

Stratégies de contrôle avancées pour systèmes de production  
sous contraintes environnementales : Politiques intégrées et  
apprentissage par renforcement

par

Hassan AARAB

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE  
COMME EXIGENCE PARTIELLE À L'OBTENTION DE  
LA MAÎTRISE AVEC MÉMOIRE EN GÉNIE DE LA PRODUCTION  
AUTOMATISÉE  
M. SC. A

MONTRÉAL, LE 19 DÉCEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE  
UNIVERSITÉ DU QUÉBEC

©Tous droits réservés

Cette licence signifie qu'il est interdit de reproduire, d'enregistrer ou de diffuser en tout ou en partie, le présent document. Le lecteur qui désire imprimer ou conserver sur un autre media une partie importante de ce document, doit obligatoirement en demander l'autorisation à l'auteur.

**PRÉSENTATION DU JURY**  
**CE MÉMOIRE A ÉTÉ ÉVALUÉ**  
**PAR UN JURY COMPOSÉ DE :**

M. Mustapha Ouhimmou, directeur de mémoire  
Département de génie des systèmes à l'École de technologie supérieure

M. Ali Gharbi, codirecteur de mémoire  
Département de génie des systèmes à l'École de technologie supérieure

Jean-Pierre Kenné, président du jury  
Département de génie mécanique à l'École de technologie supérieure

William de Paula Ferreira, membre du jury  
Département de génie des systèmes à l'École de technologie supérieure

**IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC**

**LE 17 DÉCEMBRE 2025**

**À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE**



## REMERCIEMENTS

Je désire adresser ma profonde gratitude à mes directeurs de recherche, M. Mustapha Ouhimmou et M. Ali Gharbi, pour leur disponibilité, leur patience, leurs conseils, leur encadrement rigoureux ainsi que leur soutien constant tout au long de ce parcours de recherche. Leur confiance et leur support m'ont permis de me concentrer pleinement sur mes travaux et de développer mes compétences en recherche.

Je tiens également à exprimer ma sincère reconnaissance aux membres du jury, M. Jean-Pierre Kenné et M. William de Paula Ferreira pour avoir accepté d'examiner et d'évaluer ce travail de recherche.

Mes remerciements s'adressent particulièrement à l'équipe du laboratoire de conception et contrôle des systèmes de production (LC2SP) à l'École de Technologie Supérieure (ÉTS), dont les conseils et l'expertise ont grandement contribué à l'accomplissement de ce mémoire. L'environnement de recherche stimulant du laboratoire a été essentiel à la réalisation de ces travaux.

Enfin, j'exprime ma gratitude à ma famille pour leur soutien indéfectible, ainsi qu'à mes amis et collègues qui ont rendu cette expérience académique enrichissante. Leur présence et leurs encouragements constants ont été une source de motivation inestimable.



# **Stratégies de contrôle avancées pour systèmes de production sous contraintes environnementales : Politiques intégrées et apprentissage par renforcement**

Hassan AARAB

## **RÉSUMÉ**

Cette recherche développe et valide des stratégies de contrôle pour les systèmes manufacturiers non fiables opérant sous contraintes environnementales. L'approche proposée combine deux méthodologies distinctes : la programmation dynamique stochastique combinée avec la résolution numérique pour les configurations hybrides, et l'intelligence artificielle par apprentissage par renforcement pour les systèmes simples.

La première contribution porte sur le contrôle des systèmes de production hybrides intégrant fabrication et refabrication sous contraintes environnementales. Une politique de contrôle environnemental de seuil dynamique est développée pour coordonner les activités de production tout en respectant les limites environnementales imposées. Cette stratégie intègre simultanément les coûts économiques et environnementaux tout en gérant la priorisation et la commutation entre machines selon l'état du système et les contraintes d'émissions GES. L'évaluation comparative révèle des gains substantiels par rapport aux politiques pertinentes de la littérature, tant en efficacité économique qu'en conformité environnementale.

La seconde contribution explore le potentiel de l'apprentissage par renforcement pour piloter les systèmes de base M1P1 dans un environnement stochastique. L'algorithme d'optimisation de politique proximale (PPO) est déployé avec un réglage méthodique des paramètres d'apprentissage. La validation s'effectue progressivement, commençant par une configuration basique avant d'incorporer les exigences environnementales. La comparaison qualitative atteste de la pertinence des techniques d'IA et établit une base méthodologique pour des applications ultérieures plus sophistiquées.

La synthèse de ces travaux souligne la nécessité d'intégrer les préoccupations environnementales dans l'optimisation manufacturière et révèle la synergie entre méthodologies classiques et innovations technologiques. Ces développements offrent aux gestionnaires industriels des instruments performants pour harmoniser les objectifs économiques et écologiques tout en maîtrisant les aléas opérationnels.

**Mots-clés :** programmation dynamique stochastique, apprentissage par renforcement, systèmes de production hybrides, émissions des GES, approche numérique



# **Advanced Control Strategies for Production Systems Under Environmental Constraints: Integrated Policies and Reinforcement Learning**

Hassan AARAB

## **ABSTRACT**

This research develops and validates control strategies for unreliable manufacturing systems operating under environmental constraints. The proposed approach combines two distinct methodologies: dynamic stochastic programming combined with numerical approach for hybrid configurations, and artificial intelligence through reinforcement learning for simple systems.

The first contribution focuses on controlling hybrid production systems integrating manufacturing and remanufacturing under environmental constraints. A dynamic environmental hedging point (DEHPP) control policy is developed to coordinate production activities while respecting imposed environmental limits. This strategy simultaneously integrates economic and environmental costs while managing prioritization and switching between machines according to system state and GHG emission constraints. Comparative study reveals substantial gains compared to relevant policies from literature, both in economic efficiency and environmental compliance.

The second contribution explores the potential of reinforcement learning to control basic MIP1 systems in a stochastic environment. The Proximal Policy Optimization (PPO) algorithm is deployed with methodical tuning of learning parameters. Validation is conducted progressively, starting with a basic configuration before incorporating environmental requirements. Qualitative comparison demonstrates the relevance of AI techniques and establishes a methodological foundation for more sophisticated future applications.

The synthesis of this work emphasizes the necessity of integrating environmental concerns into manufacturing optimization and reveals the synergy between classical methodologies and technological innovations. These developments offer industrial managers powerful instruments to harmonize economic and ecological objectives while mastering operational uncertainties.

**Keywords:** stochastic optimization, reinforcement learning, hybrid production systems, GHG emissions, numerical approach



## TABLE DES MATIÈRES

|                                                                                                                                   | Page |
|-----------------------------------------------------------------------------------------------------------------------------------|------|
| INTRODUCTION .....                                                                                                                | 1    |
| CHAPITRE 1    REVUE DE LITTÉRATURE.....                                                                                           | 3    |
| 1.1    Contexte et enjeux des systèmes de production hybrides face aux défis<br>environnementaux.....                             | 3    |
| 1.1.1    Contexte et motivation de la recherche.....                                                                              | 3    |
| 1.2    Enjeux environnementaux dans le secteur manufacturier .....                                                                | 4    |
| 1.2.1    Évolution des préoccupations environnementales .....                                                                     | 4    |
| 1.2.2    Cadres réglementaires et incitatifs .....                                                                                | 5    |
| 1.2.2.1    Approches réglementaires contraignantes .....                                                                          | 5    |
| 1.2.2.2    Mécanismes basés sur le marché .....                                                                                   | 5    |
| 1.2.2.3    Initiatives volontaires.....                                                                                           | 6    |
| 1.2.3    Fabrication verte : concepts et applications .....                                                                       | 7    |
| 1.3    Systèmes de production hybrides : caractéristiques et défis .....                                                          | 8    |
| 1.3.1    Définition et classification des systèmes manufacturiers .....                                                           | 8    |
| 1.3.2    Spécificités des systèmes hybrides fabrication/refabrication .....                                                       | 9    |
| 1.3.3    Logistique inverse et configurations des chaînes d’approvisionnement.....                                                | 9    |
| 1.4    Modélisation et contrôle des systèmes manufacturiers non fiables .....                                                     | 11   |
| 1.4.1    Théorie de la commande optimale stochastique .....                                                                       | 11   |
| 1.4.2    Politiques de commande à seuil critique.....                                                                             | 12   |
| 1.4.3    Extensions environnementales des politiques de contrôle.....                                                             | 13   |
| 1.5    Critique de la littérature .....                                                                                           | 15   |
| 1.6    Problématique .....                                                                                                        | 16   |
| 1.7    Objectifs du mémoire.....                                                                                                  | 17   |
| 1.8    Conclusion .....                                                                                                           | 18   |
| CHAPITRE 2    POLITIQUE DE CONTRÔLE DES SYSTÈMES DE PRODUCTION<br>HYBRIDES NON FIABLES SOUS CONTRAINTES<br>ENVIRONNEMENTALES..... | 19   |
| 2.1    Formulation du problème.....                                                                                               | 19   |
| 2.1.1    Description du système de production.....                                                                                | 19   |
| 2.1.3    Formation générale du problème d’optimisation.....                                                                       | 21   |
| 2.1.3.1    Diagramme de transition d’états .....                                                                                  | 21   |
| 2.1.3.2    Dynamique du système étudié .....                                                                                      | 23   |
| 2.1.3.3    Espace de commande admissible.....                                                                                     | 23   |
| 2.1.3.4    Fonctions de coût .....                                                                                                | 23   |
| 2.1.3.5    Conditions d’optimum .....                                                                                             | 24   |
| 2.2    Approche numérique de résolution et structure de la politique de contrôle .....                                            | 25   |
| 2.2.1    Approche numérique.....                                                                                                  | 25   |
| 2.3    Résultats de l’approche numérique.....                                                                                     | 27   |
| 2.3.1    Fonction valeur en mode 1.....                                                                                           | 27   |

|         |                                                                      |    |
|---------|----------------------------------------------------------------------|----|
| 2.3.2   | Taux de production de la machine de fabrication M1 en Mode 1 .....   | 29 |
| 2.3.3   | Taux de production de la machine de refabrication M2 en Mode 1 ..... | 31 |
| 2.4     | Structure of the production control policy .....                     | 34 |
| 2.5     | Analyse de sensibilité .....                                         | 36 |
| 2.5.1   | Coût de stockage .....                                               | 36 |
| 2.5.2   | Coût de pénurie .....                                                | 37 |
| 2.5.3   | Coût des émissions.....                                              | 37 |
| 2.5.4   | Taux de panne .....                                                  | 38 |
| 2.5.4.1 | Machine de fabrication M1 .....                                      | 39 |
| 2.5.4.2 | Machine de refabrication M2 .....                                    | 39 |
| 2.5.5   | Taux de réparation .....                                             | 40 |
| 2.5.5.1 | Machine de refabrication M2 .....                                    | 40 |
| 2.5.5.2 | Machine de fabrication M1 .....                                      | 40 |
| 2.5.6   | Taux des émissions .....                                             | 41 |
| 2.5.7   | Taux de production maximum .....                                     | 43 |
| 2.6     | Étude comparative et perspectives managériales.....                  | 44 |
| 2.6.1   | Étude comparative .....                                              | 44 |
| 2.6.2   | Analyse comparative sous variations de coût de stockage .....        | 47 |
| 2.6.3   | Analyse comparative sous variations de coût de pénurie.....          | 48 |
| 2.6.4   | Analyse comparative sous variations de coût des émissions.....       | 49 |
| 2.7     | Perspectives managériales et implémentation .....                    | 50 |
| 2.8     | Conclusion .....                                                     | 52 |

### CHAPITRE 3 APPRENTISSAGE PAR RENFORCEMENT POUR L'OPTIMISATION DU CONTRÔLE DES SYSTÈMES DE PRODUCTION M1P1 .....

|         |                                                                  |    |
|---------|------------------------------------------------------------------|----|
| 3.1     | Introduction.....                                                | 55 |
| 3.2     | Fondements théoriques .....                                      | 56 |
| 3.2.1   | Composantes fondamentales du RL.....                             | 56 |
| 3.2.2   | Cadre Mathématique : Processus de décision markovien (MDP) ..... | 57 |
| 3.2.3   | Cycle d'interaction agent - environnement.....                   | 58 |
| 3.3     | Formulation de problème.....                                     | 60 |
| 3.3.1   | Système de production étudié .....                               | 60 |
| 3.3.2   | Dynamiques du Système.....                                       | 63 |
| 3.3.3   | Espace de commande admissible.....                               | 63 |
| 3.3.4   | Fonctions de coût .....                                          | 63 |
| 3.4     | Application au système M1P1 de base .....                        | 64 |
| 3.4.1   | Conception de l'environnement RL.....                            | 64 |
| 3.4.1.1 | Architecture générale .....                                      | 64 |
| 3.4.1.2 | Paramètres du Système .....                                      | 64 |
| 3.4.1.3 | Structure économique .....                                       | 65 |
| 3.4.2   | Représentation d'état .....                                      | 65 |
| 3.4.2.1 | Espace d'observation .....                                       | 65 |
| 3.4.3   | Espace d'action et stratégies de production .....                | 66 |
| 3.4.3.1 | Actions discrètes disponibles.....                               | 66 |

|                                            |                                                                                           |     |
|--------------------------------------------|-------------------------------------------------------------------------------------------|-----|
| 3.4.4                                      | Dynamiques du Système.....                                                                | 66  |
| 3.4.5                                      | Fonction de Récompense .....                                                              | 67  |
| 3.5                                        | Résultats d'application au premier cas - M1P1 de base .....                               | 68  |
| 3.5.1                                      | Paramètres du système .....                                                               | 68  |
| 3.5.2                                      | Modèle d'apprentissage par renforcement : Optimisation de politique proximale (PPO) ..... | 70  |
| 3.5.2.1                                    | Justification du choix algorithmique.....                                                 | 70  |
| 3.5.2.2                                    | Architecture « Actor-Critic » du modèle PPO.....                                          | 71  |
| 3.5.2.3                                    | Fonctionnement de l'algorithme PPO.....                                                   | 72  |
| 3.5.3                                      | Implémentation du modèle d'apprentissage pour M1P1 de base .....                          | 73  |
| 3.5.3.1                                    | Architecture réseau du modèle PPO .....                                                   | 73  |
| 3.5.3.2                                    | Hyperparamètres utilisés pour l'apprentissage .....                                       | 75  |
| 3.5.3.3                                    | Résultats et interprétation .....                                                         | 75  |
| 3.5.3.4                                    | Analyse des facteurs d'influence sur la performance d'apprentissage .....                 | 76  |
| 3.5.3.5                                    | Facteurs d'influence identifiés .....                                                     | 77  |
| 3.5.3.6                                    | Plan d'expériences .....                                                                  | 79  |
| 3.6                                        | Extension : système M1P1 avec contraintes des émissions.....                              | 86  |
| 3.6.1                                      | Paramètres du système.....                                                                | 86  |
| 3.6.2                                      | Représentation d'état .....                                                               | 86  |
| 3.6.2.1                                    | Espace d'observation .....                                                                | 86  |
| 3.6.3                                      | Dynamiques du Système.....                                                                | 87  |
| 3.6.4                                      | Fonction de Récompense .....                                                              | 88  |
| 3.6.5                                      | Implémentation du modèle d'apprentissage pour M1P1 avec émissions des GES.....            | 88  |
| 3.6.5.1                                    | Architecture du modèle.....                                                               | 88  |
| 3.6.5.2                                    | Hyperparamètres utilisés pour l'apprentissage .....                                       | 90  |
| 3.6.6                                      | Résultats et interprétation .....                                                         | 91  |
| 3.6.7                                      | Comparaison qualitative avec les politiques de référence .....                            | 93  |
| 3.7                                        | Conclusion .....                                                                          | 94  |
| CONCLUSION .....                           |                                                                                           | 95  |
| ANNEXE I                                   | CONDITIONS D'OPTIMUM.....                                                                 | 99  |
| ANNEXE II                                  | APPROCHE DE RÉOLUTION NUMÉRIQUE DES ÉQUATIONS HJB .....                                   | 101 |
| LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES ..... |                                                                                           | 105 |



