Stage de recherche de Master 2

Optimisation et apprentissage en ligne appliqués aux problèmes logistiques du marché de Rungis.

Mots clés Apprentissage automatique en ligne, optimisation, apprentissage par renforcement, bandits.

Keywords Online machine learning, optimization, reinforcement learning, bandits.

Contexte

Le marché de Rungis est le plus grand marché de produits frais du monde, et la gestion logistique d'un tel volume de denrées est un véritable défi (il faut imaginer un chiffre d'affaires de l'ordre de 10 milliards d'euros par an). La chaîne logistique de Rungis est prise en charge par l'entreprise Califrais,

L'entreprise Califrais gère la logistique des commandes en ligne du marché, et dispose ainsi de données massives qu'elle peut exploiter. Son objectif est d'améliorer la supply chain alimentaire de Rungis par le biais de techniques issues de méthodes mathématiques à l'intersection de l'optimisation, des statistiques et de l'apprentissage automatique.

Califrais a des liens forts avec le monde de la recherche académique et collabore notamment avec le Laboratoire de Probabilités, Statistique et Modélisation (LPSM) au sein de Sorbonne Université et l'Université Paris Cité. Depuis quelques années, cette collaboration fructueuse a donné le jour à deux thèses ainsi que diverses publications dans des conférences de machine learning telles que ICML ou NeurIPS.

C'est dans le cadre de cette collaboration entre Califrais et le LPSM que nous proposons ce stage de recherche académique. Son objectif, que nous détaillons ci-dessous, est d'étudier les propriétés théoriques et empiriques des méthodes d'apprentissage séquentiel (ou d'apprentissage par renforcement) pour résoudre des problèmes logistiques tels que la gestion de stock. Son financement est garanti par l'appel à projet "Logistique 4.0" de l'Agence de la transition écologique (ADEME) remporté par Califrais. Ce stage d'un semestre a pour vocation d'aboutir en une thèse qui débuterait en Octobre 2025.

Sujet

Dans le cadre de ce stage, nous nous intéressons à la résolution de ce qu'on appelle un *problème d'inventaire*. On peut le voir comme un problème d'optimisation dynamique dans lequel un gestionnaire prend chaque jour des décisions de réapprovisionnement afin de minimiser des coûts (économiques ou écologiques) tout en essayant de satisfaire au mieux la demande du marché.

Il existe une vaste littérature sur le sujet, essentiellement issue du domaine de la recherche opérationnelle. Néanmoins, la plupart des modèles classiques supposent que tous les paramètres du problème sont connus. Par exemple, ces modèles supposent que la demande future est connue à l'avance, ou au moins qu'elle est une variable aléatoire tirée d'une distribution connue. Dans ce cas, le problème est bien posé et peut être résolu par des techniques d'optimisation déterministe ou stochastique.

Nous nous intéressons au cas contraire (et plus réaliste !) où l'on ne connait pas en avance les paramètres du problème, comme par exemple la demande qu'il y aura pour les produits dans les jours/mois à venir. Dans ce cas, le gestionnaire doit mettre en œuvre des techniques mêlant *apprentissage et optimisation*, exploitant les données du passé pour apprendre au fur et à mesure les paramètres du problème. En résumé, on s'intéresse à des problèmes *séquentiels* (système dynamique discret en temps), *non-stationnaires* (la demande n'est pas une variable aléatoire suivant une simple loi), *discrets* (la variable de décision est en général un entier) et à *large échelle* (on rappelle la taille massive du marché de Rungis).

