithme sorte de l'espace de recherche depuis un point proche des bornes. Certains algorithmes règlent ce problème en ajoutant une pénalité à l'énergie de ces candidats mais ce n'est pas une solution envisageable puisque le recuit simulé accepte avec une certaine probabilité des points sous-optimaux dans sa trajectoire. Ces candidats sont donc ignorés par l'algorithme, qui génère un nouveau candidat depuis le point courant.
Lorsqu'un candidat courant est identique à un candidat généré à une itération précédente, son énergie est déjà connue, il n'est pas nécessaire de la recalculer.
Dans la méthodologie utilisée, la condition d'arrêt de l'algorithme est le nombre d'itérations. L'espace de recherche est petit donc la notion d'équilibre thermique est ignorée et la température diminue à chaque itération pour trouver rapidement l'optimum global.

Choix des paramètres

Les paramètres du recuit simulé ont été choisis empiriquement. L'écart moyen entre deux candidats successifs est d'environ 500 €. La différence d'énergie entre deux candidats, exprimée en euros (€), est calculée en multipliant par 0.1 leur écart de score. Cette normalisation est opérée afin de réduire les ordres de grandeur entre les différentes quantités. Des instances d'optimisation de 50 itérations sont effectuées.

Température initiale. La température initiale dans l'algorithme de recuit simulé est paramétrée de manière à attribuer une probabilité de 90% d'acceptation à un candidat dont l'énergie est supérieure de 50 € à celle du candidat courant. Ce paramétrage permet une exploration plus libre de l'espace des solutions dans les premières itérations de l'algorithme, favorisant la découverte du minimum global.

Coefficient de refroidissement. Les coefficients de refroidissement ont été calculés pour une diminution linéaire puis exponentielle de la température. Ils sont choisis en fonction de la température initiale et de la température finale souhaitée.
La probabilité d'accepter une solution augmentant l'énergie de 50€ est de 0,1 % à la dernière itération pour un refroidissement linéaire. Le coefficient linéaire de l'évolution de la température β_{lin} s'obtient en divisant l'écart entre la température initiale et la température finale par le nombre d'itérations.
Pour le coefficient de refroidissement exponentiel β_{exp}, la probabilité d'accepter un écart positif de 50€ est de 0,1 % pour 10 itérations avant la fin de l'optimisation.
L'expression de la température à chaque itération d'optimisation en fonction de β_{lin} et β_{exp} est donnée par les Équations 6.3 et 6.4 respectivement.

$$\mathrm{T}^{\circ}(i) = \mathrm{T}_0^{\circ} + (\beta_{\text{lin}} \times i) \tag{6.3}$$

$$\mathrm{T}^{\circ}(i) = \mathrm{T}_0^{\circ} \beta_{\text{exp}}^{i} \tag{6.4}$$

Avec i le nombre d'itérations et T_0° la température initiale.
Les deux modes de refroidissement seront comparés dans la sous-section 6.3.1.

Taille maximale de pas. La taille de pas est la longueur de déplacement sur une variable d'optimisation d'une itération à une autre. Puisque l'espace des variables est discret, la taille d'un pas est toujours entière. Elle est choisie en utilisant un entier relatif aléatoire dans un intervalle défini par la taille maximale de pas. Cette taille peut diminuer au long des itérations pour converger plus précisément vers un optimum global. Ce choix n'a pas été fait dans ce travail. Les variables sont discrètes et l'espace de recherche est relativement restreint, les gains en précision seraient moindres par rapport à un contexte où l'espace des solutions serait continu. La température étant basse dans les dernières itérations, des pas plus petits conduisent à explorer des candidats déjà évalués, réduisant ainsi l'efficacité de l'exploration.

Les valeurs numériques de ces paramètres sont données sur la Table 6.4.

Paramètre	**Valeur**
Température initiale T_0° (€)	474.5
Coefficient de refroidissement linéaire β_{lin} (€)	−9.5
Coefficient de refroidissement exponentiel β_{exp}	0.89
Nombre d'itérations	50
Taille maximale de pas pour $\mathrm{P}_{\text{nom}}^{\text{PV}}$ (kWc)	3
Taille maximale de pas pour $\mathrm{E}_{\text{nom}}^{\text{batt}}$ (kWh)	3

Table 6.4 – Valeurs des paramètres dans la méthode de recuit simulé appliqué au dimensionnement optimal du micro-réseau.

6.3 Résultats de l'optimisation bi-niveaux du micro-réseau et analyse de la méthodologie

Les critères de dimensionnement d'un micro-réseau varient selon sa configuration. Pour l'usager, dans un micro-réseau connecté, la priorité est de maximiser la rentabilité plutôt que

l'autonomie puisque le gestionnaire du réseau peut fournir l'électricité en cas de déséquilibre. À l'inverse, l'autonomie est la préoccupation principale dans le cas d'un micro-réseau isolé puisque la rentabilité n'a pas de sens dans cette situation. Il est envisageable de dimensionner un micro-réseau isolé en fonction du coût de l'électricité si son autonomie est assurée, mais comparer ce coût au tarif du réseau central n'est pas pertinent.
Bien que les coûts d'investissement et de remplacement des unités soient les mêmes pour un micro-réseau connecté ou îloté, les coûts liés au contrôle diffèrent selon la configuration. Deux critères sont donc employés pour définir des fonctions objectif différentes :

1 **Critère économique avec échange bilatéral.** La fonction objectif considère un coût de soutirage et un revenu d'injection au réseau central dans le calcul de l'OPEX du micro-réseau. Ce critère est adapté à un micro-réseau connecté pour maximiser la rentabilité de l'installation.

2 **Critère technico-économique de soutirage unilatéral.** Ce critère est adapté à la fois à un micro-réseau connecté pour lequel l'autonomie est recherchée ou à un micro-réseau isolé. Pour un micro-réseau connecté, cela implique la considération des coûts liés à l'achat d'électricité. Dans le cas d'un micro-réseau isolé, cela revient à maximiser la capacité d'autoproduction. Dans les deux situations, le coût d'investissement est également intégré. Ce critère est adapté pour un micro-réseau situé dans une zone avec des infrastructures de réseau limité qui ne permettent pas au réseau central de garantir l'approvisionnement électrique.

L'influence de ces deux fonctions récompenses pour entraîner des agents sera évaluée et leur impact sur le score des candidats sera étudié.

La section 6.3.1 expose les dimensionnements retenus par l'algorithme d'optimisation globale pour chaque configuration (micro-réseau connecté ou isolé). La performance et la fiabilité de la méthodologie sont analysées dans la section 6.3.2 selon le temps de calcul et la capacité d'exploration.

6.3.1 Analyse des résultats d'optimisation

Pour l'ensemble des résultats présentés par la suite, les coûts sont calculés sur 10 ans. Les panneaux PV, la PAC et le compresseur du stockage hydrogène ont des durées de vie estimées à 20 ans. Les coûts d'investissement sont intégrés à 50% dans le calcul du coût total.

Dimensionnement optimal d'un micro-réseau connecté au réseau central de distribution

Deux critères économiques vont être considérés dans le dimensionnement optimal. Les coûts et revenus pris en compte dans la phase d'opération considèrent :

— soit les échanges bilatéraux avec le réseau central, c'est-à-dire le coût de soutirage et le revenu lié à l'injection respectivement de 0,2276 €/kWh et 0,1339 €/kWh,

— soit uniquement le soutirage et le coût associé (0,2276 €/kWh).

Pour chacun de ces deux problèmes d'optimisation, deux systèmes de récompense seront utilisés pour l'entraînement des agents. Le premier pénalisera uniquement l'énergie soutirée au réseau central (agent EMS A), alors que le second pénalisera à la fois l'énergie soutirée et injectée (agent EMS B). Dans les deux cas, une composante supplémentaire pour minimiser l'usure de la batterie est ajoutée au système de récompense. Par la suite, pour discriminer les systèmes de récompense employés, des couleurs sont attribuées sur les figures. Le bleu sera utilisé pour l'EMS A, et le rouge sera utilisé pour l'EMS B.

Échanges bilatéraux avec le réseau central. Les figures 6.9 et 6.10 présentent les LCE sur 10 ans respectivement pour les EMS A et EMS B. Sur chaque figure sont représentés les différents candidats rencontrés lors du processus d'optimisation. La solution optimale est indiquée en rouge.

EMS A (Figure 6.9) - Le dimensionnement optimal comprend une installation PV de 4 kWc et une batterie de capacité 6 kWh. Le LCE est de 0,1787 €/kWh. L'augmentation de la puissance PV installée augmente le coût global du système. Les coûts d'opération ont un impact significatif sur le coût global. Le plus faible coût d'opération observé est négatif et vaut −286 €. Il est associé au candidat de dimensionnement [4 kWc, 6 kWh]. Le candidat associé au dimensionnement [3 kWc, 2 kWh] a un coût d'opération de 907 €, c'est le plus élevé parmi les candidats évalués. La différence entre les deux, 1193 €, est significative par rapport aux coûts d'investissement variables (panneaux PV et batterie).

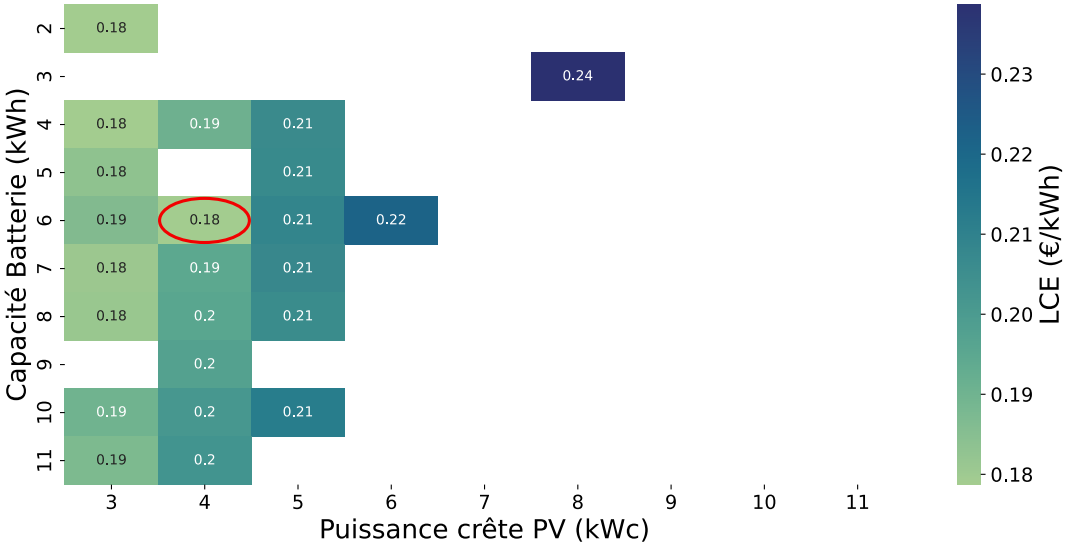

Figure 6.9 – LCE des différents dimensionnements candidats dans le cas d'échanges bilatéraux d'énergie. Le système de récompense pénalise l'énergie soutirée réseau central.

EMS B (Figure 6.10) - Le dimensionnement optimal obtenu correspond aux bornes minimales de l'espace des variables. La meilleure solution serait donc un micro-réseau avec 3 kWc de panneaux PV et 2 kWh de capacité nominale de batterie. Le LCE obtenu vaut 0,1871 €/kWh. Les augmentations de la puissance crête PV installée et de la capacité de la batterie ont tendance à augmenter le coût total du système. Comme on a pu le voir précédemment, un tel EMS favorise le stockage sur le long terme au détriment des revenus (voir section 5.1). En réduisant la capacité de la batterie, le coût d'investissement diminue et l'agent stocke moins d'énergie (voir section 5.2), qui est donc injectée au réseau central et engendre un revenu.
Le coût d'opération des micro-réseaux varie peu d'un dimensionnement à un autre en comparaison avec le coût d'investissement : la différence entre le coût d'opération le plus faible (822 € pour [4 kWc, 9 kWh]) et le plus élevé (1281 € pour [3 kWh, 2 kWh]) est de 460 €. C'est pourquoi le dimensionnement avec les plus faibles coûts d'investissement est le dimensionnement optimal dans cette configuration et avec cet EMS. Il est à noter qu'une composante du système de récompense de l'EMS B (pénalisation de l'injection) est en contradiction avec l'objectif d'optimisation des revenus. Ceci pourrait expliquer la faible variation du coût d'opération d'un dimensionnement à un autre.

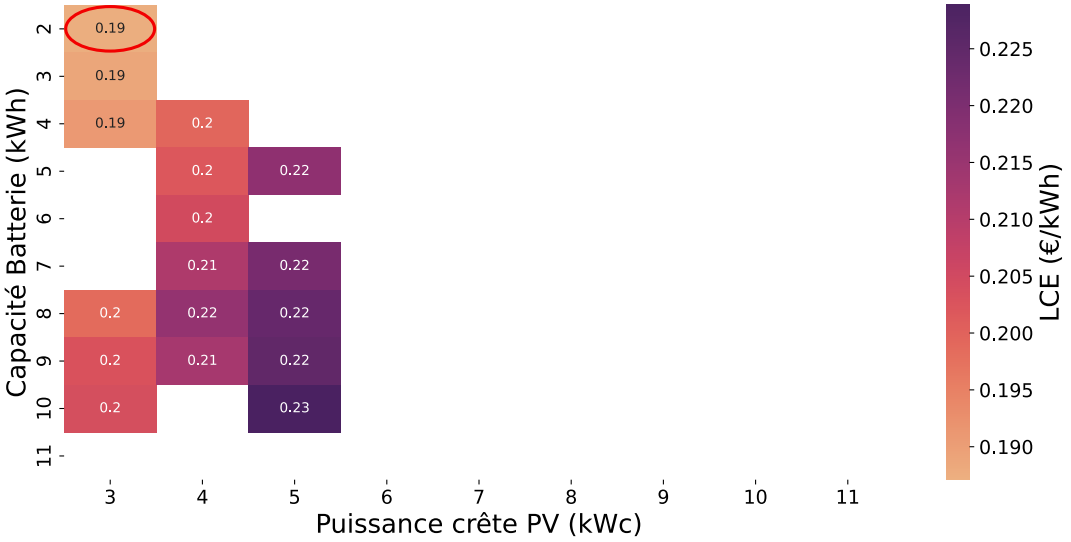

Figure 6.10 – LCE des différents dimensionnements candidats dans le cas d'échanges bilatéraux d'énergie. Le système de récompense pénalise tout échange avec le réseau central.

Les LCE calculés pour les deux EMS des micro-réseaux sont très proches.
L'EMS A injecte suffisamment d'électricité pour que le coût d'opération soit négatif. Une surface PV plus importante permet d'injecter une quantité plus élevée d'électricité. Avec une plus grande capacité de la batterie, l'EMS valorise mieux le productible PV (voir section 5.2). Les coûts d'opération dépendent fortement de l'investissement initial pour ce premier EMS. L'optimisation a convergé selon un compromis entre OPEX et CAPEX.
Avec l'EMS B, les coûts d'opération sont similaires d'un dimensionnement à un autre. L'optimisation du dimensionnement du micro-réseau a convergé en minimisant le CAPEX puisque l'OPEX varie peu. L'EMS n'utilise que très peu l'électrolyseur et le stockage hydrogène ne sert pas de stockage à long terme.
La similitude entre les deux LCE obtenus interroge sur l'intérêt d'utiliser un stockage hydrogène, qui est l'investissement le plus coûteux, dans cette configuration.

Soutirage unilatéral. Ici, seul le soutirage au réseau central est considéré dans le calcul des coûts d'opération. L'énergie produite en excès et non-stockée est perdue.

EMS A (Figure 6.11) - Comme le système de récompense employé pour l'entraînement de l'EMS A pénalise l'énergie soutirée au réseau central, et ainsi maximise le taux d'autoproduction, il est adapté pour minimiser la fonction objectif. En effet, la seule variable avec une influence sur le coût d'opération est l'électricité soutirée. Les résultats des différentes itérations de l'algorithme d'optimisation sont montrés sur la Figure 6.11. L'algorithme converge vers le dimensionnement [3 kWc, 10 kWh]. Le LCE correspondant vaut 0,2171 €/kWh. La puissance PV installée a une influence significative sur le coût global pour les valeurs explorées par l'algorithme. L'accroissement de la puissance PV installée augmente systématiquement le coût du système. Sa valeur minimale correspond à la solution optimale. Le lien entre la capacité nominale de la batterie et le coût du micro-réseau n'est pas aussi évident. À puissance PV fixe, l'augmentation de la capacité de la batterie semble diminuer le coût global du système jusqu'à un certain seuil, malgré l'augmentation concomitante du coût d'investissement. Pour une puissance PV de 3 kWc, ce seuil correspond à une capacité de 10 kWh.

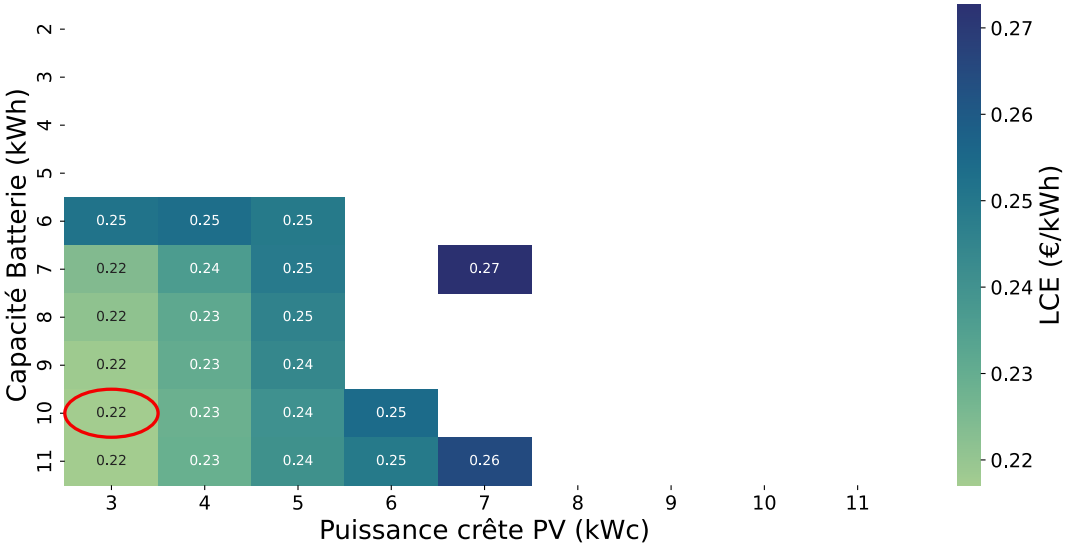

Figure 6.11 – LCE des différents dimensionnements candidats dans le cas de soutirage d'électricité. Le système de récompense pénalise l'énergie soutirée au réseau central.

EMS B (Figure 6.12) - Le système de récompense de l'EMS B est employé pour maximiser à la fois le taux d'autoproduction et d'autoconsommation (voir section 4.2). La Figure 6.12 présente les LCE obtenus par les candidats du processus d'optimisation dans la configuration d'échange unilatéral avec le réseau central. Le dimensionnement optimal obtenu est [3 kWc, 10 kWh]. Il est identique au dimensionnement optimal obtenu avec l'EMS A. La valeur de LCE correspondante, 0,2224 €/kWh, est plus élevée que celle obtenue par l'EMS A. Cela peut s'expliquer par le développement d'une stratégie de contrôle plus cohérente avec la fonction objectif qu'avec l'EMS A. Les mêmes observations sur l'influence du dimensionnement de la puissance PV et de la capacité de la batterie peuvent être faites.

