Optimisation par apprentissage du prix d'échange dans un marché hétérogène

Guénolé CHÉROT, Roman LE GOFF LATIMIER, Benjamin CAJNA, Hamid BEN AHMED
SATIE, ENS Rennes, CNRS, Bruz, France

RESUME – Les systèmes d'énergie sont susceptibles de s'organiser à l'avenir comme un système hétérogène faisant coexister des communautés énergétiques indépendantes avec un marché principal. Le prix des échanges entre communautés et marché devrait alors refléter les coûts de production, mais également les contraintes d'acheminement et les potentielles congestions locales. La présente contribution cherche à évaluer le potentiel de l'apprentissage par renforcement pour prévoir ce prix d'échange. Un cas d'étude minimaliste est introduit afin d'améliorer l'interprétabilité et la généricité des résultats obtenus. En particulier les vitesses d'apprentissage seront étudiées afin de discuter le volume de données nécessaire pour garantir un niveau de performance donné. Le transfert des algorithmes entraînés d'un cas d'étude à un autre sera également discuté.

Mots-clés – Apprentissage par renforcement, réseaux de neurones artificiels, prévision de prix, communauté énergétique, systèmes hétérogènes, congestion

1. INTRODUCTION

Du fait de la transition énergétique, les flux de puissance transitant par les systèmes électriques sont susceptibles d'augmenter significativement [1]. En effet, pour atteindre les objectifs de réduction des émissions de CO2, l'électrification des usages est un levier majeur à condition que l'électricité utilisée soit produite avec des moyens décarbonés. L'exemple des énergies renouvelables et des véhicules électriques est à ce titre illustratif. Ce développement des capacités de production et de stockage distribuées, constituées d'un ensemble de moyens de faible puissance, en très grand nombre, modifie en profondeur la structure et l'organisation du système électrique : il favorise l'émergence de nouveaux acteurs, les prosommateurs, et de nouvelles formes d'échange d'électricité. Or, l'organisation historiquement verticale et unidirectionnelle des systèmes électriques, structurée autour d'un marché de gros de l'électricité, n'est pas adaptée pour intégrer facilement les ressources énergétiques distribuées [2]. Il est donc nécessaire d'imaginer un autre modèle afin d'éviter le foisonnement de réseaux autarciques qui pourraient se développer en marge du réseau national du fait que ce dernier ne permettrait pas leur intégration. C'est ainsi que de nouveaux concepts de fonctionnement ont émergé tels que les marchés locaux de l'énergie [3], les communautés énergétiques [4] et les marchés pair à pair (P2P) d'électricité [5]. À l'échelle de l'Union européenne, les travaux législatifs ayant abouti au Clean Energy Package [6] avaient pour objectif de favoriser la décentralisation du système électrique européen en offrant un rôle actif et une autonomisation des consommateurs. C'est d'ailleurs avec ce train de mesures qu'ont été introduites les notions de communautés d'énergie renouvelable et de communautés énergétiques citoyennes. Dans ce contexte, il semble vraisemblable que les réseaux électriques deviennent des systèmes hétérogènes, où des communautés énergétiques cohabiteraient avec le système centralisé conventionnel [7].

Dans la littérature, plusieurs travaux ont abordé ce sujet. Moret et Pinson [8] ont formulé un marché communautaire où les prosommateurs sont autorisés à partager leur énergie au niveau de la communauté ou d'échanger avec l'extérieur via un tiers

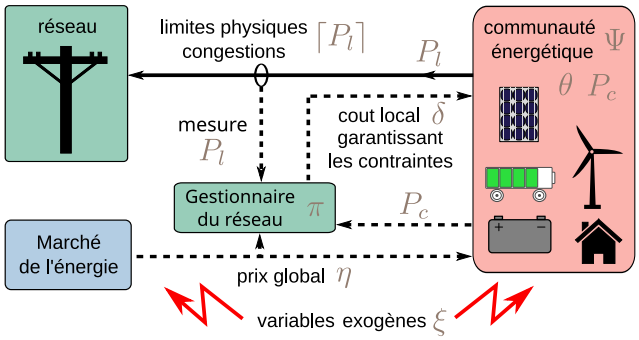

FIG. 1. Configuration considérée pour le cas d'étude : une communauté énergétique Ψ est reliée au réseau principal par une ligne dont la capacité maximale $\lceil P_l \rceil$ sera sollicitée. Le gestionnaire du réseau peut diminuer les échanges P_l dans la ligne en augmentant son coût d'utilisation δ. Il favorise ainsi les échanges intracommunautaires P_c. Le prix du marché η et des acteurs de la communauté θ varie de manière stochastique, ils peuvent être expliqués par les variables cachées exogènes ξ et par le temps t.

superviseur chargé de faire l'interface avec le marché et le gestionnaire de réseau (GR). Dans [9], Morstyn et McCulloch proposent une plateforme de marché P2P permettant aux prosommateurs d'échanger de l'énergie entre eux et avec le marché de gros. Similairement, dans [10] les auteurs introduisent une plateforme d'échange qui réalise l'interface entre les communautés de prosommateurs et les marchés de gros, et qui coordonne les décisions opérationnelles de la communauté en matière d'offre et de demande.

