

Bandits Multi-Bras Multi-Agent : application à la recharge intelligente des véhicules électriques sous contraintes réseau

Eloann Le Guern--Dall'o¹, Raphaël Féraud², Guy Camilleri³, Patrick Maillé⁴
Fabien Petit⁵, Riadh Zorgati⁶, Fatma Salem², Hamid Ben Ahmed¹, Anne Blavette¹

¹Univ Rennes, ENS Rennes, CNRS, IETR lab, Rennes, France

²Orange Innovation, Lannion, France

³IRIT, Univ. Toulouse, CNRS, Toulouse INP, Toulouse, France

⁴IMT Atlantique, Univ. Rennes, Rennes, France

⁵SRD, Poitiers, France

⁶EDF, Palaiseau, France

RÉSUMÉ – L'électrification des usages couplé aux énergies renouvelables (ENR) représente un levier majeur de décarbonation. Pour optimiser l'utilisation des ENR et gérer leur variabilité, les véhicules électriques (VE) offrent une flexibilité prometteuse. Cet article présente un système multi-agent décentralisé extensible, où chaque VE cherche les meilleurs instants de recharge pour satisfaire ses besoins de mobilité en intégrant un maximum d'énergie photovoltaïque (PV) sans congestionner le réseau électrique. Chaque VE évolue de manière autonome et apprend des signaux de son environnement, suivant un algorithme d'échantillonnage de Thompson sur un problème de bandit multi-bras. Quatre types d'algorithmes de bandit sont comparés à une référence omnisciente centralisée et une heuristique simple, dans une simulation simplifiée d'un réseau électrique avec 55 agents VE quasi-homogènes, atteignant une différence de seulement 14% à l'optimal omniscient. La faisabilité d'un tel système est démontrée par le temps de calcul nécessaire par agent par pas de temps (200 μ s) et un échange d'information limité, préservant la vie privée de l'utilisateur tout en assurant une gestion efficace des congestions.

Mots-clés– réseau intelligent, multi-agent, apprentissage par renforcement, bandits, véhicule électrique, recharge intelligente

1. INTRODUCTION

Afin de lutter contre le changement climatique, de nombreux pays se sont donnés pour objectif d'intégrer massivement des sources d'énergie renouvelable (ENR) dans leur mix énergétique. Cela se traduit par un fort taux de pénétration projeté du photovoltaïque (PV) dans les systèmes électriques. Cependant, la production PV présente un pic de production aux alentours de midi, ce qui coïncide souvent avec une faible consommation électrique, notamment en été, appelant une adaptation du système électrique pour intégrer l'excédent de puissance. Utiliser le stockage des véhicules électriques (VE) semble être une solution prometteuse. En effet, pendant le pic de production PV, les VE sont susceptibles d'être branchés au réseau électrique sur les lieux de travail, et sont ainsi disponibles pour stocker. Contribuer à l'équilibre production/consommation au niveau national/régional ne doit cependant pas se faire au détriment du réseau électrique local, en créant des congestions ou des excursions de tension trop importantes. Il est donc nécessaire de contrôler cette flotte de VE, qui, à l'échelle nationale, représente un très grand nombre d'entités. Sur la base d'un taux de pénétration modéré des ENR (scénario "N2" fourni par le gestionnaire du réseau de transport français [1]), l'excédent de puissance pendant l'été à midi en France pourrait représenter l'équivalent de plusieurs millions de VE se chargeant à une puis-

sance domestique de 6 kVA. En faisant l'hypothèse que seuls quelques acteurs du marché tels que les agrégateurs soient en mesure de fournir des services de gestion de flexibilité, la taille des flottes de VE à contrôler serait extrêmement importante. Par conséquent, il est nécessaire de développer des méthodes extensibles capables de répondre aux incertitudes en temps réel (par exemple, les congestions réseau). À l'aune de ce défi, plusieurs travaux ont développé des méthodes décentralisées, parmi lesquelles l'apprentissage par renforcement (*reinforcement learning* RL) a gagné beaucoup d'attention au cours de la dernière décennie. Le RL permet aux agents apprenants (tels que les VE) d'apprendre les meilleures actions en fonction de leurs observations d'un environnement dynamique et incertain. Parmi les approches RL, les bandits manchots représentent un outil simple mais efficace lorsqu'une action jouée ne modifie pas les distributions de récompenses des actions. Sous cette hypothèse, les méthodes de bandits peuvent converger plus rapidement en démarrage à froid que d'autres méthodes RL basées sur les processus de décision markoviens (MDP), en plus de nécessiter peu de ressource de calcul. Enfin, les bandits peuvent fournir des garanties théoriques pour certains paramètres de problème qui pourraient renforcer la confiance des acteurs de l'énergie. Bien que largement étudiés et appliqués dans l'industrie (par exemple, systèmes de recommandation [2], télécommunications [3], mise à l'échelle automatique du Cloud [4]...), ce n'est que récemment que certains travaux ont commencé à émerger sur l'application des bandits aux problèmes de systèmes électriques tels que [5, 6, 7, 8, 9]. Certains travaux considèrent le point de vue du fournisseur d'énergie, par exemple pour la conception des prix de l'électricité, indépendamment des contraintes du réseau électrique [7, 8, 9]. D'autres travaux sont plus similaires aux nôtres, car ils considèrent le point de vue des consommateurs, ou groupes de consommateurs, face à un prix fixé par un autre acteur (fournisseur d'énergie et/ou agrégateur), et souhaitant minimiser leur facture énergétique [5, 9]. Barbato *et al.* [5] considèrent l'algorithme de bandit appelé EXP3 et proposent des variantes, tandis que [9] propose une variante de l'algorithme UCB [11]. Néanmoins, aucune de ces références ne considère la congestion du réseau. Une approche connexe est adoptée par Zhao *et al.* [6], qui considèrent des capacités limitées d'équipements électriques, dans un problème de répartition économique avec réponse à la demande gérée par l'algorithme UCB. Contrairement à la plupart de ces travaux, la méthode présentée dans cet article prévient la congestion par un mécanisme fournissant des garanties et s'inspire de travaux antérieurs dans [12], avec des différences clés dans l'implémentation des algorithmes