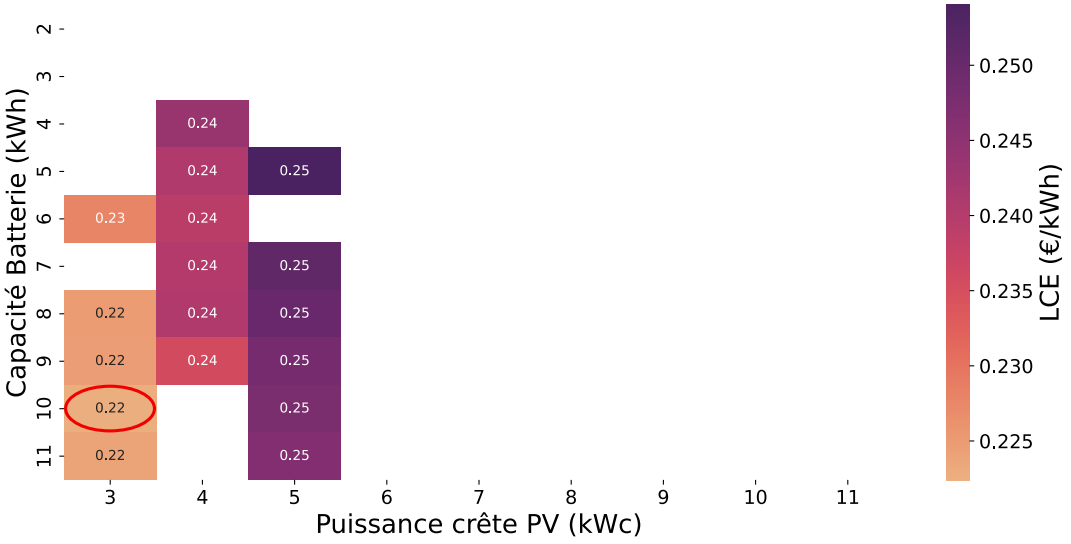

Figure 6.12 – LCE des différents dimensionnements candidats dans le cas de soutirage d'électricité. Le système de récompense pénalise tous les échanges avec le réseau central.

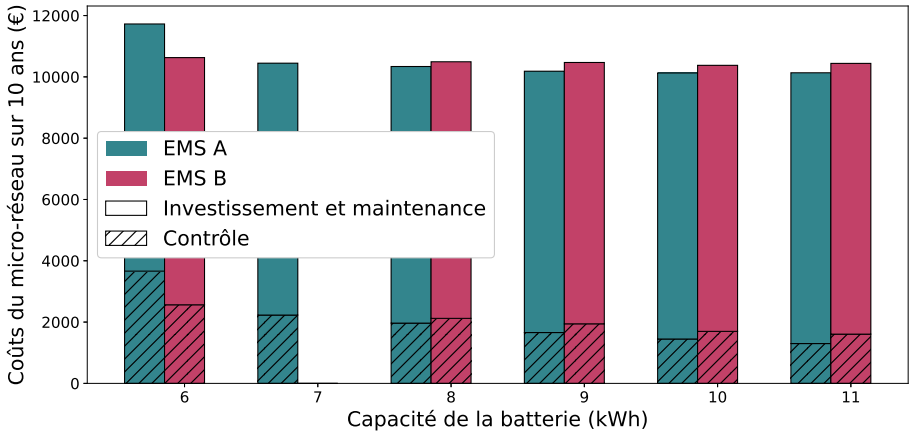

Figure 6.13 – Coûts d'opération, d'investissement et maintenance des micro-réseaux en fonction de la capacité de la batterie et de l'EMS pour une puissance crête PV de 3 kWc.

La Figure 6.13 montre les différents coûts obtenus avec une puissance PV installée fixée à 3 kWc pour les deux instances d'optimisation. Les coûts d'opération diminuent quand la capacité de la batterie augmente jusqu'à 10 kWh pour les deux EMS. Les coûts de l'EMS A sur l'environnement avec 6 kWh se révèlent sensiblement supérieurs aux autres. C'est également le seul cas où les coûts d'opération de l'EMS A sont supérieurs à ceux de l'EMS B pour le même environnement. Cela suggère que l'entraînement de l'agent n'a pas été optimal sur cet environnement. Sans considérer ce dimensionnement, les coûts totaux ne semblent pas varier significativement en fonction de la capacité de la batterie. Les coûts d'investissement et de maintenance liés à la batterie sont du même ordre de grandeur que le coût d'opération.

Conclusion

La Figure 6.14 expose les coûts des dimensionnements optimaux selon la nature de raccordement du micro-réseau au réseau central et la stratégie de l'EMS employée.

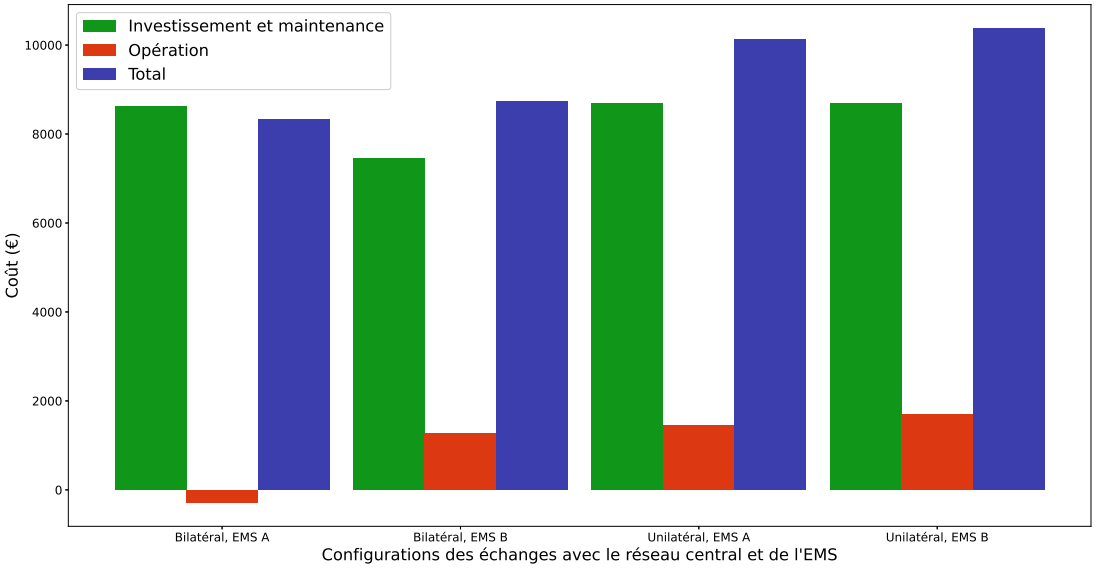

Figure 6.14 – Répartitions des coûts des dimensionnements optimaux pour chaque configuration du micro-réseau.

Le coût du système est forcément plus faible si l'injection du surplus énergétique au réseau central est valorisée financièrement. Le système de récompense le plus approprié pour ces fonctions objectif économiques est la pénalisation de l'énergie soutirée au réseau central. Les indicateurs économiques et techniques des dimensionnements retenus selon la liaison entre les réseaux figurent dans la Table 6.5.

Configuration	EMS	Puissance PV (kWc)	Capacité Batterie (kWh)	LCE (€/kWh)	τ_{autoprod}	τ_{autocons}
bilatéral	A	4	6	0.1787	0.71	0.62
bilatéral	B	3	2	0.1871	0.64	0.70
unilatéral	A	3	10	0.2171	0.88	0.87
unilatéral	B	3	10	0.2224	0.85	0.91

Table 6.5 – LCE, taux d'autoproduction et taux d'autoconsommation sur 10 ans des micro-réseaux dimensionnés selon chaque configuration.

La table montre que le contrôle des micro-réseaux dimensionnés pour le soutirage unilatéral autorise des taux d'autoproduction et d'autoconsommation plus importants. Les systèmes de stockage ont une plus grande capacité et la puissance crête des panneaux PV est minimale, il y a alors moins d'excès d'énergie ce qui augmente le taux d'autoconsommation.
La minimisation de l'énergie injectée revient à maximiser le taux d'autoproduction (voir Équation 4.38).

Dimensionnement optimal d'un micro-réseau isolé

Le micro-réseau isolé du réseau central de distribution est un système fermé. Pour son dimensionnement, le calcul des coûts d'investissement et de maintenance est identique à la sous-section précédente. En revanche, les coûts d'opération ne dépendent plus d'un tarif extérieur. Le micro-réseau n'est pas dimensionné pour rentabiliser des revenus fonctions d'échange d'électricité mais pour satisfaire les demandes. La contribution des coûts d'opération

dans la fonction objectif correspond à des préférences ou des contraintes de l'utilisateur.
Le micro-réseau est dimensionné pour minimiser les demandes non-satisfaites et les coûts d'investissement et de maintenance. En appelant Δ_{dem} cette demande non-satisfaite aussi appelée déficit énergétique, l'Équation 6.5 montre la fonction objectif du dimensionnement du micro-réseau isolé.

$$F(\text{d, n}) = \text{K}\Delta_{\text{dem}}(\text{d, n}) + \text{C}_{\text{inv}}(\text{d, n}) + \text{C}_{\text{maint}}(\text{d, n}) \tag{6.5}$$

Avec d le couple de variable de dimensionnement d'un candidat, n les caractéristiques fixes du micro-réseau, C_{inv} le coût d'investissement, C_{maint} le coût de maintenance et K un facteur pour pondérer la fiabilité par rapport au coût économique. K peut être considéré comme un tarif fictif du déficit énergétique. Il s'exprime donc en €/kWh. Ainsi, on peut retrouver la fonction objectif utilisée pour le dimensionnement avec injection unilatérale de la sous-section précédente en remplaçant K par 0,2276 €/kWh. Différentes instances d'optimisation seront initialisées avec des valeurs de K croissantes, les solutions optimales seront comparées pour chaque valeur de K.
Dans cette étude, le système de récompense utilisé pour construire la politique de l'EMS consiste en la minimisation du déficit énergétique.

Résulats. Selon les valeurs de K, l'algorithme d'optimisation converge vers des solutions différentes présentées dans la table 6.6. Ces intervalles de tarifs sont larges et englobent des tarifs significativement supérieurs à ceux auxquels serait confronté le micro-réseau si il était connecté (plus de vingt fois supérieurs). Très peu de dimensionnements optimaux différents ont été obtenus dans cet intervalle de paramètre K. Dans la plus large gamme de prix raisonnables (inférieurs à 3€/kWh), le dimensionnement sélectionné serait [3 kWc, 11 kWh].

Analyse. Le choix du dimensionnement optimal varie peu par rapport au tarif associé au déficit énergétique. Cela peut indiquer que les gains en autonomie réalisés par le choix du dimensionnement sont relativement mineurs par rapport aux autres coûts. Une fois les dimensionnements optimaux calculés, le choix doit porter sur d'autres indicateurs techniques et environnementaux pertinents comme le LPSP, le REE et la FER. La table 6.6 montre la valeur de ces indicateurs pour les dimensionnements obtenus. Le REE et la FER sont calculés par les Équations 1.2 et 1.3 (section 1.3) avec l'énergie en déficit au dénominateur (au lieu de l'énergie soutirée) puisque le micro-réseau n'est pas connecté. Plus le prix associé au déficit énergétique est élevé, plus le LPSP est faible et la FER élevée. Les variations du REE ne semblent pas directement liées à la fonction objectif.

K (€/kWh)	Puissance PV (kWc)	Capacité Batterie (kWh)	LPSP	REE	FER
0.22-0.25	3	10	0.161	0.138	0.906
0.25-3	3	11	0.145	0.154	0.915
3 et plus	5	11	0.139	0.133	0.918

Table 6.6 – Dimensionnement optimaux, LPSP, REE et FER sur 10 ans pour différents tarifs de déficit énergétique.

Conclusion. Les valeurs de LPSP obtenues demeurent significativement élevées pour un micro-réseau isolé. Des valeurs de LPSP égales à 0,2 % avaient été obtenues avec une demande énergétique moins élevée (Table 5.5 du chapitre 5) sans rechercher un dimensionnement optimal. La borne supérieure de la capacité nominale de la batterie est atteinte pour des tarifs de déficit supérieurs à 0,25 €/kWh. Cette contrainte aurait pu être allégée en augmentant la valeur de cette borne supérieure.
La méthodologie permet de réduire le risque de pannes, mais le choix du tarif de l'énergie

en déficit est trop arbitraire. Il serait pertinent dans les recherches futures de considérer les indicateurs liés à l'autonomie de l'utilisateur. En effet, la valeur minimale du LPSP obtenu est 0.139 alors qu'une valeur acceptable serait 0.02 (H. Yang et al., 2008). Une technique d'optimisation multi-critère devrait être employée car la fonction objectif aurait des composants liés à des coûts monetaires et à l'autonomie du système.

6.3.2 Évaluation de la convergence et du temps de calcul dans le processus d'optimisation

Après avoir analysé les résultats de l'optimisation, la méthodologie elle-même est évaluée. Le but est d'estimer l'efficacité de l'approche de dimensionnement par optimisation méta-heuristique au regard de ses performances en termes de temps de calcul et de sa capacité à explorer efficacement l'espace des solutions. Cette analyse est importante pour examiner la capacité d'adaptation de la méthodologie proposée.
L'analyse de la convergence de la méta-heuristique puis les temps de calculs seront exposés.

Analyse de la convergence

Les paramètres du recuit-simulé ont été présentés à la sous-section 6.3.1. Deux méthodes de refroidissement ont été envisagées : une décroissance de la température linéaire ou exponentielle. Les résultats obtenus avec les deux méthodes sont ici comparés.

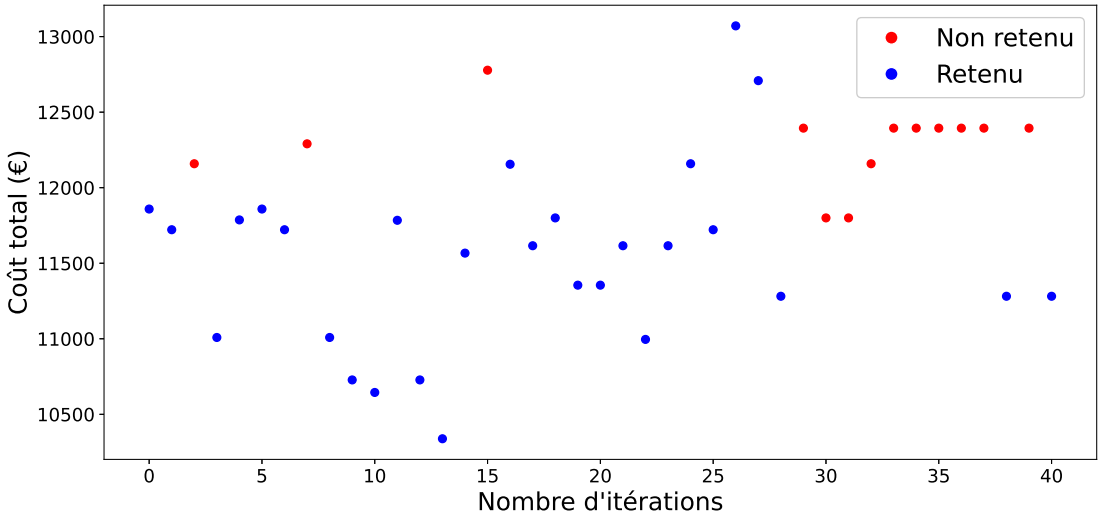

Figure 6.15 – Représentation des candidats intégrés à la trajectoire durant l'optimisation du dimensionnement avec refroidissement linéaire de la température.

Refroidissement linéaire. L'exploration de l'espace des solutions avec refroidissement linéaire du recuit simulé est visualisée sur la Figure 6.15. L'axe des abscisses représente le nombre d'itérations pour lesquels le candidat proposé est compris dans le domaine de recherche $\mathcal{I}$ (voir section 6.2). L'axe des ordonnées est le score de chaque candidat pour une fonction objectif économique avec une configuration unilatérale. La couleur de chaque point diffère selon que le candidat représenté a été intégré à la trajectoire du recuit simulé ou non. Le bleu est associé aux candidats retenus et le rouge aux candidats non-retenus. La fréquence de candidats retenus ayant une énergie plus élevée que le

point de la trajectoire est très élevée avant l'itération 29. Elle diminue significativement après et les scores ne s'améliorent plus. Le coût le plus faible a été observé à l'itération 13 et correspond au dimensionnement [3 kWc, 8 kWh]. Il est de 10 339 €. En acceptant un candidat avec une énergie bien plus élevée dans sa trajectoire lors de l'itération suivante, l'algorithme quitte une région à haut potentiel (voir Figure 6.11). Il converge vers une solution sous-optimale lorsque la température devient faible. La dernière solution retenue est [3 kWc, 2 kWh] dont le coût est 11 281 €.

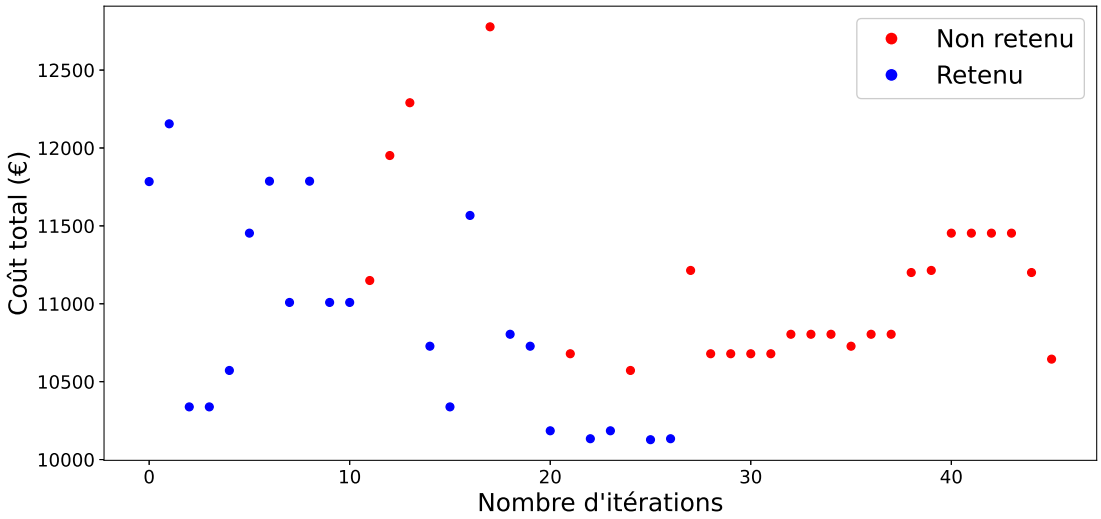

Figure 6.16 – Représentation des candidats intégrés à la trajectoire durant l'optimisation du dimensionnement avec refroidissement exponentiel de la température.

Refroidissement exponentiel. La même visualisation est construite et analysée pour une diminution exponentielle de la température en fonction des itérations sur la Figure 6.16. La fréquence de candidats acceptés alors que leur énergie est supérieure au point de la trajectoire diminue vers la moitié de l'instance d'optimisation. Aucun candidat avec une énergie supérieure n'est retenu au cours des 20 dernières itérations. L'écart minimal des scores entre ces candidats et le dernier point de la trajectoire est de l'ordre de 500€, ce qui justifie le taux d'acceptation nul à ce stade de l'instance d'optimisation. L'optimisation a convergé vers le dimensionnement [3 kWc, 11 kWh], dernier point de la trajectoire. Le score associé est 10 134 € à l'itération 26. Son coût est plus élevé qu'à l'itération précédente, qui est le minima observé. L'optimum est associé au dimensionnement [3 kWc, 10 kWh] dont le score est 10 128 €.
La convergence est plus rapide que pour le refroidissement linéaire. C'est pourquoi le refroidissement exponentiel du recuit simulé a été choisi.