## LISTE DES TABLEAUX

|             | Page                                                                                |
|-------------|-------------------------------------------------------------------------------------|
| Tableau 2.1 | États du système.....21                                                             |
| Tableau 2.2 | Données d'exemple numérique – Coûts .....26                                         |
| Tableau 2.3 | Données d'exemple numérique – paramètres .....26                                    |
| Tableau 2.4 | Performance des politiques pour différentes valeurs du coût de<br>stockage.....48   |
| Tableau 2.5 | Performance des politiques pour différentes valeurs du coût de<br>pénurie.....48    |
| Tableau 2.6 | Performance des politiques pour différentes valeurs du coût des<br>émissions.....49 |
| Tableau 3.1 | Espace d'actions admissibles.....66                                                 |
| Tableau 3.2 | paramètres du système .....68                                                       |
| Tableau 3.3 | Hyperparamètres du modèle .....75                                                   |
| Tableau 3.4 | Facteurs étudiés.....79                                                             |
| Tableau 3.5 | Matrice des expériences .....80                                                     |
| Tableau 3.6 | Résultats du plan d'expérience .....81                                              |
| Tableau 3.7 | paramètres du système .....86                                                       |
| Tableau 3.8 | Hyperparamètre du modèle.....90                                                     |



## LISTE DES FIGURES

|             | Page                                                                                         |
|-------------|----------------------------------------------------------------------------------------------|
| Figure 2.1  | Schéma bloc du système étudié .....19                                                        |
| Figure 2.2  | Diagramme de transition d'états .....22                                                      |
| Figure 2.3  | Fonction valeur en Mode 1 .....28                                                            |
| Figure 2.4  | Taux de production de la machine de fabrication M1 - Mode 1- vue 3D.....30                   |
| Figure 2.5  | Taux de production de la machine de fabrication M1 - Mode 1- projection sur $(x, e)$ .....30 |
| Figure 2.6  | Taux de production de la machine M2 - Mode 1 – vue 3D .....31                                |
| Figure 2.7  | Taux de production de la machine M2 - Mode 1 - projection sur $(x, e)$ ...32                 |
| Figure 2.8  | Taux de production intégré des machines M1 et M2.....33                                      |
| Figure 2.9  | Impact du coût de stockage $c$ +sur les machines M1 et M2 .....37                            |
| Figure 2.10 | Impact du coût des émissions $ce$ sur les machines M1 et M2.....38                           |
| Figure 2.11 | Impact du taux de panne $q_{110} = q_{13}$ sur les machines M1 et M2 .....39                 |
| Figure 2.12 | Impact de $q_{201} = q_{21}$ sur les machines M1 et M2.....40                                |
| Figure 2.13 | Impact de $\theta_1$ sur les machines M1 et M2.....42                                        |
| Figure 2.14 | Impact de $\theta_2$ sur les machines M1 et M2.....42                                        |
| Figure 2.15 | Impact de $U1max$ sur les machines M1 et M2 .....43                                          |
| Figure 2.16 | Impact de $U2max$ sur les machines M1 et M2 .....44                                          |
| Figure 2.17 | Fonction valeur pour les politiques comparées.....46                                         |
| Figure 2.18 | Logigramme d'implémentation .....51                                                          |
| Figure 3.1  | Cycle d'interaction Agent-Environnement .....58                                              |
| Figure 3.2  | Diagramme des états du système étudié. ....61                                                |
| Figure 3.3  | Système de production étudié .....62                                                         |

|             |                                                                                               |    |
|-------------|-----------------------------------------------------------------------------------------------|----|
| Figure 3.4  | Architecture réseau de neurones du critique .....                                             | 74 |
| Figure 3.5  | Architecture réseau de neurones de l'acteur .....                                             | 74 |
| Figure 3.6  | Politique de contrôle pour M1P1 appris par l'agent PPO .....                                  | 76 |
| Figure 3.7  | Évolution de l'écart d'optimalité pour une longueur d'entraînement<br>de 2M pas de temps..... | 78 |
| Figure 3.8  | Évolution de l'écart d'optimalité pour une longueur d'entraînement<br>de 1M pas de temps..... | 78 |
| Figure 3.9  | Analyse des effets principaux .....                                                           | 83 |
| Figure 3.10 | Analyse des effets d'interaction.....                                                         | 84 |
| Figure 3.11 | Architecture réseau de neurones du critique .....                                             | 89 |
| Figure 3.12 | Architecture réseau de neurones de l'acteur .....                                             | 90 |
| Figure 3.13 | Politique de contrôle appris par le modèle PPO .....                                          | 91 |
| Figure 3.14 | Tranches de politique de contrôle à différents niveaux d'émissions .....                      | 92 |

## LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

|                    |                                                                                                                       |
|--------------------|-----------------------------------------------------------------------------------------------------------------------|
| $u_1(.)$           | Taux de production de la machine de fabrication M1                                                                    |
| $u_2(.)$           | Taux de production de la machine de refabrication M2                                                                  |
| $x(t)$             | Niveau de stock des produits fabriqués                                                                                |
| $e(t)$             | Niveau des émissions                                                                                                  |
| $\varepsilon(t)$   | Processus stochastique du système hybride                                                                             |
| $\varepsilon_1(t)$ | Processus stochastique de la machine de fabrication M1                                                                |
| $\varepsilon_2(t)$ | Processus stochastique de la machine de fabrication M2                                                                |
| $x_i^+$            | Stock positif = $\max(0, x(t))$ pour $x(t) > 0$ de la machine $i \in \{1,2\}$                                         |
| $x_i^-$            | Stock négatif = $\max(0, -x(t))$ pour $x(t) < 0$ de la machine $i \in \{1,2\}$                                        |
| $d_c$              | Taux de demande des clients                                                                                           |
| $U_{max}^1$        | Taux de production maximal de la machine de fabrication                                                               |
| $U_{max}^2$        | Taux de production maximal de la machine de refabrication                                                             |
| $q_{ij}$           | Taux de transition du mode $i$ vers $j$ pour le processus $\varepsilon(t)$ avec $i, j \in \{1,2,3,4\}$                |
| $q_{ij}^k$         | Taux de transition du mode $i$ vers $j$ pour le processus $\varepsilon(t)$ avec $i, j \in \{0,1\}$ et $k \in \{1,2\}$ |
| $c^+$              | Coût de stockage des produits                                                                                         |
| $c^-$              | Coût de pénurie des produits                                                                                          |
| $c_e$              | Coût des émissions                                                                                                    |
| $c_{p1}$           | Coût de production de la machine de fabrication M1                                                                    |
| $c_{p2}$           | Coût de production de la machine de refabrication M2                                                                  |
| $\rho$             | Taux d'actualisation                                                                                                  |

|                              |                                     |
|------------------------------|-------------------------------------|
| <b><math>L</math></b>        | Limite des émissions                |
| <b><math>\theta_1</math></b> | Taux d'émissions pour la machine M1 |
| <b><math>\theta_2</math></b> | Taux d'émissions pour la machine M2 |
| <b><math>HPP</math></b>      | Hedging Point Policy                |
| <b><math>PPO</math></b>      | Proximal Policy Optimization        |
| <b><math>EHPP</math></b>     | Environmental Hedging Point Policy  |
| <b><math>HJB</math></b>      | Hamilton-Jacobi-Bellman             |
| <b><math>RL</math></b>       | Reinforcement Learning              |

## INTRODUCTION

L'évolution récente du secteur industriel au niveau mondial révèle une transformation majeure des priorités manufacturières. Les entreprises ne peuvent plus se contenter de l'optimisation de leurs indicateurs financiers traditionnels. Elles doivent désormais intégrer des critères environnementaux contraignants dans leurs processus décisionnels. Ce changement s'accélère sous l'influence de réglementations internationales renforcées, de pressions sociétales croissantes et d'objectifs climatiques ambitieux fixés par les accords internationaux.

Les systèmes de production modernes opèrent ainsi dans un environnement à contraintes multiples où l'efficacité économique doit coexister avec la conformité environnementale. Cette dualité crée des défis opérationnels inévitables, particulièrement pour les systèmes soumis aux aléas de fonctionnement. Ces aléas de fonctionnement (pannes et réparations aléatoires) constituent un défi majeur pour l'optimisation des systèmes de production, défi qui se complexifie encore avec l'ajout de contraintes environnementales.

Les systèmes hybrides de fabrication/refabrication émergent comme une solution prometteuse pour concilier performance économique et responsabilité environnementale. En combinant production neuve et valorisation de produits refabriqués, ces systèmes s'inscrivent dans la logique d'économie circulaire tout en offrant des avantages économiques potentiels. Toutefois, la gestion de ces systèmes hybrides soulève des questions opérationnelles complexes. La coordination entre fabrication et refabrication nécessite des arbitrages sophistiqués tenant compte des coûts relatifs, des impacts environnementaux différenciés et de la disponibilité des machines. Les gestionnaires doivent déterminer les priorités de production selon l'état du système et les contraintes environnementales.

Parallèlement aux développements analytiques traditionnels ainsi qu'à l'application des méthodes numériques pour des cas complexes, l'intelligence artificielle, notamment l'apprentissage par renforcement, ouvre de nouvelles perspectives pour le contrôle de systèmes complexes dans un environnement stochastique. Ces techniques permettent l'apprentissage de politiques de contrôle par interaction directe avec l'environnement, sans

nécessiter de modélisation exhaustive des dynamiques du système. L'apprentissage par renforcement présente un intérêt particulier pour les systèmes de production où la complexité croissante des contraintes rend les solutions analytiques ou numériques difficiles à établir. Ces méthodes offrent une adaptabilité naturelle aux changements d'environnement et une capacité d'apprentissage continu qui pourrait révolutionner l'approche traditionnelle du contrôle des systèmes de production.

Cette situation crée une opportunité unique de recherche : développer des solutions analytiques avancées pour les systèmes hybrides et valider l'utilité de l'intelligence artificielle sur des systèmes plus simples, établissant ainsi les bases méthodologiques pour une extension future pour des systèmes plus complexes.

L'objectif de ce mémoire consiste donc à développer des politiques de contrôle pour les systèmes de production soumis à des pannes et des réparations aléatoires et des contraintes environnementales, en explorant à la fois les approches analytiques et numériques pour les systèmes hybrides et les méthodes d'intelligence artificielle pour les systèmes simples.

Ce mémoire s'articule en trois chapitres suivant une progression logique. Le chapitre 1 présente une revue critique de la littérature, définissant précisément la problématique et les objectifs de recherche. Le chapitre 2 développe l'approche analytique et la résolution numérique pour le contrôle des systèmes hybrides de fabrication/refabrication avec des contraintes environnementales. Il couvre la modélisation mathématique du problème de contrôle, la résolution numérique, le développement de la politique environnementale de seuil de couverture dynamique et la validation de l'approche proposée par une analyse de sensibilité et une étude comparative. Le chapitre 3 explore l'application de l'apprentissage par renforcement aux systèmes M1P1. Cette partie détaille la formulation du problème, la présentation de l'approche utilisée, l'implémentation de l'algorithme PPO, et l'extension du problème simple en intégrant les contraintes environnementales.

## **CHAPITRE 1**

### **REVUE DE LITTÉRATURE**

#### **1.1 Contexte et enjeux des systèmes de production hybrides face aux défis environnementaux**

##### **1.1.1 Contexte et motivation de la recherche**

Le secteur manufacturier traverse aujourd'hui une période de transformation sans précédent. D'une part, les entreprises font face aux défis économiques traditionnels - optimisation des coûts, satisfaction des demandes fluctuantes, intensification de la concurrence mondiale. D'autre part, elles doivent répondre à des exigences environnementales de plus en plus pressantes, renforcées par les récentes crises climatiques et sanitaires. Les défis récents, notamment ceux révélés par la pandémie de COVID-19, ont mis en évidence la nécessité de repenser les chaînes d'approvisionnement vers plus de résilience et de durabilité (Dolgui et al., 2020; Ivanov & Dolgui, 2021).

Cette dualité éconómico-environnementale redéfinit fondamentalement l'approche de conception, de gestion et de contrôle des systèmes de production. Les entreprises ne peuvent plus se contenter d'optimiser uniquement leurs performances économiques ; elles doivent intégrer simultanément les impacts environnementaux dans leurs décisions opérationnelles. Cette évolution s'accompagne d'un durcissement des réglementations, notamment avec le Green Deal européen qui impose de nouveaux standards de durabilité aux industries (Fetting, 2020).

Dans cette dynamique, les systèmes hybrides de fabrication/refabrication émergent comme une réponse prometteuse à cette double contrainte. En combinant la production traditionnelle à partir de matières premières brutes avec la refabrication de produits usagés, ces systèmes s'inscrivent naturellement dans l'économie circulaire. Ils offrent des perspectives intéressantes de réduction des déchets, de diminution de la consommation de ressources et de baisse des émissions polluantes. Cependant, leur mise en œuvre soulève des défis opérationnels considérables.

La complexité de ces systèmes hybrides provient de multiples sources qui s'entrecroisent : les aléas techniques traditionnels (pannes d'équipements, temps de réparation variables), les quantités d'émissions variables d'une machine à l'autre, et la coordination entre les différentes composantes du système. Cette combinaison de contraintes multiples crée un environnement décisionnel particulièrement complexe.

C'est dans ce contexte que la théorie de la commande optimale stochastique trouve toute sa pertinence. Elle offre un cadre mathématique rigoureux pour développer des politiques de contrôle qui optimisent simultanément les performances économiques et environnementales, tout en tenant compte de l'ensemble des incertitudes du système. Notre travail de recherche s'inscrit précisément dans cette démarche d'optimisation intégrée.

## **1.2 Enjeux environnementaux dans le secteur manufacturier**

### **1.2.1 Évolution des préoccupations environnementales**

L'évolution des préoccupations environnementales dans le secteur manufacturier reflète une prise de conscience progressive, mais accélérée des impacts industriels sur l'écosystème planétaire. Historiquement, les entreprises ont privilégié la maximisation de leurs profits économiques, souvent au détriment des considérations environnementales (Kang et al., 2018). Cette approche, longtemps tolérée, devient aujourd'hui indéfendable face à l'urgence climatique et aux attentes sociétales croissantes.

Le cadre conceptuel proposé par (Elkington, 1998), connu sous le nom de "Triple Bottom Line", a marqué un tournant en proposant d'évaluer les performances des entreprises selon trois dimensions équivalentes : économique, environnementale et sociale. Cette approche holistique a progressivement gagné en légitimité, particulièrement depuis les Accords de Paris sur le climat en 2015 (*Paris Agreement COP21*, 2015), et s'est renforcée avec les récentes initiatives telles que l'Agenda 2030 des Nations Unies pour le développement durable (*UN Agenda for Sustainable Development Web*, 2015).

Pour approfondir la compréhension de cette évolution conceptuelle, les travaux de (Carter & Rogers, 2008) offrent des perspectives complémentaires sur l'intégration de la durabilité dans

les chaînes d'approvisionnement. Ces recherches montrent que la transition vers un modèle industriel durable ne relève plus d'un choix stratégique, mais d'une nécessité opérationnelle.

### **1.2.2 Cadres réglementaires et incitatifs**

La réponse aux défis environnementaux s'organise autour de trois grandes catégories d'instruments, chacune avec ses spécificités et ses domaines d'application privilégiés.

#### **1.2.2.1 Approches réglementaires contraignantes**

Les approches dites de "command and control" constituent l'arsenal traditionnel des politiques environnementales. Elles se caractérisent par l'imposition de normes techniques précises, de limites d'émission strictes ou de technologies spécifiques, sans recours aux mécanismes de marché. Ces instruments permettent une intervention directe et rapide des autorités publiques (Lee & Yik, 2004), particulièrement efficaces lorsqu'il s'agit d'interdire des pratiques dangereuses ou d'imposer des standards minimaux.

Cependant, cette approche prescriptive présente des limitations importantes. Sa rigidité peut freiner l'innovation en imposant des solutions techniques spécifiques, potentiellement rapidement obsolètes. De plus, elle peut s'avérer coûteuse tant pour les entreprises (coûts de mise en conformité) que pour les autorités (coûts de contrôle et de sanction) (SHIN & CHEN, 2000). Malgré ces limitations, cette approche conserve sa pertinence dans certains contextes spécifiques.

#### **1.2.2.2 Mécanismes basés sur le marché**

Face aux limites des approches contraignantes, les mécanismes basés sur le marché ont gagné en popularité en raison de leur flexibilité et de leur efficacité économique. Ces instruments s'appuient sur les signaux de prix pour orienter les comportements vers des pratiques plus respectueuses de l'environnement.