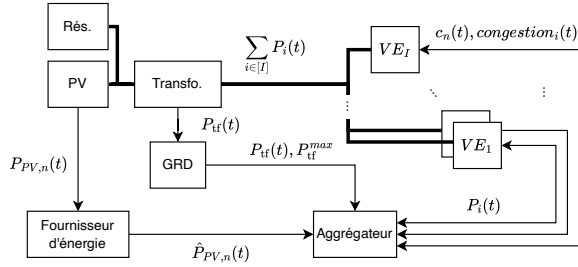


FIG. 1. Architecture du système. Les lignes en gras sont des connexions physiques tandis que les flèches sont des communications réseau. L'agrégateur reçoit tous les signaux pertinents du système puis envoie des signaux pour influencer le comportement des VE. Les flèches bidirectionnelles indiquent que les VE reçoivent les signaux de prix et de congestion de l'agrégateur, tandis que l'agrégateur reçoit la puissance de charge instantanée des VE.

de bandits et la formulation du réseau, avec trois nouveaux algorithmes utilisant l'échantillonnage de Thompson et des informations contextuelles hétérogènes via l'échantillonnage de Thompson linéaire [13]. Bien que ces méthodes permettent un apprentissage en ligne, multi-agents, évolutif, décentralisé et sans connaissance de l'environnement, ainsi qu'une garantie de prévention de la congestion avec le mécanisme susmentionné, il est important de noter que cela modifie le cadre original de l'échantillonnage de Thompson, car les actions peuvent avoir une influence significative sur les distributions de récompenses.

La suite de cet article est organisée comme suit : la section II détaille la méthodologie et les modèles utilisés, la section III présente l'étude de cas, tandis que la section IV décrit les résultats. Enfin, la section V conclut et présente les orientations de recherche futures.

2. MÉTHODOLOGIE ET MODÉLISATION

2.1. Formulation générale du problème

Le problème considéré dans cet article porte sur la recharge intelligente d'une flotte de véhicule électrique sur un lieu de travail, où les propriétaires de VE arrivent le matin et repartent le soir. Durant la journée, leur VE est connecté au réseau et peut se charger, avec pour objectif de minimiser leur facture d'électricité (i.e. de choisir les meilleurs moments auxquels se charger). Ils tentent également d'atteindre leur état de charge cible avant leur départ, tout en évitant de congestionner le réseau.

Le scénario inclut un fournisseur d'électricité, un gestionnaire de réseau de distribution (GRD), un agrégateur et des véhicules électriques. L'agrégateur envoie aux propriétaires de VE des signaux de congestion en temps réel qu'il reçoit du GRD. Il peut également envoyer les prix de l'électricité du fournisseur d'énergie. Chaque VE est un agent qui apprendra en ligne les meilleurs instants pour se charger en fonction du signal envoyé par l'agrégateur (prix de l'électricité et signaux de congestion). La Fig. 1 décrit l'architecture du système.

Le réseau électrique se compose de I véhicules électriques connectés à un transformateur tf . Un réseau simple (sans pertes et sans puissance réactive) est considéré, où le flux de puissance active au transformateur est égal à la somme des puissances actives prélevées par chaque véhicule électrique. Chaque véhicule électrique $i \in \{1, \dots, I\}$ possède une puissance nominale $P_i^{\max}$ et une capacité $E_i^{\max}$. Aucune injection véhicule-réseau (V2G) n'est considéré ici. À chaque instant discret t , nous désignons par $E_i(t)$ la quantité d'énergie dans la batterie de chaque VE i , par $P_i(t) \in [0, P_i^{\max}]$ sa puissance de charge instantanée, et son état de charge $SoC_i = \frac{E_i(t)}{E_i^{\max}}$. Avec une résolution temporelle Δ_t , le niveau d'énergie $E_i(t)$ évolue selon :

$$E_i(t + \Delta_t) = E_i(t) + P_i(t)\Delta_t \quad (1)$$

où l'état de charge ne doit pas dépasser $SoC_i^{\max}$.

Il est important de mentionner que pour éviter la congestion dans le réseau électrique, un signal de congestion spécifique est supposé être envoyé en temps réel par l'agrégateur à un sous-ensemble des VE en cours de charge, qui sont identifiés comme responsables de la congestion. Ce sous-ensemble est choisi aléatoirement selon une distribution uniforme, de sorte que l'annulation de leur charge réduise la demande globale en dessous du seuil de congestion : le nombre de VE contactés est donc le ratio de la demande excédentaire divisé par la puissance nominale de 7 kW commune à tous les VE. Le sous-ensemble peut changer chaque fois qu'une congestion est détectée ; les VE recevant le signal de congestion arrêtent instantanément de se charger, tandis que les autres peuvent continuer à se charger normalement.