Cependant, bien que ces études proposent différentes organisations pour le fonctionnement d'un système électrique où interagissent des communautés énergétiques et le système conventionnel, elles ne prennent pas en compte les contraintes physiques du réseau électrique. Or, la gestion d'un tel système hétérogène nécessite l'élaboration de nouvelles règles de gestion, en particulier à propos de l'interaction entre un marché centralisé, une communauté énergétique et le gestionnaire du réseau – de transport ou de distribution [11]. En effet, le prix d'échange entre la communauté et le réseau principal doit tenir compte non seulement des coûts de production, mais également des contraintes d'acheminement telles que les congestions ou les limites de tension. Par conséquent, l'exploration approfondie de la relation des échanges d'énergie entre les communautés et le marché de gros est un sujet important. Récemment, Faria [12] à proposé d'intégrer le GR à un marché P2P soit en pénalisant les échanges sources de congestions, soit en incitant les acteurs via un marché de flexibilité.

Ce problème peut naturellement être résolu par une approche d'optimisation sous contrainte, potentiellement distribuée entre le gestionnaire du réseau et la communauté énergétique [13]. Cependant, il serait alors nécessaire de prévoir des échanges d'information : soit les fonctions objectif des acteurs de la communauté, soit a minima un échange itératif des variables duales de l'optimisation [14]. Le déploiement opérationnel d'une telle démarche serait donc confronté à des difficultés réglementaires

et techniques. Il semblerait préférable que le gestionnaire du réseau puisse annoncer à la communauté le prix de l'énergie pour chaque pas de temps d'un horizon temporel afin de respecter ses contraintes physiques [15].

La structure de ce problème appelle donc à se tourner vers les méthodes d'apprentissage automatique telles que les réseaux de neurones ou l'apprentissage par renforcement. Une littérature riche et dynamique est actuellement consacrée à ces questions : résolution de l'optimal power flow (OPF) [16], prévision des prix d'un marché de l'énergie [17], régulation des congestions d'un réseau [18]. Le positionnement de la présente contribution est d'investiguer le potentiel des méthodes d'apprentissage pour la construction d'outils réglementaires dans les systèmes hétérogènes. Sous quelles conditions est-il envisageable de les utiliser pour fixer le prix d'échange entre un marché central et une communauté énergétique ? De quelle manière peut-on garantir la performance de la régulation obtenue ? La résolution de ces questions est indispensable en amont d'un déploiement réel. L'enjeu du volume de données nécessaires à un apprentissage efficace est donc particulièrement crucial. En effet, la collecte de données réelles du fonctionnement d'une communauté représente un délai incompressible préalable à la mise en place de la gestion par apprentissage.

Sans changement du cadre réglementaire, la multiplication des communautés énergétiques semble inévitable. Celles-ci ne seront pas nécessairement constituées d'acteurs proches géographiquement et pourront donc avoir de multiples points de connexion avec le réseau [19]. Dans le cadre de cette contribution, nous considérons que ce type de communauté peut être vu comme une somme de communautés ayant un unique point de connexion et qu'elles peuvent donc être traitées séparément. Dans ce cas, faudra-t-il entraîner un algorithme de gestion pour chacune d'entre elles ? Les résultats de l'une seront-ils transposables à l'autre ?

Cette contribution se concentrera principalement sur la détermination des conditions requises pour un apprentissage efficace. Nous discuterons des métriques appropriées pour quantifier la qualité des données [20]. Les volumes de données nécessaires et la vitesse d'apprentissage seront particulièrement pris en compte. Afin de traiter les questions de similitude entre communautés, la faisabilité d'un apprentissage par transfert sera également investiguée [21].

2. CAS D'ÉTUDE

La figure 1 présente le cas d'étude. Une communauté énergétique Ψ est interconnectée au réseau principal en un seul point. Elle accueille des productions renouvelables distribuées, des consommations non flexibles ainsi que des consommations flexibles. Son interconnexion avec le réseau principal est une ligne dont la contrainte de puissance maximale $\lceil P_l \rceil$ sera sollicitée. Le prix d'échange de la communauté avec le marché extérieur doit donc être ajusté afin de garantir le respect de la capacité de cette ligne.

Les acteurs au sein de la communauté sont tirés au sort à partir d'une base de données. Elle contient des séries temporelles, pas de temps minute, de puissances échangées par des foyers, des véhicules électriques (VE) et des panneaux photovoltaïques (PV). Le nombre de chaque type d'acteur est spécifié à la création de la communauté. Un prix θ – fixé pour toute la simulation et suivant une loi normale $\mathcal{N}(\mu, \sigma)$ – est ensuite associé à chaque acteur. Le nombre d'acteurs, et les valeurs de μ et σ sont donnés tableau 1.

TABLEAU 1. Valeurs utilisées pour la création du cas test.

	μ	σ	Nombre d'acteurs
Foyer	0.25	0.10	40
VE	0.15	0.05	10
PV	0.08	0.02	30

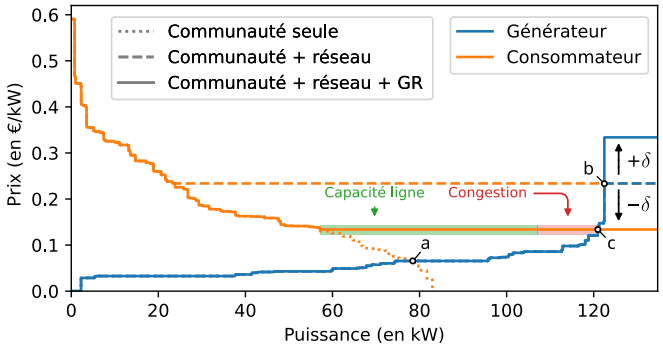

FIG. 2. Influence du coût d'utilisation de la ligne δ sur l'ordre de mérite de la communauté. Les ordres de mérite des consommateurs et des générateurs sont respectivement tracés en orange et en bleu. Ils sont modifiés en fonction des acteurs considérés : communauté, réseau et GR. ⓐ, ⓑ et ⓒ représentent les différents points d'équilibre en fonction des acteurs considérés. Les zones verte et rouge représentent respectivement la limite de puissance $\lceil P_l \rceil$ et le surplus de puissance $||\lceil P_l \rceil - |P_l||$ dans l'interconnexion.