Temps de calcul et intérêt du transfert de politique

Cette sous-section présente le temps de calcul de l'algorithme d'optimisation et l'influence de l'utilisation d'apprentissage par renforcement hors ligne.

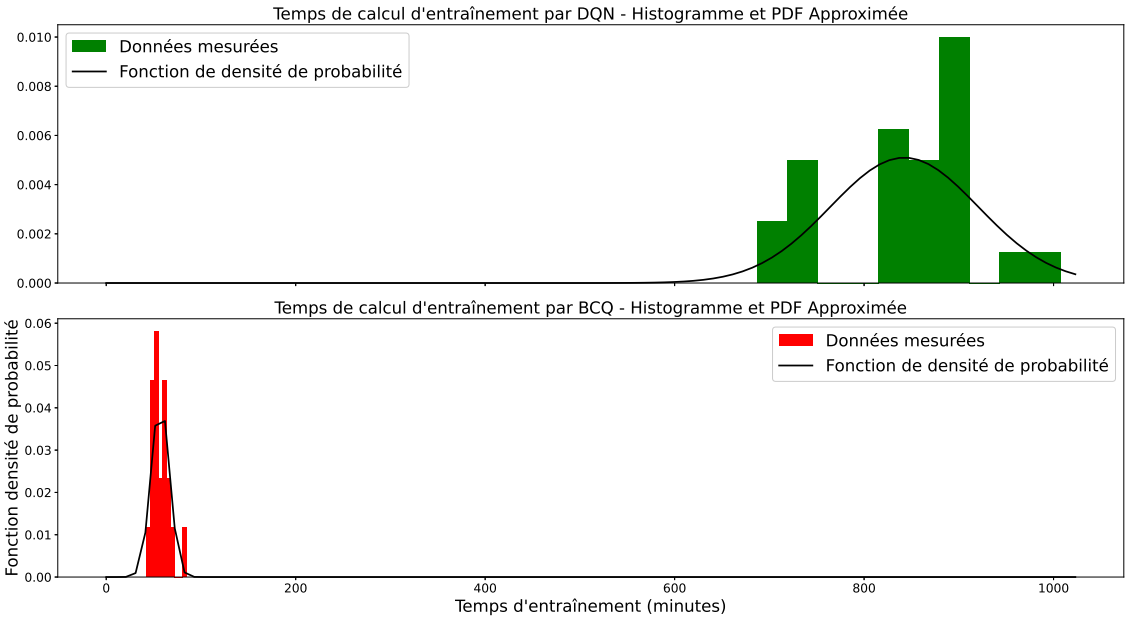

Figure 6.17 – Histogramme des temps d'entraînement des agents de DQN (en haut) et de BCQ (en bas) et approximation des fonctions densité de probabilité.

Temps de calcul et entraînement. La durée moyenne d'une instance d'optimisation avec la méthodologie présentée est de 4 jours et 10 heures. C'est l'entraînement des agents qui occupe l'essentiel du temps de calcul.

Lorsqu'un candidat est proposé par l'algorithme d'optimisation et qu'il a été préalablement déjà proposé, l'entraînement d'un agent n'est pas nécessaire et le score est simplement ré-utilisé. En revanche, si un nouveau candidat est associé à un dimensionnement inconnu de l'algorithme, le score lié au contrôle est évalué après l'entraînement d'un agent sur un horizon temporel simulé de 10 ans. Si le candidat admet au moins un voisin (voir section 6.1), l'entraînement est effectué avec un algorithme de BCQ. Dans le cas contraire, DQN est employé pour son entraînement.

L'entraînement hors ligne d'un agent avec BCQ se déroule en deux étapes. La mémoire de relecture de l'agent à entraîner est remplie par un ou des agents déjà entraînés appelés démonstrateurs, qui interagissent avec son environnement selon leur politique de contrôle. Un facteur aléatoire modifie les actions prises par les démonstrateurs pour augmenter la diversité des séquences échantillonnées. De ce fait, plusieurs épisodes d'interactions sont requis. Dans un second temps, l'agent à entraîner échantillonne les séquences d'actions de sa mémoire de relecture pour mettre à jour sa stratégie de contrôle sans interagir directement avec son environnement.

Puisque la mémoire de relecture d'un agent a une taille fixe, définie à la section 5.1, son temps de remplissage est fixe et dure 30 minutes. La seconde étape a une durée qui dépend de la condition d'arrêt de l'entraînement (la patience) et dure en moyenne 27 minutes. La durée moyenne de l'entraînement d'un agent par la méthodologie de transfert de politique développée est donc de 57 minutes.

La durée de l'entraînement d'un agent par l'algorithme de DQN est très variable et dépend du temps nécessaire pour atteindre le seuil de patience. En moyenne, le temps d'entraînement d'un de ces agents est de 14 heures et 2 minutes.

L'approximation de fonctions densité de probabilité du temps d'entraînement des agents selon l'algorithme est montrée sur la Figure 6.17.

La durée d'un entraînement par BCQ est en moyenne 14.75 fois plus courte que celle

d'un entraînement par DQN. Le gain de temps obtenu grâce à la méthodologie dépend du nombre d'agents entraînés par BCQ.

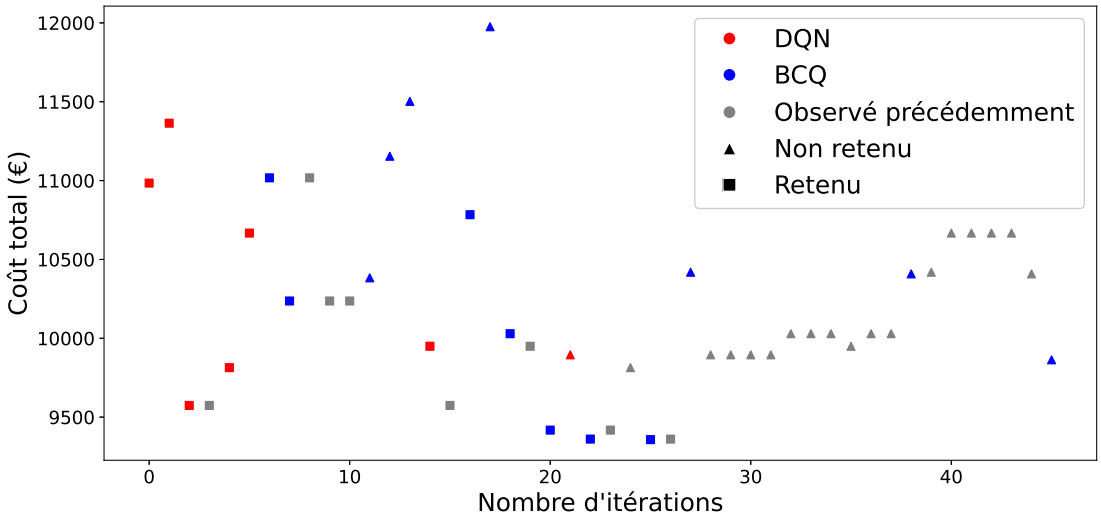

Figure 6.18 – Représentation des candidats intégrés à la trajectoire durant l'optimisation du dimensionnement selon les algorithmes utilisés pour construire la politique de contrôle de l'EMS.

Fréquence d'activation des algorithmes d'entraînement. La méthodologie développée réduit le temps des calculs sans besoin de les paralléliser. L'efficacité dépend de la fréquence d'utilisation de l'algorithme de BCQ pendant l'optimisation bi-niveaux. La Figure 6.18 montre la répartition des algorithmes utilisés pour entraîner les agents pendant le processus d'optimisation. Lorsqu'un candidat est associé à un dimensionnement qui a déjà été observé, aucun agent n'est entraîné, les scores sont réutilisés. Sur cette instance, 14 agents ont étés entraînés par BCQ et 7 par DQN. En utilisant les moyennes de temps de calcul déterminées auparavant, le temps de calcul avec la méthodologie développée est 2.64 fois plus rapide que si les entraînements avaient uniquement été effectués avec DQN.
On observe que la fréquence des entraînements effectués avec BCQ augmente au long des itérations d'optimisation. Dans cette étude, l'espace de recherche est restreint, le nombre d'itérations requis pour atteindre une solution est relativement bas. Cependant, la méthodologie proposée permettrait des économies de temps bien plus significatives si la fréquence des entraînements avec BCQ était accrue.
De nombreux dimensionnements déjà fournis par l'algorithme sont de nouveaux proposés à plusieurs reprises. Ils peuvent permettre de sortir d'une région à faible potentiel si ils sont intégrés à la trajectoire. En revanche, leur proposition est inadaptée pour une recherche locale en fin d'optimisation. Une liste taboue pourrait alors éviter cette situation afin de privilégier la découverte d'autres solutions locales.

6.4 Conclusion

L'utilisation combinée d'optimisation méta-heuristique et d'apprentissage par renforcement permet de dimensionner le micro-réseau malgré de nombreux facteurs aléatoires. L'approche d'optimisation bi-niveaux traite le contrôle comme un sous-problème d'optimisation. La

fonction objectif globale dépend des résultats de ce sous-problème. Une méta-heuristique est adaptée à la résolution du problème global en explorant l'espace de manière efficace afin de ne pas converger dans une région sous-optimale. Parmi les méta-heuristiques, une approche basée sur la trajectoire d'exploration de l'espace des solutions a été choisie. Le recuit simulé est utilisé en raison de sa faible complexité de calcul. Ces paramètres ont été sélectionnés empiriquement, en privilégiant un refroidissement exponentiel. Le besoin d'une exploration efficace de l'espace des solutions est crucial car le calcul des coûts dans la sous-boucle d'optimisation est long.
Le temps d'apprentissage étant conséquent, une méthode de transfert de politique a été développée. L'utilisation de l'apprentissage par renforcement hors ligne a permis d'apprendre des stratégies de contrôle des unités sans interaction aléatoire avec la simulation du micro-réseau. L'apprentissage des stratégies s'effectue selon l'observation des décisions prises par un modèle entraîné sur d'autres dimensionnements à des itérations antérieures. Avec cette méthode, le temps d'entraînement a été divisé par 5 sur un horizon temporel de simulation de 1 an et par 14 sur un horizon d'entraînement de 6 ans par rapport à l'utilisation d'algorithme d'apprentissage par renforcement classique. La qualité de la politique de contrôle apprise ne s'en trouve pas pour autant dégradée.
En combinant le recuit simulé pour une exploration plus efficace de l'espace des solutions et le transfert de politique entre les dimensionnements explorés, le gain en temps de calcul est significatif. La fréquence d'apprentissage hors ligne augmente tout au long du processus d'optimisation. La méthodologie est appliquée à un espace de recherche restreint et le dimensionnement optimal aurait pu être trouvé par recherche exhaustive. Cependant, l'objectif de cette recherche s'étend à la conception d'une méthodologie efficace et capable de s'adapter à des problèmes aux dimensions supérieures.
L'optimisation ne converge pas vers les mêmes solutions selon la configuration de la fonction objectif ou du système de récompense employé pour développer les politiques de contrôle. Il a été observé que lorsque la fonction objectif et le système de récompense sont en partie contradictoires, alors la solution trouvée consiste à minimiser la taille des unités pour réduire les coûts d'investissement. L'augmentation de la capacité de la batterie est privilégiée par rapport à celle de la surface des panneaux PV dans les autres configurations. La méthodologie pourrait être testée dans une région ou l'ensoleillement est plus faible pour vérifier la consistance de cette observation. Les bornes maximales de la capacité nominale de la batterie pourraient être augmentées pour dimensionner un micro-réseau isolé du réseau central.
Les puissances d'opération de la PAC et de l'électrolyseur ont été considérées comme fixes dans ce travail. Ce choix limite l'intérêt du stockage hydrogène, en particulier lorsque la demande nette est forte ou que la capacité de la batterie est faible. Une piste de recherche future pourrait consister à les intégrer dans les variables de dimensionnement et d'entraîner les agents à choisir leur puissance d'opération. Des algorithmes compatibles avec un espace d'actions continu pourraient être utilisés comme DDPG, SAC, TRPO et PPO. Puisque BCQ fonctionne en espace d'actions continu, le transfert de politique pourrait être envisagé.

Conclusions générales et perspectives

Conclusions générales

L'objectif de cette thèse est de développer un outil méthodologique qui aborde conjointement le dimensionnement et le contrôle d'un micro-réseau électrique avec des sources d'énergie renouvelable. Le micro-réseau étudié est composé de panneaux photovoltaïques, d'une batterie lithium-ion, et d'un stockage hydrogène comprenant un électrolyseur et une pile à combustible. Des données sont utilisées pour la modélisation de la production photovoltaïque et de la demande énergétique, et des modèles de comportement dynamique sont employés pour la modélisation des systèmes de stockage. L'apprentissage par renforcement est sélectionné pour l'élaboration des stratégies de contrôle. Il est implémenté progressivement sur des simulations de plus en plus complexes du micro-réseau avec un horizon temporel de simulation passant de 1 à 10 ans. La mise en œuvre de cet algorithme a pour objectif de réduire le coût de l'électricité en phase d'opération dans un environnement dans lequel les variables aléatoires ne sont pas prédites et le comportement dynamique des unités n'est pas explicitement formulé. L'efficacité de la politique de contrôle obtenue est évaluée selon le coût nivelé de l'énergie avec des données différentes que lors de l'élaboration de la stratégie de contrôle. Une analyse de la stratégie de contrôle est menée selon différents indicateurs économiques et techniques. La stratégie de contrôle est ensuite implémentée dans un processus de dimensionnement optimal du micro-réseau à deux niveaux. Un optimiseur de type méta-heuristique détermine la valeur optimale des variables de dimensionnement du micro-réseau, à savoir la puissance nominale des panneaux photovoltaïques et la capacité nominale de la batterie, pour réduire les coûts du système. L'apprentissage par renforcement est employé à plus bas niveau pour déterminer les coûts d'opération. Les temps de calcul sont conséquents en raison de l'apprentissage des agents, diverses techniques sont employées pour le réduire. Ainsi, afin de réduire significativement les temps de calcul, les agents sont entraînés sur un horizon temporel plus court et une méthode de transfert de politique de contrôle entre les différents agents est proposée. La méthodologie de dimensionnement bi-niveaux est évaluée selon sa capacité à explorer l'espace des solutions et son temps de calcul. À l'issue de ces travaux de recherche, nos principales conclusions sont proposées dans les paragraphes suivants.

Analyse de l'impact de différents facteurs sur l'apprentissage d'une politique de contrôle par DQN dans le cadre du contrôle séquentiel du stockage hydrogène dans un micro-réseau avec génération d'énergie renouvelable.

Les hyper-paramètres exercent une influence importante sur la garantie et la vitesse de convergence, ainsi que sur la qualité de la politique développée. Bien que la problématique du contrôle séquentiel du stockage hydrogène d'un micro-réseau à un pas horaire et sur un horizon d'au moins un an ait été présentée dans de nombreux travaux, l'analyse de sensibilité des hyper-paramètres n'a jamais été exposée.
Au cours de son entraînement, l'agent apprend à stocker de plus en plus d'hydrogène en été pour l'hiver. Le taux de remplissage de la réserve d'hydrogène augmente progressivement au cours des itérations d'entraînement jusqu'à un certain seuil. Ce seuil dépend du système de récompense choisi pour l'agent. Si l'agent minimise l'énergie soutirée au réseau central, ce seuil sera plus bas que si l'agent minimise tout échange. L'agent apprend à ne pas se servir du stockage hydrogène lorsqu'il est entraîné pour maximiser les gains économiques liés à l'achat et à la vente d'électricité au réseau central avec un tarif fixe, ce qui remet en question l'intérêt d'un tel stockage dans ces conditions.
Pour une simulation d'un an, le taux de remplissage initial du stockage hydrogène influence fortement la politique apprise. Un EMS favorisera une utilisation plus fréquente de l'électrolyseur si le taux de remplissage initial d'hydrogène est faible.
Si quatre états observables suffisent à garantir une politique de contrôle optimale sur un environnement simple, ils ne sont plus suffisants pour assurer la convergence vers une politique optimale dans un environnement plus complexe. Le contrôle du stockage hydrogène, dans un environnement dans lequel la dégradation de la batterie est modélisée et l'horizon temporel est plus long, nécessite au moins deux états supplémentaires pour être performant : le niveau de remplissage du stockage hydrogène et le niveau de dégradation de la batterie. La dégradation de la batterie est un état observable important pour contrôler efficacement le stockage hydrogène dans un environnement où la batterie vieillit selon son utilisation. Le niveau de remplissage du réservoir d'hydrogène est instauré pour permettre à l'agent de généraliser l'apprentissage d'une politique sur un environnement de simulation plus long. L'augmentation du nombre de dimensions de l'espace d'état impose la réévaluation des hyper-paramètres pour garantir une convergence stable de l'entraînement.
Dans le cas de simulations plus simples de l'environnement, certains paramètres tels que le facteur d'actualisation, la fréquence de validation et la taille des batchs ne conditionnent pas la convergence, mais influencent sa vitesse. En revanche, les valeurs de ces paramètres conditionnent la convergence de l'entraînement dans les environnements de simulation complexifiés.

Influence de la dégradation de la batterie et de l'horizon temporel d'entraînement d'un agent d'apprentissage par renforcement sur la politique de contrôle développée.

Deux modèles de dégradation de la batterie sont développés avec des approches empiriques issues de la littérature : un modèle de dégradation linéaire selon le nombre de cycles d'utilisation et un modèle non-linéaire selon la profondeur de décharge. Le coefficient linéaire de dégradation du premier modèle est ajusté pour suivre un comportement similaire au second modèle pour une même utilisation de la batterie. Malgré cette correction, le mode de dégradation de la batterie a une influence significative sur la politique apprise par l'agent. Un agent entraîné sur un modèle de dégradation voit sa performance considérablement affectée lorsqu'il est testé dans un environnement qui utilise un autre modèle de dégradation.
L'horizon temporel de simulation est aussi à considérer lorsqu'un agent est entraîné. Le déploiement sur un système réel d'un EMS entraîné pour contrôler un micro-réseau sur une courte période est à proscrire. En effet, la capacité d'un agent à généraliser sa politique de

contrôle sur un horizon temporel long est insuffisante si son espace d'état ne comprend pas toute les variables utilisées. Malgré l'ajustement des états observables et des hyper-paramètres, l'agent ne peut pas appliquer une stratégie de contrôle efficace sur 10 ans si son entraînement se limite à une période inférieure à 6 ans. En revanche, les travaux menés ont permis de développer un politique similaire pour des agents entraînés sur des horizons de 6 et 10 ans. La stratégie optimale obtenue montre que le taux de remplissage maximal annuel de l'hydrogène diminue avec le temps. La dégradation de la batterie conduit l'agent à combler la capacité dégradée de stockage à court terme par l'utilisation de la pile à combustible, même lorsque la production photovoltaïque est forte. Ainsi, les pics d'hydrogène stocké l'été sont de plus en plus faibles et le micro-réseau perd en autonomie lors des périodes de plus faible production.

Intérêt de l'apprentissage par renforcement hors ligne dans l'élaboration d'une politique de contrôle pour l'optimisation bi-niveaux du dimensionnement du micro-réseau.