Dans le cas où un VE i est empêché de charger en raison d'une congestion, il doit modifier son plan de charge initial, et peut donc re-sélectionner d'autres instants pendant la journée pour obtenir son état de charge souhaité au départ. À chaque instant, il calcule ses $m_i(t)$ instants de charge restants requis pour atteindre son état de charge souhaité au départ SoC_i^* comme :

$$m_i(t) = \left\lceil \frac{[SoC_i^* - SoC_i(t)]^+ E_i^{\max}}{P_i^{\max}} \right\rceil \quad (2)$$

avec $\lceil \cdot \rceil$ la partie entière, et $[\cdot]^+ = \max(0, \cdot)$ la partie positive.

Le prix de l'électricité est considéré comme variant à chaque instant de la journée (chacun séparé par une durée Δ_t) et fortement influencé par la production PV, résultant ainsi en des prix plus bas pendant la journée, généralement autour de midi. Un prix d'électricité normalisé $c_n(t)$ pour chaque instant t de chaque jour n s'exprime alors comme :

$$c_n(t) = 1 - \frac{P_{PV,n}(t)}{\bar{P}}$$

avec $\bar{P}$ une borne supérieure de la production PV (par exemple, la puissance totale installée), garantissant que $c_n(t) \in [0, 1]$ pour tout jour n et pour tout instant t dans la journée.

2.2. Formulation en bandits multi-bras

Les bandits multi-bras (MAB) s'inscrivent dans le cadre théorique plus général de l'apprentissage par renforcement [11]. Le problème MAB consiste à trouver le meilleur ensemble de bras (actions) parmi K bras disponibles (c'est-à-dire, ici, les instants de charge). Chaque bras individuel $k \in [1, \dots, K] = [K]$ est associé à une récompense r_k qui suit une distribution inconnue $\mathcal{D}_k$. Un scalaire binaire $k_{t_i} \in \{0, 1\}$ est utilisé pour désigner un bras de base joué à l'instant $t \in [T]$ par le VE i (1 indiquant que le VE se charge, tandis que 0 indique qu'il ne se charge pas). Pour plus de clarté, l'indice i est omis dans le reste du document et nous nous référons par exemple à k_t comme le bras de base joué à l'instant $t \in [T]$ par un VE donné. Chaque VE inclura un MAB (pour les MAB combinatoires) ou plusieurs (pour les MAB indépendants). La séquence de bras de base pendant une itération (ici, un jour) $n \in [N]$ est un vecteur $\mathbf{s}_n = [k_{1,n}, k_{2,n}, \dots, k_{T,n}]$ tel que $\mathbf{s}_n \in \{0, 1\}^T$, et représente la séquence des instants de charge planifiés, appelée super-bras.

Les bandits multi-bras répondent au problème du dilemme entre exploration et exploitation : à chaque itération, on peut en effet choisir d'explorer de nouvelles solutions (potentiellement meilleures) ou d'exploiter des solutions déjà identifiées comme bonnes. L'objectif est de gérer ce compromis afin de minimiser le pseudo-regret $R(N)$ après N itérations, c'est-à-dire la différence entre la séquence optimale d'actions et les actions effectivement sélectionnées, qui peut s'exprimer comme :

$$R(N) = \sum_{n=1}^N (\mu_{s^*} - \mu_{\mathbf{s}_n}) \quad (3)$$

avec $\mathbf{s}^* = \arg \max_{\mathbf{s} \in \{0,1\}^T, \|\mathbf{s}\|_1 = m(t)} \mu_{\mathbf{s}}$ le super-bras optimal pour tout jour, $\mu_{\mathbf{s}^*}$ sa récompense moyenne associée, $\mu_{\mathbf{s}_n}$ la récompense espérée du super-bras effectivement sélectionné pour le jour n , et $m(t)$ les instants de charge requis (Éq. (2)). Le super-bras optimal fait référence à la séquence d'actions qui est optimale (en espérance) et sera appliquée chaque jour par un VE donné.

L'idée centrale de la formulation s'inspire de [12]. Elle consiste à discrétiser le temps durant une journée en T instants, parmi lesquels les bandits multi-bras doivent apprendre à sélectionner à chaque instant t les meilleurs $m(t)$ instants pour se charger. Chaque VE inclurait donc un bandit multi-bras (dans le cadre combinatoire) ou plusieurs (dans le cadre indépendant). Les MAB cherchent à apprendre la meilleure politique de charge (c'est-à-dire, sélectionner les meilleurs instants pour se charger) afin de minimiser le coût de charge du VE, tout en satisfaisant l'état de charge souhaité au départ. Puisque les congestions peuvent empêcher la charge, les VE voudront également éviter de provoquer des congestions sur le réseau électrique. Le cadre théorique de l'apprentissage par renforcement implique que le ou les agents apprenants reçoivent une récompense de l'environnement qui dépend en partie de leurs actions. La récompense instantanée r_t dépend du prix de l'électricité $c_n(t)$ à chaque instant t et jour n et de la réception d'un signal de congestion. Cette récompense est calculée par chaque VE qui se charge à un instant donné t et s'exprime comme :

$$\begin{cases} r_t = 1 - c_n(t) & \text{si } \text{congestion}(t) = \text{false} \\ r_t = 0 & \text{sinon} \end{cases} \quad (4)$$

Cet article présente les performances de différents types de bandits qui sont listés ci-dessous et qui seront décrits dans les sections suivantes :

- MAB indépendants (Algorithme 1)
- MAB contextuels indépendants (Algorithme 2)
- MAB combinatoires (Algorithme 3)
- MAB contextuels combinatoires (Algorithme 3)

Ces algorithmes ont été comparés à une référence simple et à un MILP omniscient qui sont décrits dans la section suivante.