La communauté est donc caractérisée par un ensemble de N_a agents, dont la puissance évolue avec le temps et dont le prix d'achat ou de vente θ est fixé. Les variations de puissance de chaque agent sont régies par des phénomènes exogènes ξ – l'heure de la journée, la saison, la température, etc. – qui ne sont pas renseignés dans la base de données. Cette formulation offre deux difficultés d'intérêt : d'une part, les puissances des acteurs de la communauté sont privées, elles doivent donc être estimées. D'autre part, le prix δ imposé par le GR modifie ces puissances, ce qui rend la tâche de prévision plus complexe.

Dans une démarche de science ouverte et reproductible, l'ensemble des données utilisées sont publiquement accessibles : [22] pour la consommation des foyers et les productions éoliennes, [23] pour les consommations de véhicules électriques et EPEX [1] pour le prix du marché de l'énergie. Les algorithmes d'apprentissages sont issus de l'implémentation proposée par Stable-Baselines3 [24] avec le jeu de paramètres par défaut. Les codes sources sont ouverts et accessibles sur GitLab [2].

Les simulations ont été conduites en séparant les 365 jours de la base de données en deux groupes : 90 % des jours ont servi à l'entraînement, 10 % pour le test. Les données d'entraînement peuvent être vues plusieurs fois par l'algorithme au cours de l'apprentissage.

3. MÉTHODE

3.1. *Calcul de la puissance échangée*

La figure 2 présente l'ordre de mérite de la communauté et comment il est affecté par l'évolution du prix d'utilisation de la ligne. Les offres de consommation (en orange) sont classées par ordre croissant du prix d'achat – plus un consommateur est prêt à acheter cher, plus il a de chance de voir sa demande satisfaite. Les offres de production sont classées par ordre décroissant – les producteurs les moins chers seront sélectionnés en premier. Trois ordres de mérite sont représentés : ⓐ Communauté seule ⓑ échangeant avec le réseau extérieur sans contraintes de capacité ⓒ incluant le coût imposé par le GR. Chaque ordre de mérite est associé à un point d'équilibre (couple prix/puissance échangée) représenté par un cercle sur la figure. Dans le cas ⓐ aucune puissance ne transite dans l'interconnexion, environ 80 kW sont échangés dans la communauté. Dans le cas ⓑ, le réseau de puissance infinie propose d'acheter ou de vendre l'énergie au prix de 0.25 €/kW : tous les consommateurs (resp. producteur) souhaitant acheter (resp. vendre) à un prix inférieur (resp. supérieur) ne seront pas sélectionnés dans l'ordre de mérite et n'échangeront aucune puissance. Dans ce cas, environ

1. https://ewoken.github.io/epex-spot-data/
2. https://gitlab.com/satie.sete/sge_rl

100 kW seront exportés de la communauté vers le réseau. Dans le cas ⓒ, le GR impose un coût δ d'utilisation de la ligne. Du point de vue de la communauté, ce coût diminue (resp. augmente) le prix d'échange avec les consommateurs (resp. producteurs) du réseau extérieur. Plus de puissance est échangée dans la communauté et donc moins dans l'interconnexion : la congestion (représenté par un rectangle rouge) diminue. En augmentant encore le coût, la congestion diminuerait jusqu'à devenir absente. La puissance échangée serait inférieure ou égale à la capacité de la ligne, représentée en vert.

L'objectif du GR est de maintenir l'intégrité du réseau tout en minimisant son impact sur les échanges de puissance. On observe que la puissance transitant dans la ligne décroît de façon monotone quand le coût augmente. Un GR omniscient – connaissant les préférences de chaque acteur et donc capable de calculer l'ordre de mérite – peut donc aisément calculer le coût optimal d'utilisation de la ligne δ_*.

En pratique, deux facteurs limitent le calcul de δ_*. D'une part, il existe une latence entre la mesure des grandeurs du réseau et l'envoi d'un nouveau coût. Les algorithmes proposés seront donc comparés à un GRD "avec retard" qui appliquera au temps $t + \Delta t$ la stratégie optimale $\delta_*^{[t]}$ calculé en t. D'autre part, les préférences des acteurs sont souvent inconnues. Il est alors impossible de calculer l'ordre de mérite et le δ_* associé. Seuls la puissance transitant dans l'interconnexion et le prix du marché de l'énergie sont connus. Dans ce contexte, la prévision du coût d'échange est bien plus complexe, nous utiliserons donc une approche d'apprentissage par renforcement pour déterminer la stratégie optimale.