Un entraînement est nécessaire pour chaque nouveau dimensionnement parcouru lors des itérations de la boucle principale de l'optimisation bi-niveaux du dimensionnement du micro-réseau. L'algorithme de DQN met en moyenne 14 heures de temps de calcul avant l'arrêt des entraînements avec un horizon temporel de 6 ans.
Le temps de calcul nécessaire à l'apprentissage est réduit en utilisant l'apprentissage hors ligne. Lorsque au moins un agent est entraîné sur un dimensionnement proche de celui de l'itération actuelle, il interagit avec l'environnement de l'agent à entraîner afin de lui fournir des séquences sur lesquelles apprendre. Cet agent, appelé démonstrateur, prend une proportion définie d'actions aléatoires pendant ce processus. L'agent apprend sans interagir avec l'environnement mais avec les échantillons fournis par les interactions du démonstrateur grâce à l'algorithme de batch constrained Q-learning. Cet algorithme permet de favoriser l'apprentissage d'une stratégie proche de celle du démonstrateur tout en privilégiant les actions qui maximisent l'espérance des récompenses perçues à long terme.
Un apprentissage avec cette méthodologie a permis de diviser le temps de calcul par 14.75 par rapport à l'utilisation de l'algorithme de DQN. Le temps alloué aux interactions entre le ou les démonstrateurs et l'environnement est fixe et le temps d'apprentissage dépend d'un seuil de patience, comme pour le DQN. Les stratégies de contrôle développées par les agents ainsi entraînés fournissent des résultats comparables aux résultats obtenus par DQN.

Évaluation de la méthodologie d'optimisation bi-niveaux pour le dimensionnement des panneaux photovoltaïques et de la batterie d'un micro-réseau.

L'intégration de l'apprentissage par renforcement dans le dimensionnement optimal du micro-réseau permet de prendre en compte les erreurs liées à l'application d'une stratégie de contrôle développée dans un environnement différent de l'environnement sur lequel elle est déployée. En effet, le calcul des coûts d'opération s'effectue avec des données différentes des données d'entraînement de l'agent. Les coûts d'opération calculés par cette méthode sont, de ce fait, plus fiables que ceux obtenus grâce à des méthodes déterministes comme l'optimisation hors ligne.
Un algorithme d'optimisation méta-heuristique est utilisé pour explorer efficacement l'espace des solutions. Une optimisation basée sur une trajectoire est privilégiée afin d'explorer séquentiellement les différents dimensionnements possibles, facilitant ainsi le déclenchement d'un transfert de politique. En particulier, l'algorithme de recuit simulé est retenu pour sa faible complexité comme l'espace des solutions est réduit.
Des optimisations sont effectuées avec la méthodologie proposée avec deux fonctions objectif

différentes pour le calcul du coût d'opération du micro-réseau. Une fonction objectif considère un revenu généré par l'injection d'électricité alors que l'autre non. Pour chaque cas, deux systèmes de récompense sont définis pour l'élaboration des politiques de contrôle.
Lorsque l'électricité injectée génère un revenu, le dimensionnement optimal n'est pas le même selon le système de récompense de l'agent. Une solution, avec des équipements dont les dimensions sont minimales est trouvée lorsqu'une composante du système de récompense de l'agent recherche la minimisation de l'injection d'électricité. Un dimensionnement intermédiaire de 4 kWc de panneaux photovoltaïques et d'une capacité de batterie de 6 kWh est retenu dans le cas où l'EMS n'est pénalisé que pour l'énergie soutirée. Ce même dimensionnement des équipements est optimal pour les deux systèmes de récompense lorsque le coût d'opération ne dépend que de l'énergie soutirée. Ce dimensionnement est de 3 kWc pour les panneaux photovoltaïques et 10 kWh pour la batterie. Les dimensionnements retenus offrent des coûts nivelés de l'énergie compétitifs par rapport au tarif actuel de l'électricité (0,2276 €/kWh).
Le dimensionnement d'un micro-réseau non-connecté au réseau central de distribution est réalisé en associant le coût de soutirage de l'électricité à un coût variable lié au déficit énergétique. Pour une faible augmentation de ce coût, la capacité optimale de la batterie augmente pour atteindre la borne maximale fixée (11 kWh). Le dimensionnement optimal inclut 5 kWc de panneaux photovoltaïques pour des augmentations plus prononcées de ce coût. L'autonomie augmente avec le prix associé au déficit énergétique mais reste globalement trop faible pour un micro-réseau isolé (probabilité de panne de 14 % au minimum).
La méthodologie proposée converge rapidement vers les minima globaux. La probabilité pour un agent d'être entraîné avec l'algorithme de BCQ augmente au long des itérations. Le temps de calcul nécessaire à l'optimisation est divisé par 2.6. De nombreux dimensionnements déjà observés le sont à nouveau au cours du processus d'optimisation, ce qui pourrait nuire à la découverte de l'optimum global dans les dernières itérations.

Perspectives

La méthodologie bi-niveau proposée dans cette thèse offre une nouvelle approche pour le dimensionnement et le contrôle de micro-réseaux électriques. Toutefois, des questions restent en suspens pour son application et méritent une exploration plus approfondie.

Vers une modélisation plus complète du problème de contrôle. Le problème de contrôle, bien que complexifié au cours de ces travaux de recherche, contient toujours des éléments simplifiés. Les consignes calculées par l'EMS ne s'appliquent qu'au stockage hydrogène alors que la batterie est pourtant une unité contrôlable. Le contrôle conjoint des deux unités de stockage pourrait être une voie intéressante pour examiner la stratégie de préservation de la batterie. L'espace d'action serait alors continu et en deux dimensions. L'algorithme de DDPG est adapté et pourrait être envisagé. Des modèles de vieillissement de l'électrolyseur et de la pile à combustible pourraient être intégrés à la simulation dynamique de l'environnement.

Enrichissement des critères et caractéristiques de l'optimisation du dimensionnement. Le problème de dimensionnement ne comporte que deux variables dans la version proposée. Avec des consignes continues, l'électrolyseur et la pile à combustible peuvent être ajoutés aux variables de dimensionnement.
Les bornes des variables de notre problème d'optimisation, notamment en ce qui concerne la capacité de la batterie, doivent être adaptées en fonction du profil de consommation. Cette nécessité s'est avérée particulièrement évidente lors du dimensionnement d'un micro-réseau isolé à forte consommation, où la borne supérieure de la capacité de la batterie est atteinte. La probabilité de panne demeurait significative pour ce cas de figure. Des études approfondies sur l'influence du profil de consommation et de la localisation du micro-réseau sur le dimensionnement optimal mériteraient d'être conduites. De plus, une optimisation de type multi-critères pourrait être menée pour le problème du dimensionnement d'un micro-réseau

isolé car les coûts d'investissement et le déficit énergétiques sont des objectifs dont la priorité dépend des utilisateurs.

Développement d'une heuristique autour du transfert de politique. Bien que les méthodes proposées pour accélérer le processus d'optimisation soient efficaces, le temps de calcul reste conséquent. L'augmentation des dimensions de l'espace de solution pour accroître le réalisme des systèmes modélisés posera un problème de temps de calcul. La parallélisation des calculs, combinée à l'emploi d'une méta-heuristique évolutive, a été envisagée mais demande plus de puissance de calcul et serait ralentie par la variance observée dans le temps d'apprentissage des agents. Un axe de réflexion pourrait concerner le développement d'une heuristique incluant des entraînements parallélisés en augmentant la fréquence de l'entraînement hors ligne des agents. Cette heuristique pourrait consister à établir plusieurs trajectoires d'optimisation en parallèle, en instaurant une liste d'attente entre les candidats et les démonstrateurs.

A

Annexe

A.1 Apprentissage par renforcement inverse

Dans l'Équation 2.52, $w \in \mathbb{R}^n$ est le vecteur de poids associé aux caractéristiques des états pour un jeu de démonstrations donné. La fonction valeur d'état s'exprime alors avec les Équations (A.1, A.2 et A.3). La comparaison des politiques se fait comme avec l'Équation 2.25. Puisque le jeu de données vient du démonstrateur, l'état initial S_0 est le seul état forcément commun à chaque trajectoire. Seule la valeur de cet état peut être un critère pour comparer des politiques.

$$\mathrm{V}^{\pi}(\mathrm{S}_0) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t R(\mathrm{S}_t)\right] \tag{A.1}$$

$$= \mathbb{E}_{\tau\sim\pi}\left[w^k \sum_{t=0}^{T}\gamma^t x(\mathrm{S}_t)\right] \tag{A.2}$$

$$= w^k \mu(\pi) \tag{A.3}$$

Avec $\mu(\pi)$ la fréquence pondérée amortie des caractéristiques des états sous la politique π. Autrement dit, μ est une fonction qui décompte le nombre d'observations des caractéristiques des états, amorti par le nombre de pas de temps avant de les observer. Puisqu'il n'y a pas de système de récompense, c'est la fréquence d'observation d'états sous certaines politiques qui dirige l'apprentissage de l'agent. Ainsi, en utilisant l'Équation 2.25 et en faisant l'hypothèse que la politique montrée par le démonstrateur est optimale, il est possible pour l'agent d'inférer un vecteur de poids optimal w^{*k} (Abbeel et al., 2004) qui vérifie l'Équation A.4 :

$$w^{*k}\mu(\pi^*) \geq w^{*k}\mu(\pi), \forall \pi \neq \pi^* \tag{A.4}$$

avec π^* la politique optimale selon le jeu de démonstrations. Il s'agit donc de trouver une fonction récompense telle que la politique affichée par l'expert soit meilleure que n'importe quelle autre politique. Les valeurs associées aux états selon la distribution d'états observés sous la politique de l'expert doivent être supérieures aux valeurs de n'importe quelle autre distribution d'états $\mu(\pi)$ avec le vecteur w^*. Si $\|\mu(\pi) - \mu(\pi^*)\|_1 \leq \varepsilon$, alors l'Équation A.5 est vérifiée.

$$\forall w, \|w\|_\infty < 1, \quad \mid w^k\mu(\pi) - w^k\mu(\pi^*) \mid \leq \varepsilon \tag{A.5}$$

Si cette condition est vérifiée, alors la politique de l'expert peut être imitée par l'agent. En revanche, le calcul de $\mu(\pi)$ requiert d'avoir accès à un modèle de transition et donc de connaître $P^a_{ss'}$ pour chaque état. De plus, plusieurs politiques sont optimales selon le système de récompense inféré, car il prend en compte le décompte des observations de caractéristiques d'états et non les trajectoires. Plusieurs trajectoires peuvent mener au même décompte de caractéristiques d'états. Des algorithmes développés plus tard comme GAIL (Ho et al., 2018) ou l'IRL avec maximisation de l'entropie (Ziebart et al., 2008) permettent d'utiliser l'IRL sans modèle. De plus, ils lèvent le problème des nombreuses potentielles politiques optimales.

A.2 Schémas des flux de puissance hebdomadaire

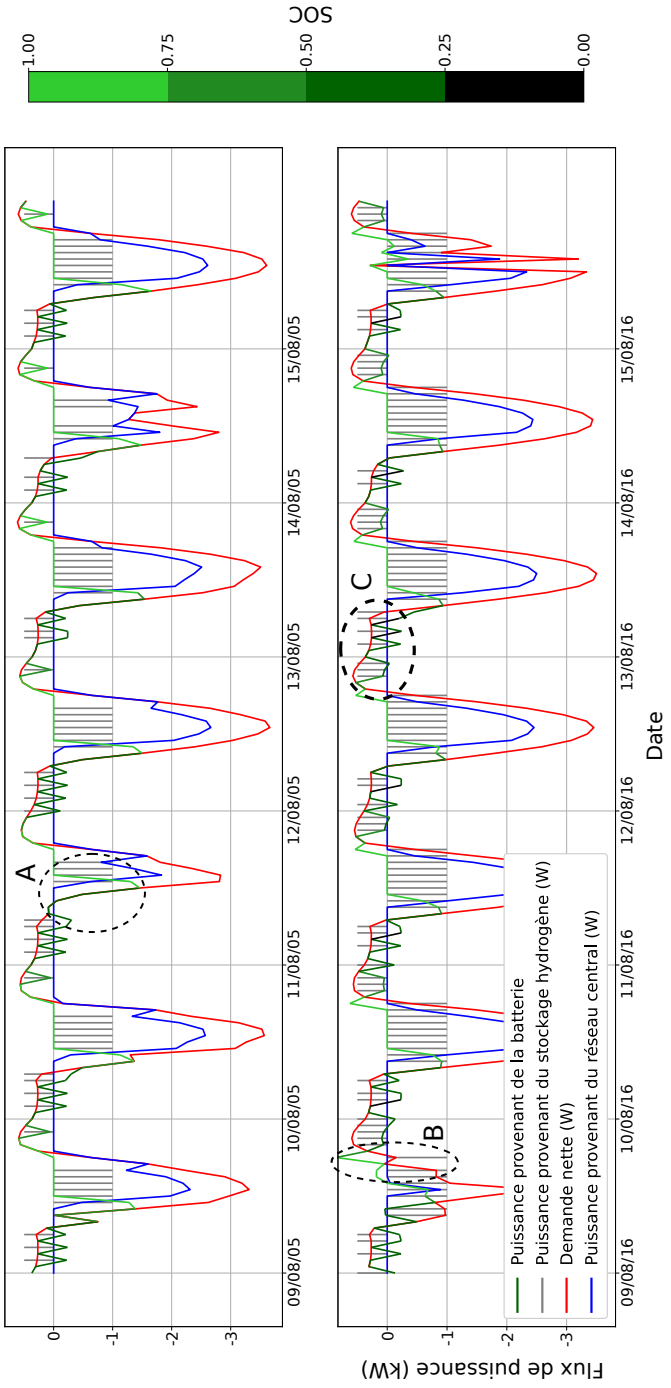

Figure A.1 – Flux de puissance du micro-réseau pour une semaine d'août en début (2005) et en fin (2016) de simulation.

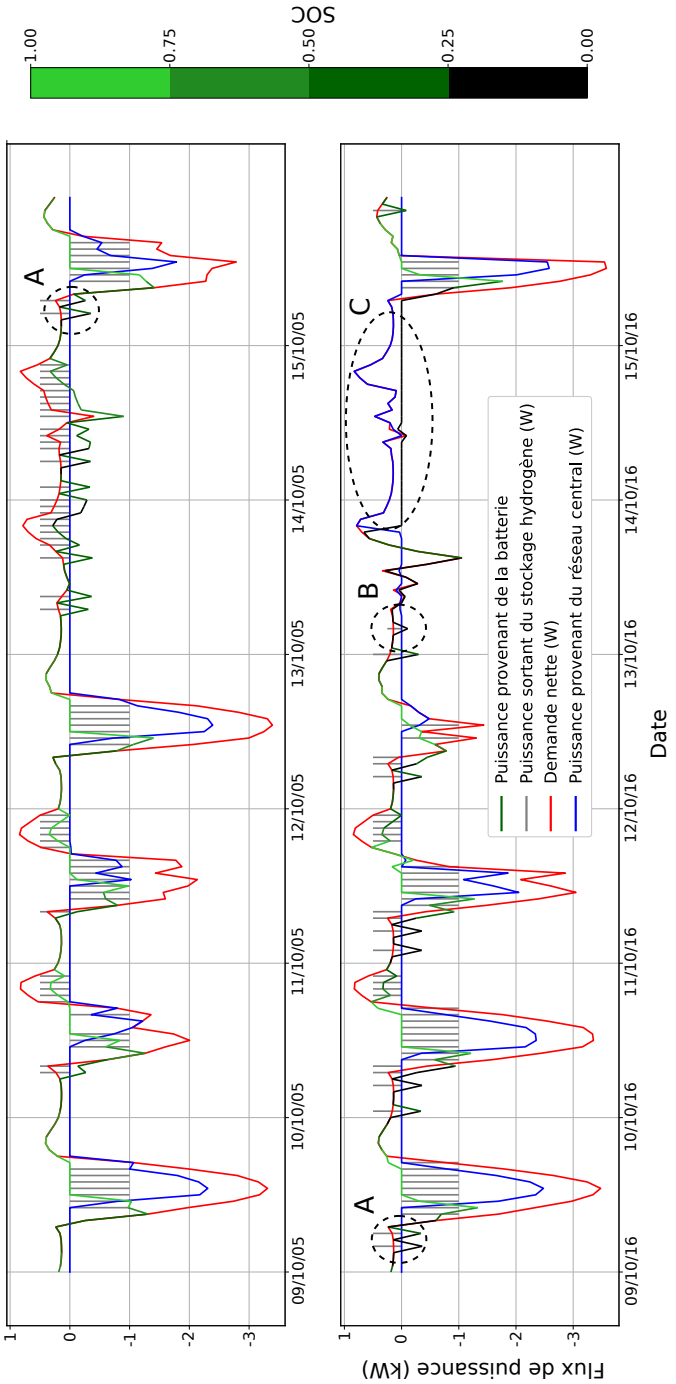

Figure A.2 – Flux de puissance du micro-réseau pour une semaine d'octobre en début (2005) et en fin (2016) de simulation.

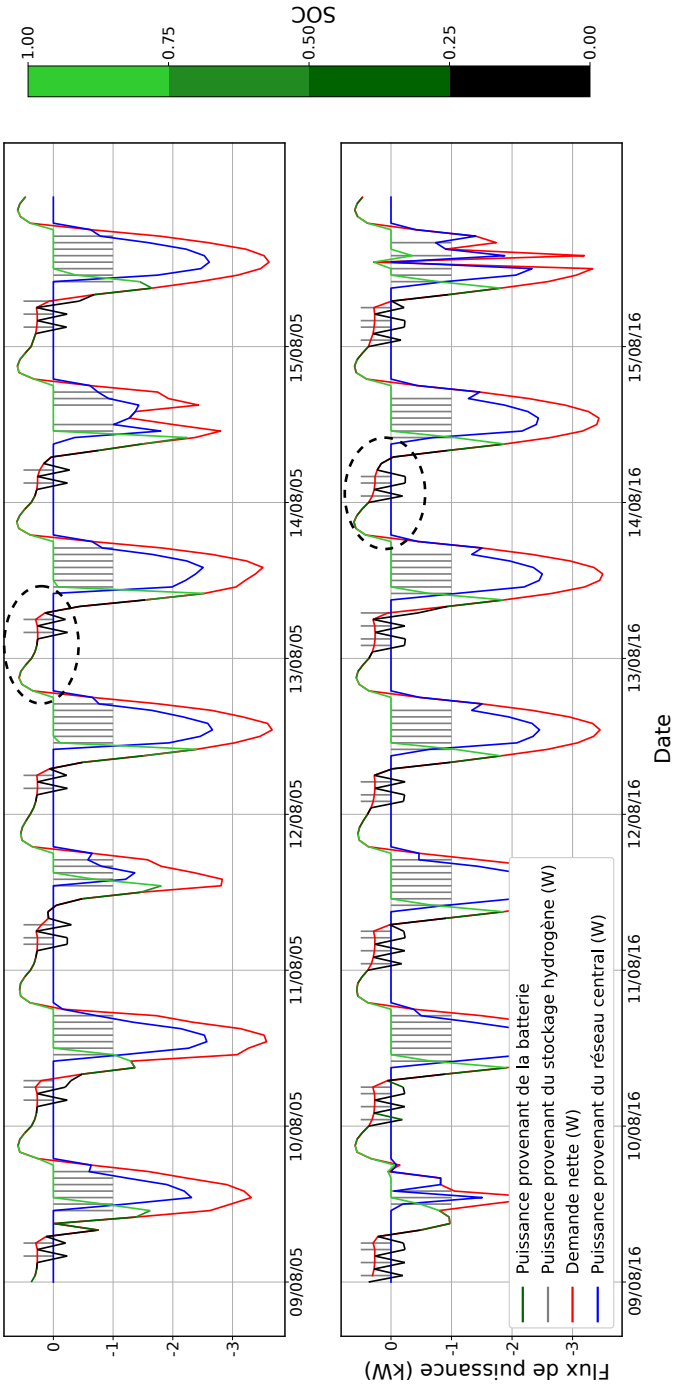

Figure A.3 – Flux de puissance du micro-réseau pour une semaine d'août en début (2005) et en fin (2016) de simulation. L'environnement inclut un remplacement des batteries usées.