On se tournera donc vers les domaines de l'apprentissage séquentiel (Online Learning [7]) et de l'apprentissage par renforcement (Reinforcement Learning [8]), qui fournissent un cadre théorique et des algorithmes permettant de résoudre notre problmème. Par exemple, la littérature classique sur les bandits [3, 1] propose des algorithmes simples et adaptés aux problèmes séquentiels, discrets et non-stationnaires (dits adversariaux) avec des garanties théoriques. C'est le cas, par exemple de Exp3 (Exponential-weight algorithm for Exploration and Exploitation). Néanmoins, on observe que cette littérature n'inclut pas la

notion d'état qui est centrale dans la plupart des problèmes d'inventaires. D'autre part, l'apprentissage par renforcement inclut bien cette notion, mais généralement les algorithmes proposés passent difficilement à l'échelle, et leur analyse théorique dépend généralement d'hypothèses de stationnarité dont on ne veut pas.

Objectifs et déroulé du stage

Ce stage a des objectifs à la fois théoriques et méthodologiques, qui auront pour but de développer les compétences théoriques de l'étudiant·e sur l'analyse de problèmes d'optimisation en ligne et d'apprentissage, et des compétences techniques en programmation.

En ce qui concerne l'analyse théorique, l'étudiant·e commencera par étudier la littérature sur les bandits adversariaux et les algorithmes tels que Exp3 : quelles performances théoriques peut-on espérer et sous quelles hypothèses. Dans un deuxième temps, on cherchera à se rapprocher du cadre qui nous intéresse en essayant de modifier cette analyse classique en incorporant une notion de variable d'état. Pour cela, on s'inspirera de techniques prometteuses et plus récentes, à la croisée de l'apprentissage séquentiel et du contrôle optimal, qui ont été récemment proposées [2, 5, 4]. Enfin, une dernière piste plus exploratoire consistera à combiner des techniques issues de la littérature sur les bandits et le reinforcment learning, à l'instar de ce qui est proposé [6]. Concernant le volet méthodologique, l'étudiant·e implémentera différentes méthodes en comparant leurs performances sur des données *réelles*.

Pour les deux volets, on commencera par consiuderer des problèmes d'inventaires très simplifiés (par exemple sans notion d'état) puis progressivement nous rajouterons de la complexité en nous rapprochant de modèles plus réalistes.

Conditions

Ce stage est réalisé dans le cadre d'une collaboration entre Califrais et le LPSM.

- Lieux du stage : en partie dans les locaux de Califrais et en partie au LPSM. Les deux se situent à Paris et sont accessibles en transports en commun.
- Durée : 5 à 6 mois entre février et septembre 2025.
- Profil requis : M2 recherche en mathématiques ou troisième année d'école d'ingénieurs. Compétences en programmation (Python) requises.

Le stage sera supervisé par

- Massil HIHAT (Califrais) : massil.hihat@califrais.fr
- Adeline FERMANIAN (Califrais) :adeline.fermanian@califrais.fr
- Guillaume GARRIGOS (LPSM, Université Paris Cité) : garrigos@lpsm.paris

Candidature

Pour candidater, envoyer

- un CV,
- une lettre de motivation,
- les relevés de notes de Master,

par mail à Massil Hihat, Adeline Fermanian et Guillaume Garrigos.

Références

[1] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1) :1–122, 2012.

[2] Jihun Kim and Javad Lavaei. Online bandit nonlinear control with dynamic batch length and adaptive learning rate. *arXiv preprint arXiv :2410.03230*, 2024.

[3] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[4] Yingying Li, James A Preiss, Na Li, Yiheng Lin, Adam Wierman, and Jeff S Shamma. Online switching control with stability and regret guarantees. In *Learning for Dynamics and Control Conference*, pages 1138–1151. PMLR, 2023.

[5] Yiheng Lin, James A Preiss, Emile Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online adaptive policy selection in time-varying systems : No-regret via contractive perturbations. *Advances in Neural Information Processing Systems*, 36, 2024.

[6] Bianca Marin Moreno, Margaux Brégère, Pierre Gaillard, and Nadia Oudjane. Metacurl : Non-stationary concave utility reinforcement learning. *arXiv preprint arXiv :2405.19807*, 2024.

[7] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv :1912.13213*, 2019.

[8] Richard S Sutton and Andrew G Barto. *Reinforcement learning : An introduction*. MIT press, 2018.