3.2. *Apprentissage*

Le formalisme de l'apprentissage par renforcement vise à résoudre les problèmes de prise de décisions séquentielles, il est donc parfaitement adapté ici. Le schéma d'apprentissage est donné en figure 4. À chaque instant t, le GR doit choisir un coût d'utilisation de la ligne $\delta^{[t+1]}$ basé sur un certain nombre de variables explicatives noté $O^{[t]}$ mesuré au pas de t : le prix du marché européen $\eta^{[t]}$, le coût de l'interconnexion $\delta^{[t]}$, la puissance transitant par l'interconnexion $P_l^{[t]}$, la puissance totale échangée dans la communauté $P_c^{[t]}$ et le temps t. L'observation est ensuite normée entre -1 et 1 pour faciliter l'apprentissage. La stratégie, aussi nommée politique, se note $\pi(O^{[t]}) = \delta^{[t]}$. Pour l'améliorer, l'agent par renforcement se base sur un unique signal scalaire nommé récompense R, dont il doit maximiser l'espérance de la somme. Le choix de cette fonction est primordial, car la stratégie optimale π_* en découle directement.

L'objectif du GR étant de maintenir l'intégrité du réseau tout en minimisant son impact sur les échanges de puissance, nous avons choisi la forme décrite équation (1) , où $\lceil\cdot\rceil$ est l'opérateur $max(\cdot)$. Tant qu'il n'y a pas de congestion, la récompense est proportionnelle aux échanges dans la ligne. Les congestions sont pénalisées par une récompense négative, proportionnelle à l'amplitude de celle-ci. Ces deux termes sont pondérés par $\alpha \in [0, 1]$ dont l'influence sera discutée section 4.3.

$$R = \begin{cases} (1-\alpha)\cdot|P_l| & \text{si } |P_l| < \lceil P_l \rceil \\ -\alpha\cdot\left(|P_l| - \lceil P_l \rceil\right) & \text{sinon} \end{cases} \quad (1)$$

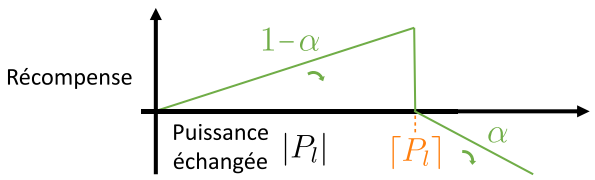

FIG. 3. Forme de la fonction récompense. Quand α augmente la puissance échangée est moins récompensée et la congestion plus pénalisée.

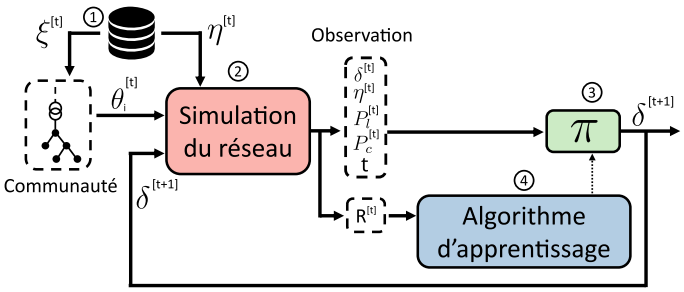

FIG. 4. Apprentissage par renforcement du coût d'utilisation de la ligne. ① Les variables exogènes à l'instant t – noté $\xi^{[t]}$– issues d'une base de données sont transmises à la communauté énergétique qui donne en retour les prix $\theta_i^{[t]}$ de chaque acteur. ② Ces prix $\theta_i^{[t]}$, accompagnés du prix du réseau $\eta^{[t]}$ et du coût d'utilisation de la ligne $\delta^{[t]}$ permettent de calculer l'ordre de mérite et d'en déduire le vecteur d'observation $O^{[t]}$. ③ L'observation est transmise au GR pour qu'il déduise le prochain prix $\delta^{[t+1]}$. ④ En parallèle, un algorithme d'apprentissage améliore les performances du contrôleur grâce à un signal de récompense pénalisant les violations des contraintes.

L'essor récent de l'apprentissage par renforcement a permis l'émergence de nombreux algorithmes. Nous en évaluerons deux parmi les plus performants. L'algorithme PPO pour *Proximal Policy Optimization* [25] est un algorithme basé sur l'optimisation directe de la politique. L'idée principale consiste à écrêter le gradient lors de l'amélioration de la politique. Ainsi la nouvelle politique est proche de l'ancienne, ce qui permet de stabiliser l'apprentissage. L'algorithme SAC pour *Soft Actor Critic* [26] se base sur l'apprentissage de la fonction état-action Q_π, un estimateur de la performance de la politique. Contrairement à PPO, cet algorithme stocke les expériences passées dans une mémoire, ce qui lui permet d'être plus efficace pour les problèmes imposant un nombre d'interactions limité avec l'environnement.

4. RÉSULTATS ET DISCUSSION

4.1. *Séries temporelles*

L'évolution des grandeurs d'intérêt au cours du temps est représentée figure 5. Sans GR (en violet) P_l dépasse la valeur maximale autorisée $\lceil P_l \rceil$ entre 9 h et 15 h. Cela se traduit par une récompense négative, qui vient pénaliser les congestions. À l'opposé, la stratégie optimale (en vert) respecte toujours la contrainte en augmentant le coût d'utilisation de la ligne. Enfin, l'agent SAC (en orange) respecte la contrainte 90 % du temps, même si quelques dépassements d'une amplitude maximale de $2 \cdot \lceil P_l \rceil$ sont présents. Le coût est bien prédit lors des phases de congestions, et surestimé sinon. Ce comportement est connu de l'algorithme SAC : les actions (ici δ) sont échantillonnées à partir d'un modèle gaussien initialement centré en zéro. L'apprentissage vise à modifier cette distribution, mais les valeurs extrêmes sont difficilement atteintes. Ce problème pourrait être résolu en réalisant une translation de l'espace d'action de façon à centrer l'action $\delta = 0$.