La taxation carbone représente l'un des instruments les plus directs. Introduite pour la première fois en Norvège en 1991 (Gottinger, 1995), elle consiste à appliquer un prix aux

émissions de carbone, internalisant ainsi le coût environnemental dans les décisions économiques. Cette approche a depuis été adoptée par de nombreux pays avec des modalités variables : l'Australie, la Suède, la Norvège, l'Afrique du Sud, le Mexique et la province canadienne de Colombie-Britannique ont développé leurs propres systèmes de tarification carbone (Varsei et al., 2014). L'avantage principal de cette approche réside dans sa simplicité conceptuelle et sa capacité à générer des revenus publics pouvant être réinvestis dans la transition énergétique.

Les systèmes de plafonnement et d'échange (cap-and-trade) constituent une alternative sophistiquée qui combine fixation d'objectifs environnementaux et mécanismes de marché. Ces mécanismes de marché ont démontré leur efficacité, notamment avec le programme américain de réduction des émissions de SO<sub>2</sub> qui a permis une baisse de 36% des émissions entre 1990 et 2004 (Schmalensee & Stavins, 2013).

Le principe consiste à définir un plafond global d'émissions, puis à distribuer ou vendre aux enchères des quotas aux entreprises. Celles qui parviennent à émettre moins que leur allocation peuvent vendre leurs quotas excédentaires, créant ainsi une incitation économique à la réduction des émissions. Le Système d'échange de quotas d'émission de l'Union européenne (EU ETS), mis en place depuis 2005, constitue le plus grand marché du carbone au monde et couvre environ 40% des émissions de gaz à effet de serre de l'UE.

L'efficacité comparative de ces mécanismes de marché a été largement documentée. Selon certaines études, ils permettent des économies de coûts comprises entre 15% et 90% par rapport aux approches réglementaires traditionnelles (Hoen et al., 2014). Cette efficience supérieure s'explique par leur capacité à laisser les entreprises choisir les moyens les plus économiques de réduire leurs émissions, favorisant ainsi l'innovation et l'adoption de technologies propres.

### **1.2.2.3 Initiatives volontaires**

Les approches volontaires représentent une troisième voie qui s'appuie sur l'engagement spontané des entreprises plutôt que sur la contrainte réglementaire. Ces initiatives, souvent structurées autour de normes internationales ou de programmes sectoriels, connaissent un

développement important depuis les années 1991 et sont utilisées pour améliorer l'efficacité environnementale et la réduction des émissions (Brouhle et al., 2009).

L'attrait de ces approches réside dans leur capacité à intégrer les considérations environnementales au cœur de la stratégie d'entreprise (Corbett & Kirsch, 2001). Elles permettent aux entreprises de réduire certains coûts opérationnels et d'augmenter leurs parts de marche (Blackman & Boyd, 2002). De plus, elles peuvent contribuer à améliorer les relations avec les parties prenantes, notamment les consommateurs et les investisseurs, de plus en plus sensibles aux enjeux environnementaux (Arora & Cason, 1996). Un avantage particulier des approches volontaires réside dans leur capacité à stimuler l'innovation (Wallace, 2017).

### **1.2.3 Fabrication verte: concepts et applications**

L'émergence du concept de fabrication verte (green manufacturing) traduit l'évolution des mentalités industrielles vers une approche plus intégrée de la performance environnementale. La définition proposée par (Melnik & Smith, 1996) reste aujourd'hui une référence : "un système qui intègre la conception des produits et des processus avec les problématiques de planification et de contrôle de la production, dans le but d'identifier, de quantifier, d'évaluer et de gérer le flux des déchets environnementaux, avec l'objectif de réduire l'impact environnemental tout en maximisant l'efficacité des ressources."

Cette définition met l'accent sur l'approche systémique nécessaire à la mise en œuvre d'une production véritablement durable. Il ne s'agit plus seulement d'ajouter des équipements de dépollution en fin de processus, mais de repenser l'ensemble de la chaîne de valeur. Cette approche s'articule autour de trois piliers fondamentaux : l'engagement de fournisseurs dans des pratiques durables, l'utilisation de matériaux écologiques, et le déploiement de technologies à faible impact environnemental (Barange & Agarwal, 2016).

Cependant, la transition vers la fabrication verte reste un défi majeur pour de nombreux secteurs industriels. Dans l'électronique, l'industrie des pâtes et papiers, ou encore la plasturgie, les systèmes de production traditionnels génèrent encore d'importantes quantités d'émissions de carbone (Huisingh et al., 2015).

### **1.3 Systèmes de production hybrides : caractéristiques et défis**

#### **1.3.1 Définition et classification des systèmes manufacturiers**

La complexité croissante des environnements industriels contemporains nécessite une approche nuancée de la classification des systèmes manufacturiers. Un système de production peut être défini comme un réseau d'interactions complexes entre ressources matérielles (actifs de production) et ressources humaines (opérateurs, gestionnaires, dirigeants), orchestrées pour transformer des intrants (matières premières, énergie, information) en extrants (produits finis, services) répondant aux exigences spécifiques des clients. Ces ressources interagissent par des flux physiques (produits) et des flux d'informations (quantité, qualité, planification de production) (Stevenson et al., 2012).

Cette transformation s'opère dans un environnement caractérisé par l'interaction de multiples sources d'incertitude qui se renforcent mutuellement : défaillances imprévisibles d'équipements, volatilité de la demande client, perturbations dans les chaînes d'approvisionnement, évolutions réglementaires soudaines, etc. L'intégration récente des préoccupations environnementales ajoute une dimension supplémentaire de complexité, à laquelle les entreprises doivent faire face dans leurs systèmes de décisions.

Les systèmes manufacturiers peuvent être appréhendés selon plusieurs axes de classification qui reflètent leurs caractéristiques opérationnelles distinctes :

Le volume de production constitue un premier critère discriminant. La production unitaire se caractérise par la fabrication de produits spécifiques, souvent sur mesure, nécessitant une forte flexibilité et des compétences artisanales. La production par lot (Bouslah et al., 2013) représente un compromis entre flexibilité et efficacité, permettant d'adapter la production à une variété de demandes tout en bénéficiant d'économies d'échelle partielles. La production en série (Sethi et al., 1997) privilégie l'efficacité et la standardisation, optimisant les coûts unitaires au détriment de la flexibilité.

Le mode de gestion des flux est un deuxième critère. La production en flux poussé (push) s'appuie sur des prévisions de demande pour anticiper la production, créant des stocks tampons pour absorber les incertitudes. La production en flux tiré (pull)(Lavoie et al.,

2010) synchronise la production sur la demande réelle, minimisant les stocks, mais exigeant une réactivité et une fiabilité accrues. Le choix entre ces deux approches influence directement la gestion des stocks, les délais de livraison, et la capacité d'adaptation aux variations de demande.

La nature des flux distingue les systèmes selon les caractéristiques du processus de fabrication. Les systèmes à flux continu caractérisent les industries de processus (chimie, raffinerie, sidérurgie) où la matière circule en permanence. Les systèmes à flux discret concernent la fabrication de produits distincts (automobile, électronique).

### **1.3.2 Spécificités des systèmes hybrides fabrication/refabrication**

Les systèmes hybrides de fabrication/refabrication représentent une évolution particulièrement innovante qui répond simultanément aux contraintes économiques et environnementaux contemporains. Ces systèmes intègrent de manière coordonnée deux logiques de fabrication complémentaires : la fabrication traditionnelle transforme des matières premières vierges en produits neufs, tandis que la refabrication reconditionne des produits usagés pour restaurer des performances fonctionnelles comparables à celles des produits neufs.

Cette approche hybride s'inscrit naturellement dans les principes de l'économie circulaire en créant des boucles de valeur qui prolongent la durée de vie des produits et réduisent la pression sur les ressources naturelles. Les bénéfices potentiels sont substantiels : réduction pouvant atteindre 70% de la consommation de matières premières, diminution jusqu'à 50% des coûts de production, et baisse significative des émissions de gaz à effet de serre (Wang et al., 2013). L'étude menée par (Francie et al., 2015) dans le secteur des consommables pour imprimantes illustre concrètement ces avantages.

### **1.3.3 Logistique inverse et configurations des chaînes d'approvisionnement**

L'efficacité des systèmes hybrides repose fondamentalement sur une organisation optimale des flux de produits depuis leur utilisation initiale jusqu'à leur réintégration dans le cycle

productif. Cette problématique relève du domaine de la logistique inverse, conceptualisée par (Lambert & Riopel, 2003) comme "le processus de planification, d'implantation, et de contrôle de l'efficience, de la rentabilité des matières premières, des en-cours de production, des produits finis, et l'information pertinente du point d'utilisation jusqu'au point d'origine dans le but de reprendre ou générer de la valeur ou pour en disposer de la bonne façon tout en assurant une utilisation efficace et environnementale des ressources mises en œuvre."

Cette définition souligne la complexité multidimensionnelle de la logistique inverse qui doit optimiser simultanément les flux physiques, informationnels et financiers, tout en intégrant les contraintes environnementales et réglementaires.

Les configurations organisationnelles de ces systèmes peuvent être caractérisées selon deux modèles principaux :

- **La logistique inverse en boucle ouverte**

Représente une configuration décentralisée où les produits en fin de vie ne retournent pas nécessairement vers leur producteur initial. Dans cette configuration, les produits récupérés peuvent être traités par d'autres acteurs économiques spécialisés dans la valorisation des déchets ou la refabrication. Cette approche présente l'avantage de la flexibilité et peut conduire à des spécialisations sectorielles efficaces. Les matériaux récupérés peuvent être réutilisés pour créer des produits similaires à ceux d'origine ou être détournés vers d'autres applications, maximisant ainsi les opportunités de valorisation.

L'industrie électronique illustre parfaitement cette logique, particulièrement dans le secteur des téléphones mobiles. L'étude de (Geyer & Jackson, 2004) sur les opérations de logistique inverse dans ce secteur, menée au Royaume-Uni et aux États-Unis, démontre la viabilité économique de cette approche décentralisée.

- **La logistique inverse en boucle fermée**

Correspond à une configuration intégrée où les produits usagés retournent vers leur fabricant d'origine pour être réintégrés dans son propre cycle de production. Cette approche présente des avantages significatifs en termes de contrôle qualité, de cohérence des processus, et d'optimisation des ressources. Cependant, elle exige une coordination complexe entre les flux directs (du producteur vers le consommateur) et les flux inverses (du consommateur vers le producteur). La complexité de cette configuration provient notamment de la nécessité

d'intégrer les informations et les flux entre les deux chaînes. Les décisions de production doivent tenir compte simultanément des demandes de produits neufs et des disponibilités de produits à refabrication. Cette intégration informationnelle est cruciale pour optimiser l'allocation des ressources entre fabrication et refabrication. Les travaux de (Jorjani et al., 2004) dans l'industrie électronique illustrent cette complexité. Leur recherche sur l'allocation optimale des composants électroniques démontés met en évidence les défis de coordination entre les opérations de démontage, d'évaluation, et de réassemblage.

## **1.4 Modélisation et contrôle des systèmes manufacturiers non fiables**

### **1.4.1 Théorie de la commande optimale stochastique**

La théorie de la commande optimale stochastique constitue le fondement mathématique nécessaire pour aborder la complexité des systèmes manufacturiers évoluant dans des environnements incertains. Cette approche vise à déterminer des politiques de contrôle qui optimisent un critère de performance donné (minimisation des coûts, maximisation des profits, etc.) tout en tenant compte des diverses sources d'incertitudes qui affectent le système.

L'architecture conceptuelle de cette théorie a été établie par les travaux pionniers de (Rishel, 1975), qui a formulé les conditions nécessaires et suffisantes d'optimalité pour les problèmes de commande dans un contexte stochastique. S'appuyant sur le principe du maximum de Pontryagin et les techniques de programmation dynamique, ces travaux ont abouti au développement des équations d'Hamilton-Jacobi-Bellman (HJB). Ces équations aux dérivées partielles caractérisent la fonction valeur optimale et permettent, une fois résolues, de déterminer la politique de contrôle optimale.

L'application de ces concepts théoriques aux systèmes manufacturiers a été initiée par (Olsder & Suri, 1980), qui ont modélisé des systèmes de production sujets à des pannes et réparations aléatoires sous forme d'un processus markovien. Cette transposition a ouvert la voie à une littérature riche et diversifiée sur l'optimisation des systèmes manufacturiers stochastiques.

Face à la complexité croissante des problèmes industriels réels, diverses approches de résolution ont été développées. L'intelligence artificielle, explorée par (BASNET & and MIZE, 1995), offre des outils pour traiter des systèmes complexes où les approches analytiques traditionnelles atteignent leurs limites. Les méthodes heuristiques (Thesen, 1999) proposent des solutions pragmatiques pour des problèmes de grande dimension, tandis que la simulation (Kenne & Gharbi, 2000) permet de valider et d'affiner les politiques de contrôle dans des environnements réalistes.

Les développements récents de ces méthodes intègrent les avancées de l'intelligence artificielle, notamment l'apprentissage par renforcement pour le contrôle de systèmes stochastiques. Il existe plusieurs aspects des systèmes manufacturiers qui présentent des propriétés markoviennes, ce qui rend l'apprentissage par renforcement basé sur les processus de décision markoviens (MDP) particulièrement adapté comme outil de modélisation pour des stratégies conjointes exploitant l'information d'état du système (Zhang et al., 2024).

#### **1.4.2 Politiques de commande à seuil critique**

Dans l'univers des systèmes manufacturiers non fiables, les politiques de commande à seuil critique (Hedging Point Policy - HPP) ont acquis une position centrale en raison de leur efficacité démontrée et de leur relative simplicité conceptuelle. Introduites par (Kimemia & Gershwin, 1983), ces politiques reposent sur le principe de maintenir un niveau de stock de sécurité optimal, appelé seuil critique ou point de couverture, pour faire face aux périodes d'indisponibilité des équipements de production.

La logique derrière ces politiques est intuitive : pendant les périodes de disponibilité des équipements, le système produit non seulement pour satisfaire la demande courante, mais aussi pour constituer un stock de sécurité qui permettra de continuer à satisfaire la demande des clients pendant les périodes de panne. L'optimisation porte sur la détermination du niveau optimal de ce stock de sécurité, qui représente un compromis entre les coûts de stockage et les coûts de rupture.

La structure opérationnelle de ces politiques se caractérise par sa simplicité :

- Si le niveau de stock est inférieur au seuil critique : production au taux maximal pour reconstituer rapidement le stock de sécurité.
- Si le niveau de stock égale le seuil critique : production exactement au taux de la demande pour maintenir le niveau optimal.
- Si le niveau de stock dépasse le seuil critique : arrêt de la production pour éviter des coûts de stockage excessifs.

Akella & Kumar (1986) ont démontré l'optimalité de cette politique pour un système composé d'une seule machine produisant un seul type de produit avec une demande constante. Leur contribution majeure réside dans la résolution analytique explicite des équations HJB, permettant de déterminer la valeur optimale du seuil critique qui minimise le coût total sur un horizon infini.

Pour des systèmes plus complexes (plusieurs machines, plusieurs produits, demandes stochastiques), la résolution analytique devient généralement impossible, nécessitant le recours à des méthodes numériques. Boukas & Haurie (1990) ont développé une approche innovante basée sur la méthode de (Kushner & Dupuis, 1992) pour résoudre numériquement les équations HJB dans ces configurations complexes.

#### **1.4.3 Extensions environnementales des politiques de contrôle**

L'intégration des préoccupations environnementales dans les politiques de contrôle manufacturier représente un développement relativement récent de la littérature. Cette évolution répond à la fois aux pressions réglementaires croissantes et à la prise de conscience des enjeux de durabilité.

Ben-Salem et al. (2015) ont proposé une extension fondamentale des politiques à seuil critique traditionnelles avec le développement de la politique environnementale à seuil critique (Environmental Hedging Point Policy - EHPP). Cette approche innovante élargit l'espace des variables d'état en intégrant non seulement le niveau d'inventaire, mais aussi le niveau d'émissions cumulées du système. La politique de contrôle proposée détermine le taux de production en fonction de ces deux variables d'état, avec l'objectif de minimiser

simultanément les coûts économiques (stockage, pénurie) et environnementaux (émissions, pénalités).