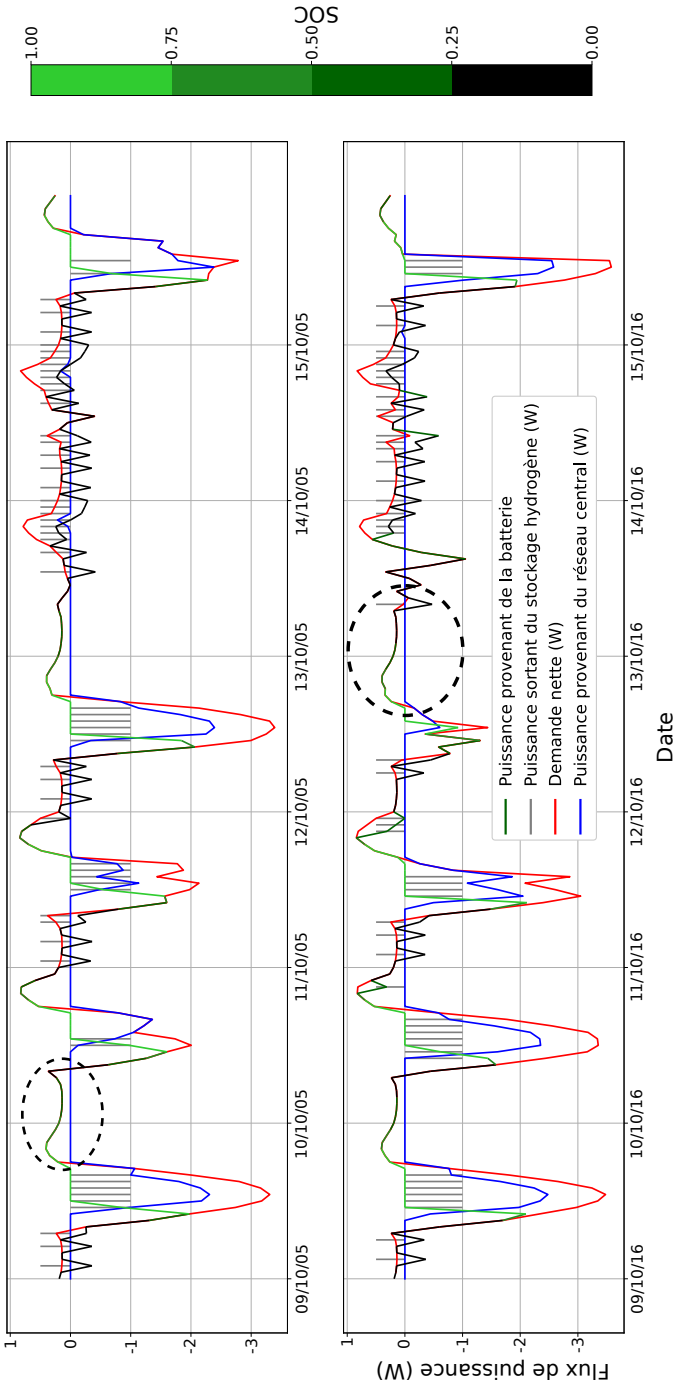

Figure A.4 – Flux de puissance du micro-réseau pour une semaine d'octobre en début (2005) et en fin (2016) de simulation. L'environnement inclut un remplacement des batteries usées.

Table des figures

Liste des tableaux

Liste des symboles

Modélisation du micro-réseau

C	€	Coût ... 75
C_{maint}	€	Coût de maintenance d'une unité du micro-réseau du micro-réseau ... 74
C_{remp}	€	Coût de remplacement associé aux unités du micro-réseau ... 74
C_{inv}	€	Coût d'investissement ... 74
C_{ope}	€	Coût d'opération du micro-réseau ... 74
C_{inje}	€/kWh	Revenus d'injection d'électricité au réseau central de distribution ... 75
C_{sout}	€/kWh	Coûts de soutirage d'électricité au réseau central de distribution ... 75
E	kW h	Capacité ... 75
E_{inje}	kWh	Énergie injectée au réseau central de distribution ... 75
E_{sout}	kWh	Énergie soutirée au réseau central de distribution ... 75
$\mathcal{D}$	-	Espace des variables dimensionnables des modules ... 60
$\mathcal{N}$	-	Espace des variables non-dimensionnables des modules ... 60
η	kWh/kg	Rendement de conversion d'électricité en masse d'hydrogène pour un électrolyseur et de masse d'hydrogène en électricité pour une pile à combustible ... 63
m	kg	Masse ... 63
D^{net}	kWh	Demande électrique nette ... 65
$dist^{sols}$	jour	Distance temporelle au soltice d'été ... 70
δ	kWh	Capacité de la batterie dégradée ... 92
μ	-	Rendement ... 65
P	W	Puissance ... 60

Apprentissage par renforcement

$\tau_{autocons}$	-	Taux d'autoconsommation ... 72
$\tau_{autoprod}$	-	Taux d'autoproduction ... 72
α	-	Taux d'apprentissage ... 22
ε	-	Probabilité d'exploration ... 29
N_{cible}	-	Nombre d'itérations avant mise-à-jour du Q-réseau cible ... 30
γ	-	Facteur d'actualisation ... 22
J	-	Espérance à maximiser en apprentissage par renforcement ... 21
L	-	Valeur de la fonction coût d'un réseau de neurones lors de sa mise à jour ... 31

A_t	-	Action prise par l'agent à l'instant t 19
S_t	-	État observables de l'environnement à l'instant t 19
$\mathcal{S}$	-	Espace d'état de l'environnement 20
$\mathcal{A}$	-	Espace d'action de l'agent 20
R_t	-	Récompense perçue par l'agent à l'instant t 20
$P^a_{ss'}$	-	Probabilité de transition d'un état s à un état s' suite à l'action a 20
π	-	Politique de l'agent 20
V	-	Valeur d'un état 21
Q	-	Valeur d'un couple action-état 23
$\hat{V}$	-	Estimation de la valeur d'un état 21
$\hat{Q}$	-	Estimation de la valeur d'un couple action-état 23
G_t	-	Retour amorti à partir de l'instant t 23
τ	-	Trajectoire 21
τ_{BCQ}	-	Seuil de BCQ 37
N^{ep}	-	Numéro de l'épisode en cours dans un processus d'entraînement 68
$N^{ep}_{\max}$	-	Nombre maximal d'épisodes dans un processus d'entraînement 68
t_{eval}	-	Nombre d'itérations dans l'apprentissage avant évaluation de la politique apprise par un agent 68
N_{patience}	-	Nombre d'itérations de validation pour lesquelles les scores n'ont pas dépassés le meilleur score obtenu 69
ε_{dec}	-	Nombre d'itérations d'entraînement pendant lesquels le taux d'exploration décroît 81
τ_ε	-	Taux de décroissance de l'exploration sur les itérations d'entraînement 81

Exposants

$^{\text{batt}}$	-	Indice se référant à la batterie électrochimique 62
$^{\text{PV}}$	-	Indice se référant à la génération photovoltaïque 62
$^{\text{elec}}$	-	Indice se référant à l'électrolyseur 60
^{H_2}	kW	Indice se référant au stockage hydrogène 71
$^{\text{PAC}}$	-	Indice se référant à la pile à combustible 60
$^{\pi}$	-	Variable dépendant d'une politique π 21
$^{\text{comp}}$	-	Indice se référant au compresseur en sortie d'électrolyse de l'eau 75
$^{\text{grid}}$	-	Indice se référant au réseau central de distribution 64

Indices

$_\theta$	-	Paramétrage du réseau de neurones associé 30
$_{\text{max}}$	-	Valeur maximale de la variable associée 60
$_{\text{nom}}$	-	Valeur nominale de la variable associée 75
$_{\text{dech}}$	-	Valeur liée à la décharge de l'unité associée 62
$_{\text{charg}}$	-	Valeur liée à la charge de l'unité associée 62
$_{\text{lin}}$	-	Caractère linéaire de la variable associée 92
$_{\text{nl}}$	-	Caractère non-linéaire de la variable associée 93

Acronymes

AC Alternating current – Courant alternatif 8
BCQ Batch constraint Q-learning 37
CAPEX Capital Expenditure – Dépenses de capital 74
CVC Système de chauffage, climatisation et ventilation 49
DC Direct current – Courant continu 8
DDPG Deep deterministic policy gradient 34
DoD Depth of discharge – Profondeur de décharge 45
DQN Deep Q-learning 30
ELF Equivalent loss factor – Facteur de perte équivalent 16
EMS Energy management system – Système de gestion de l'énergie 1
EV Electric Vehicle – Véhicules électriques 8
FER Fraction d'énergie renouvelable 16
GRD Gestionnaire de réseau de distribution 7
IRL Inverse reinforcement learning – Apprentissage par renforcement inverse 35
LCE Levelized cost of energy – Coût nivelé de l'énergie 15
LOH Level of hydrogen – Taux de remplissage du réservoir d'hydrogène 63
LPSP Loss of load supply probability – Probabilité de perte de charge 16
LSTM Long short term memory 50
MADDPG Multi-agent deep deterministic policy gradient 43
MDP Markov decision process – Processus de décision de Markov 19
MPC Model predictive control – Commande prédictive 14
OPEX Operating Expenditure – Dépenses d'exploitation 74
PAC Pile à combustible 7
PCC Point de couplage commun 8
PPO Proximal policy optimization 34
PV Photovoltaïque 7
REE Ratio d'excès d'énergie 16
RL Reinforcement learning – Apprentissage par renforcement 19
SOC State of charge – État de charge de la batterie 41
TRPO Trust region policy optimization 34

Bibliographie

(Abbeel et al., 2004) P. Abbeel and A. Y. Ng. "Apprenticeship Learning via Inverse Reinforcement Learning". In: *Twenty-First International Conference on Machine Learning - ICML '04*. Twenty-First International Conference. Banff, Alberta, Canada: ACM Press, 2004, p. 1 (cf. p. 147).

(Abdel-hamed et al., 2019) A. M. Abdel-hamed, K. Ellissy, A. R. Adly, and H. Abdelfattah. "Optimal Sizing and Design of Isolated Micro-Grid Systems". In: *International Journal of Environmental Science & Sustainable Development* 4.3 (Dec. 30, 2019), p. 1 (cf. p. 17).

(Ahmad Khan et al., 2016) A. Ahmad Khan, M. Naeem, M. Iqbal, S. Qaisar, and A. Anpalagan. "A Compendium of Optimization Objectives, Constraints, Tools and Algorithms for Energy Management in Microgrids". In: *Renewable and Sustainable Energy Reviews* 58 (May 2016), pp. 1664–1683 (cf. p. 13).

(Ahmad et al., 2023) S. Ahmad, M. Shafiullah, C. B. Ahmed, and M. Alowaifeer. "A Review of Microgrid Energy Management and Control Strategies". In: *IEEE Access* 11 (2023), pp. 21729–21757 (cf. p. 13).

(Ali et al., 2021) K. H. Ali, M. Sigalo, S. Das, E. Anderlini, A. A. Tahir, and M. Abusara. "Reinforcement Learning for Energy-Storage Systems in Grid-Connected Microgrids: An Investigation of Online vs. Offline Implementation". In: *Energies* 14.18 (Sept. 9, 2021), p. 5688 (cf. p. 71).

(Allwyn et al., 2023) R. G. Allwyn, A. Al-Hinai, and V. Margaret. "A Comprehensive Review on Energy Management Strategy of Microgrids". In: *Energy Reports* 9 (Dec. 2023), pp. 5565–5591 (cf. p. 12).

(Altin et al., 2023) N. Altin, S. E. Eyimaya, and A. Nasiri. "Multi-Agent-Based Controller for Microgrids: An Overview and Case Study". In: *Energies* 16.5 (Mar. 3, 2023), p. 2445 (cf. p. 15).

(Amrouche et al., 2015) S. O. Amrouche, D. Rekioua, and T. Rekioua. "Overview of Energy Storage in Renewable Energy Systems". In: *2015 3rd International Renewable and Sustainable Energy Conference (IRSEC)*. 2015 3rd International Renewable and Sustainable Energy Conference (IRSEC). Marrakech: IEEE, Dec. 2015, pp. 1–6 (cf. p. 9).

(An et al., 2015) L. N. An, T. Quoc-Tuan, B. Seddik, and N. Van-Linh. "Optimal Sizing of a Grid-Connected Microgrid". In: *2015 IEEE International Conference on Industrial Technology (ICIT)*. 2015 IEEE International Conference on Industrial Technology (ICIT). Seville: IEEE, Mar. 2015, pp. 2869–2874 (cf. p. 16).

(Andrew Y. Ng et al., 2000) Andrew Y. Ng et Stuart Russel. "Algorithms for Inverse Reinforcement Learning". In: (2000) (cf. p. 35).

(Ardakani et al., 2010) F. J. Ardakani, G. Riahy et M. Abedi. "Optimal Sizing of a Grid-Connected Hybrid System for North-West of Iran-case Study". In: *2010 9th International Conference on Environment and Electrical Engineering*. 2010 9th International Conference on Environment and Electrical Engineering. Prague, Czech Republic: IEEE, 2010, p. 29-32 (cf. p. 16).

(Atia et al., 2016) R. Atia et N. Yamada. "Sizing and Analysis of Renewable Energy and Battery Systems in Residential Microgrids". In: *IEEE Transactions on Smart Grid* 7.3 (mai 2016), p. 1204-1213 (cf. p. 17).

(Battula et al., 2021) A. R. Battula, S. Vuddanti, and S. R. Salkuti. "Review of Energy Management System Approaches in Microgrids". In: *Energies* 14.17 (Sept. 2, 2021), p. 5459 (cf. p. 12, 14).

(Bellman, 1958) R. Bellman. "Dynamic Programming and Stochastic Control Processes". In: *Information and Control* 1.3 (Sept. 1958), pp. 228–239 (cf. p. 15).

(Bergstra et al., 2012) J. Bergstra and Y. Bengio. "Random Search for Hyper-Parameter Optimization". In: *JMLR.org* 13 (Jan. 3, 2012), pp. 281–305 (cf. p. 83).

(Berlink et al., 2015) H. Berlink, N. Kagan, and A. H. Reali Costa. "Intelligent Decision-Making for Smart Home Energy Management". In: *Journal of Intelligent & Robotic Systems* 80.S1 (Dec. 2015), pp. 331–354 (cf. p. 45).

(Bharadwaj et al., 2018) D. R. Bharadwaj, S. K. R. Danda, K. Narayanam, and S. Bhatnagar. "A Unified Decision Making Framework for Supply and Demand Management in Microgrid Networks". In: *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)* (Oct. 2018), pp. 1–7. arXiv: 1711.05078 (cf. p. 51).

(Bharath et al., 2019) K. R. Bharath, M. Krishnan Mithun, and P. Kanakasabapathy. "A Review on DC Microgrid Control Techniques, Applications and Trends". In: *International Journal of Renewable Energy Research* (v9i3 2019) (cf. p. 12).

(Bian et al., 2020) H. Bian, X. Tian, J. Zhang, and X. Han. "Deep Reinforcement Learning Algorithm Based on Optimal Energy Dispatching for Microgrid". In: *2020 5th Asia Conference on Power and Electrical Engineering (ACPEE)*. 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE). Chengdu, China: IEEE, June 2020, pp. 169–174 (cf. p. 46).

(Boyd et al., 2008) S. Boyd, L. Xiao, A. Mutapcic, and J. Mattingley. "Notes on Decomposition Methods". In: (Apr. 13, 2008) (cf. p. 53).

(Bratko et al., 1995) I. Bratko, T. Urbančič, and C. Sammut. "Behavioural Cloning: Phenomena, Results and Problems". In: *IFAC Proceedings Volumes* 28.21 (Sept. 1995), pp. 143–149 (cf. p. 35).

(Brockman et al., 2016) G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. *OpenAI Gym*. June 5, 2016. arXiv: 1606.01540 [cs]. URL: http://arxiv.org/abs/1606.01540. preprint (cf. p. 67).

(Cany et al., 2016) C. Cany, C. Mansilla, P. Da Costa, G. Mathonnière, T. Duquesnoy, and A. Baschwitz. "Nuclear and Intermittent Renewables: Two Compatible Supply Options? The Case of the French Power Mix". In: *Energy Policy* 95 (Aug. 2016), pp. 135–146 (cf. p. 7).

(Cao et al., 2019) T. Cao, Z. Shen, and G. Zhang. "LSTM-Aided Reinforcement Learning for Energy Management in Microgrid with Energy Storage and EV Charging". In: *2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*. 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN). Shenzhen, China: IEEE, Dec. 2019, pp. 13–18 (cf. p. 45, 46).

(Casado et al., 2022) A. Casado, S. Pérez-Peló, J. Sánchez-Oro, and A. Duarte. "A GRASP Algorithm with Tabu Search Improvement for Solving the Maximum Intersection of K-Subsets Problem". In: *Journal of Heuristics* 28.1 (Feb. 2022), pp. 121–146 (cf. p. 124).

(Chebabhi et al., 2023) Ardjouna. Chebabhi, Ilyes. Tegani, A. D. Benhamadouche, and Okba. Kraa. "Optimal Design and Sizing of Renewable Energies in Microgrids Based on Financial Considerations a Case Study of Biskra, Algeria". In: *Energy Conversion and Management* 291 (Sept. 2023), p. 117270 (cf. p. 16).

(Chen et al., 2018) T. Chen and W. Su. "Local Energy Trading Behavior Modeling With Deep Reinforcement Learning". In: *IEEE Access* 6 (2018), pp. 62806–62814 (cf. p. 46).

(Chen et al., 2019) T. Chen and S. Bu. "Realistic Peer-to-Peer Energy Trading Model for Microgrids Using Deep Reinforcement Learning". In: *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe). Bucharest, Romania: IEEE, Sept. 2019, pp. 1–5 (cf. p. 51).

(Chen et al., 2022) T. Chen, S. Bu, X. Liu, J. Kang, F. R. Yu et Z. Han. "Peer-to-Peer Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning". In: *IEEE Transactions on Smart Grid* 13.1 (jan. 2022), p. 715-727 (cf. p. 53).

(Cicirelli et al., 2020) F. Cicirelli, A. F. Gentile, E. Greco, A. Guerrieri, G. Spezzano, and A. Vinci. "An Energy Management System at the Edge Based on Reinforcement Learning". In: *2020 IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. 2020 IEEE/ACM 24th International Symposium on Distributed Simulation and Real Time Applications (DS-RT). Prague, Czech Republic: IEEE, Sept. 2020, pp. 1–8 (cf. p. 49).

(Citepa, 2023) Citepa. *Gaz à Effet de Serre et Polluants Atmosphériques. Bilan Des Émissions En France de 1990 à 2022*. 2023 (cf. p. 1).

(Collet et al., 2007) P. Collet et J. Rennard. "Stochastic Optimization Algorithms:" in: *Handbook of Research on Nature-Inspired Computing for Economics and Management*. Sous la dir. J.-P. Rennard. IGI Global, 2007, p. 28-44 (cf. p. 15).