4.2. *Vitesse d'apprentissage*

La figure 6 montre l'évolution de la récompense totale moyenne Eq. (2) en fonction du temps d'entraînement.

$$\overline{R_{tot}} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=0}^{T_{fin}} R_{\pi_i}^{[t]}\right) \quad (2)$$

où N est le nombre d'algorithmes entraîné, T_{fin} est le temps d'une simulation et R_π est la récompense obtenue par l'agent suivant la politique π. L'apprentissage étant stochastique, il est nécessaire d'évaluer chaque algorithme plusieurs fois – 100 entraînements par algorithme ici. Cinq stratégies sont évaluées :

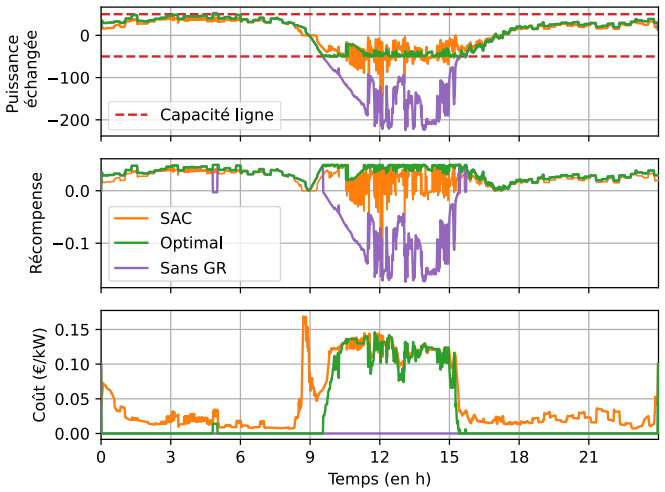

FIG. 5. Évolution de la puissance échangée P_l, de la récompense R et du coût δ en fonction du temps pour trois types d'agent : contrôle optimal, agent SAC entraîné et sans GR ($\delta = 0$).

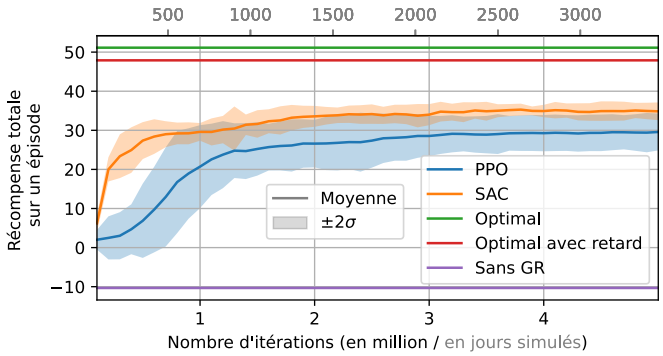

FIG. 6. Récompense moyenne obtenue au cours de l'entraînement par les algorithmes SAC et PPO comparés à trois stratégies déterministes : pas de gestion ($\delta = 0$), gestion optimale et gestion optimale sans connaissance du futur. Le nombre d'itérations est donné en million (bas de la figure) et en jours simulé (haut de la figure). Le calcul des écarts types σ permet d'afficher les intervalles à $\pm 2\sigma$ en transparence.

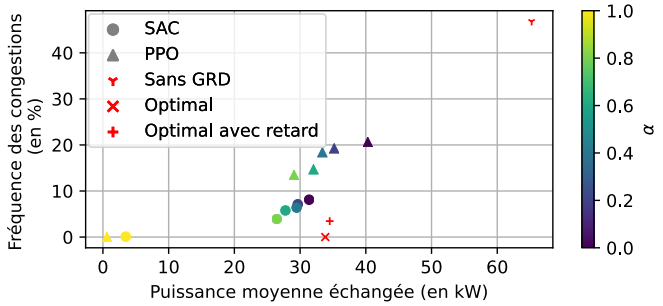

FIG. 7. Front de Pareto obtenu en faisant varier α entre 0 et 1. La couleur des points indique la valeur des paramètres α. Chaque rond (resp. triangle) représente la moyenne de cinq entraînements de l'algorithme SAC (resp. PPO). Les performances de trois stratégies ne nécessitant pas d'apprentissage (optimal, optimal avec retard et sans gestion) sont données à titre de comparaison.

d'une part SAC et PPO dont les performances évoluent au cours de l'entraînement et d'autre part les stratégies optimale, optimale avec retard et sans GR (voir section 3.1) qui constituent une référence.

La figure 6 permet d'une part d'apprécier la quantité de données nécessaire à l'apprentissage et d'autre part de comparer les vitesses de convergence des algorithmes. Elle ne donne pas d'indications directes concernant la capacité des algorithmes à respecter les contraintes, car la récompense R est une métrique agrégée (voir section 4.3).