2.2.1. Référence

L'algorithme de référence de base consiste à charger dès l'arrivée du VE au temps d'arrivée t_i^a avec un état de charge SoC_i^a . Dans le cas où le VE reçoit un signal de congestion ($\text{congestion} = \text{true}$), le VE arrête de se charger et tente de se charger à nouveau à l'instant suivant jusqu'à atteindre le temps de départ t_i^b ou l'état de charge souhaité SoC_i^* .

2.2.2. Programmation Linéaire en Nombres Entiers Mixtes (MILP)

Dans cet algorithme, l'agrégateur omniscient contrôle tous les VE pour minimiser les coûts totaux de charge sous contrainte physique et réseau, avec une connaissance complète du système et des variables futures. Il est utilisé comme borne supérieure théorique, *ie*, la solution optimale a posteriori.

2.2.3. MAB indépendants

Dans les MAB indépendants, la distribution des récompenses de chaque bras $\mathcal{N}(\mu_k, \nu_k)$ est supposée être indépendante des autres. Cela conduit à utiliser un échantillonnage de Thompson pour chaque instant $t \in [T]$ de la journée. La récompense moyenne $\mu_{\mathbf{s}}$ du super-bras $\mathbf{s}$ est la somme de toutes les récompenses moyennes des bras de base sélectionnés : $\mu_{\mathbf{s}} = \sum_{t=1}^T \mathbf{s}(t) \mu_t$, donc $\mu_{\mathbf{s}} = \mathbf{s}^T \boldsymbol{\mu}$ avec $\boldsymbol{\mu} = [\mu_1, \dots, \mu_T]$.

Le choix de l'échantillonnage de Thompson gaussien est approprié pour une récompense continue (Éq. 4) [13]. Ainsi, la récompense de chaque bras de base r_t est donnée par une distribution gaussienne $\mathcal{N}(\mu_t, \sigma_t^2)$ et la distribution a priori pour

le paramètre inconnu θ_t est également gaussien avec une valeur moyenne estimée $\hat{\theta}_t$ et une variance σ_t^2 (voir Algorithme 1).

Il est important de noter qu'à chaque pas de temps, une re-sélection intra-journalière (ligne 11) du super-bras peut être effectuée parmi un ensemble autorisé de super-bras $\mathcal{S}(t)$ (ligne 10), qui exclut les instants passés $\mathbf{s}_o$ (ligne 11) et inclut uniquement les super-bras avec $m(t)$ instants de charge (ligne 9).

Algorithm 1 MAB Indépendant

Entrée : $\forall t \in [T], \hat{\theta}_t := 0, \sigma_t^2 := 1, \forall k \in \{0, 1\}, N_{k_t}(0) := 0$

- 1: **Pour** $n = 1, \dots, N$ **faire**
- 2: $E_i(0) := \text{SoC}_i^a \times E_i^{\max}$
- 3: $\mathbf{s}_o := \emptyset$
- 4: **Pour** $t = 1, \dots, T$ **faire**
- 5: Échantillonner $\hat{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \sigma_t^2)$
- 6: $\tilde{\boldsymbol{\mu}}[t] := \hat{\theta}_t$
- 7: **Fin Pour**
- 8: **Pour** $t = 1, \dots, T$ **faire**
- 9: Calculer le nombre d'instants de charge $m(t)$ pour atteindre SoC_i^* selon l'Éq. (2)
- 10: Définir l'ensemble $\mathcal{S}(t) := \{0, 1\}^{T-t+1}$
- 11: Choisir $\mathbf{s}_t := \arg \max_{\mathbf{s} \in \mathcal{S}(t), \|\mathbf{s}\|_1 = m(t)} [\mathbf{s}_o, \mathbf{s}]^T \tilde{\boldsymbol{\mu}}$
- 12: $\mathbf{s}_o := [\mathbf{s}_o, k_t]$
- 13: **Si** $k_t = 1$ **Alors**
- 14: Appliquer la gestion de congestion
- 15: Observer la récompense r_t selon l'Éq. (4)
- 16: $N_{k_t}(t) := N_{k_t}(t-1) + 1$
- 17: $\hat{\theta}_t := \frac{r_t + N_{k_t}(t-1)\hat{\theta}_{t-1}}{N_{k_t}(t)}$
- 18: $\sigma_t^2 := \frac{1}{N_{k_t}(t)}$
- 19: **Fin Si**
- 20: **Fin Pour**
- 21: **Fin Pour**

2.2.4. MAB contextuels indépendants

Avec l'hypothèse d'indépendance précédente, l'information contextuelle pendant le jour n pour chaque bras de base k_t avec le vecteur de contexte $\mathbf{x}_{n,t}$ de dimension d est ajoutée (par exemple, le prix prédit de cet instant ou le jour de la semaine). L'hypothèse linéaire donne :

$$\forall n \in [N], \forall t \in [T], \mu_t = \mathbf{x}_{n,t} \boldsymbol{\theta}_t. \quad (5)$$

En utilisant l'échantillonnage de Thompson linéaire [13], la récompense de chaque bras de base r_t est supposée être donnée par une distribution gaussienne $\mathcal{N}(\mu_t, \sigma^2)$ et la distribution a priori pour le paramètre inconnu $\boldsymbol{\theta}_t$ est également gaussien $\mathcal{N}(\hat{\boldsymbol{\theta}}_t, \sigma^2 A_t^{-1})$. Le MAB contextuel indépendant est décrit par l'Algorithme 2.