(2019). *Coûts et rentabilités du grand photovoltaïque en métropole continentale*. Commission de régulation de l'énergie, 28 fév. 2019, p. 46 (cf. p. 74).

(Dahiru et al., 2023) A. T. Dahiru, D. Daud, C. W. Tan, Z. T. Jagun, S. Samsudin, and A. M. Dobi. "A Comprehensive Review of Demand Side Management in Distributed Grids Based on Real Estate Perspectives". In: *Environmental Science and Pollution Research* 30.34 (Jan. 18, 2023), pp. 81984–82013 (cf. p. 13).

(De Brabandere et al., 2007) K. De Brabandere, K. Vanthournout, J. Driesen, G. Deconinck, and R. Belmans. "Control of Microgrids". In: *2007 IEEE Power Engineering Society General Meeting*. 2007 IEEE Power Engineering Society General Meeting. Tampa, FL, USA: IEEE, June 2007, pp. 1–7 (cf. p. 11).

(Dewey, 2014) D. Dewey. "Reinforcement Learning and the Reward Engineering Principle". In: *2014 AAAI Spring Symposium Series*. 2014 (cf. p. 44).

(Diaf et al., 2008) S. Diaf, G. Notton, M. Belhamel, M. Haddadi, and A. Louche. "Design and Techno-Economical Optimization for Hybrid PV/Wind System under Various Meteorological Conditions". In: *Applied Energy* 85.10 (Oct. 2008), pp. 968–987 (cf. p. 74).

(Dimeas et al., 2010) A. L. Dimeas and N. D. Hatziargyriou. "Multi-Agent Reinforcement Learning for Microgrids". In: *IEEE PES General Meeting*. Energy Society General Meeting. Minneapolis, MN: IEEE, July 2010, pp. 1–8 (cf. p. 42).

(Domínguez-Barbero et al., 2020) D. Domínguez-Barbero, J. García-González, M. A. Sanz-Bobi, and E. F. Sánchez-Úbeda. "Optimising a Microgrid System by Deep Reinforcement Learning Techniques". In: *Energies* 13.11 (June 2, 2020), p. 2830 (cf. p. 45).

(Dreher et al., 2022) A. Dreher, T. Bexten, T. Sieker, M. Lehna, J. Schütt, C. Scholz, and M. Wirsum. "AI Agents Envisioning the Future: Forecast-based Operation of Renewable Energy Storage Systems Using Hydrogen with Deep Reinforcement Learning". In: *Energy Conversion and Management* 258 (Apr. 2022), p. 115401 (cf. p. 43).

(Du et al., 2020) Y. Du and F. Li. "Intelligent Multi-Microgrid Energy Management Based on Deep Neural Network and Model-Free Reinforcement Learning". In: *IEEE Transactions on Smart Grid* 11.2 (Mar. 2020), pp. 1066–1076 (cf. p. 53).

(Dufo-López et al., 2007) R. Dufo-López, J. L. Bernal-Agustín, and J. Contreras. "Optimization of Control Strategies for Stand-Alone Renewable Energy Systems with Hydrogen Storage". In: *Renewable Energy* 32.7 (June 2007), pp. 1102–1126 (cf. p. 63, 75).

(Ebell et al., 2019) N. Ebell, M. Gutlein, and M. Pruckner. "Sharing of Energy Among Cooperative Households Using Distributed Multi-Agent Reinforcement Learning". In: *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe). Bucharest, Romania: IEEE, Sept. 2019, pp. 1–5 (cf. p. 52).

(Eimer et al., 2023) T. Eimer, M. Lindauer, and R. Raileanu. *Hyperparameters in Reinforcement Learning and How To Tune Them*. June 2, 2023. arXiv: 2306.01324 [cs]. URL: http://arxiv.org/abs/2306.01324. preprint (cf. p. 83).

(Elena et al., 1996) P. Elena, R. Irina, and R. Dechter. "Value Iteration and Policy Iteration Algorithms for Markov Decision Problem". In: (1996) (cf. p. 26).

(Erazo-Caicedo et al., 2022) D. Erazo-Caicedo, E. Mojica-Nava, and J. Revelo-Fuelagán. "Model Predictive Control for Optimal Power Flow in Grid-Connected Unbalanced Microgrids". In: *Electric Power Systems Research* 209 (Aug. 2022), p. 108000 (cf. p. 14).

(Ernst et al., 2005) D. Ernst, P. Geurts, and L. Wehenkel. "Tree-Based Batch Mode Reinforcement Learning". In: *Journal of Machine Learning Research* 6.18 (2005), pp. 503–556 (cf. p. 36).

(Espinar et al., 2011) B. Espinar and D. Mayer. "The Role of Energy Storage for Mini-Grid Stabilization". In: (July 2011) (cf. p. 10).

(Fallahifar et al., 2023) R. Fallahifar and M. Kalantar. "Optimal Planning of Lithium Ion Battery Energy Storage for Microgrid Applications: Considering Capacity Degradation". In: *Journal of Energy Storage* 57 (Jan. 2023), p. 106103 (cf. p. 46, 92, 93).

(Fang et al., 2019) X. Fang, J. Wang, G. Song, Y. Han, Q. Zhao, and Z. Cao. "Multi-Agent Reinforcement Learning Approach for Residential Microgrid Energy Scheduling". In: *Energies* 13.1 (Dec. 25, 2019), p. 123 (cf. p. 46, 47, 50).

(Fang et al., 2020) X. Fang, J. Wang, C. Yin, Y. Han, and Q. Zhao. "Multiagent Reinforcement Learning With Learning Automata for Microgrid Energy Management and Decision Optimization". In: *2020 Chinese Control And Decision Conference (CCDC)*. 2020 Chinese Control And Decision Conference (CCDC). Hefei, China: IEEE, Aug. 2020, pp. 779–784 (cf. p. 46, 50, 52).

(Feo et al., 1989) T. A. Feo and M. G. Resende. "A Probabilistic Heuristic for a Computationally Difficult Set Covering Problem". In: *Operations Research Letters* 8.2 (Apr. 1989), pp. 67–71 (cf. p. 124).

(Foruzan et al., 2018) E. Foruzan, L.-K. Soh, and S. Asgarpoor. "Reinforcement Learning Approach for Optimal Distributed Energy Management in a Microgrid". In: *IEEE Transactions on Power Systems* 33.5 (Sept. 2018), pp. 5749–5758 (cf. p. 42).

(France, 2022) É. de France. *Update on the Stress Corrosion Phenomenon and Adjustment of 2022 French Nuclear Output Estimate*. 18 mai 2022. URL: https://www.edf.fr/sites/groupe/files/epresspack/3045/PR-EDF1.pdf (cf. p. 7).

(François-Lavet, 2017) V. François-Lavet. "Contributions to Deep Reinforcement Learning and Its Applications in Smartgrids". ULiège - Université de Liège, Sept. 11, 2017. 117 pp. (cf. p. 16).

(François-Lavet et al., 2016) V. François-Lavet, D. Taralla, D. Ernst, and R. Fonteneau. "Deep Reinforcement Learning Solutions for Energy Microgrids Management". In: *European Workshop on Reinforcement Learning (EWRL 2016)* (2016), p. 7 (cf. p. 47, 70, 71).

(Fujimoto et al., 2019) S. Fujimoto, D. Meger, and D. Precup. *Off-Policy Deep Reinforcement Learning without Exploration*. Aug. 9, 2019. arXiv: arXiv:1812.02900. URL: http://arxiv.org/abs/1812.02900. preprint (cf. p. 37, 38, 117).

(Fujimoto et al., 2018) S. Fujimoto, H. van Hoof, and D. Meger. *Addressing Function Approximation Error in Actor-Critic Methods*. Oct. 22, 2018. arXiv: arXiv:1802.09477. URL: http://arxiv.org/abs/1802.09477. preprint (cf. p. 34).

(Gamarra et al., 2015) C. Gamarra and J. M. Guerrero. "Computational Optimization Techniques Applied to Microgrids Planning: A Review". In: *Renewable and Sustainable Energy Reviews* 48 (Aug. 2015), pp. 413–424 (cf. p. 14).

(F. Gao et al., 2019) F. Gao, R. Kang, J. Cao, and T. Yang. "Primary and Secondary Control in DC Microgrids: A Review". In: *Journal of Modern Power Systems and Clean Energy* 7.2 (Mar. 2019), pp. 227–242 (cf. p. 12).

(G. Gao et al., 2020) G. Gao, Y. Wen, X. Wu, and R. Wang. *Distributed Energy Trading and Scheduling among Microgrids via Multiagent Reinforcement Learning*. July 8, 2020. arXiv: 2007.04517 [cs, eess]. URL: http://arxiv.org/abs/2007.04517. preprint (cf. p. 53).

(K. Gao et al., 2021) K. Gao, T. Wang, C. Han, J. Xie, Y. Ma, and R. Peng. "A Review of Optimization of Microgrid Operation". In: *Energies* 14.10 (May 14, 2021), p. 2842 (cf. p. 12).

(García et al., 1989) C. E. García, D. M. Prett, and M. Morari. "Model Predictive Control: Theory and Practice—A Survey". In: *Automatica* 25.3 (May 1989), pp. 335–348 (cf. p. 14).

(Gazijahani et al., 2018) F. S. Gazijahani et J. Salehi. "Robust Design of Microgrids With Reconfigurable Topology Under Severe Uncertainty". In: *IEEE Transactions on Sustainable Energy* 9.2 (avr. 2018), p. 559-569 (cf. p. 17).

(Giraldez Miner et al., 2018) J. I. Giraldez Miner, F. Flores-Espino, S. MacAlpine, and P. Asmus. *Phase I Microgrid Cost Study: Data Collection and Analysis of Microgrid Costs in the United States*. NREL/TP–5D00-67821, 1477589. Oct. 9, 2018, NREL/TP–5D00-67821, 1477589 (cf. p. 41).

(Glover, 1986) F. Glover. "Future Paths for Integer Programming and Links to Artificial Intelligence". In: *Computers & Operations Research* 13.5 (Jan. 1986), pp. 533–549 (cf. p. 123).

(Guo et al., 2022) C. Guo, X. Wang, Y. Zheng, and F. Zhang. "Real-Time Optimal Energy Management of Microgrid with Uncertainties Based on Deep Reinforcement Learning". In: *Energy* 238 (Jan. 2022), p. 121873 (cf. p. 42).

(Haarnoja et al., 2018) T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*. Aug. 8, 2018. arXiv: arXiv:1801.01290. URL: http://arxiv.org/abs/1801.01290. preprint (cf. p. 34).

(Hancke et al., 2022) R. Hancke, T. Holm, and Ø. Ulleberg. "The Case for High-Pressure PEM Water Electrolysis". In: *Energy Conversion and Management* 261 (June 2022), p. 115642 (cf. p. 60, 63, 75).

(Harifi et al., 2021) S. Harifi, J. Mohammadzadeh, M. Khalilian, and S. Ebrahimnejad. "Giza Pyramids Construction: An Ancient-Inspired Metaheuristic Algorithm for Optimization". In: *Evolutionary Intelligence* 14.4 (Dec. 2021), pp. 1743–1761 (cf. p. 122, 123).

(Harrold et al., 2021) D. J. B. Harrold, J. Cao, and Z. Fan. *Data-Driven Battery Operation for Energy Arbitrage Using Rainbow Deep Reinforcement Learning*. 2021. arXiv: 2106.06061 [cs]. URL: http://arxiv.org/abs/2106.06061. preprint (cf. p. 47).

(Hasanvand et al., 2020) S. Hasanvand, M. Rafiei, M. Gheisarnejad, and M.-H. Khooban. "Reliable Power Scheduling of an Emission-Free Ship: Multiobjective Deep Reinforcement Learning". In: *IEEE Transactions on Transportation Electrification* 6.2 (June 2020), pp. 832–843 (cf. p. 45, 47).

(Hesse et al., 2017) H. Hesse, M. Schimpe, D. Kucevic, and A. Jossen. "Lithium-Ion Battery Storage for the Grid—A Review of Stationary Battery Storage System Design Tailored for Applications in Modern Power Grids". In: *Energies* 10.12 (Dec. 11, 2017), p. 2107 (cf. p. 10).

(Hinton, 2012) G. E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines". In: *Neural Networks: Tricks of the Trade*. Ed. by G. Montavon, G. B. Orr, and K.-R. Müller. Vol. 7700. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 599–619 (cf. p. 83).

(A. Hirsch et al., 2018) A. Hirsch, Y. Parag, and J. Guerrero. "Microgrids: A Review of Technologies, Key Drivers, and Outstanding Issues". In: *Renewable and Sustainable Energy Reviews* 90 (July 2018), pp. 402–411 (cf. p. 9).

(M. J. Hirsch et al., 2007) M. J. Hirsch, C. N. Meneses, P. M. Pardalos, and M. G. C. Resende. "Global Optimization by Continuous Grasp". In: *Optimization Letters* 1.2 (Jan. 17, 2007), pp. 201–212 (cf. p. 125).

(Ho et al., 2018) J. Ho and S. Ermon. "Generative Adversarial Imitation Learning". In: (2018), p. 9 (cf. p. 148).

(Hu et al., 2020) R. Hu and A. Kwasinski. "Energy Management for Isolated Renewable-Powered Microgrids Using Reinforcement Learning and Game Theory". In: *2020 22nd European Conference on Power Electronics and Applications (EPE'20 ECCE Europe)*. 2020 22nd European Conference on Power Electronics and Applications (EPE'20 ECCE Europe). Lyon, France: IEEE, Sept. 2020, P.1–P.9 (cf. p. 48).

(C. Huang et al., 2022) C. Huang, H. Zhang, L. Wang, X. Luo et Y. Song. "Mixed Deep Reinforcement Learning Considering Discrete-continuous Hybrid Action Space for Smart Home Energy Management". In: *Journal of Modern Power Systems and Clean Energy* 10.3 (mai 2022), p. 743-754 (cf. p. 50).

(X. Huang et al., 2021) X. Huang, D. Zhang, and X. Zhang. "Energy Management of Intelligent Building Based on Deep Reinforced Learning". In: *Alexandria Engineering Journal* 60.1 (Feb. 2021), pp. 1509–1517 (cf. p. 45, 50).

(Hurtado et al., 2018) L. A. Hurtado, E. Mocanu, P. H. Nguyen, M. Gibescu, and R. I. G. Kamphuis. "Enabling Cooperative Behavior for Building Demand Response Based on Extended Joint Action Learning". In: *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS* 14.1 (2018), p. 10 (cf. p. 49).

(Hussain et al., 2022) A. Hussain, V.-H. Bui, and H.-M. Kim. "Deep Reinforcement Learning-Based Operation of Fast Charging Stations Coupled with Energy Storage System". In: *Electric Power Systems Research* 210 (Sept. 2022), p. 108087 (cf. p. 46).

(International Energy Agency, 2021) International Energy Agency. *France 2021 Energy Policy Review*. IEA Energy Policy Reviews. OECD, Dec. 7, 2021 (cf. p. 7).

(International Energy Agency et al., 2021) International Energy Agency and Réseau de Transport d'Electricité. *Conditions and Requirements for the Technical Feasibility of a Power System with a High Share of Renewables in France Towards 2050*. OECD, Mar. 16, 2021 (cf. p. 7).

(Ji et al., 2019) Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang. "Real-Time Energy Management of a Microgrid Using Deep Reinforcement Learning". In: *Energies* 12.12 (June 15, 2019), p. 2291 (cf. p. 42).

(Ji et al., 2021) Y. Ji, J. Wang, J. Xu, and D. Li. "Data-Driven Online Energy Scheduling of a Microgrid Based on Deep Reinforcement Learning". In: *Energies* 14.8 (2021), p. 2120 (cf. p. 46, 49).

(Johnen et al., 2021) M. Johnen, S. Pitzen, U. Kamps, M. Kateri, P. Dechent, and D. U. Sauer. "Modeling Long-Term Capacity Degradation of Lithium-Ion Batteries". In: *Journal of Energy Storage* 34 (Feb. 2021), p. 102011 (cf. p. 92, 93).

(Justo et al., 2013) J. J. Justo, F. Mwasilu, J. Lee, and J.-W. Jung. "AC-microgrids versus DC-microgrids with Distributed Energy Resources: A Review". In: *Renewable and Sustainable Energy Reviews* 24 (Aug. 2013), pp. 387–405 (cf. p. 9).

(Kamal et al., 2022) F. Kamal and B. Chowdhury. "Model Predictive Control and Optimization of Networked Microgrids". In: *International Journal of Electrical Power & Energy Systems* 138 (June 2022), p. 107804 (cf. p. 14).

(Kanakadhurga et al., 2022) D. Kanakadhurga and N. Prabaharan. "Demand Side Management in Microgrid: A Critical Review of Key Issues and Recent Trends". In: *Renewable and Sustainable Energy Reviews* 156 (Mar. 2022), p. 111915 (cf. p. 13).

(Kandari et al., 2022) R. Kandari, N. Neeraj, and A. Micallef. "Review on Recent Strategies for Integrating Energy Storage Systems in Microgrids". In: *Energies* 16.1 (Dec. 27, 2022), p. 317 (cf. p. 9).

(Kanwar et al., 2015) A. Kanwar and D. I. H. Rodrıguez. "A Comparative Study of Optimization- and Rule-Based Control for Microgrid Operation". In: (Jan. 2015) (cf. p. 14).

(Katoch et al., 2021) S. Katoch, S. S. Chauhan, and V. Kumar. "A Review on Genetic Algorithm: Past, Present, and Future". In: *Multimedia Tools and Applications* 80.5 (Feb. 2021), pp. 8091–8126 (cf. p. 15).

(Kebede et al., 2022) A. A. Kebede, T. Kalogiannis, J. Van Mierlo, and M. Berecibar. "A Comprehensive Review of Stationary Energy Storage Devices for Large Scale Renewable Energy Sources Grid Integration". In: *Renewable and Sustainable Energy Reviews* 159 (May 2022), p. 112213 (cf. p. 10).

(Kennedy et al., 1995) J. Kennedy et R. Eberhart. "Particle Swarm Optimization". In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. ICNN'95 - International Conference on Neural Networks. T. 4. Perth, WA, Australia: IEEE, 1995, p. 1942-1948 (cf. p. 15).

(Kim et al., 2016) B.-G. Kim, Y. Zhang, M. van der Schaar, and J.-W. Lee. "Dynamic Pricing and Energy Consumption Scheduling With Reinforcement Learning". In: *IEEE Transactions on Smart Grid* 7.5 (Sept. 2016), pp. 2187–2198 (cf. p. 46, 50).

(Kiran et al., 2022) M. Kiran and M. Ozyildirim. *Hyperparameter Tuning for Deep Reinforcement Learning Applications*. Jan. 26, 2022. arXiv: 2201.11182 [cs]. URL: http://arxiv.org/abs/2201.11182. preprint (cf. p. 83).

(Kirkpatrick et al., 1983) S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. "Optimization by Simulated Annealing". In: *Science* 220.4598 (May 13, 1983), pp. 671–680 (cf. p. 76, 123).

(Kofinas et al., 2018a) P. Kofinas, A. Dounis, and G. Vouros. "Fuzzy Q-Learning for Multi-Agent Decentralized Energy Management in Microgrids". In: *Applied Energy* 219 (June 2018), pp. 53–67 (cf. p. 44, 47).

(Kofinas et al., 2018b) P. Kofinas, G. Vouros, and A. I. Dounis. "Energy Management in Solar Microgrid via Reinforcement Learning Using Fuzzy Reward". In: *Advances in Building Energy Research* 12.1 (Jan. 2, 2018), pp. 97–115 (cf. p. 46).