Analysons les performances de chaque stratégie. Une des pires stratégies possibles consistent à ne rien faire $\delta = 0$, ce qui conduit à une récompense moyenne de -10 là où les stratégies optimale et optimale avec retard obtiennent respectivement 50.5 et 49. Tout au long de l'apprentissage, l'algorithme SAC obtient une récompense moyenne strictement supérieure à PPO. Sa récompense moyenne asymptotique est également meilleure d'environ $5\,\%$. SAC est également plus rapide à converger : deux millions de pas de temps lui sont nécessaires, ce qui correspond de 2 à 4 ans simulés. Ces résultats sont bien connus de la littérature : la mémoire de SAC lui permet d'apprendre en utilisant moins de données en contrepartie d'un coût de calcul plus important.

4.3. Influence des paramètres de la récompense

La figure 7 décrit l'influence de α sur les performances du contrôle, il varie entre 0 et 1 (voir figure 3). Cinq entraînements sont réalisés pour chaque algorithme et chaque valeur de α. La moyenne des performances est représentée. En abscisse, la puissance échangée doit être maximisée, en ordonnée, la fréquence des congestions doit être minimisée. Le point de fonctionnement optimal se situe donc dans le coin inférieur droit.

La politique optimale (x) domine l'ensemble des autres solutions et ne conduit à aucune congestion. La politique optimale avec retard (+) conduit à l'échange de plus de puissance – 1 kW en moyenne – au détriment de congestions plus fréquentes. Enfin la politique sans GR conduit aux échanges les plus importants ainsi qu'à une violation de la capacité dans $48\,\%$ du temps.

Étudions maintenant en détail les algorithmes d'apprentissages et leurs sensibilités au paramètre α. Pour $\alpha = 1$, la puissance échangée n'est pas récompensée, la récompense maximale vaut donc $\lceil R \rceil = 0$. Dans ce cas, la stratégie optimale consiste à imposer $\delta = \lceil \delta \rceil$. C'est ce que nous observons dans le coin inférieur gauche de la figure 7 : la puissance moyenne échangée est presque nulle et aucune congestion n'est observée. Pour $\alpha = 0$, les congestions ne sont pas pénalisées. Cependant, la stratégie optimale ne consiste pas à imposer $\delta = 0$ car cela conduirait fréquemment à des congestions : $P_l > \lceil P_l \rceil$. Or la récompense maximale est atteinte lorsque $P_l = \lceil P_l \rceil$. Le signal de récompense proposé équation 1 conduit donc à la maximisation des échanges tout en minimisant les contraintes pour tout α différent de 0. Pour $\alpha = 0.16$, l'algorithme PPO échange en moyenne $29\,\text{kW}$ ce qui conduit à des congestions dans $14\,\%$ du temps. Pour $\alpha = 0.32$, la puissance augmente légèrement ($32\,\text{kW}$) et les congestions sont plus nombreuses ($15\,\%$ du temps). De manière générale, pour SAC comme pour PPO, l'augmentation de α va de pair avec l'augmentation de la puissance échangée et des congestions. Le choix de sa valeur est donc essentiel, car elle permet au GR de fixer les paramètres de l'apprentissage en fonction de son aversion aux risques. Enfin, notons que SAC domine une partie des solutions donnée par PPO. Cela confirme la supériorité de SAC pour ce cas d'application.

Notons que la forme de la fonction récompense n'est pas discutée ici –seul le paramètre α évolue. Nous avons suivi les règles suivantes pour créer la fonction récompense : i) Dans la zone sans congestion, il faut que la récompense augmente avec la puissance. Si elle était décroissante ou constante, l'algorithme serait récompensé à limiter le flux dans la ligne, ce qui va à l'encontre du rôle du GR. ii) Dans la zone avec congestion, il faut que la récompense diminue avec la puissance, car les congestions de grandes amplitudes sont les plus dangereuses pour le réseau. Cela étant, au lieu d'une relation linéaire en R et P_l nous aurions pu choisir une relation polynomiale, exponentielle, etc. Cela aurait modifié les performances de l'algorithme, sans modifier les conclusions générales de cet article.

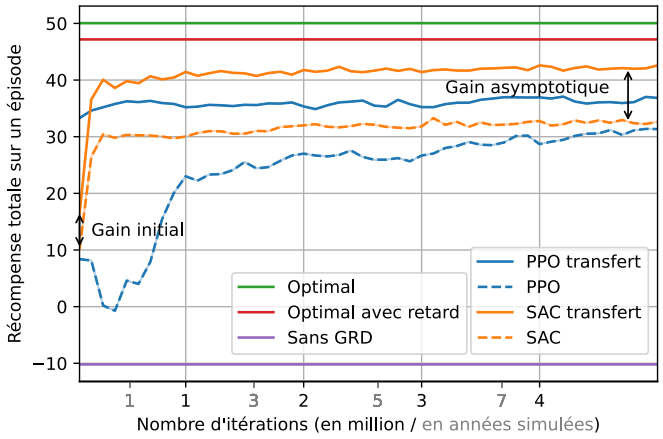

FIG. 8. Performance de l'apprentissage par transfert comparé à un entraînement simple. Les performances de trois stratégies ne nécessitant pas d'apprentissage (optimal, optimal avec retard et sans gestion) sont données à titre de comparaison. Les gains initial et asymptotiques sont indiqués pour l'algorithme SAC.

4.4. Transfert de connaissances

Comme nous l'avons vu figure 6, l'apprentissage nécessite de nombreuses interactions avec l'environnement. Durant cette phase, l'agent prend des décisions non optimales qui pourraient avoir de graves conséquences sur le réseau. Or le partage d'algorithme pré-entraîné pourrait grandement accélérer le temps d'apprentissage tout en réduisant les risques liés à l'exploration.