Cette extension conceptuelle nécessite une reformulation complète du problème d'optimisation. L'espace d'état bidimensionnel (inventaire, émissions) crée des régions de décision plus complexes que dans le cas traditionnel unidimensionnel. La politique optimale peut prescrire l'arrêt de la production même si le niveau d'inventaire est faible, si le niveau d'émissions approche des seuils critiques.

Le développement de ces approches s'est enrichi de contributions spécialisées pour différentes configurations industrielles. Afshar-Bakeshloo et al. (2018) ont développé des politiques adaptées aux systèmes comprenant des machines à différents niveaux d'émission, permettant aux entreprises d'optimiser l'utilisation de technologies plus ou moins polluantes selon les contraintes environnementales.

Les travaux de Hajej et al. (2017a, 2017b) ont exploré l'intégration des décisions de maintenance dans ce cadre environnemental. Leurs recherches démontrent que les politiques optimales doivent coordonner les décisions de production, de maintenance et de sous-traitance pour minimiser l'impact environnemental global. Cette approche holistique reconnaît que l'optimisation environnementale ne peut se limiter aux seules décisions de production, mais doit englober l'ensemble des décisions opérationnelles.

Dans cette perspective Entezaminia et al. (2020, 2021) ont développé des politiques spécifiquement adaptées aux systèmes hybrides de fabrication et refabrication sous contraintes environnementales. Leurs travaux montrent que l'intégration des activités de refabrication peut considérablement améliorer les performances environnementales globales du système, mais nécessite des politiques de contrôle sophistiquées pour optimiser l'allocation des ressources entre fabrication et refabrication. Plus récemment, Kenné et al. (2024) ont abordé la problématique de planification de production dans des systèmes manufacturiers non fiables en intégrant les contraintes d'émissions de gaz à effet de serre. Leur approche utilise la théorie du contrôle optimal stochastique et la résolution numérique pour développer une politique de contrôle basée sur des seuils multiples.

## 1.5 Critique de la littérature

Dans le domaine des systèmes hybrides de production, plusieurs configurations intégrant fabrication et refabrication ont été étudiées dans la littérature du contrôle optimal. Les travaux d'Entezaminia et al. (2020, 2021) explorent l'intégration des activités de fabrication et de refabrication sous contraintes environnementales, établissant des bases solides pour l'optimisation de ces systèmes complexes. Ces recherches démontrent les avantages potentiels de la refabrication en termes de réduction des émissions et de conservation des ressources. Cependant, malgré ces avancées, la plupart des études s'appuient sur l'adaptation des cadres HPP ou EHPP existants plutôt que sur le développement de politiques de contrôle entièrement nouvelles adaptées aux systèmes hybrides sujets aux pannes opérant sous réglementations environnementales. En plus, les aspects critiques tels que la priorisation et la commutation entre machines de fabrication et de refabrication restent sous-explorés dans ces modèles.

En ce qui concerne l'intégration perspective discutée précédemment, deux gaps clés émergent dans la littérature sur les systèmes hybrides environnementaux : premièrement, l'absence de politiques de contrôle novatrices - bien que la politique environnementale à seuil critique (EHPP) ait été étendue pour incorporer les taxes carbone et les réglementations de plafonnement et d'échange, aucun cadre de contrôle intégré et optimal n'a été développé pour les systèmes de production hybrides qui traitent simultanément les pannes aléatoires de machines et les émissions de GES. Deuxièmement, l'exploration insuffisante de la priorisation et la commutation des machines - peu d'études investiguent la priorisation et la commutation dynamique entre machines de fabrication et de refabrication, particulièrement dans le contexte de comment ces décisions impactent à la fois les niveaux d'émissions et les coûts de production. Cette lacune limite l'applicabilité des approches existantes aux configurations industrielles réalistes où la coordination optimale entre modes de production devient cruciale pour la performance économique-environnementale.

En ce qui concerne les méthodologies de développement des politiques de contrôle pour les systèmes manufacturiers, la littérature existante repose sur deux approches traditionnelles qui

présentent des limitations. D'une part, les approches analytiques et numériques dominent le domaine, comme illustré par les travaux fondamentaux d' Akella & Kumar (1986) et les extensions récentes de (Kenné et al., 2024) qui s'appuient sur la résolution numérique des équations d'Hamilton-Jacobi-Bellman. Ces méthodes, bien qu'offrant des garanties d'optimalité théorique, deviennent computationnellement prohibitives pour des systèmes complexes et souffrent du problème de la dimensionnalité. D'autre part, une seconde catégorie de travaux ((Ben-Salem et al. 2015), (Afshar-Bakeshloo et al., 2018)) pose directement des politiques de contrôle (souvent inspirées des HPP classiques) et les valide empiriquement en utilisant la simulation et les méthodologies de surface de réponse (RSM). Cette approche, bien que pragmatique, ne garantit pas l'optimalité des politiques proposées et limite l'exploration de l'espace des solutions possibles.

À notre connaissance, aucun travail n'a démontré la capacité de l'apprentissage par renforcement à découvrir de manière autonome des politiques de contrôle manufacturier qui rivalisent avec les solutions analytiques optimales, tout en offrant la flexibilité nécessaire pour s'adapter à des environnements opérationnels stochastiques et des systèmes de complexité industrielle réaliste. Cette lacune méthodologique limite l'évolutivité des approches actuelles vers des applications industrielles sophistiquées où les hypothèses simplificatrices des modèles analytiques ne tiennent plus.

## 1.6 Problématique

Dans ce contexte, les questions suivantes émergent naturellement :

- Comment les systèmes de production hybrides peuvent-ils intégrer efficacement la planification de la production, les considérations environnementales et la fiabilité opérationnelle pour minimiser les coûts totaux tout en respectant les contraintes environnementales ? Plus spécifiquement, comment les décideurs peuvent-ils déterminer les taux de production optimaux pour les machines de fabrication et de refabrication tout en maintenant l'efficience des coûts sous contraintes environnementales et en assurant la satisfaction de la demande ?

- Comment définir une stratégie de priorisation efficace entre les machines de fabrication et de refabrication selon les niveaux de stock, les émissions générées et la disponibilité des équipements ? Comment les gestionnaires peuvent-ils équilibrer les arbitrages entre la machine économiquement efficiente, mais néfaste pour l'environnement et la machine écologique, mais plus coûteuse ? Dans quelles conditions une machine devrait-elle avoir la priorité sur l'autre, et comment le système peut-il basculer efficacement entre elles pour optimiser la performance globale ?
- Comment les nouvelles approches et méthodes d'intelligence artificielle (l'apprentissage par renforcement comme exemple) peuvent contribuer à concevoir des politiques de contrôle pour des systèmes de production non fiables ?
- Comment les nouvelles approches et méthodes d'intelligence artificielle, notamment l'apprentissage par renforcement, peuvent-elles contribuer à concevoir des politiques de contrôle adaptatives et robustes pour les systèmes de production simples non fiables opérant sous contraintes environnementales ?

Ce mémoire vise à répondre à ces questions en proposant une nouvelle politique dynamique de contrôle adaptée aux systèmes de production hybrides soumis à des contraintes environnementales strictes et en examinant aussi le potentiel de l'apprentissage par renforcement dans le contrôle des systèmes de production opérant dans un environnement stochastique.

## **1.7 Objectifs du mémoire**

L'objectif général de cette étude est de développer des politiques de contrôle pour les systèmes de production soumis à des pannes et réparations aléatoires et des contraintes environnementales, en explorant à la fois les approches numériques pour les systèmes hybrides et les méthodes d'intelligence artificielle pour les systèmes simples. Plus spécifiquement, les objectifs de recherche incluent :

### **1. Développer une politique de contrôle intégrée pour les systèmes hybrides.**

Élaborer une politique optimale de production pour les systèmes de fabrication/refabrication intégrant simultanément les coûts de stockage, de pénurie, de production et d'émissions tout en gérant la priorisation et la commutation entre machines.

### **2. Exploiter l'apprentissage par renforcement pour les systèmes simples.**

Développer et implémenter des approches d'apprentissage par renforcement pour le contrôle de systèmes de production non fiables, simples et sous contraintes environnementales afin d'examiner l'utilité de ces nouvelles approches dans ce domaine de recherche.

## **1.8 Conclusion**

Ce chapitre a présenté le contexte de ce travail de recherche, les concepts fondamentaux relatifs aux systèmes de production sous contraintes environnementales, ainsi que la terminologie définissant le cadre général de cette recherche. La revue effectuée révèle deux lacunes importantes dans la littérature existante. D'une part, les approches de contrôle pour les systèmes hybrides fabrication/refabrication intègrent insuffisamment les contraintes environnementales dans leurs mécanismes décisionnels. D'autre part, l'application de l'apprentissage par renforcement aux systèmes de production reste largement exploratoire, avec peu de validations méthodiques par rapport aux solutions analytiques optimales. En nous appuyant sur cette analyse critique, nous avons formulé notre problématique de recherche autour du développement de politiques de contrôle intégrées pour les systèmes hybrides et de l'examen de l'utilité de l'intelligence artificielle pour les systèmes simples. Finalement, les objectifs de recherche ont été définis selon cette approche duale.

## CHAPITRE 2

### POLITIQUE DE CONTRÔLE DES SYSTÈMES DE PRODUCTION HYBRIDES NON FIABLES SOUS CONTRAINTES ENVIRONNEMENTALES

#### 2.1 Formulation du problème

##### 2.1.1 Description du système de production

Le système analysé est un système de production hybride intégrant des processus de fabrication et de refabrication. Il comprend deux machines, M1 et M2, qui peuvent subir des pannes et des réparations aléatoires. La machine M1 est dédiée à la fabrication, tandis que la machine M2 s'occupe de la refabrication. Les deux machines émettent des gaz à effet de serre à des taux différents. Voici le schéma bloc de ce système :

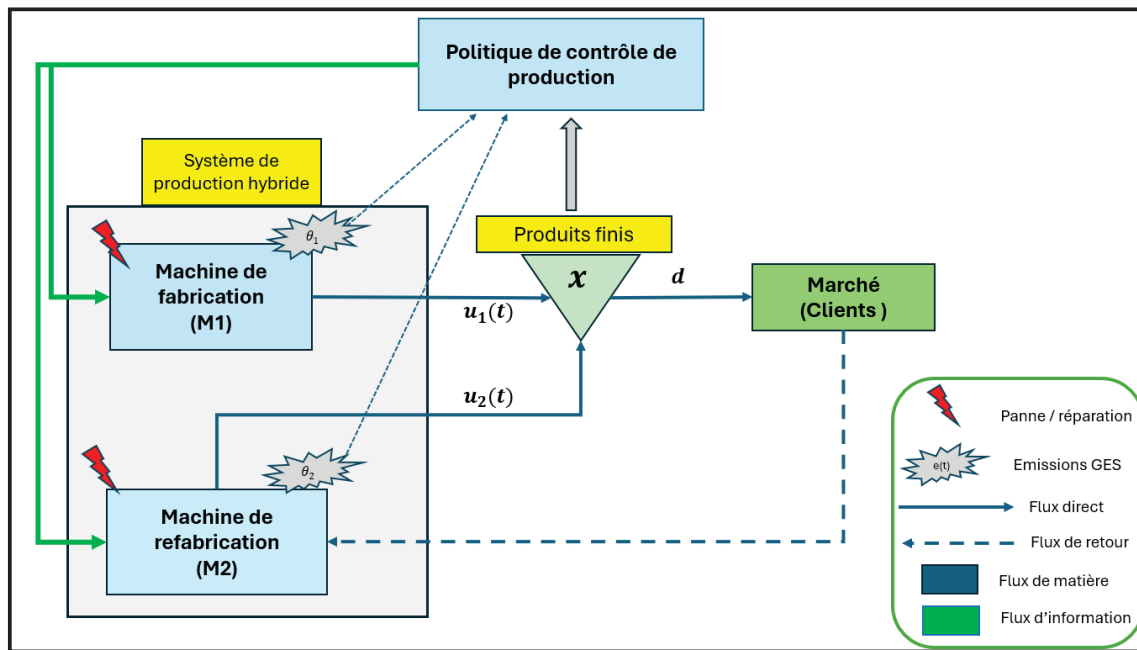


Figure 2.1 Schéma bloc du système étudié

Pour le reste de ce chapitre, nous considérons les notations suivantes :

- $u_1(.)$  : taux de production de la machine de fabrication M1
- $u_2(.)$  : taux de production de la machine de refabrication M2
- $x_1(t)$  : niveau de stock des produits fabriqués
- $x_2(t)$  : niveau de stock des produits retournées
- $e(t)$  : niveau des émissions

- $\varepsilon(t)$ : processus stochastique du système hybride
- $\varepsilon_1(t)$ : processus stochastique de la machine de fabrication M1
- $\varepsilon_2(t)$ : processus stochastique de la machine de fabrication M2
- $x^+$ : stock positif =  $\max(0, x(t))$  pour  $x(t) > 0$  de la machine ,  $i \in \{1,2\}$
- $x^-$ : stock négatif =  $\max(0, -x(t))$  pour  $x(t) < 0$  de la machine ,  $i \in \{1,2\}$
- $d$  : taux de demande des clients
- $U_{max}^1$  : taux de production maximal de la machine de fabrication
- $U_{max}^2$  : taux de production maximal de la machine de refabrication
- $q_{ij}$ : taux de transition du mode  $i$  vers  $j$  pour le processus  $\varepsilon(t)$  avec  $i, j \in \{1,2,3,4\}$
- $q_{ij}^k$ : taux de transition du mode  $i$  vers  $j$  pour le processus  $\varepsilon_1(t)$  avec  $i, j \in \{0,1\}$  et  $k \in \{1,2\}$
- $c^+$ : coût de stockage des produits
- $c^-$  : coût de pénurie des produits
- $c_e$ : coût des émissions
- $c_{p1}$ : coût de production de la machine de fabrication M1
- $c_{p2}$ : coût de production de la machine de refabrication M2
- $\rho$  : taux d'actulisation
- $L$  : limite des émissions
- $\theta_1$ : taux d'émissions pour la machine M1
- $\theta_2$ : taux d'émissions pour la machine M2
- $MTBF$  : temps moyen entre les pannes
- $MTTR$  : temps moyen de réparation

### 2.1.2 Hypothèse de modélisation

Les hypothèses considérées pour la modélisation de ce problème sont les suivantes :

- Les emplacements des clients sont connus et fixes.
- Les quantités retournées sont des fonctions connues de la demande initiale des clients.
- La demande des clients est connue et son taux est constant dans le temps.
- La qualité des produits refabriqués est identique à celle des produits fabriqués.
- Le taux de production maximal pour chaque machine est connu.

- Les deux machines M1 et M2 sont sujettes à des pannes et réparations aléatoires.
- La localisation des fournisseurs est connue et la matière première est toujours disponible pour les deux machines.
- Les coûts des matériaux, de la fabrication, de la capacité non utilisée, des pénuries, du transport, du stockage, du recyclage, de la refabrication et de l'élimination sont connus pour chaque emplacement.
- Le coût de pénurie dépend de la quantité en pénurie et du temps (valeur moyenne en dollars par produit/unité de temps).
- Le coût de stockage dépend du niveau moyen des stocks (valeur moyenne en dollars par produit/unité de temps).
- Les machines émettent des émissions nocives uniquement lorsqu'elles sont en production. Aucune émission n'est produite lorsqu'elles sont en panne.