2.2.5. MAB combinatoires

Ici, l'hypothèse d'indépendance est abandonnée et remplacée par l'hypothèse que la récompense de chaque bras de base est linéairement corrélée avec les autres (deux instants très proches pourraient être fortement corrélés du fait de la production PV). Exploiter cette dépendance peut conduire à de meilleures performances par rapport aux MAB indépendants (qu'ils soient sans contexte ou contextuels). Avec $\mathbf{s} \in \mathcal{S}$ le vecteur combinatoire de dimension T et le paramètre inconnu $\boldsymbol{\theta}$ de dimension T :

$$\forall \mathbf{s} \in \mathcal{S}, \mu_{\mathbf{s}} = \mathbf{s}^T \boldsymbol{\theta}. \quad (6)$$

Algorithm 2 MAB Contextuel Indépendant

Entrée : $\forall t \in [T], A_t := I_{d,d}, \mathbf{b}_t := 0_d, \hat{\boldsymbol{\theta}}_t := 0_d, \sigma > 0$

- 1: **Pour** $n = 1, \dots, N$ **faire**
- 2: $\mathbf{s}_o := \emptyset$
- 3: $E_i(0) = \text{SoC}_i^a \times E_i^{\max}$
- 4: **Pour** $t = 1, \dots, T$ **faire**
- 5: Observer le contexte $\mathbf{x}_{n,t}$
- 6: Échantillonner $\boldsymbol{\theta}_t \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_t, \sigma^2 A_t^{-1})$
- 7: Calculer $\tilde{\boldsymbol{\mu}}_t := \mathbf{x}_{n,t}^\top \boldsymbol{\theta}_t$
- 8: $\tilde{\boldsymbol{\mu}}[t] := \tilde{\boldsymbol{\mu}}_t$
- 9: **Fin Pour**
- 10: **Pour** $t = 1, \dots, T$ **faire**
- 11: Calculer le nombre d'instants de charge $m(t)$ pour atteindre SoC_i^* selon l'Éq. (2)
- 12: Définir l'ensemble $\mathcal{S}(t) := \{0, 1\}^{T-t+1}$
- 13: Choisir $\mathbf{s}_t := \arg \max_{\mathbf{s} \in \mathcal{S}(t), \|\mathbf{s}\|_1 = m(t)} [\mathbf{s}_o, \mathbf{s}]^\top \tilde{\boldsymbol{\mu}}$
- 14: $\mathbf{s}_o := [\mathbf{s}_o, k_t]$
- 15: **Si** $k_t = 1$ **Alors**
- 16: Appliquer la gestion de congestion
- 17: Observer la récompense r_t selon l'Éq. (4)
- 18: *Mettre à jour la distribution des bras de base*
 $\mathcal{N}(\hat{\boldsymbol{\theta}}_t, \sigma^2 A_t^{-1})$:
- 19: $A_t := A_t + \mathbf{x}_{n,t}^\top \mathbf{x}_{n,t}$
- 20: $\mathbf{b}_t := \mathbf{b}_t + r_t$
- 21: $\hat{\boldsymbol{\theta}}_t := \mathbf{b}_t^\top A_t^{-1}$
- 22: **Fin Si**
- 23: **Fin Pour**
- 24: **Fin Pour**

En utilisant l'échantillonnage de Thompson linéaire [13], la récompense de chaque super-bras r_s est supposée être donnée par une distribution gaussienne $\mathcal{N}(\mathbf{s}^\top \boldsymbol{\theta}, \sigma^2)$ et la distribution a priori pour le paramètre inconnu $\boldsymbol{\theta}$ est également gaussien $\mathcal{N}(\hat{\boldsymbol{\theta}}, \sigma^2 A^{-1})$. L'Algorithme 3 présente le pseudocode pour le MAB combinatoire à échantillonnage de Thompson (sans contexte et contextuel). Le cas où $d = 0$ et donc $\mathbf{x}_n = \emptyset$ correspond au MAB combinatoire à échantillonnage de Thompson (c'est-à-dire sans contexte).

2.2.6. MAB contextuels combinatoires

L'information contextuelle est ajoutée sous la forme d'un vecteur de contexte $\mathbf{x}_n$ de dimension d concaténé avec le vecteur super-bras $\mathbf{s}$, comme décrit ci-dessous. Le MAB contextuel combinatoire est décrit par l'Algorithme 3.

$$\forall n \in [N], \forall \mathbf{s} \in \mathcal{S}, \mu_s = [\mathbf{s}, \mathbf{x}_n]^\top \boldsymbol{\theta} \quad (7)$$

3. ÉTUDE DE CAS

La simulation est exécutée sur $N = 400$ jours, chaque VE sélectionnant son ensemble de bras/super-bras pour le jour donné. Une re-sélection intra-journalière est appliquée en cas de congestion. La résolution temporelle est de $\Delta_t = 15$ minutes, avec $T = 96$ instants par jour. Les jours de télétravail sont les jours [2, 4] dans les jours de la semaine [1, 2, ..., 7]. Aucun jour de week-end n'est considéré (*ie*, les VE sont connectés au réseau le week-end). Les simulations sont exécutées sur 5 tirages par algorithme et leur moyenne est présentée dans les résultats, ainsi que l'intervalle de confiance.