La littérature de l'apprentissage par transfert [21] propose des dizaines d'approches dont les performances sont très variables en fonction du problème posé. Certaines méthodes sont spécifiques à l'apprentissage par renforcement [27, 28]. Elles se basent principalement sur l'existence d'une ou plusieurs politiques expertes π_X – issue d'un entraînement préalable – pour entraîner la politique par transfert π_T. Les performances de π_T et π_0 – politique sans transfert – sont caractérisées par quatre métriques : le gain de récompense à l'initialisation, le gain de temps pour atteindre un certain seuil, le gain de récompense asymptotique et le regret exprimé équation (3). π_* étant la politique optimale, il est toujours positif.

$$\mathcal{R}(T, \pi_T, \pi_*) = \sum_{t=0}^{T} R^{[t]}_{\pi_*} - R^{[t]}_{\pi_T} \tag{3}$$

La méthode la plus intuitive, nommée transfert de politique, consiste à utiliser une politique experte π_X pour initialiser la politique par transfert π_T : $\pi_T^{[0]} = \pi_X$. Cette approche permet en général un gain de performance à l'initialisation, mais ne change peu la courbe d'apprentissage et les performances asymptotiques. De plus, elle ne tire pas parti de la présence possible de plusieurs politiques expertes.

L'apprentissage par démonstration consiste à utiliser un groupe de politiques expertes et d'apprendre en sélectionnant celles commettant les erreurs de prévisions les plus faibles. Cette approche sera explorée dans une prochaine contribution.

La figure 8 présente les performances du transfert de politique. Chacune des N politiques expertes $\pi_{X,i}$ a été entraînée sur une communauté énergétique différente Ψ_i. Comme décrit section 2, elles diffèrent par les agents qui les composent. Elles servent ensuite à initialiser un agent devant apprendre dans la communauté Ψ_T : $\pi_{T,i}^{[0]} = \pi_{X,i}$. Six agents par transfert sont ainsi entraînés. Leurs performances seront comparées à six agents π_0 entraînés à partir de zéro sur la communauté Ψ_T. Pour simplifier la lecture, seul le meilleur agent de chaque catégorie est représenté.

Pour PPO, le gain initial est de 25 ce qui est très encourageant, car cela signifie que π_T prend des décisions proches de l'optimum dès le début de l'entraînement. Les risques de congestions sont donc limités lors des premières interactions avec le nouvel environnement. Le regret vaut $\mathcal{R} = 7.3 \cdot 10^7$. Le gain de performance asymptotique est inférieur à 5 et il aurait peut-être été nul si π_0 avait été entraîné plus longtemps. Pour SAC, le gain initial est faible, mais l'apprentissage est plus rapide. Le regret vaut $\mathcal{R} = 4.9 \cdot 10^7$. Le gain asymptotique est d'environ 10. Cela peut s'expliquer par le fait que π_T explore son environnement plus efficacement que π_0. Il est donc moins sensible aux minimums locaux.

5. CONCLUSION

La présente contribution a permis d'évaluer différentes techniques d'apprentissage afin de prévoir le prix optimal entre une communauté énergétique et le marché principal, en prenant en compte la capacité de la ligne reliant la communauté au réseau principal. La méthode *Soft Actor Critic* s'est avérée plus efficace que la méthode *Proximal policy optimization* tant en termes de vitesse d'apprentissage que de performances asymptotiques. Nous montrons que l'apprentissage du coût peut être réalisé dans des temps raisonnables – de l'ordre d'un à deux ans simulé – et qu'un mécanisme simple comme le transfert de politique permet d'accélérer significativement le temps de convergence tout en minimisant le regret. Cela confirme l'intérêt du partage d'information entre communautés. De plus, la fonction de récompense proposée peut être modifiée par le GRD en fonction de son aversion aux risques.

Les perspectives de cette étude sont les suivantes : d'une part, l'apprentissage doit être robustifié pour prévenir les coûts pouvant mener à des congestions trop importantes. Cela peut être réalisé en changeant la forme de la fonction récompense, en intégrant un superviseur dit "pessimiste" empêchant l'exploration d'état risqué, ou en améliorant l'apprentissage par transfert, notamment grâce à l'apprentissage par démonstration. D'autre part, la prévision du coût doit être généralisée à l'ensemble des lignes du réseau. Le formalisme de l'optimal power flow nous renseigne sur l'existence des prix nodaux, permettant de contrôler parfaitement le réseau et d'atteindre le point de fonctionnement optimal. La prévision de ces prix serait une généralisation de la présente contribution qui permettrait de prendre en compte les contraintes de tensions qui sont les contraintes prédominantes au sein des réseaux de distribution.

Pour conclure, ce cas d'étude volontairement minimaliste a été développé afin de privilégier l'interprétabilité des résultats et d'aboutir à des règles potentiellement généralisables. Dans une démarche de science reproductible et ouverte, les codes développés, ainsi que les données utilisées sont publiquement accessibles sur un dépôt GitLab [3].

6. RÉFÉRENCES

[1] H. R. Galiveeti, A. K. Goswami, and N. B. Dev Choudhury, "Impact of plug-in electric vehicles and distributed generation on reliability of distribution systems," *Eng. Sci. Technol. an Int. J.*, vol. 21, pp. 50–59, feb 2018.