### 2.1.3 Formation générale du problème d'optimisation

#### 2.1.3.1 Diagramme de transition d'états

Étant donné que les machines de fabrication M1 et de refabrication M2 sont sujettes à des pannes et réparations aléatoires et les réparations des machines sont parfaites. Les variables aléatoires qui décrivent l'état des machines M1 et M2 sont notées  $\varepsilon_1(t)$  et  $\varepsilon_2(t)$  respectivement avec des valeurs dans  $B_1 = \{0,1\}$  et  $B_2 = \{0,1\}$ . Le système suit donc un processus de Markov à temps continu - état discret  $\{\varepsilon(t)\}$  avec des valeurs dans  $B = B_1 \times B_2 = \{1,2,3,4\}$

Soit  $\varepsilon(t)$  le processus stochastique qui gouverne le système étudié. Le diagramme ci-dessous représente le diagramme de transition d'états :

Avec :

Tableau 2.1 États du système

| État du système $\varepsilon(t)$ | État de la machine 1 $\varepsilon_1(t)$ | État de la machine 2 $\varepsilon_2(t)$ |
|----------------------------------|-----------------------------------------|-----------------------------------------|
| 1                                | 1                                       | 1                                       |
| 2                                | 1                                       | 0                                       |
| 3                                | 0                                       | 1                                       |
| 4                                | 0                                       | 0                                       |

Et  $q_{12} = q_{34} = q_{10}^2, q_{21} = q_{43} = q_{01}^2, q_{13} = q_{24} = q_{10}^1, q_{31} = q_{42} = q_{01}^1$

On note  $Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \\ q_{41} & q_{42} & q_{43} & q_{44} \end{pmatrix}$  la matrice de transition de ce système. Avec  $q_{ij}$  est le

taux de transition de l'état  $i$  vers l'état  $j$ . et qui vérifie les conditions suivantes :

$$\begin{cases} q_{ij} \geq 0 & \text{si } i \neq j \\ q_{ii} = - \sum_{j \neq i} q_{ij} \end{cases} \quad (2.1)$$

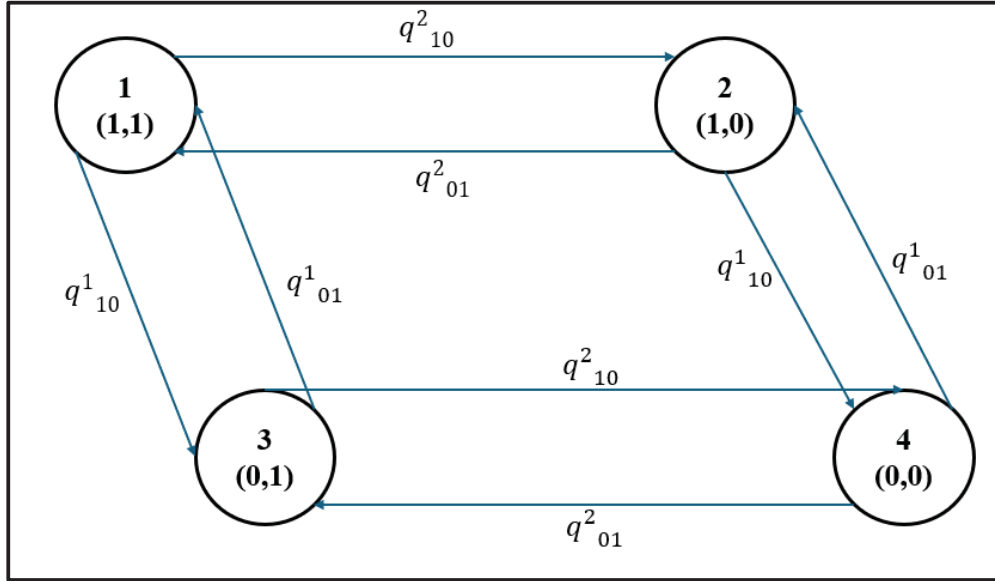


Figure 2.2 Diagramme de transition d'états

De plus, les probabilités de transitions sont données par :

$$P(\varepsilon(t + \delta t) = j / \varepsilon(t) = i) = \begin{cases} q_{ij}(\cdot) \cdot \delta t + o(\delta t) & \text{si } i \neq j \\ 1 + q_{ii}(\cdot) \cdot \delta t + o(\delta t) & \text{si } i = j \end{cases} \quad (2.2)$$

### 2.1.3.2 Dynamique du système étudié

Le système étudié suit deux dynamiques principales. La première concerne le stock dans le système. La seconde concerne les émissions générées par les deux machines composant notre système.

La dynamique de stock est décrite par l'équation suivante :

$$\dot{x}(t) = u_1(t) + u_2(t) - d, x(0) = x \quad (2.3)$$

La dynamique des émissions est décrite par l'équation suivante :

$$\dot{e}(t) = \theta_1 \cdot u_1(t) + \theta_2 \cdot u_2(t) \text{ pour } t \in [T_i, T_{i+1}] \text{ et } e(T_i) = 0, i = 0, \dots, \infty \quad (2.4)$$

Les autorités compétentes fixent un plafond des émissions pendant une période  $T_i$ .

### 2.1.3.3 Espace de commande admissible

L'espace des commandes admissibles est noté  $A(\varepsilon)$ . Il est défini par :

$$A(\varepsilon) = \{(u_1(.), u_2(.)) / 0 \leq u_1(.) \leq U^1_{max}; 0 \leq u_2(.) \leq U^2_{max}\} \quad (2.5)$$

pour tout  $\varepsilon \in \{1, 2, 3, 4\}$ .

### 2.1.3.4 Fonctions de coût

- **Fonction de coût instantanée :**

La fonction de coût instantanée  $g(., e)$  est composée de quatre composantes comme suit :

$$\begin{aligned} g(\varepsilon, x, e) = & \text{coût de stockage} + \text{coût de pénurie} \\ & + \text{coût des émissions} + \text{coût de non qualité} \\ & + \text{coût de production} \end{aligned} \quad (2.6)$$

Avec :

- $\text{coût de stockage} = c^+ \cdot x^+$
- $\text{coût de pénurie} = c^- \cdot x^-$
- $\text{coût de production} = c_{p1} \cdot u_1(t) \cdot \varepsilon_1(t) + c_{p2} \cdot u_2(t) \cdot \varepsilon_2(t)$
- $\text{coût des émissions} :$

Le coût des émissions à la fin d'une période  $T_i$  est défini comme suit :

$$\text{coût des émissions}(T_i) = c_e * \max(0, e(t_{T_i}) - L) ; i = 0, \dots, \infty \quad (2.7)$$

Avec L est la limite fixée par les autorités compétentes au-dessus de laquelle des pénalités s'appliquent.

- **Fonction de coût total actualisé :**

$$J(\varepsilon, x, e) = E \left\{ \int_0^\infty e^{-\rho t} g(.) dt + \sum_{i=1}^N c_e * \max(0, e(t_{T_i}) - L) \mid x(0) = x, e(T_i) = 0, \varepsilon(0) = \varepsilon \right\} \quad (2.8)$$

Sujet aux contraintes sur  $x(t), u_1(.), u_2(.), \varepsilon(t), e(t)$

- **Fonction Valeur :**

La fonction valeur actualisée pour notre problème est :

$$v(\varepsilon, x, e) = \inf_{(u_1, u_2) \in A(\alpha)} J(\varepsilon, x, e) \quad (2.9)$$

La fonction valeur dépend de deux variables d'état à savoir : le niveau de stock  $x$  et le niveau des émissions  $e$  ainsi que l'état des machines de production.

- **Probabilités limites :**

Les probabilités limites  $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$  des modes  $i \in \{1, 2, 3, 4\}$  dans le régime permanent sont l'unique solution du système suivant :

$$\begin{cases} \pi * Q = 0 \\ \sum_{i=1}^4 \pi_i = 1 \end{cases} \quad (2.10)$$

Note système hybride doit satisfaire la condition de faisabilité suivant sur un horizon infini :

$$(U_{max}^1 + U_{max}^2). \pi_1 + U_{max}^1 * \pi_2 + U_{max}^2 \pi_3 \geq d_c \quad (2.11)$$

### 2.1.3.5 Conditions d'optimum

Dans cette sous-section nous développons les équations d'HJB (Hamilton-Jacobi-Bellman) qui correspondent aux conditions d'optimum. À l'optimum l'équation (2.9) peut être approchée par l'équation suivante :

$$\begin{aligned}
\rho v(\varepsilon, x, e) = & \min_{(u_1, u_2) \in A(\varepsilon)} \left\{ g(\varepsilon, x, e) + \frac{\partial v(\varepsilon, x, e)}{\partial x} \cdot (u_1 + u_2 - d) \right. \\
& + \frac{\partial v(\varepsilon, x, e)}{\partial e} \cdot (\theta_1 \cdot u_1 + \theta_1 \cdot u_2) + \sum_{\beta \in B} q_{\alpha\beta} \cdot v(\beta, x, e) \\
& \left. + \sum_{i=1}^N c_e * \max(0, e(t_{Ti}) - L) \right\}
\end{aligned} \tag{2.12}$$

Le développement des équations de HJB pour les différents modes est disponible en Annexe I.

La complexité du problème rend la résolution analytique des équations d'HJB presque impossible, ce qui nécessite l'utilisation de méthodes numériques. Dans la section suivante, nous présenterons l'approche numérique adoptée ainsi que les résultats obtenus.

## 2.2 Approche numérique de résolution et structure de la politique de contrôle

### 2.2.1 Approche numérique

L'objectif principal consiste à trouver la structure de la politique de contrôle optimale applicable à notre système hybride, dans le but de minimiser l'ensemble des coûts engendrés, notamment les coûts de stockage et de pénurie, les coûts associés aux émissions des GES, ainsi que les coûts de fabrication et de refabrication. Cette démarche s'appuie sur les développements théoriques proposés par (Kushner & Dupuis, 1992) concernant les algorithmes de résolution des problèmes de programmation dynamique. Les quatre équations issues de la discrétisation des équations HJB sont présentées en détail dans l'Annexe II.

Dans le cadre de l'approche de contrôle proposée pour la production, les commandes possibles pour les machines de fabrication et de refabrication sont définies comme suit :

#### a. Taux de production maximum $U^1_{max}$ ou $U^2_{max}$

Cette commande impose un fonctionnement à capacité maximal pour l'ensemble des équipements de fabrication et de refabrication. Cette stratégie d'exploitation est généralement mise en œuvre lors de situations critiques où les niveaux de stock atteignent des seuils particulièrement bas, nécessitant une mobilisation complète des ressources productives afin de satisfaire la demande et d'éviter l'accumulation de commandes en backlog.

**b. Taux de production complémentaire :**

Dans cette configuration, une machine fonctionne à un taux équivalent à l'écart entre le taux de demande et la capacité maximale de production de l'autre machine (qu'il s'agisse de la machine de fabrication ou de refabrication).

**c. Arrêt de production :**

L'équipement est mis en mode arrêt afin d'éviter les phénomènes de surstockage et les coûts d'inventaire excessifs lorsque les seuils d'inventaire critiques sont atteints, ou lorsque les contraintes environnementales imposent de privilégier l'utilisation de l'autre machine.

Ces différentes modalités de commande permettent au système de s'adapter de manière dynamique aux fluctuations des niveaux de stock, aux exigences environnementales et à la disponibilité des équipements, assurant ainsi un équilibre optimal entre efficacité économique, impact environnemental et fiabilité opérationnelle.

Dans l'exemple numérique présenté ci-dessous, nous exposons l'ensemble des données (paramètres et structures de coûts) utilisées pour la résolution des équations HJB par le biais de l'approche numérique développée :

Tableau 2.2 Données d'exemple numérique – Coûts

| $c^+$ | $c_e$ | $c^-$ | $c_{man}$ | $c_{rem}$ |
|-------|-------|-------|-----------|-----------|
| 1     | 50    | 350   | 150       | 20        |

Tableau 2.3 Données d'exemple numérique – paramètres

| $q^1_{10}$ | $q^1_{01}$ | $q^2_{10}$ | $q^2_{01}$ | $d$ | $\theta_1$ | $\theta_2$ | $L$ |
|------------|------------|------------|------------|-----|------------|------------|-----|
| 0,01       | 0,16       | 0,12       | 0,16       | 5   | 1.5        | 3          | 150 |

| $\rho$ | $x_{inf}$ | $x_{sup}$ | $e_{inf}$ | $e_{sup}$ | $h_x$ | $h_e$ | $U^1_{max}$ | $U^2_{max}$ |
|--------|-----------|-----------|-----------|-----------|-------|-------|-------------|-------------|
| 0,1    | -10       | 200       | 0         | 210       | 2.5   | 2.5   | 3.5         | 2.5         |

Dans cet exemple, il convient de souligner que le système de production hybride ne peut satisfaire la demande clientèle que lorsque les deux machines (fabrication et refabrication) fonctionnent de manière conjointe ; aucune des machines ne peut répondre à la demande de façon autonome.

Une grille de calcul étendue est définie avec des pas de discrétisation  $h_x$  et  $h_e$  pour les variables d'état. Elle est donnée par les intervalles suivants :

$$G_{xe} = G_x * G_e = \{(-10 < x < 200), (0 < e < 210)\} \quad (2.13)$$

## 2.3 Résultats de l'approche numérique

Cette sous-section expose les résultats numériques issus de la résolution des équations HJB, incluant la fonction de valeur optimale ainsi que la politique de contrôle pour le système de production hybride étudié. Plus précisément, nous présenterons les taux de production optimaux de chaque machine pour les différents états du système, caractérisés par les niveaux d'inventaire et les émissions de gaz à effet de serre. L'ensemble de ces résultats a été obtenu en utilisant le logiciel MATLAB dans sa version 2024.

### 2.3.1 Fonction valeur en mode 1

Dans cette configuration (Mode 1), les deux machines sont disponibles. La Figure 2.3 illustre la fonction de valeur correspondante pour ce scénario. La fonction de valeur pour le Mode 1, où les deux machines sont opérationnelles, représente le coût optimal attendu dans le temps, selon le niveau d'inventaire actuel et les émissions de GES. L'axe horizontal  $x$  correspond au niveau d'inventaire, l'axe  $y$  représente le niveau des émissions de GES, tandis que l'axe vertical (axe  $z$ ) illustre la fonction de valeur pour le mode 1.

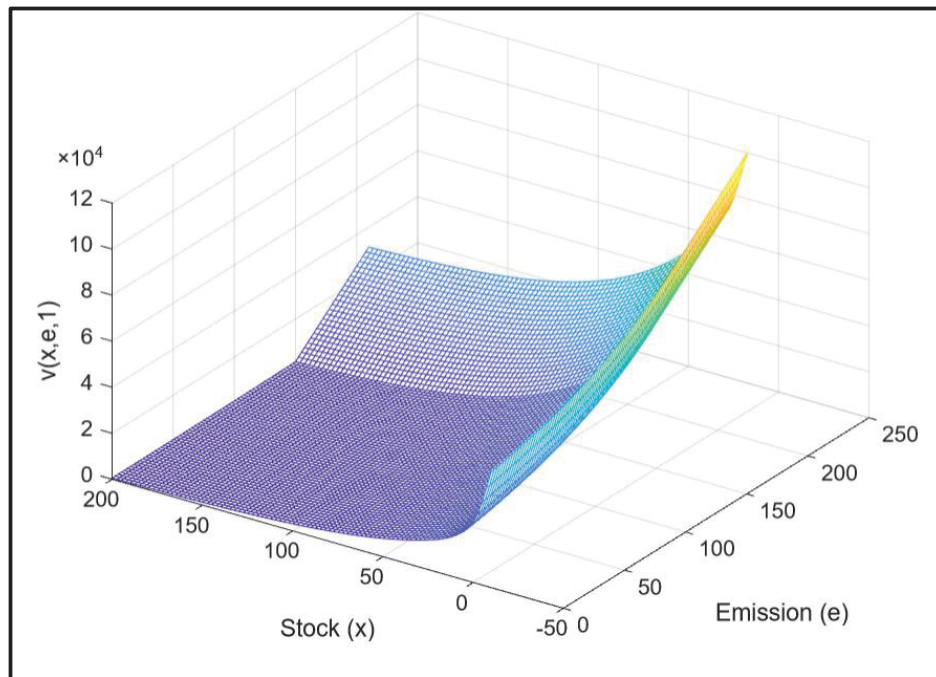


Figure 2.3 Fonction valeur en Mode 1

La forme de la fonction de valeur (axe vertical) peut s'expliquer par les observations suivantes :

Lorsque le système maintient un niveau d'inventaire élevé, il bénéficie d'un tampon qui lui permet de répondre à la demande future sans nécessiter une production immédiate. Ce tampon diminue l'urgence d'intensifier la production, ce qui réduit les coûts associés à l'opération de fabrication de refabrication. Par ailleurs, des niveaux d'inventaire importants permettent d'éviter les coûts de pénurie qui peuvent survenir lorsque la demande clientèle dépasse le stock disponible. En minimisant ces coûts de pénalité, la fonction de valeur globale, qui représente le coût total attendu, diminue.