(Kolodziejczyk et al., 2021) W. Kolodziejczyk, I. Zoltowska, and P. Cichosz. "Real-Time Energy Purchase Optimization for a Storage-Integrated Photovoltaic System by Deep Reinforcement Learning". In: *Control Engineering Practice* 106 (Jan. 2021), p. 104598 (cf. p. 44).

(Konda et al., 1999) V. R. Konda and J. N. Tsitsiklis. "Actor-Critic Algorithms". In: (1999) (cf. p. 33).

(Kozlov et al., 2020) A. N. Kozlov, N. V. Tomin, D. N. Sidorov, E. E. S. Lora, and V. G. Kurbatsky. "Optimal Operation Control of PV-Biomass Gasifier-Diesel-Hybrid Systems Using Reinforcement Learning Techniques". In: *Energies* 13.10 (May 21, 2020), p. 2632 (cf. p. 43, 45).

(Kuznetsova et al., 2013) E. Kuznetsova, Y.-F. Li, C. Ruiz, E. Zio, G. Ault, and K. Bell. "Reinforcement Learning for Microgrid Energy Management". In: *Energy* 59 (Sept. 2013), pp. 133–146 (cf. p. 45).

(Le et al., 2023) T. S. Le, T. N. Nguyen, D.-K. Bui, and T. D. Ngo. *Optimal Sizing of Renewable Energy Storage: A Techno-Economic Analysis of Hydrogen, Battery and Hybrid Systems Considering Degradation and Seasonal Storage | Elsevier Enhanced Reader*. URL: https://reader.elsevier.com/reader/sd/pii/S0306261923001812?token=2947E905A7BA90445B2719A4877D9A0C64FFEFA828D75CD82C21B0021124D1418E0B8447214F6E664C1B8A101F3F8&originRegion=eu-west-1&originCreation=20230403152923 (cf. p. 92, 94).

(Lebrouhi et al., 2022) B. E. Lebrouhi, E. Schall, B. Lamrani, Y. Chaibi, and T. Kousksou. "Energy Transition in France". In: *Sustainability* 14.10 (May 11, 2022), p. 5818 (cf. p. 1).

(H. Lee et al., 2023) H. Lee and J. Romero. *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.* Intergovernmental Panel on Climate Change (IPCC), July 25, 2023 (cf. p. 1).

(S. Lee et al., 2019) S. Lee and D.-H. Choi. "Reinforcement Learning-Based Energy Management of Smart Home with Rooftop Solar Photovoltaic System, Energy Storage System, and Home Appliances". In: *Sensors* 19.18 (Sept. 12, 2019), p. 3937 (cf. p. 50).

(J.-W. Lee et al., 2021) J.-W. Lee, M.-K. Kim, and H.-J. Kim. "A Multi-Agent Based Optimization Model for Microgrid Operation with Hybrid Method Using Game Theory Strategy". In: *Energies* 14.3 (Jan. 25, 2021), p. 603 (cf. p. 15).

(W. Lee et al., 2022) W. Lee, M. Chae, and D. Won. "Optimal Scheduling of Energy Storage System Considering Life-Cycle Degradation Cost Using Reinforcement Learning". In: *Energies* 15.8 (Apr. 11, 2022), p. 2795 (cf. p. 46).

(Lei et al., 2021) L. Lei, Y. Tan, G. Dahlenburg, W. Xiang et K. Zheng. "Dynamic Energy Dispatch Based on Deep Reinforcement Learning in IoT-Driven Smart Isolated Microgrids". In: *IEEE Internet of Things Journal* 8.10 (mai 2021), p. 7938-7953 (cf. p. 44).

(Leo et al., 2014) Leo, Milton et Kaviya. "Multi Agent Reinforcement Learning Based Distributed Optimization of Solar Microgrid". In: *2014 IEEE International Conference on Computational Intelligence and Computing Research* (2014) (cf. p. 45).

(Levent et al., 2021) T. Levent, P. Preux, G. Henri, R. Alami, P. Cordier, and Y. Bonnassieux. "The Challenge of Controlling Microgrids in the Presence of Rare Events with Deep Reinforcement Learning". In: *IET Smart Grid* 4.1 (Feb. 2021), pp. 15–28 (cf. p. 44, 45).

(Levent et al., 2019) T. Levent, P. Preux, E. le Pennec, J. Badosa, G. Henri, and Y. Bonnassieux. "Energy Management for Microgrids: A Reinforcement Learning Approach". In: *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe). Bucharest, Romania: IEEE, Sept. 2019, pp. 1–5 (cf. p. 46).

(B. Li et al., 2017) B. Li, R. Roche, and A. Miraoui. "Microgrid Sizing with Combined Evolutionary Algorithm and MILP Unit Commitment". In: *Applied Energy* 188 (Feb. 2017), pp. 547–562 (cf. p. 17, 75).

(C. Li et al., 2021) C. Li and I. E. Grossmann. "A Review of Stochastic Programming Methods for Optimization of Process Systems Under Uncertainty". In: *Frontiers in Chemical Engineering* 2 (Jan. 28, 2021), p. 622241 (cf. p. 122).

(F.-D. Li et al., 2012) F.-D. Li, M. Wu, Y. He, and X. Chen. "Optimal Control in Microgrid Using Multi-Agent Reinforcement Learning". In: *ISA Transactions* 51.6 (Nov. 2012), pp. 743–751 (cf. p. 42, 50).

(Liashchynskyi et al., 2019) P. Liashchynskyi and P. Liashchynskyi. *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS*. Dec. 12, 2019. arXiv: 1912.06059 [cs, stat]. URL: http://arxiv.org/abs/1912.06059. preprint (cf. p. 83).

(Lillicrap et al., 2019) T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. "Continuous Control with Deep Reinforcement Learning". July 5, 2019 (cf. p. 34).

(W. Liu et al., 2017) W. Liu, P. Zhuang, Y. Liu, H. Liang, Z. Huang, and J. Peng. "Cooperative Neural Fitted Learning for Distributed Energy Management in Microgrids via Wireless Networks". In: *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall). Toronto, ON: IEEE, Sept. 2017, pp. 1–5 (cf. p. 46, 52).

(Y. Liu et al., 2020) Y. Liu, D. Zhang et H. B. Gooi. "Optimization Strategy Based on Deep Reinforcement Learning for Home Energy Management". In: *CSEE Journal of Power and Energy Systems* 6.3 (sept. 2020), p. 572-582 (cf. p. 45, 46, 50).

(Lopes et al., 2019) J. A. P. Lopes, A. G. Madureira, and C. Moreira. "A View of Microgrids". In: *Advances in Energy Systems*. Ed. by P. D. Lund, J. Byrne, R. Haas, and D. Flynn. 1st ed. Wiley, Mar. 18, 2019, pp. 149–166 (cf. p. 8).

(R. Lu et al., 2019) R. Lu, S. H. Hong, and M. Yu. "Demand Response for Home Energy Management Using Reinforcement Learning and Artificial Neural Network". In: *IEEE Transactions on Smart Grid* 10.6 (Nov. 2019), pp. 6629–6639 (cf. p. 50).

(X. Lu et al., 2019) X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor. "Reinforcement Learning-Based Microgrid Energy Trading With a Reduced Power Plant Schedule". In: *IEEE Internet of Things Journal* 6.6 (Dec. 2019), pp. 10728–10737 (cf. p. 51).

(Maheshwari et al., 2018) A. Maheshwari, M. Heck, and M. Santarelli. "Cycle Aging Studies of Lithium Nickel Manganese Cobalt Oxide-Based Batteries Using Electrochemical Impedance Spectroscopy". In: *Electrochimica Acta* 273 (May 2018), pp. 335–348 (cf. p. 92).

(Maheshwari et al., 2020) A. Maheshwari, N. G. Paterakis, M. Santarelli, and M. Gibescu. "Optimizing the Operation of Energy Storage Using a Non-Linear Lithium-Ion Battery Degradation Model". In: *Applied Energy* 261 (Mar. 2020), p. 114360 (cf. p. 92).

(Mathew et al., 2020) A. Mathew, A. Roy, and J. Mathew. "Intelligent Residential Energy Management System Using Deep Reinforcement Learning". In: *IEEE Systems Journal* 14.4 (Dec. 2020), pp. 5362–5372. arXiv: 2005.14259 (cf. p. 50).

(B. V. Mbuwir et al., 2020) B. V. Mbuwir, D. Geysen, F. Spiessens, and G. Deconinck. "Reinforcement Learning for Control of Flexibility Providers in a Residential Microgrid". In: *IET Smart Grid* 3.1 (Feb. 2020), pp. 98–107 (cf. p. 42).

(B. V. Mbuwir et al., 2019) B. V. Mbuwir, F. Spiessens, and G. Deconinck. "Distributed Optimization of Energy Flows in Microgrids Based on Dual Decomposition". In: *IFAC-PapersOnLine* 52.4 (2019), pp. 500–505 (cf. p. 46, 53).

(B. Mbuwir et al., 2017) B. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck. "Battery Energy Management in a Microgrid Using Batch Reinforcement Learning". In: *Energies* 10.11 (Nov. 12, 2017), p. 1846 (cf. p. 71).

(Mirjalili et al., 2014) S. Mirjalili, S. M. Mirjalili, and A. Lewis. "Grey Wolf Optimizer". In: *Advances in Engineering Software* 69 (Mar. 2014), pp. 46–61 (cf. p. 15).

(Mladenović et al., 1997) N. Mladenović and P. Hansen. "Variable Neighborhood Search". In: *Computers & Operations Research* 24.11 (Nov. 1997), pp. 1097–1100 (cf. p. 124).

(Mnih et al., 2016) V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. *Asynchronous Methods for Deep Reinforcement Learning*. June 16, 2016. arXiv: arXiv:1602.01783. URL: http://arxiv.org/abs/1602.01783. preprint (cf. p. 33, 34).

(Mnih et al., 2013) V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. "Playing Atari with Deep Reinforcement Learning". Dec. 19, 2013 (cf. p. 30).

(Mohamed et al., 2013) A. Mohamed and O. Mohammed. "Real-Time Energy Management Scheme for Hybrid Renewable Energy Systems in Smart Grid Applications". In: *Electric Power Systems Research* 96 (Mar. 2013), pp. 133–143 (cf. p. 14).

(Mori et al., 2021) M. Mori, M. Gutiérrez, and P. Casero. "Micro-Grid Design and Life-Cycle Assessment of a Mountain Hut's Stand-Alone Energy System with Hydrogen Used for Seasonal Storage". In: *International Journal of Hydrogen Energy* 46.57 (Aug. 2021), pp. 29706–29723 (cf. p. 16).

(Mourshed et al., 2015) M. Mourshed, S. Robert, A. Ranalli, T. Messervey, D. Reforgiato, R. Contreau, A. Becue, K. Quinn, Y. Rezgui, and Z. Lennard. "Smart Grid Futures: Perspectives on the Integration of Energy and ICT Services". In: *Energy Procedia* 75 (Aug. 2015), pp. 1132–1137 (cf. p. 8).

(Nagapurkar et al., 2019) P. Nagapurkar and J. D. Smith. "Techno-Economic Optimization and Social Costs Assessment of Microgrid-Conventional Grid Integration Using Genetic Algorithm and Artificial Neural Networks: A Case Study for Two US Cities". In: *Journal of Cleaner Production* 229 (Aug. 2019), pp. 552–569 (cf. p. 74).

(Nakabi et al., 2021) T. A. Nakabi and P. Toivanen. "Deep Reinforcement Learning for Energy Management in a Microgrid with Flexible Demand". In: *Sustainable Energy, Grids and Networks* 25 (Mar. 2021), p. 100413 (cf. p. 50).

(Nie et al., 2020) H. Nie, Y. Chen, Y. Xia, S. Huang, and B. Liu. "Optimizing the Post-Disaster Control of Islanded Microgrid: A Multi-Agent Deep Reinforcement Learning Approach". In: *IEEE Access* 8 (2020), pp. 153455–153469 (cf. p. 48).

(Onen et al., 2022) P. S. Onen, G. Mokryani, and R. H. A. Zubo. "Planning of Multi-Vector Energy Systems with High Penetration of Renewable Energy Source: A Comprehensive Review". In: *Energies* 15.15 (Aug. 5, 2022), p. 5717 (cf. p. 41).

(Ouramdane et al., 2021) O. Ouramdane, E. Elbouchikhi, Y. Amirat, and E. Sedgh Gooya. "Optimal Sizing and Energy Management of Microgrids with Vehicle-to-Grid Technology: A Critical Review and Future Trends". In: *Energies* 14.14 (July 10, 2021), p. 4166 (cf. p. 12).

(Pecenak et al., 2020) Z. K. Pecenak, M. Stadler, P. Mathiesen, K. Fahy, and J. Kleissl. "Robust Design of Microgrids Using a Hybrid Minimum Investment Optimization". In: *Applied Energy* 276 (Oct. 2020), p. 115400 (cf. p. 17).

(Père et al., 2022) V. Père, F. Baillon, M. Milhe, and J.-L. Dirion. "The Impact of Reward Shaping in Reinforcement Learning for Agent-based Microgrid Control". In: *Computer Aided Chemical Engineering*. Vol. 51. Elsevier, 2022, pp. 1459–1464 (cf. p. 88).

(Perera et al., 2021) M. K. Perera, K. T. M. U. Hemapala et W. D. A. S. Wijayapala. "Grid Dependency Minimization of a Microgrid Using Single and Multi Agent Reinforcement Learning". In: *2021 IEEE Region 10 Symposium (TENSYMP)*. 2021 IEEE Region 10 Symposium (TENSYMP). Août 2021, p. 1-8 (cf. p. 49).

(Qazi et al., 2020) H. S. Qazi, N. Liu, and T. Wang. "Coordinated Energy and Reserve Sharing of Isolated Microgrid Cluster Using Deep Reinforcement Learning". In: *2020 5th Asia Conference on Power and Electrical Engineering (ACPEE)*. 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE). Chengdu, China: IEEE, June 2020, pp. 81–86 (cf. p. 53).

(Qin et al., 2021) Z. Qin, D. Liu, H. Hua et J. Cao. "Privacy Preserving Load Control of Residential Microgrid via Deep Reinforcement Learning". In: *IEEE Transactions on Smart Grid* 12.5 (sept. 2021), p. 4079-4089 (cf. p. 49).

(D. Qiu et al., 2022) D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac. "Safe Reinforcement Learning for Real-Time Automatic Control in a Smart Energy-Hub". In: *Applied Energy* 309 (Mar. 2022), p. 118403 (cf. p. 43).

(X. Qiu et al., 2016) X. Qiu, T. A. Nguyen, and M. L. Crow. "Heterogeneous Energy Storage Optimization for Microgrids". In: *IEEE Transactions on Smart Grid* 7.3 (May 2016), pp. 1453–1461 (cf. p. 47).

(Rashidi et al., 2021) R. Rashidi, A. Hatami, and M. Abedini. "Multi-Microgrid Energy Management through Tertiary-Level Control: Structure and Case Study". In: *Sustainable Energy Technologies and Assessments* 47 (Oct. 2021), p. 101395 (cf. p. 13).

(2023). *Règlement Du Parlement Européen et Du Conseil*. 19 avr. 2023 (cf. p. 1).

(Ren et al., 2023) H. K. Ren, M. Ashtine, M. McCulloch, and D. Wallom. "An Analytical Method for Sizing Energy Storage in Microgrid Systems to Maximize Renewable Consumption and Minimize Unused Storage Capacity". In: *Journal of Energy Storage* 68 (Sept. 2023), p. 107735 (cf. p. 17).

(Riedmiller, 2005) M. Riedmiller. "Neural Fitted Q Iteration – First Experiences with a Data Efficient Neural Reinforcement Learning Method". In: *Machine Learning: ECML 2005*. Ed. by J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo. Red. by D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, and G. Weikum. Vol. 3720. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 317–328 (cf. p. 37).

(Rios et al., 2022) M. A. Rios and A. Garces. "An Optimization Model Based on the Frequency Dependent Power Flow for the Secondary Control in Islanded Microgrids". In: *Computers & Electrical Engineering* 97 (Jan. 2022), p. 107617 (cf. p. 12).

(Rodriguez-Martinez et al., 2023) O. F. Rodriguez-Martinez, F. Andrade, C. A. Vega-Penagos, and A. C. Luna. "A Review of Distributed Secondary Control Architectures in Islanded-Inverter-Based Microgrids". In: *Energies* 16.2 (Jan. 12, 2023), p. 878 (cf. p. 12).

(Rohit et al., 2017) A. K. Rohit and S. Rangnekar. "An Overview of Energy Storage and Its Importance in Indian Renewable Energy Sector: Part II – Energy Storage Applications, Benefits and Market Potential". In: *Journal of Energy Storage* 13 (Oct. 2017), pp. 447–456 (cf. p. 9).

(Ross et al., 2010) S. Ross, G. J. Gordon, and J. A. Bagnell. "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning". In: (2010) (cf. p. 35).

(Roy et al., 2021) A. Roy, J.-C. Olivier, F. Auger, B. Auvity, E. Schaeffer, S. Bourguet, J. Schiebel, and J. Perret. "A Combined Optimization of the Sizing and the Energy Management of an Industrial Multi-Energy Microgrid: Application to a Harbour Area". In: *Energy Conversion and Management: X* 12 (Dec. 2021), p. 100107 (cf. p. 17).

(Rummery, G.A et al., 1994) Rummery, G.A et Niranjan, M. "On-Line Q-Learning Using Connectionist Systems". In: (1994) (cf. p. 29).

(Salehi et al., 2022) N. Salehi, H. Martinez-Garcia, G. Velasco-Quesada, and J. M. Guerrero. "A Comprehensive Review of Control Strategies and Optimization Methods for Individual and Community Microgrids". In: *IEEE Access* 10 (2022), pp. 15935–15955 (cf. p. 11).

(Salehmin et al., 2022) M. N. I. Salehmin, T. Husaini, J. Goh, and A. B. Sulong. "High-Pressure PEM Water Electrolyser: A Review on Challenges and Mitigation Strategies towards Green and Low-Cost Hydrogen Production". In: *Energy Conversion and Management* 268 (Sept. 2022), p. 115985 (cf. p. 75).

(Samadi et al., 2020) E. Samadi, A. Badri, and R. Ebrahimpour. "Decentralized Multi-Agent Based Energy Management of Microgrid Using Reinforcement Learning". In: *International Journal of Electrical Power & Energy Systems* 122 (Nov. 2020), p. 106211 (cf. p. 46).

(Sarwar et al., 2022) S. Sarwar, D. Kirli, M. M. C. Merlin, and A. E. Kiprakis. "Major Challenges towards Energy Management and Power Sharing in a Hybrid AC/DC Microgrid: A Review". In: *Energies* 15.23 (Nov. 23, 2022), p. 8851 (cf. p. 12).

(Schaul et al., 2016) T. Schaul, J. Quan, I. Antonoglou, and D. Silver. "Prioritized Experience Replay". Feb. 25, 2016 (cf. p. 32, 80).

(Schulman, 2016) J. Schulman. "Optimizing Expectations: From Deep Reinforce- Ment Learning to Stochastic Computation Graphs". 2016 (cf. p. 32, 33).

(Schulman et al., 2017a) J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. "Trust Region Policy Optimization". Apr. 20, 2017 (cf. p. 34).

(Schulman et al., 2017b) J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. "Proximal Policy Optimization Algorithms". Aug. 28, 2017 (cf. p. 34).