[2] X. Jin, Q. Wu, and H. Jia, "Local flexibility markets : Literature review on concepts, models and clearing methods," *Appl. Energy*, vol. 261, p. 114387, mar 2020.

[3] F. Teotia and R. Bhakar, "Local energy markets : Concept, design and operation," *2016 Natl. Power Syst. Conf. NPSC 2016*, feb 2017.

[4] S. Moroni, V. Alberti, V. Antoniucci, and A. Bisello, "Energy communities in the transition to a low-carbon future : A taxonomical approach and some policy dilemmas," *Journal of Environmental Management*, vol. 236, pp. 45–53, 4 2019.

[5] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin, "Peer-to-peer and community-based markets : A comprehensive review," apr 2019.

[6] D.-G. for Energy (European Commission), *Clean energy for all Europeans*. Publications Office of the European Union, 2019.

3. https://gitlab.com/satie.sete/sge_rl

[7] S. Kerscher and P. Arboleya, "The key role of aggregators in the energy transition under the latest European regulatory framework," *Int. J. Electr. Power Energy Syst.*, vol. 134, p. 107361, jan 2022.

[8] F. Moret and P. Pinson, "Energy collectives : A community and fairness based approach to future electricity markets," *IEEE Transactions on Power Systems*, vol. 34, pp. 3994–4004, 9 2019.

[9] T. Morstyn and M. D. McCulloch, "Multiclass energy management for peer-to-peer energy trading driven by prosumer preferences," *IEEE Transactions on Power Systems*, vol. 34, pp. 4005–4014, 9 2019.

[10] J. M. Zepter, A. Lüth, P. del Granado, and R. Egging, "Prosumer integration in wholesale electricity markets : Synergies of peer-to-peer trade and residential storage," *Energy and Buildings*, vol. 184, pp. 163–176, 2 2019.

[11] I. Bouloumpasis, D. Steen, and L. A. Tuan, "Congestion Management using Local Flexibility Markets : Recent Development and Challenges," in *Proc. 2019 IEEE PES Innov. Smart Grid Technol. Eur. ISGT-Europe 2019*, Institute of Electrical and Electronics Engineers Inc., sep 2019.

[12] A. S. Faria, T. Soares, T. Orlandini, C. Oliveira, T. Sousa, P. Pinson, and M. Matos, "P2P market coordination methodologies with distribution grid management," *Sustain. Energy, Grids Networks*, p. 101075, may 2023.

[13] A. Kargarian, J. Mohammadi, J. Guo, S. Chakrabarti, M. Barati, G. Hug, *et al.*, "Toward Distributed/Decentralized DC Optimal Power Flow Implementation in Future Electric Power Systems," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 2574–2594, 2018.

[14] R. L. G. Latimier, G. Chérot, and H. B. Ahmed, "Online learning for distributed optimal control of an electric vehicle fleet," *Electric Power Systems Research*, vol. 212, p. 108330, 2022.

[15] A. Heydari, M. Majidi Nezhad, E. Pirshayan, D. Astiaso Garcia, F. Keynia, and L. De Santoli, "Short-term electricity price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm," *Appl. Energy*, vol. 277, p. 115503, nov 2020.

[16] M. Chatzos, T. W. Mak, and P. V. Hentenryck, "Spatial Network Decomposition for Fast and Scalable AC-OPF Learning," *IEEE Trans. Power Syst.*, vol. 37, pp. 2601–2612, jul 2022.

[17] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron, "Forecasting day-ahead electricity prices : A review of state-of-the-art algorithms, best practices and an open-access benchmark," *Appl. Energy*, vol. 293, p. 116983, jul 2021.

[18] R. Henry and D. Ernst, "Gym-ANM : Reinforcement learning environments for active network management tasks in electricity distribution systems," *Energy AI*, vol. 5, p. 100092, sep 2021.

[19] V. Z. Gjorgievski, S. Cundeva, and G. E. Georghiou, "Social arrangements, technical designs and impacts of energy communities : A review," *Renew. Energy*, vol. 169, pp. 1138–1156, may 2021.

[20] C. Shi, R. Wan, R. Song, W. Lu, and L. Leng, "Does the markov decision process fit the data : Testing for the markov property in sequential decision making," in *37th Int. Conf. Mach. Learn. ICML 2020*, vol. PartF16814, pp. 8766–8776, International Machine Learning Society (IMLS), feb 2020.

[21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, *et al.*, "A Comprehensive Survey on Transfer Learning," *Proc. IEEE*, vol. 109, pp. 43–76, jan 2021.

[22] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J. Albrecht, and Others, "Smart* : An open data set and tools for enabling research in sustainable homes," *SustKDD, August*, vol. 111, no. 112, p. 108, 2012.

[23] A. L. Sørensen, K. B. Lindberg, I. Sartori, and I. Andresen, "Residential electric vehicle charging datasets from apartment buildings," *Data Br.*, vol. 36, p. 107105, jun 2021.

[24] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann, "Stable Baselines3." \url{https ://github.com/DLR-RM/stable-baselines3}, 2019.

[25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv*, vol. 1707.06347, jul 2017.

[26] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic : Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 5, pp. 2976–2989, International Machine Learning Society (IMLS), jan 2018.

[27] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains : A survey," *J. Mach. Learn. Res.*, vol. 10, pp. 1633–1685, dec 2009.

[28] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer Learning in Deep Reinforcement Learning : A Survey," *arXiv Prepr.*, vol. abs/2009.0, sep 2020.