Le nombre d'agents véhicules électriques est de $I = 55$. Dans un premier temps, tous les agents véhicules électriques sont homogènes en termes d'heure d'arrivée (8h), heure de dé-

Algorithm 3 MAB Contextuel Combinatoire

Entrée : $A := I_{T+d, T+d}, \mathbf{b} := 0_{T+d}, \hat{\boldsymbol{\theta}} := 0_{T+d}, \sigma > 0$

- 1: **Pour** $n = 1, \dots, N$ **faire**
- 2: $r_n := 0$
- 3: $\mathbf{s}_n := 0_T$
- 4: Observer le contexte $\mathbf{x}_n$
- 5: Échantillonner $\boldsymbol{\theta} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \sigma^2 A^{-1})$
- 6: **Pour** $t = 1, \dots, T$ **faire**
- 7: $\mathbf{s}_o := \emptyset$
- 8: Calculer le nombre d'instants de charge $m(t)$ pour atteindre SoC_i^* selon l'Éq. (2)
- 9: Définir l'ensemble $\mathcal{S}(t) := \{0, 1\}^{T-t+1}$
- 10: Choisir $\mathbf{s}_t := \arg \max_{\mathbf{s} \in \mathcal{S}(t), \|\mathbf{s}\|_1 = m(t)} [\mathbf{s}_o, \mathbf{s}, \mathbf{x}_n]^\top \tilde{\boldsymbol{\theta}}$
- 11: $\mathbf{s}_o := [\mathbf{s}_o, k_t]$
- 12: **Si** $k_t = 1$ **Alors**
- 13: Appliquer la gestion de congestion
- 14: Observer la récompense r_t selon l'Éq. (4)
- 15: $r_n = r_n + r_t$
- 16: **Fin Si**
- 17: **Fin Pour**
- 18: *Mettre à jour la distribution des bras de base et super-bras*
 $\mathcal{N}(\hat{\boldsymbol{\theta}}, \sigma^2 A^{-1})$:
- 19: $A := A + [\mathbf{s}_o, \mathbf{x}_n]^\top [\mathbf{s}_o, \mathbf{x}_n]$
- 20: $\mathbf{b} := \mathbf{b} + r_n$
- 21: $\hat{\boldsymbol{\theta}} := \mathbf{b}^\top A^{-1}$
- 22: **Fin Pour**

part (18h), état de charge à l'arrivée (0,7) et état de charge souhaité au départ ($\geq 0,8$). Ce cas d'étude est conçu pour permettre une analyse simple, sans affecter la méthodologie. Tous les VE se chargent à pleine puissance (aucune modulation de puissance n'est considérée ici). Les véhicules électriques partagent également la même puissance nominale ($P^{\max} = 7$ kW) et capacité énergétique ($E^{\max} = 52$ kWh). Quatre-vingt pour cent des VE sont en télétravail les jours de télétravail [2,4]. Ces VE sont supposés se charger à domicile et leur puissance consommée n'est pas considérée dans cette étude qui se concentre sur la charge intelligente sur le lieu de travail.

Tous les VE sont connectés au transformateur dimensionné à un tiers de la puissance nominale de chaque VE multipliée par le nombre de VE, soit $7 \text{ kW} \times 55/3 = 128 \text{ kW}$. Aucune autre charge n'est présente dans le réseau par souci de simplicité dans cette analyse préliminaire. Ainsi, pendant les jours de télétravail, lorsque seulement 20% des VE sont présents sur les lieux de travail, aucune congestion ne peut se produire même s'ils se chargent tous simultanément. Ils peuvent donc exploiter pleinement le réseau pendant les périodes de bas prix durant ces jours.

En ce qui concerne la production PV, une série temporelle d'irradiance solaire de 60 jours collectée par le NREL à Hawaï [14] est utilisée. Cette série temporelle est répétée afin de produire une série temporelle de 400 jours.

Deux types de vecteurs de contexte ont été considérés : a) le vecteur de contexte inclut uniquement les prix de l'électricité journalier du jour n , b) le vecteur de contexte inclut à la fois les prix de l'électricité journalier du jour n et une variable booléenne pour indiquer si le jour est un jour de télétravail ou non. Les dimensions des vecteurs de contexte pour l'Algorithme (2) sont $d = 1$ (prix seul) ou $d = 2$ (prix et type de jour). Pour l'Algorithme (3), les dimensions du vecteur de contexte sont soit $d = T$ (prix seul) ou $d = T + 1$ (prix et type de jour).

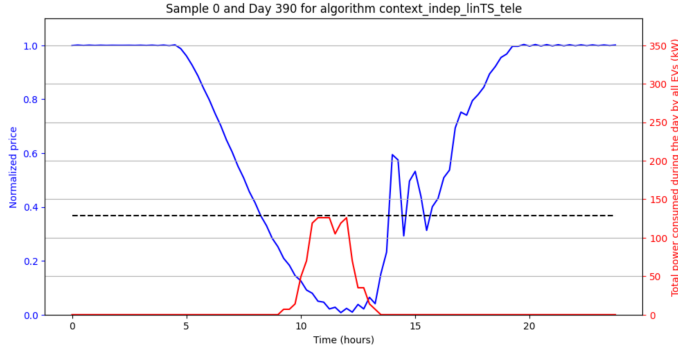


FIG. 2. Profil de prix (bleu) et puissance totale consommée par tous les VE (rouge) pour le Jour 390 de l'entraînement (Échantillon 0). La ligne pointillée noire représente la limite de congestion.