De plus, avec un inventaire suffisant, le système gagne en flexibilité de production, lui permettant d'optimiser entre fabrication et refabrication selon la rentabilité et les taux d'émission. Cette flexibilité opérationnelle, combinée à la capacité d'alterner entre les deux machines, contribue à réduire le coût global attendu en permettant au système d'optimiser les taux de production et de minimiser les émissions.

À l'inverse, lorsque les niveaux d'inventaire sont faibles, le système subit une pression accrue pour augmenter la production afin de satisfaire la demande et éviter les coûts de pénurie. Cette urgence peut entraîner des taux de production plus élevés, conduisant souvent à des émissions de GES accrues, particulièrement lorsque les deux machines fonctionnent à des capacités élevées (taux de production maximaux). Avec des niveaux d'inventaire réduits,

le système dispose également de moins d'options pour équilibrer production et émissions, nécessitant parfois l'utilisation de la machine de refabrication, qui génère davantage de émissions des GES.

En résumé, un inventaire plus élevé permet au système de fonctionner de manière plus efficace en réduisant les taux de production, évitant les coûts de pénurie et privilégiant la machine à moindres émissions. Par conséquent, le coût global attendu (fonction de valeur) diminue. Cependant, avec un inventaire faible, le système doit réagir de manière plus agressive à la demande, entraînant des taux de production accrus et une plus grande dépendance envers la machine à émissions élevées, ce qui accentue la sensibilité aux coûts liés aux GES.

Dans ce système, le coût d'inventaire demeure relativement modeste comparé aux coûts plus significatifs, tels que les pénalités de pénurie, les coûts de production et les émissions de GES. Le maintien de niveaux d'inventaire élevés garantit une flexibilité essentielle qui aide à optimiser la production et les émissions, ayant un impact plus important sur la fonction de valeur. En conséquence, le coût de possession joue un rôle mineur dans l'influence de la fonction de valeur globale. Cependant, nous ne pouvons maintenir des niveaux d'inventaire élevés indéfiniment, car cela conduirait à une escalade des coûts d'inventaire. Pour équilibrer cet aspect, un niveau de couverture sera défini pour optimiser l'inventaire, garantissant la flexibilité tout en contrôlant les coûts de possession.

### **2.3.2 Taux de production de la machine de fabrication M1 en Mode 1**

Les figures ci-dessous présentent le taux de production de la machine de fabrication M1 en mode 1. L'axe horizontal x représente le niveau d'inventaire, l'axe y représente le niveau des émissions de GES, tandis que l'axe vertical (axe z) correspond au taux de production de la machine M1.

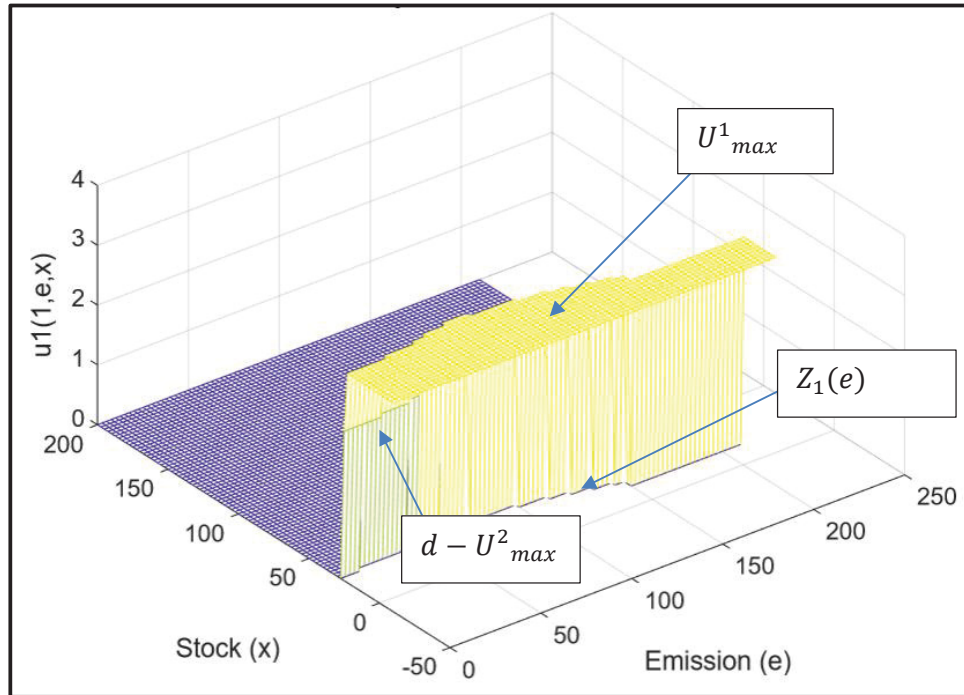


Figure 2.4 Taux de production de la machine de fabrication M1 - Mode 1- vue 3D

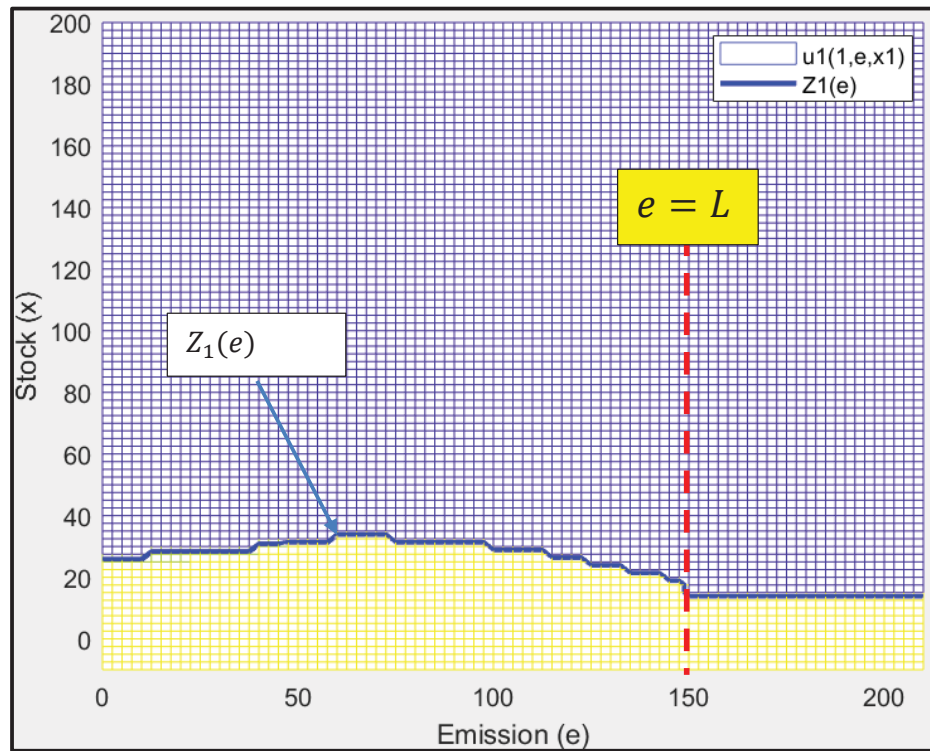


Figure 2.5 Taux de production de la machine de fabrication M1 - Mode 1- projection sur  $(x, e)$

Dans ces figures, qui illustrent la variation du taux de production de la machine de fabrication M1, il est visible que la politique de contrôle fonctionne comme une "politique de seuil de couverture" avec un niveau de couverture dynamique qui s'ajuste selon les niveaux d'émissions générés. Ce niveau de couverture variable est noté  $Z_1(e)$ . De plus, nous pouvons observer que  $Z_1(e)$  demeure constant lorsque le niveau d'émission  $e$  dépasse le seuil  $L$ . Le système atteint un seuil de saturation au niveau de sa flexibilité opérationnelle et se stabilise autour d'une politique établie où les niveaux d'inventaire sont régulés selon un seuil déterminé afin de préserver l'équilibre entre les pénalités de rupture de stock et les coûts liés aux émissions. Cette interprétation s'applique également au comportement de la machine de refabrication.

### 2.3.3 Taux de production de la machine de refabrication M2 en Mode 1

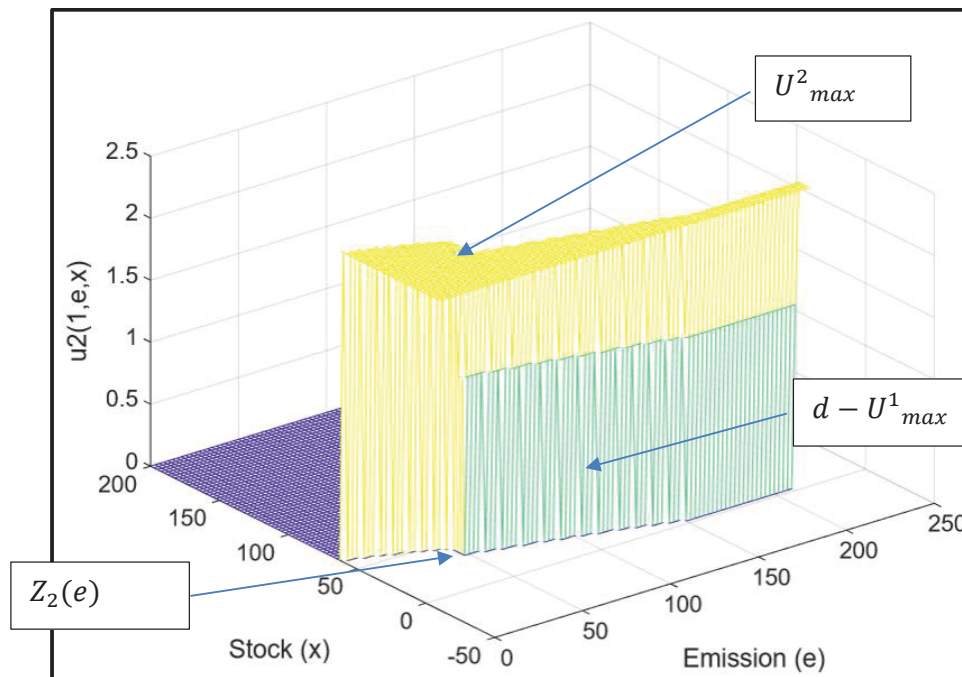


Figure 2.6 Taux de production de la machine M2 - Mode 1 – vue 3D

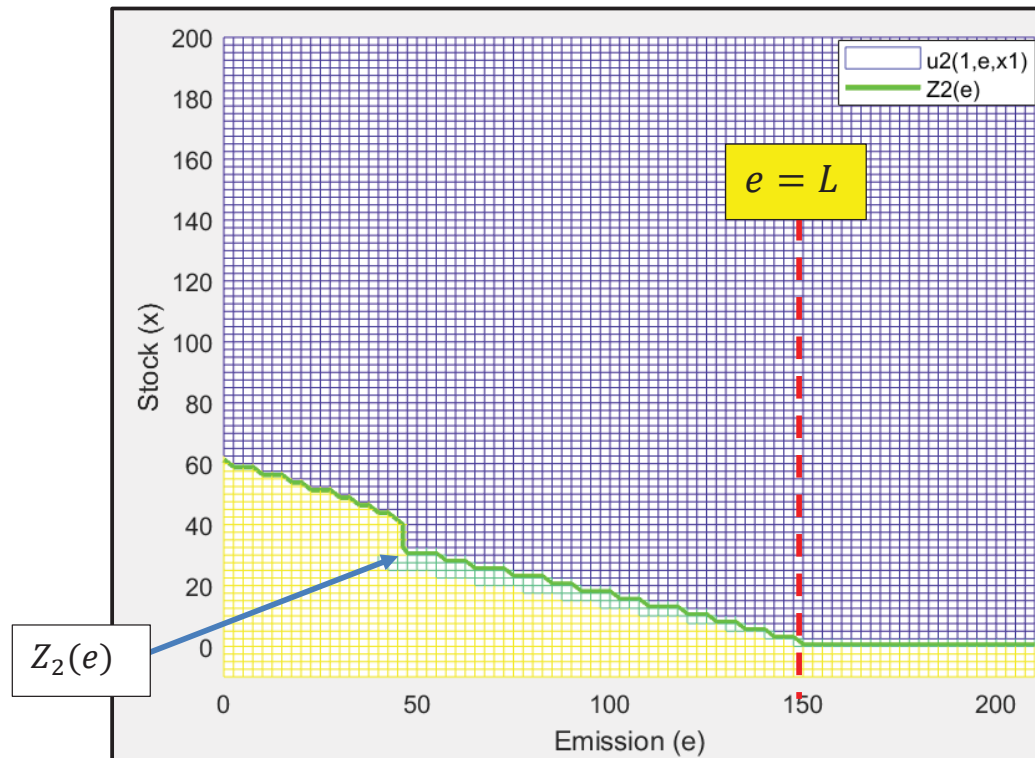


Figure 2.7 Taux de production de la machine M2 - Mode 1 - projection sur  $(x, e)$

Une observation similaire s'applique à la machine de refabrication M2. Les figures illustrent la variation du taux de production de la machine M2, montrant que sa politique de contrôle suit également une "politique de seuil de couverture" avec un niveau de couverture dynamique qui s'ajuste en fonction des niveaux d'émissions produits. Ce niveau de couverture variable est noté  $Z_2(e)$ . En outre, il peut être observé que  $Z_2(e)$  reste constant lorsque le niveau d'émission  $e$  dépasse le seuil  $L$ .

- **Taux de production intégré**

Dans cette sous-section, nous combinons les figures précédentes des taux de production pour démontrer l'intégration des machines M1 et M2, mettant l'accent sur leur fonctionnement coordonné pour satisfaire la demande clientèle. Les figures suivantes illustrent la superposition des deux contours,  $Z_1(e)$  (bleu) et  $Z_2(e)$  (vert), qui représentent respectivement les niveaux de couverture des machines de fabrication et de refabrication.

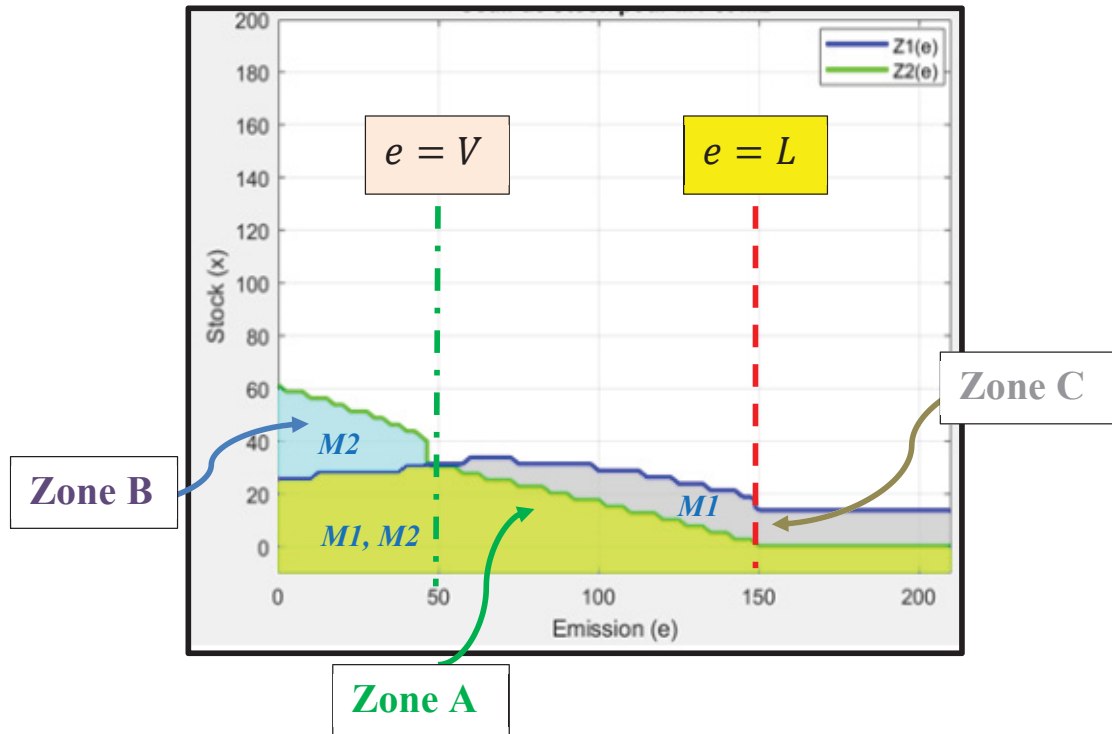


Figure 2.8 Taux de production intégré des machines M1 et M2

Dans la figure ci-dessus, trois zones de production distinctes peuvent être identifiées. Dans chaque zone, les deux machines peuvent soit fonctionner soit demeurer inactives.