(Schwaegerl et al., 2013) C. Schwaegerl and L. Tao. "The Microgrids Concept". In: *Microgrids*. Ed. by N. Hatziargyriou. Chichester, United Kingdom: John Wiley and Sons Ltd, Dec. 6, 2013, pp. 1–24 (cf. p. 8).

(Seger et al., 2023) P. V. Seger, R. Rigo-Mariani, P.-X. Thivel, and D. Riu. "A Storage Degradation Model of Li-ion Batteries to Integrate Ageing Effects in the Optimal Management and Design of an Isolated Microgrid". In: *Applied Energy* 333 (Mar. 2023), p. 120584 (cf. p. 46).

(Shahgholian, 2021) G. Shahgholian. "A Brief Review on Microgrids: Operation, Applications, Modeling, and Control". In: *International Transactions on Electrical Energy Systems* 31.6 (June 2021) (cf. p. 11).

(Shang et al., 2020) Y. Shang, W. Wu, J. Guo, Z. Ma, W. Sheng, Z. Lv, and C. Fu. "Stochastic Dispatch of Energy Storage in Microgrids: An Augmented Reinforcement Learning Approach". In: *Applied Energy* 261 (Mar. 2020), p. 114423 (cf. p. 46).

(Sheikhi et al., 2016) A. Sheikhi, M. Rayati, and A. M. Ranjbar. "Demand Side Management for a Residential Customer in Multi-Energy Systems". In: *Sustainable Cities and Society* 22 (Apr. 2016), pp. 63–77 (cf. p. 47, 49).

(Shojaeighadikolaei et al., 2020) A. Shojaeighadikolaei, A. Ghasemi, K. R. Jones, A. G. Bardas, M. Hashemi, and R. Ahmadi. *Demand Responsive Dynamic Pricing Framework for Prosumer Dominated Microgrids Using Multiagent Reinforcement Learning*. Sept. 22, 2020. arXiv: 2009.10890 [cs, eess]. URL: http://arxiv.org/abs/2009.10890. preprint (cf. p. 50).

(Shuai et al., 2021) H. Shuai et H. He. "Online Scheduling of a Residential Microgrid via Monte-Carlo Tree Search and a Learned Model". In: *IEEE Transactions on Smart Grid* 12.2 (mar. 2021), p. 1073-1087 (cf. p. 42, 45, 46).

(Siarry et al., 1997) P. Siarry and G. Berthiau. "FITTING OF TABU SEARCH TO OPTIMIZE FUNCTIONS OF CONTINUOUS VARIABLES". In: *International Journal for Numerical Methods in Engineering* 40.13 (July 15, 1997), pp. 2449–2457 (cf. p. 124).

(Sidorov et al., 2020) D. Sidorov, D. Panasetsky, N. Tomin, D. Karamov, A. Zhukov, I. Muftahov, A. Dreglea, F. Liu, and Y. Li. "Toward Zero-Emission Hybrid AC/DC Power Systems with Renewable Energy Sources and Storages: A Case Study from Lake Baikal Region". In: *Energies* 13.5 (Mar. 6, 2020), p. 1226 (cf. p. 47).

(Smith et al., 2018) S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. *Don't Decay the Learning Rate, Increase the Batch Size*. Feb. 23, 2018. arXiv: 1711.00489 [cs, stat]. URL: http://arxiv.org/abs/1711.00489. preprint (cf. p. 84).

(Stegherr et al., 2022) H. Stegherr, M. Heider, and J. Hähner. "Classifying Metaheuristics: Towards a Unified Multi-Level Classification System". In: *Natural Computing* 21.2 (June 2022), pp. 155–171 (cf. p. 122).

(Stigka et al., 2014) E. K. Stigka, J. A. Paravantis, and G. K. Mihalakakou. "Social Acceptance of Renewable Energy Sources: A Review of Contingent Valuation Applications". In: *Renewable and Sustainable Energy Reviews* 32 (Apr. 2014), pp. 100–106 (cf. p. 17).

(Sun et al., 2017) Q. Sun, D. Wang, D. Ma, and B. Huang. "Multi-Objective Energy Management for We-Energy in Energy Internet Using Reinforcement Learning". In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2017 IEEE Symposium Series on Computational Intelligence (SSCI). Honolulu, HI: IEEE, Nov. 2017, pp. 1–6 (cf. p. 43).

(Šúri et al., 2005) M. Šúri, T. A. Huld, and E. D. Dunlop. "PV-GIS: A Web-Based Solar Radiation Database for the Calculation of PV Potential in Europe". In: *International Journal of Sustainable Energy* 24.2 (June 2005), pp. 55–67 (cf. p. 83).

(Sutton et al., 1995) R. S. Sutton and A. G. Barto. "Reinforcement Learning: An Introduction". In: (1995), p. 352 (cf. p. 15, 27).

(Sutton et al., 1999) R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: (1999) (cf. p. 33).

(Totaro et al., 2021) S. Totaro, I. Boukas, A. Jonsson, and B. Cornélusse. "Lifelong Control of Off-Grid Microgrid with Model-Based Reinforcement Learning". In: *Energy* 232 (Oct. 2021), p. 121035 (cf. p. 44).

(Townsend et al., 2022) A. Townsend and R. Gouws. "A Comparative Review of Lead-Acid, Lithium-Ion and Ultra-Capacitor Technologies and Their Degradation Mechanisms". In: *Energies* 15.13 (July 5, 2022), p. 4930 (cf. p. 10).

(Transport d'Électricité, 2023) R. de Transport d'Électricité. *BILAN ÉLECTRIQUE 2022 - Rapport Complet*. RTE, 16 fév. 2023 (cf. p. 7).

(Upadhyay et al., 2014) S. Upadhyay and M. Sharma. "A Review on Configurations, Control and Sizing Methodologies of Hybrid Energy Systems". In: *Renewable and Sustainable Energy Reviews* 38 (Oct. 2014), pp. 47–63 (cf. p. 15-17).

(Van Hasselt et al., 2015) H. van Hasselt, A. Guez, and D. Silver. "Deep Reinforcement Learning with Double Q-learning". Dec. 8, 2015 (cf. p. 32, 37, 80).

(Vandoorn et al., 2013) T. Vandoorn, J. De Kooning, B. Meersman, and L. Vandevelde. "Review of Primary Control Strategies for Islanded Microgrids with Power-Electronic Interfaces". In: *Renewable and Sustainable Energy Reviews* 19 (Mar. 2013), pp. 613–628 (cf. p. 12).

(Venayagamoorthy et al., 2016) G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi. "Dynamic Energy Management System for a Smart Microgrid". In: *IEEE Transactions on Neural Networks and Learning Systems* 27.8 (Aug. 2016), pp. 1643–1656 (cf. p. 45, 46).

(C. Wang et al., 2020) C. Wang, S. Mei, Q. Dong, R. Chen, and B. Zhu. "Coordinated Load Shedding Control Scheme for Recovering Frequency in Islanded Microgrids". In: *IEEE Access* 8 (2020), pp. 215388–215398 (cf. p. 48).

(N. Wang et al., 2019) N. Wang, W. Xu, W. Shao, and Z. Xu. "A Q-Cube Framework of Reinforcement Learning Algorithm for Continuous Double Auction among Microgrids". In: *Energies* 12.15 (July 26, 2019), p. 2891 (cf. p. 52).

(Z. Wang et al., 2015) Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas. "Dueling Network Architectures for Deep Reinforcement Learning". 2015 (cf. p. 32, 80).

(Watkins, Christopher J. C. H. et al., 1992) Watkins, Christopher J. C. H. et Dayan, Peter. "Q-Learning". In: (1992) (cf. p. 29).

(Xiao et al., 2017) X. Xiao, C. Dai, Y. Li, C. Zhou, and L. Xiao. "Energy Trading Game for Microgrids Using Reinforcement Learning". In: *Game Theory for Networks*. Ed. by L. Duan, A. Sanjab, H. Li, X. Chen, D. Materassi, and R. Elazouzi. Vol. 212. Cham: Springer International Publishing, 2017, pp. 131–140 (cf. p. 51).

(Xiong et al., 2022) L. Xiong, Y. Tang, S. Mao, H. Liu, K. Meng, Z. Dong et F. Qian. "A Two-Level Energy Management Strategy for Multi-Microgrid Systems With Interval Prediction and Reinforcement Learning". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 69.4 (avr. 2022), p. 1788-1799 (cf. p. 53).

(Xu et al., 2020) X. Xu, Z. Xu, R. Zhang, S. Chai, and J. Li. "Data-driven-based Dynamic Pricing Method for Sharing Rooftop Photovoltaic Energy in a Single Apartment Building". In: *IET Generation, Transmission & Distribution* 14.24 (Dec. 2020), pp. 5720–5727 (cf. p. 50).

(H. Yang et al., 2008) H. Yang, W. Zhou, L. Lu, and Z. Fang. "Optimal Sizing Method for Stand-Alone Hybrid Solar–Wind System with LPSP Technology by Using Genetic Algorithm". In: *Solar Energy* 82.4 (Apr. 2008), pp. 354–367 (cf. p. 17, 109, 136).

(J. Yang et al., 2022) J. Yang, Z. Sun, W. Hu, and L. Steinmeister. "Joint Control of Manufacturing and Onsite Microgrid System via Novel Neural-Network Integrated Reinforcement Learning Algorithms". In: *Applied Energy* 315 (June 2022), p. 118982 (cf. p. 48).

(Y. Yang et al., 2019) Y. Yang, J. Hao, Y. Zheng, and C. Yu. "Large-Scale Home Energy Management Using Entropy-Based Collective Multiagent Deep Reinforcement Learning Framework". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}. Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 630–636 (cf. p. 52).

(Y. Ye et al., 2020) Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward. "Model-Free Real-Time Autonomous Control for a Residential Multi-Energy System Using Deep Reinforcement Learning". In: *IEEE Transactions on Smart Grid* 11.4 (July 2020), pp. 3068–3082 (cf. p. 43).

(Z. Ye et al., 2022) Z. Ye, Y. Gao et N. Yu. "Learning to Operate an Electric Vehicle Charging Station Considering Vehicle-Grid Integration". In: *IEEE Transactions on Smart Grid* 13.4 (juil. 2022), p. 3038-3048 (cf. p. 47).

(Yoldas et al., 2020) Y. Yoldas, S. Goren, and A. Onen. "Optimal Control of Microgrids with Multi-stage Mixed-integer Nonlinear Programming Guided Q-learning Algorithm". In: *Journal of Modern Power Systems and Clean Energy* 8.6 (2020), pp. 1151–1159 (cf. p. 42, 46).

(Yu et al., 2020) L. Yu, W. Xie, D. Xie, Y. Zou, D. Zhang, Z. Sun, L. Zhang, Y. Zhang, and T. Jiang. "Deep Reinforcement Learning for Smart Home Energy Management". In: *IEEE Internet of Things Journal* 7.4 (Apr. 2020), pp. 2751–2762 (cf. p. 46, 49).

(C. Zhang et al., 2019) C. Zhang, S. R. Kuppannagari, C. Xiong, R. Kannan, and V. K. Prasanna. "A Cooperative Multi-Agent Deep Reinforcement Learning Framework for Real-Time Residential Load Scheduling". In: *Proceedings of the International Conference on Internet of Things Design and Implementation*. IoTDI '19: International Conference on Internet-of-Things Design and Implementation. Montreal Quebec Canada: ACM, Apr. 15, 2019, pp. 59–69 (cf. p. 50).

(G. Zhang et al., 2022) G. Zhang, W. Hu, D. Cao, Z. Zhang, Q. Huang, Z. Chen, and F. Blaabjerg. "A Multi-Agent Deep Reinforcement Learning Approach Enabled Distributed Energy Management Schedule for the Coordinate Control of Multi-Energy Hub with Gas, Electricity, and Freshwater". In: *Energy Conversion and Management* 255 (Mar. 2022), p. 115340 (cf. p. 43).

(Q. Zhang et al., 2020a) Q. Zhang, K. Dehghanpour, Z. Wang, and Q. Huang. "A Learning-based Power Management for Networked Microgrids Under Incomplete Information". In: *IEEE Transactions on Smart Grid* 11.2 (Mar. 2020), pp. 1193–1204. arXiv: 1810.01758 (cf. p. 53).

(Q. Zhang et al., 2020b) Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao. "Multi-Agent Safe Policy Learning for Power Management of Networked Microgrids". Oct. 27, 2020 (cf. p. 46).

(S. Zhang et al., 2021) S. Zhang, S. Nandakumar, Q. Pan, E. Yang, R. Migne et L. Subramanian. "Benchmarking Reinforcement Learning Algorithms on Island Microgrid Energy Management". In: *2021 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia)*. 2021 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia). Déc. 2021, p. 1-5 (cf. p. 43).

(T. Zhang et al., 2022) T. Zhang, D. Yue, L. Yu, C. Dou et X. Xie. "Joint Energy and Workload Scheduling for Fog-Assisted Multimicrogrid Systems: A Deep Reinforcement Learning Approach". In: *IEEE Systems Journal* (2022), p. 1-12 (cf. p. 53).

(Zhou et al., 2019) S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang. "Artificial Intelligence Based Smart Energy Community Management: A Reinforcement Learning Approach". In: *CSEE Journal of Power and Energy Systems* (2019) (cf. p. 51).

(Zhu et al., 2022) D. Zhu, B. Yang, Y. Liu, Z. Wang, K. Ma, and X. Guan. "Energy Management Based on Multi-Agent Deep Reinforcement Learning for a Multi-Energy Industrial Park". In: *Applied Energy* 311 (Apr. 2022), p. 118636 (cf. p. 43).

(Ziebart et al., 2008) B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. "Maximum Entropy Inverse Reinforcement Learning". In: (2008) (cf. p. 148).

Table des matières

Abstract

Contributions to the control and sizing of microgrids by reinforcement learning : application to systems with renewable generation and hybrid battery-hydrogen storage

Combining photovoltaic panels with an electrochemical battery reduces the daily phase difference between electricity production and demand in a microgrid. For long-term electricity storage, the combined use of an electrolyzer, hydrogen storage and a fuel cell offers the possibility of conserving electricity produced in summer to meet increased winter demand. Optimal real-time control of microgrid storage units is hampered by random data and the non-linear dynamic behavior of the units over long time horizons.

This work presents a methodology for sizing and controlling a microgrid comprising photovoltaic electricity production, a lithium-ion battery and hydrogen storage, based on economic, environmental and technical objectives. The sizing of the units in a microgrid establishes their constraints of use, while the criteria to be optimized for its sizing (such as the cost of energy, the rate of self-consumption, the probability of breakdowns) depend on the management of these units. This interdependence justifies the development of a sizing methodology coupled with long-term energy management algorithm.

The management of a microgrid is influenced by random variables such as demand and the energy produced at any given time. Reinforcement learning is a sequential decision-making methodology based on a dynamic model of the system that can adapt its strategy to random data. As a first step, a reinforcement learning control methodology is adopted by integrating non-linearities such as the aging of the storage system. Reinforcement learning enabled the energy management system to maintain an effective unit control policy with respect to the targeted criteria. This effectiveness is maintained despite different data and a longer time horizon than those on which the model was built. The control strategies developed suggest that the advantages of long-term electricity storage depend on the characteristics of the microgrid, and in particular on the amplitude of demand and the capacity of the battery. The study shows that a compromise must be found between the economic profitability of the microgrid and the guarantee of its autonomy.

A bi-level optimization method is developed to achieve optimal unit sizing and energy management. The control of the microgrid by reinforcement learning forms the inner loop, while unit sizing is carried out using a simulated-annealing algorithm in the main loop. Particular attention is paid to minimizing computing time, by developing a method for transferring control policy from one iteration of the main loop to another. Offline reinforcement learning has been used to learn unit control strategies without random interaction with the microgrid simulation. The strategies are learned by observing the control decisions made by a model trained on other sizing in previous iterations. The calculation time is reduces by over 50% and the quality of the control policy learned is not affected. The results are analyzed in regard to the objectives considered, the control strategy and the data incorporated into the microgrid simulation.

Keywords : Design, Energy management, Microgrid, Optimization, Reinforcement Learning, Renewable energy

Résumé

Contributions au contrôle et au dimensionnement des micro-réseaux par apprentissage par renforcement : application aux systèmes avec production renouvelable et stockage hybride batterie-hydrogène

Le couplage de panneaux photovoltaïques avec une batterie électrochimique permet d'atténuer le déphasage journalier entre production et demande d'électricité dans un micro-réseau. Pour le stockage d'électricité à long-terme, l'utilisation conjointe d'un électrolyseur, d'un stockage hydrogène et d'une pile à combustible offre en principe la possibilité de conserver l'électricité produite en été pour les besoins accrus de l'hiver. Le contrôle optimal en temps réel des unités de stockage des micro-réseaux est entravé par les données aléatoires et le comportement dynamique non-linéaire des unités sur des horizons temporels longs.
Ce travail présente une méthodologie de dimensionnement et pilotage d'un micro-réseau comprenant une production d'électricité photovoltaïque, une batterie lithium-ion et un stockage hydrogène selon des objectifs économiques, environnementaux et techniques. Le dimensionnement des unités d'un micro-réseau établit leurs contraintes d'utilisation, tandis que les critères à optimiser pour son dimensionnement (comme le coût de l'énergie, le taux d'autoconsommation, la probabilité de pannes) dépendent de la gestion de ces unités. Cette interdépendance justifie le développement d'une méthodologie de dimensionnement couplée au contrôle à long-terme du micro-réseau.
La gestion d'un micro-réseau est influencée par des grandeurs aléatoires telles que la demande et l'énergie produite à chaque instant. L'apprentissage par renforcement est une méthodologie de prise de décision séquentielle s'appuyant sur un modèle dynamique du système et pouvant adapter sa stratégie à des données aléatoires. Dans un premier temps, une méthodologie de contrôle par apprentissage par renforcement est adoptée en intégrant des non-linéarités telles que le vieillissement du système de stockage. L'apprentissage par renforcement a permis de maintenir une politique de contrôle des unités efficace au regard des critères ciblés. Cette efficacité est maintenue malgré des données différentes et un horizon temporel plus long que ceux sur lesquels le modèle a été construit. Les stratégies de contrôle développées suggèrent que l'intérêt du stockage à long-terme de l'électricité dépend des caractéristiques du micro-réseau et en particulier de l'amplitude de la demande et de la capacité de la batterie. L'étude montre qu'un compromis doit être trouvé entre la rentabilité économique du micro-réseau et la garantie de son autonomie.
Une méthode d'optimisation bi-niveaux est développée afin de parvenir à un dimensionnement des équipements avec contrôle optimal. La gestion du micro-réseau par apprentissage par renforcement en constitue la boucle interne, tandis que le dimensionnement des unités est réalisé via un algorithme de recuit-simulé dans la boucle principale. Une attention particulière est accordée à la minimisation du temps de calcul, grâce au développement d'une méthode de transfert de politique de contrôle d'une itération de la boucle principale à une autre. L'utilisation de l'apprentissage par renforcement hors ligne a permis d'apprendre des stratégies de contrôle des unités sans interaction aléatoire avec la simulation du micro-réseau. L'apprentissage des stratégies s'effectue selon l'observation des décisions de contrôle prises par un modèle entraîné sur d'autres dimensionnements à des itérations antérieures. Le temps de calcul est plus de deux fois plus court et la qualité de la politique de contrôle apprise n'est pas affectée. Les résultats sont analysés au regard des objectifs considérés, de la stratégie de contrôle et des données intégrées à la simulation du micro-réseau.
Mots-clés : Apprentissage par renforcement, Dimensionnement, Énergie renouvelable, Gestion énergétique, Micro-réseau, Optimisation