4. RÉSULTATS ET DISCUSSION

Par conception, la méthode empêche les congestions durables. Lorsqu'une congestion est détectée, un signal est envoyé à un sous-ensemble de VE en charge pour qu'ils arrêtent immédiatement de se charger, éliminant ainsi la congestion pendant les simulations. Tous les VE (100%) atteignent leur état de charge cible (80%) avant le départ comme requis. La Fig. 2 présente les profils de prix (bleu) et la consommation totale de puissance des VE (rouge) pour le Jour 390, montrant que la charge des VE se concentre autour de midi lorsque la production PV atteint son pic et que les prix de l'électricité baissent. La Fig. 3 illustre la récompense moyenne cumulée à travers tous les 55 VE pour chaque algorithme, avec des résultats moyens sur 5 échantillons pour les algorithmes basés sur les bandits. Les algorithmes sont étiquetés comme « baseline », « combi_linTS » (bandits combinatoires), « context_combi_linTS » (bandits combinatoires contextuels avec contexte de prix), « context_combi_linTS_tele » (incluant les contextes de type de jour), « context_indep_linTS » (bandits indépendants contextuels avec contexte de prix), « context_indep_linTS_tele » (incluant les contextes de type de jour), et « TS_indep » (bandits indépendants). Les performances de la référence évoluent au fil du temps principalement en raison des motifs saisonniers dans les données solaires sur la période de 60 jours sélectionnée, où la production solaire augmente progressivement depuis des valeurs initialement faibles jusqu'à des sorties significativement plus élevées.

Seuls deux algorithmes performant en dessous de la référence : les bandits indépendants sans contexte et le bandit combinatoire. La sous-performance des bandits indépendants, l'outil le plus simple dans la boîte à outils des bandits, était attendue. Cependant, la faible performance des bandits combinatoires est surprenante, car les bandits combinatoires basés sur l'échantillonnage de Thompson démontrent typiquement de hautes performances lorsque les récompenses correspondent directement aux bras sélectionnés. Dans ce cas, la gestion de congestion et la re-sélection intra-journalière font que les récompenses sont associées à des bras différents de ceux initialement sélectionnés, créant un décalage avec le cadre théorique. Cela conduit à une perte de performance significative par rapport à la référence. Il convient de noter que l'étude de cas actuelle favorise particulièrement la stratégie de référence (charger dès que possible) : les temps d'arrivée coïncident avec l'augmentation de la production PV (baisse des prix de l'électricité), tandis que les temps de départ s'alignent avec la diminution de la production (augmentation des prix). Dans des scénarios de charge domestique, où charger immédiatement (en début de soirée) coïnciderait avec des prix d'électricité plus élevés, la référence performerait probablement considérablement moins bien. Le meilleur algorithme performe 14% en dessous de la politique optimale, déterministe et centralisée (MILP) - un compromis acceptable

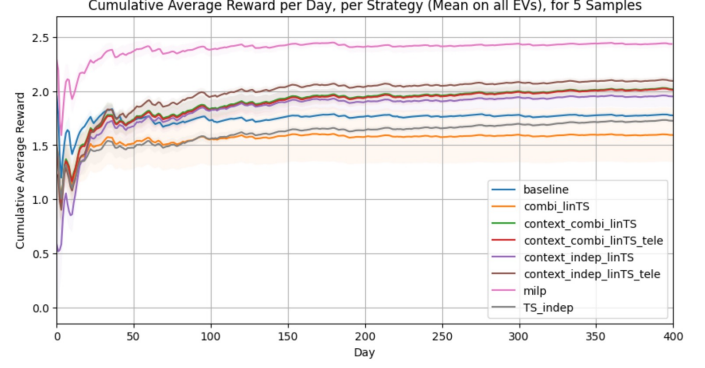


FIG. 3. Récompense moyenne cumulée sur tous les VE (pour 5 échantillons et un intervalle de confiance de 95%)

considérant que notre algorithme est décentralisé (améliorant l'extensibilité et préservant la vie privée) et gère les incertitudes que l'approche MILP n'aborde pas ici.

L'ajout de contexte(s) s'avère significativement pertinent. Alors que tous les algorithmes sans contexte (indépendants ou combinatoires) performant en dessous de la référence pendant la période de simulation, l'ajout de contexte apporte des améliorations substantielles. Les bandits indépendants contextuels montrent des améliorations allant de 9% (avec seulement le contexte de prix) à 17% (avec les contextes de prix et de type de jour) par rapport à la référence. De même, les bandits combinatoires s'améliorent de 26% lorsque le contexte est introduit, comparés à leurs homologues sans contexte. Cela se traduit par une amélioration de 14% par rapport à la référence, atténuant les effets négatifs de la modification du cadre théorique de l'échantillonnage de Thompson.

Au-delà de la valeur asymptotique des algorithmes, la vitesse de convergence doit également être considérée. Il est important de rappeler que dans ce problème, un démarrage à froid est considéré (c'est-à-dire que les bandits n'ont aucune information sur l'environnement, ce qui préserve la vie privée). Bien que cela puisse augmenter le temps de convergence, la Fig. 3 montre que tous les algorithmes tendent à atteindre au moins 95% de leur valeur asymptotique après environ 150 jours (voir Table 1). Les valeurs atteintes par les bandits contextuels (dépassant la référence, comme mentionné précédemment) atteignent déjà au moins 85% de leur valeur asymptotique après seulement 50 jours d'entraînement, tandis qu'au moins 77,8% de leur valeur asymptotique est atteinte après seulement 25 jours d'entraînement. Cette convergence rapide tend à diminuer le coût de l'entraînement, c'est-à-dire le coût associé à la performance dégradée pendant les premières phases d'entraînement en raison de l'exploration nécessaire de l'algorithme à la recherche de meilleures solutions. Ce coût s'applique bien sûr si l'entraînement est effectué dans la vie réelle, par opposition à l'entraînement effectué sur un simulateur. Comparé à l'entraînement en vie réelle, l'entraînement sur simulateur peut nécessiter des ressources informatiques plus importantes tout en introduisant plus d'incertitudes.