- **Zone A** (zone verte) : Dans cette zone, les deux machines collaborent pour satisfaire la demande clientèle.
- **Zone B** (zone bleue) : Lorsque le niveau d'inventaire se situe entre  $Z_2(e)$  et  $Z_1(e)$ , et que le niveau d'émissions demeure sous la limite volontaire  $e = V$ , la priorité est accordée à la machine de refabrication M2, tandis que la machine de fabrication M1 reste inactive. Ceci s'explique par le fait que la machine de refabrication présente des coûts de production inférieurs, et les émissions restent dans des niveaux réduits, permettant l'utilisation de la machine à émissions plus élevées (M2) sans préoccupation majeure.
- **Zone C** (zone grise) : Lorsque le niveau d'inventaire chute en dessous de  $Z_1(e)$ , dépassant  $Z_2(e)$ , et que les émissions excèdent le seuil  $e = V$ , la machine de fabrication M1 passe en priorité, tandis que la machine de refabrication M2 reste inactive. Dans ce cas, malgré son efficacité économique, la machine M2 ne peut être privilégiée en raison du niveau élevé d'émissions, ce qui nécessite l'utilisation de la machine de fabrication à moindres émissions M1.

On constate que  $Z_1(e)$  est inférieur a  $Z_2(e)$  lorsque  $e < V$ , et supérieur a  $Z_2(e)$  lorsque  $e > V$ . En outre,  $Z_1(e)$  et  $Z_2(e)$  deviennent constant lorsque  $e \geq L$ .

## 2.4 Structure of the production control policy

En s'appuyant sur les résultats numériques et l'analyse graphique présentés ci-dessus, nous développons une politique de contrôle de production structurée en zones opérationnelles distinctes. Chaque zone est caractérisée par des seuils d'inventaire spécifiques, la disponibilité des équipements de production, ainsi que les contraintes d'émissions, garantissant une intégration dynamique et une adaptation continue des cadences de fabrication et de refacturation.

$$\begin{aligned}
& \circ \quad e \leq V: \\
& \quad u(1, x, e) = \begin{cases} U^1_{max} & \text{if } x < Z_1(e) \text{ et } \alpha_1(t) = 1 \\ d - U^2_{max} & \text{if } x = Z_1(e) \text{ et } \alpha_1(t) = 1 \\ 0 & \text{si } x > Z_1(e) \text{ ou } \alpha_1(t) = 0 \end{cases} \\
& \quad u(2, x, e) = \begin{cases} U^2_{max} & \text{if } x \leq Z_2(e) \text{ et } \alpha_2(t) = 1 \\ 0 & \text{if } x > Z_2(e) \text{ ou } \alpha_2(t) = 0 \end{cases} \\
& \quad \text{Avec } Z_1(e) < Z_2(e) \\
& \circ \quad V < e < L: \\
& \quad u(1, x, e) = \begin{cases} U^1_{max} & \text{if } x \leq Z_1(e) \text{ et } \alpha_1(t) = 1 \\ 0 & \text{if } x > Z_1(e) \text{ ou } \alpha_1(t) = 0 \end{cases} \\
& \quad u(2, x, e) = \begin{cases} U^2_{max} & \text{if } x < Z_2(e) \text{ et } \alpha_2(t) = 1 \\ d - U^1_{max} & \text{if } x = Z_2(e) \text{ et } \alpha_2(t) = 1 \\ 0 & \text{if } x > Z_2(e) \text{ ou } \alpha_2(t) = 0 \end{cases} \\
& \quad \text{Avec } Z_1(e) > Z_2(e) \\
& \circ \quad e \geq L: \\
& \quad u(1, x, e) = \begin{cases} U^1_{max} & \text{if } x \leq Z_3 \text{ et } \alpha_1(t) = 1 \\ 0 & \text{if } x > Z_3 \text{ ou } \alpha_1(t) = 0 \end{cases} \\
& \quad u(2, x, e) = \begin{cases} U^2_{max} & \text{if } x < Z_4 \text{ et } \alpha_2(t) = 1 \\ d - U^1_{max} & \text{if } x = Z_4 \text{ et } \alpha_2(t) = 1 \\ 0 & \text{if } x > Z_4 \text{ ou } \alpha_2(t) = 0 \end{cases} \\
& \quad \text{Avec } Z_3 > Z_4
\end{aligned} \tag{2.14}$$

Chaque machine est gouvernée par son propre seuil dynamique basé sur les émissions selon un politique de type « Hedging Point Policy HPP », comme décrit ci-dessous :

- Lorsque  $e < V$  :

La machine de fabrication produit à son taux maximum  $U^1_{max}$  jusqu'à ce que le niveau d'inventaire atteigne le niveau de seuil  $Z_1(e)$ , puis elle continue de produire à  $d - U^2_{max}$  pour soutenir la machine de refabrication afin d'atteindre le taux de demande. Au-delà de ce seuil, la machine de fabrication est arrêtée. La machine de refabrication produit à taux maximum  $U^2_{max}$  jusqu'à ce que le niveau d'inventaire atteigne le niveau de seuil  $Z_2(e)$ , puis elle est arrêtée. Il convient de noter que  $Z_1(e)$  et  $Z_2(e)$  sont des seuils dynamiques qui varient en fonction du niveau d'émissions produit et  $Z_1(e) < Z_2(e)$ . Ceci peut s'expliquer par le fait que le coût de la machine de refabrication est inférieur et le niveau d'émissions permet de davantage produire en utilisant cette machine.

- Lorsque  $V < e < L$  :

La machine de fabrication produit à son taux maximum  $U^1_{max}$  jusqu'à ce que le niveau d'inventaire atteigne le niveau de seuil  $Z_1(e)$ . Au-delà de ce seuil, la machine de fabrication est arrêtée. La machine de refabrication produit à taux maximum  $U^2_{max}$  jusqu'à ce que le niveau d'inventaire atteigne le niveau de seuil  $Z_2(e)$ , puis elle continue de produire à  $d - U^1_{max}$  pour soutenir la machine de refabrication afin d'atteindre le taux de demande, puis elle est arrêtée. Dans cette section du niveau d'émissions, nous remarquons que  $Z_1(e) > Z_2(e)$ , ce qui peut s'expliquer par le fait que le niveau d'émissions est plus élevé et la machine M2 ne peut plus être privilégiée.

- Lorsque  $e > L$  :

Les niveaux de seuils deviennent constants avec les mêmes taux de production pour les deux machines que la section précédente du niveau d'émissions.

La politique proposée peut être actualisée pour être plus cohérente et refléter l'aspect physique du système de production. Puisqu'une machine ne peut constituer seule un stock, une fois que le niveau de stock se trouve dans la Zone A (où les deux machines travaillent ensemble), il ne peut atteindre les Zones B ou C ( $d > U^1_{max} > U^2_{max}$ ). Une amélioration consisterait à produire avec les deux machines pour constituer le stock jusqu'à ce qu'il atteigne  $Z_2(e)$  lorsque  $e < V$  et  $Z_1(e)$  lorsque  $e > V$ . Ensuite, la machine de fabrication est arrêtée lorsque  $e < V$ , laissant la machine de refabrication produire seule (Zone B) et ralentissant la baisse du niveau d'inventaire ( $\dot{x} = U^2_{max} - d < 0$ ). Lorsque  $e > V$ , c'est la machine de fabrication qui continue de produire, ralentissant la diminution du niveau d'inventaire ( $\dot{x} = U^1_{max} - d < 0$ ).

Donc, la politique peut être actualisée comme suit :

- Dans la **Zone A** :
  - $e \leq V$ :
 
$$\begin{aligned} u(1, x, e) &= \{ U^1_{max} && \text{jusqu'à } x = Z_2(e) \} \\ u(2, x, e) &= \{ U^2_{max} && \text{jusqu'à } x = Z_2(e) \} \end{aligned}$$
  - $V < e < L$ :
 
$$\begin{aligned} u(1, x, e) &= \{ U^1_{max} && \text{jusqu'à } x = Z_1(e) \} \\ u(2, x, e) &= \{ U^2_{max} && \text{jusqu'à } x = Z_1(e) \} \end{aligned}$$
  - $e \geq L$ :
 
$$\begin{aligned} u(1, x, e) &= \{ U^1_{max} && \text{jusqu'à } x = Z_3 \} \\ u(2, x, e) &= \{ U^2_{max} && \text{jusqu'à } x = Z_4 \} \end{aligned}$$

(2.15)

La politique reste la même comme décrit dans l'équation 2.14.

## 2.5 Analyse de sensibilité

Cette section développe une analyse de sensibilité visant à évaluer la robustesse de la politique de contrôle proposée ainsi que sa réactivité aux variations des paramètres système et des coûts. Nous examinons particulièrement l'impact des modifications du coût de stockage  $c^+$ , du coût de pénurie  $c^-$ , et du coût des émissions  $c_e$ . Concernant les paramètres système, nous analysons les effets des changements du taux d'émissions et des taux de panne des machines de fabrication et de refabrication.

Cette analyse s'appuie sur des représentations graphiques qui superposent les taux de production des deux machines selon différentes valeurs des coûts et paramètres considérés. Ces visualisations permettent d'observer l'impact individuel de chaque coût et paramètre sur le taux de production de chaque machine et d'évaluer leur effet sur la performance collaborative des deux équipements. Dans les figures suivantes, nous présentons les niveaux de seuil d'inventaire en fonction des niveaux d'émissions. La ligne continue représente le niveau de seuil pour la machine de fabrication, tandis que la ligne pointillée indique le niveau de seuil pour la machine de refabrication.

### 2.5.1 Coût de stockage

Comme illustré dans la Figure 2.9, lorsque le coût de stockage  $c^+$  est faible, le système permet des niveaux d'inventaire plus élevés, résultant en des valeurs de seuil élevées  $Z_1(e)$  et

$Z_2(e)$  pour les machines de fabrication et de refabrication, respectivement. À l'inverse, lorsque le coût de stockage augmente, le système réagit en abaissant ces niveaux de seuil pour éviter des coûts de possession excessifs. Cette augmentation du coût de stockage affecte également l'interaction entre les deux machines : le point de basculement entre machines se déplace vers le bas, signifiant que la machine de refabrication a moins de priorité relativement à la machine de fabrication. En conséquence, la Zone B se rétrécit, car le niveau de seuil  $Z_2(e)$ , qui définit cette zone, diminue. Cette réduction limite la capacité du système à tirer pleinement avantage du coût de production inférieur de la machine de refabrication.

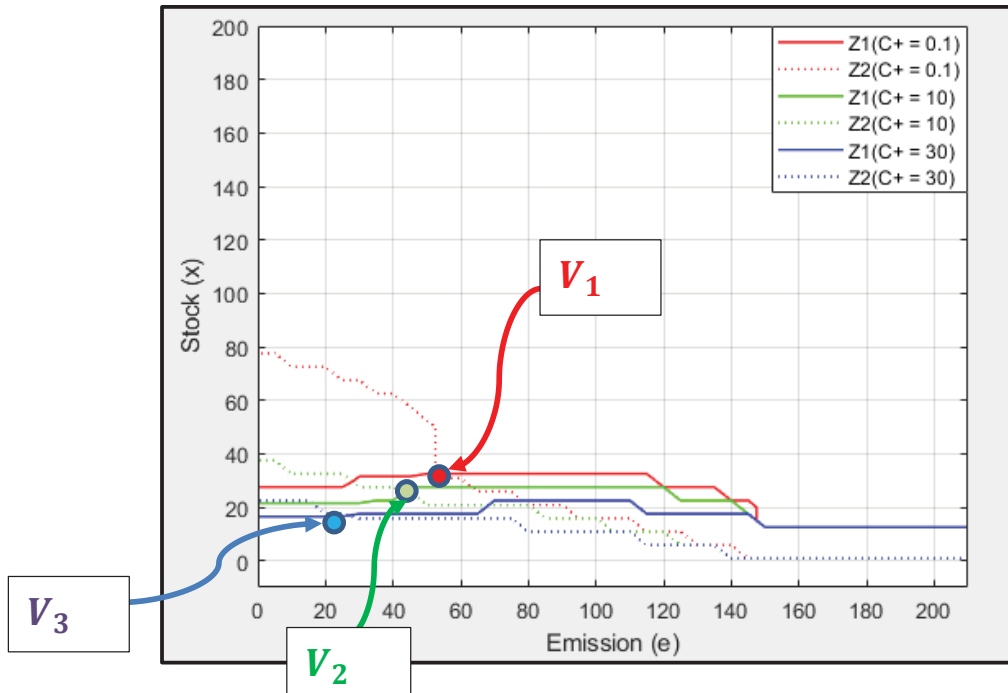


Figure 2.9 Impact du coût de stockage  $c^+$  sur les machines M1 et M2

### 2.5.2 Coût de pénurie

Une analyse similaire a été menée concernant le coût de pénurie, révélant un comportement inverse à celui du coût de possession sur les niveaux de seuil de couverture.

### 2.5.3 Coût des émissions

Comme présenté dans la Figure 2.10, lorsque le coût des émissions  $c_e$  est faible, le système établit un seuil plus élevé  $Z_2(e)$  pour la machine de refabrication, lui permettant de tirer avantage de ses coûts de production inférieurs. Cela signifie que la machine de refabrication peut fonctionner plus fréquemment à son taux maximum, réduisant la dépendance envers la

machine de fabrication. Ceci résulte en un point de basculement  $V$  plus élevé en termes de niveau d'émissions, comparé aux cas où le coût des émissions est plus important. À l'inverse, lorsque le coût des émissions augmente, le système abaisse le seuil  $Z_2(e)$  pour limiter l'utilisation de la machine de refabrication à émissions plus élevées. Cet ajustement résulte en une augmentation du seuil de la machine de fabrication  $Z_1(e)$ , déplaçant davantage de charge de production vers la machine à émissions moindres. Par conséquent, la Zone B, où la machine de refabrication a la priorité, se rétrécit, tandis que la Zone C, où la machine de fabrication seule est privilégiée en raison d'émissions élevées, s'étend pour accommoder l'augmentation du coût des émissions et équilibrer la charge de production et émissions.

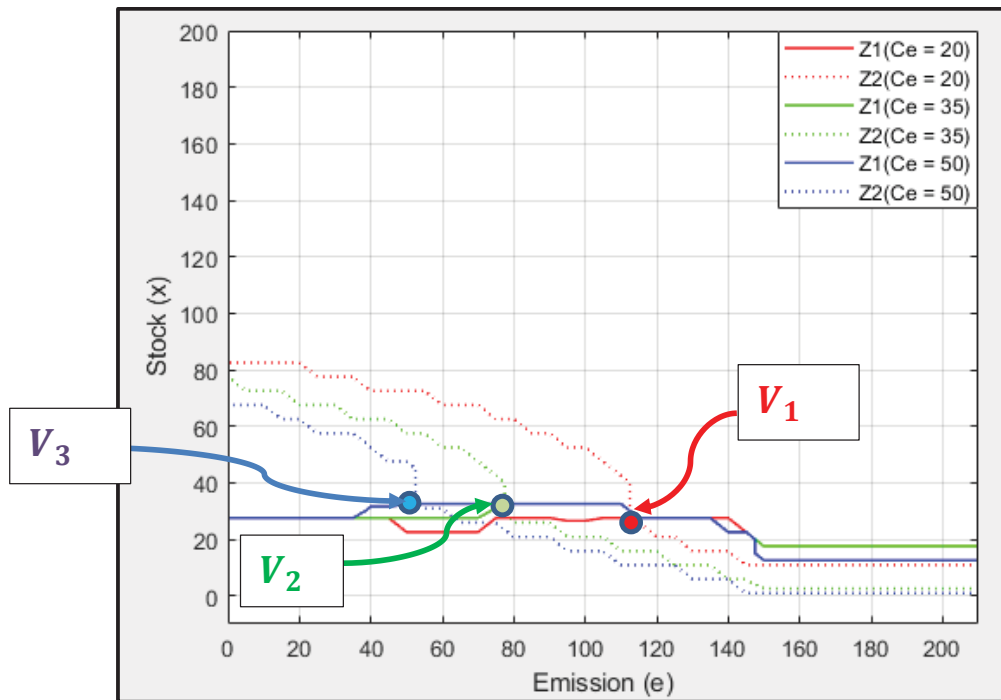


Figure 2.10 Impact du coût des émissions  $c_e$  sur les machines M1 et M2

#### 2.5.4 Taux de panne

Dans les figures suivantes, nous présentons l'analyse de sensibilité pour les taux de panne variables de chaque machine au sein du système de production hybride. Nous nous concentrons sur les taux de panne, car ils indiquent la fiabilité des machines de fabrication et de refabrication, ce qui impacte directement la performance globale du système. Le taux de panne de la machine de fabrication correspond au taux de transition du mode 1 (où les deux équipements sont opérationnels) vers le mode 3 (caractérisé par l'indisponibilité de la machine de fabrication). De la même manière, le taux de panne de la machine de

refabrication s'exprime par le taux de transition du mode 1 vers le mode 2 (où la machine de refabrication devient indisponible).