Ces résultats doivent être nuancés au regard de la nature relativement homogène des agents. Bien que cette étude montre des performances sur un ensemble d'agents bien conçu (approvisionnement énergétique suffisant) avec des préférences déterministes et une hétérogénéité avec les jours de télétravail, des études supplémentaires permettront d'éclaircir le comportement du système avec des préférences stochastiques et une quantité d'énergie moindre, apportant une plus grande hétérogénéité et donc des préoccupations sur l'équité de ce système.

TABLEAU 1. Récompense moyenne cumulée (% de la valeur au Jour 400)

	Nombre de jours		
	25	50	150
Baseline	95,9	96,9	99,3
MILP	92,2	97,1	99,6
MAB Combinatoire	93,0	95,0	99,6
MAB Comb. Context. (prix)	81,2	85,8	95,9
MAB Comb. Context. (prix, jour)	80,6	85,6	95,6
MAB Indep. Context. (prix)	79,9	85,7	97,4
MAB Indep. Context (prix, jour)	77,8	86,6	97,4
MAB Indep.	83,7	85,6	94,5

5. CONCLUSIONS

Cet article a permis d'évaluer 6 algorithmes basés sur la méthodes des bandits manchots, pour la charge intelligente décentralisée de VE avec des contraintes de mobilité et de réseau. Les algorithmes contextuels performant mieux que la références jusqu'à 17%, avec une convergence rapide atteignant 77,8% de la valeur asymptotique en 25 jours. L'écart minimum avec l'optimum théorique d'un agent omniscient et centralisé est de 14%. Cette approche décentralisée permet un déploiement pratique en distribuant le calcul à travers les agents VE, l'algorithme le plus intensif en calcul ne nécessitant que 200 μ s par VE par pas de temps. Le système maintient la confidentialité en ne partageant que des informations minimales - puissance instantanée et nombre d'agents - pour la gestion de congestion. Les travaux futurs aborderont une version entièrement décentralisée sans partage d'informations privées à l'agrégateur, et exploreront différentes approches d'agrégateur pour maximiser la consommation d'énergie renouvelable, tout en étudiant une plus grande hétérogénéité dans les préférences des VE pour mieux refléter les scénarios de charge du monde réel.

6. RÉFÉRENCES

- [1] RTE, "Energy Pathways to 2050 - Key results", Technical report, 2021.
- [2] L. Li, W. Chu, J. Langford, R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation", in Proc. 19th international conference on World wide web, 2010.
- [3] H. Dakdouk, R. Féraud, N. Varsier, P. Maillé, R. Laroche, "Massive multi-player multi-armed bandits for IoT networks : An application on LoRa networks", Ad Hoc Networks, 2023.
- [4] D. Delande, P. Stolf, R. Féraud, J.M. Pierson, A. Bottaro, "Horizontal scaling in cloud using contextual bandits" European Conference on Parallel Processing, 2021.
- [5] A. Barbato, L. Chen, F. Martignon, S. Paris, "A Multi-Armed Bandit Formulation for Distributed Appliances Scheduling in Smart Grids", in Proc. IEEE GreenCom conference, Taipei, Taiwan, 2014.
- [6] Z. Zhao, A. L. Liu, Y. Chen, "Electricity Demand Response under Real-Time Pricing : A Multi-armed Bandit Game", in Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, USA, 2018.
- [7] M. Brégère, "Stochastic bandit algorithms for demand side management", PhD thesis, Université Paris-Saclay, France, 2020.
- [8] S. Jain, S. Gujar, "A Multiarmed Bandit Based Incentive Mechanism for a Subset Selection of Customers for Demand Response in Smart Grids", in Proc. of the AAAI Conf. on Artificial Intelligence, 2020.
- [9] S. Chandekar, S. Jain, S. Gujar, "A Novel Demand Response Model and Method for Peak Reduction in Smart Grids", in Proc. IJCAI conference, Macao, China, 2023.
- [10] E. Spyrou, V. Kappatos, A. Tromaras, A. Anagnostopoulou, "A Multi-armed Bandit Approach for Electricity Reward of Smart Homes in a Smart Power Grid", 2024.
- [11] R. Sutton, A. Barto, Reinforcement Learning : An Introduction, 2nd ed., MIT Press, 2018.

- [12] S. Zafar, R. Féraud, A. Blavette, G. Camilleri, H. Ben Ahmed, "Decentralized Smart Charging of Large-Scale EVs using Adaptive Multi-Agent Multi-Armed Bandits", in Proc. CIRED conference, Rome, Italy, 2023.
- [13] S. Agrawal, N. Goyal, "Thompson Sampling for Contextual Bandits with Linear Payoffs," in Proc. PMLR conference, 2013, revised in Feb. 2014 (arXiv website)
- [14] NREL, Oahu Solar Measurement Grid, dataset, accessed 07/01/2025.