#### 2.5.4.1 Machine de fabrication M1

Pour la machine de fabrication, comme illustrés dans la figure 2.11, nous observons que lorsque sa fiabilité est élevée (c'est-à-dire un taux de panne faible), le système permet un niveau d'inventaire modéré. Ceci se reflète dans des niveaux de seuil inférieurs  $Z_1(e)$  et  $Z_2(e)$ , ainsi qu'une Zone A plus petite. À l'inverse, lorsque la fiabilité est faible (c'est-à-dire un taux de panne élevé), le système répond en augmentant le niveau d'inventaire, ce qui élève les niveaux de seuil  $Z_1(e)$  et  $Z_2(e)$ , conduisant à une expansion de la Zone A.

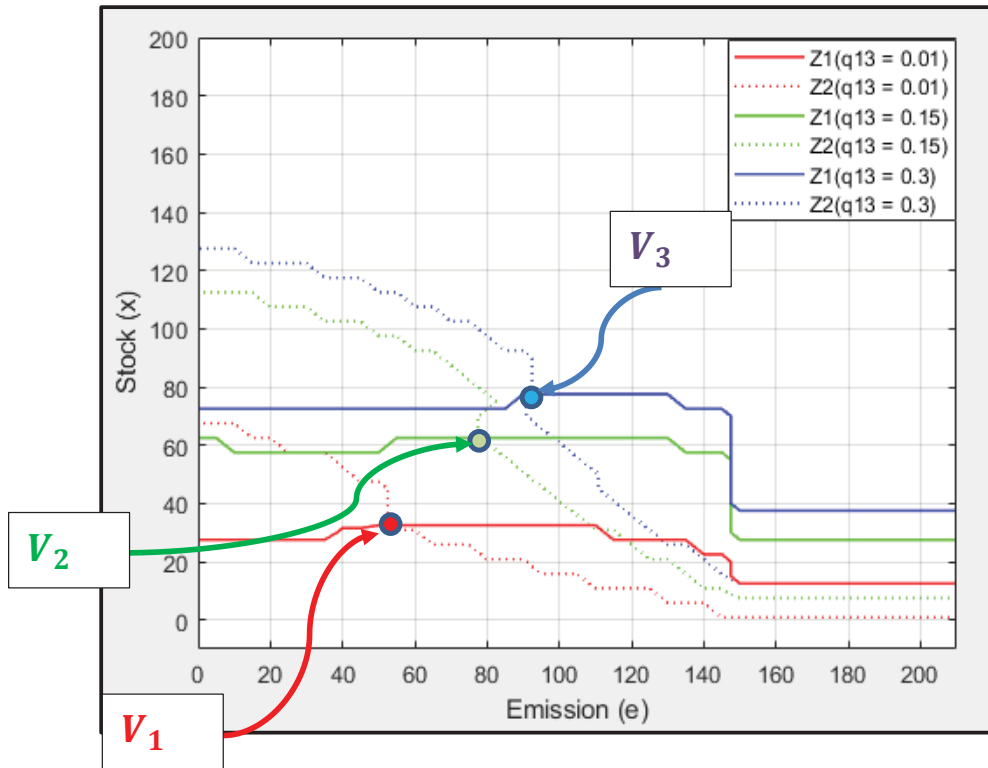


Figure 2.11 Impact du taux de panne  $q_{13}^1 = q_{13}$  sur les machines M1 et M2

#### 2.5.4.2 Machine de refabrication M2

La même observation s'applique à la machine de refabrication : lorsque la fiabilité est élevée (c'est-à-dire que le taux de panne est faible), le système maintient un niveau de stock modéré afin de minimiser les coûts de stockage inutiles. À l'inverse, lorsque la fiabilité est faible, le système augmente les niveaux de stock afin d'atténuer le risque de pénurie.

## 2.5.5 Taux de réparation

### 2.5.5.1 Machine de refabrication M2

Comme l'illustre la figure 2.12, lorsque le taux de réparation de la machine M2 est faible (c'est-à-dire que les réparations prennent plus de temps), le système réagit en augmentant les niveaux de couverture pour les deux machines afin d'éviter d'éventuelles pénuries. À l'inverse, lorsque le taux de réparation est élevé (c'est-à-dire que les réparations sont rapides), le système réduit les niveaux de stock afin d'éviter des coûts de stockage excessifs.

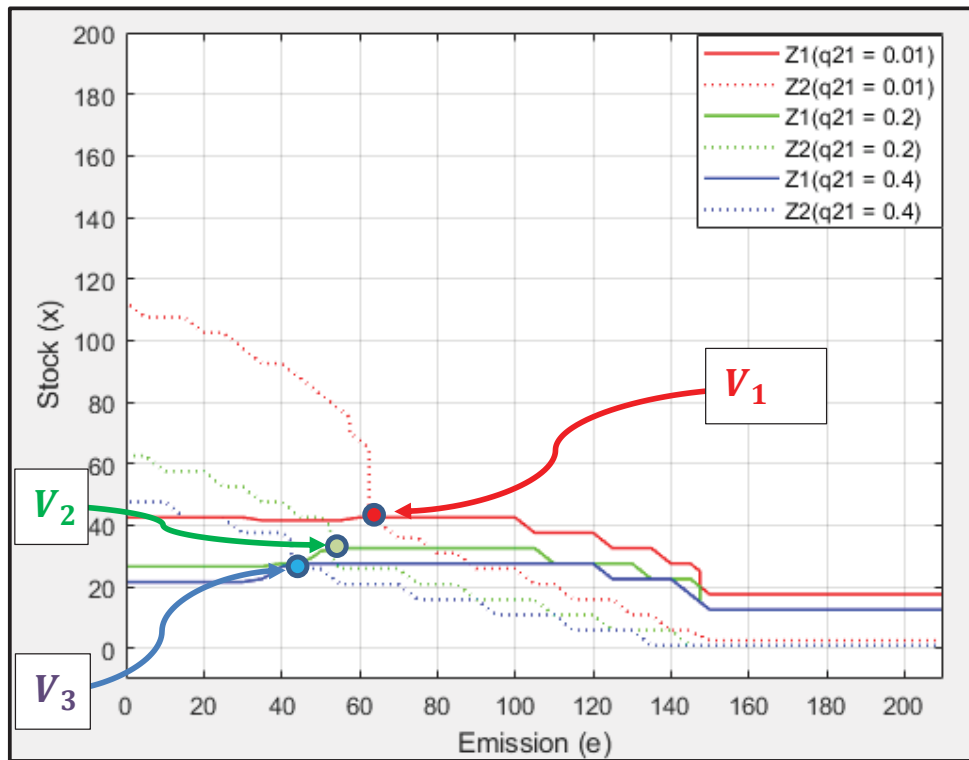


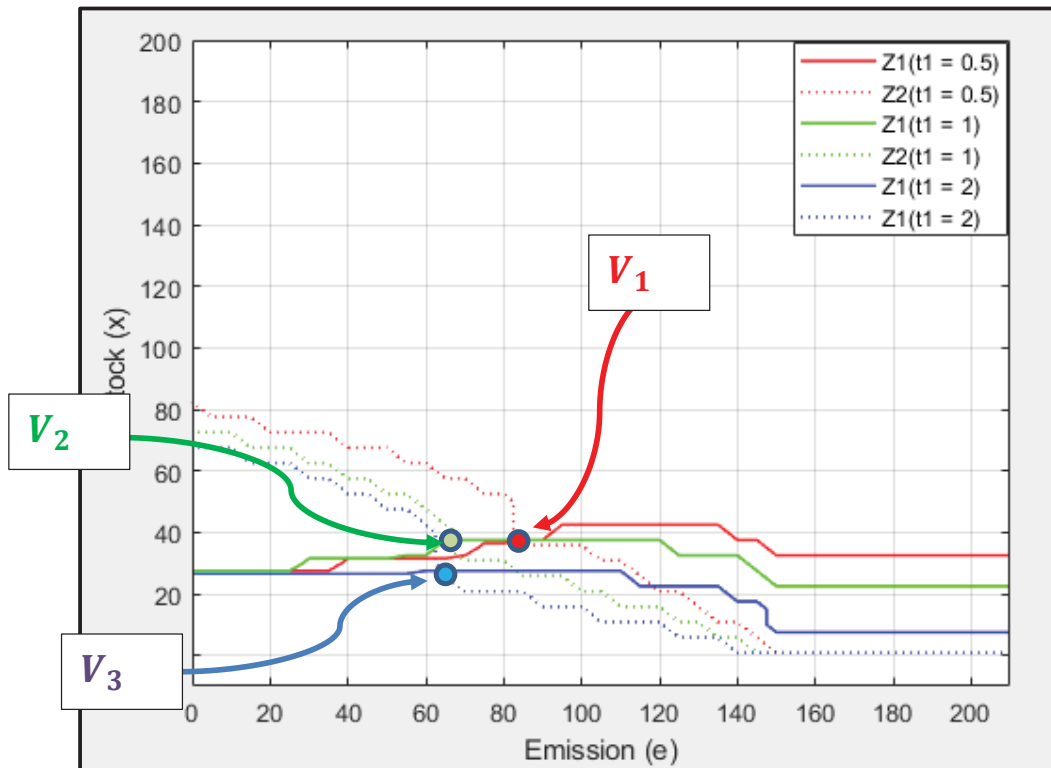
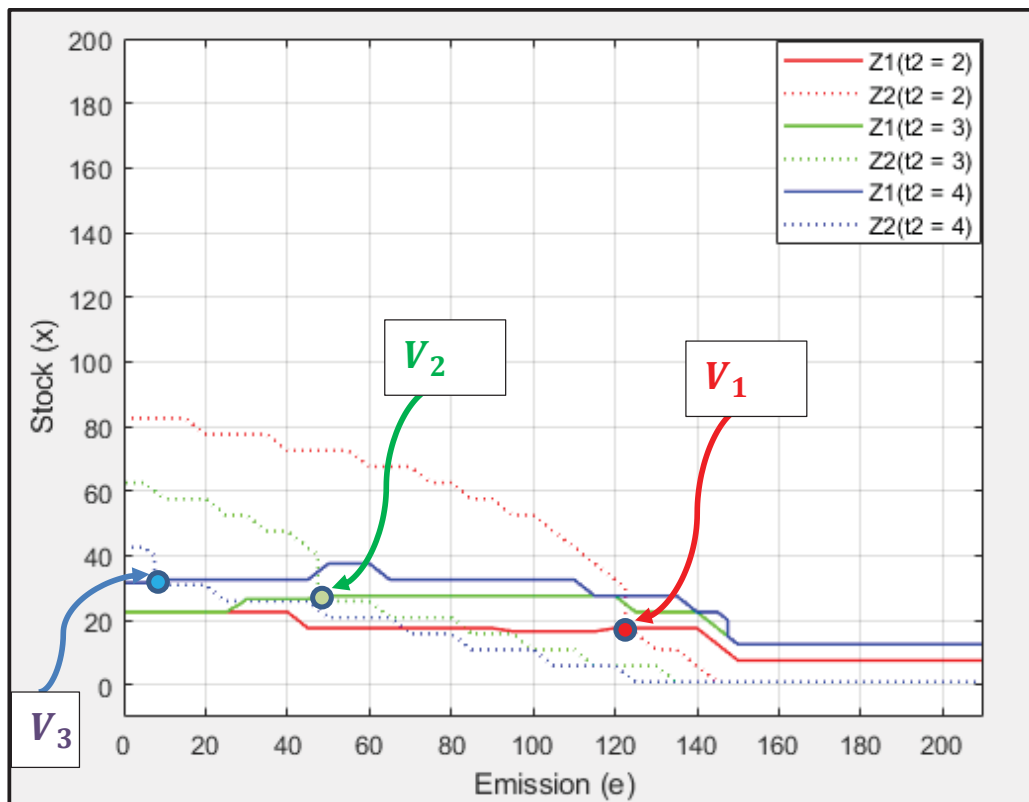
Figure 2.12 Impact de  $q^2_{01} = q_{21}$  sur les machines M1 et M2

### 2.5.5.2 Machine de fabrication M1

Il en est de même pour la machine M1, où un faible taux de réparation augmente le niveau de couverture, tandis qu'un taux de réparation élevé le réduit afin d'éviter des coûts de possession excessifs.

### 2.5.6 Taux des émissions

Comme dans la sous-section précédente où nous avons présenté l'analyse de sensibilité pour le taux de panne, dans cette sous-section, nous examinons l'impact de l'indice d'émissions sur le système de production hybride. Dans la figure ci-dessous, nous observons que lorsque l'indice d'émissions pour la machine de fabrication est faible, elle produit moins d'émissions. Cela lui permet de produire davantage tout en permettant également à la machine de refabrication une capacité additionnelle pour fonctionner et bénéficier de ses coûts de production inférieurs, car le risque de dépasser les limites d'émissions diminue. À l'inverse, lorsque l'indice d'émissions augmente, la machine de refabrication est attribuée moins de capacité opérationnelle, avec la priorité se déplaçant vers la machine de fabrication. Ce passage est évident dans un point de basculement plus bas en termes de niveaux d'émissions. Pour la machine de refabrication, un indice d'émissions plus élevé déplace la priorité vers la machine de fabrication, comme indiqué par la contraction de la Zone B, un point de basculement plus bas en termes de niveau d'émissions, un niveau de seuil réduit  $Z_2(e)$  pour la machine de refabrication, et un niveau de seuil élevé  $Z_1(e)$  pour la machine de fabrication. Cependant, lorsque l'indice d'émissions diminue, la machine M2 se voit allouer davantage de capacité opérationnelle pour tirer avantage de ses coûts de production inférieurs.

Figure 2.13 Impact de  $\theta_1$  sur les machines M1 et M2Figure 2.14 Impact de  $\theta_2$  sur les machines M1 et M2

### 2.5.7 Taux de production maximum

Dans cette sous-section, nous analysons l'influence des taux de production maximums des machines de fabrication et de refabrication sur l'architecture de la politique de contrôle.

Comme l'illustre la figure 2.15, nous constatons que lorsque le taux de production maximum de la machine de fabrication  $U^1_{max}$  est faible, le système compense cette limitation en autorisant un niveau d'inventaire plus élevé, rehaussant ainsi les seuils critiques pour les deux machines M1 et M2 afin de prévenir les ruptures de stock. À l'inverse, lorsque  $U^1_{max}$  s'accroît, le système réagit en réduisant les seuils d'inventaire, étant donné que M1 peut désormais constituer du stock plus rapidement, diminuant ainsi le risque de pénuries. Dans ce contexte, la priorité se déplace vers la minimisation des coûts de possession excessifs. Cette dynamique se reflète également dans le point de basculement, qui diminue tant au niveau des stocks que des émissions pour maîtriser les coûts de possession et privilégier l'utilisation de la machine à faibles émissions M1. Les mêmes observations s'appliquent à la machine de refabrication, comme le démontre la figure 2.16, concernant l'impact du taux de production maximum  $U^2_{max}$  sur les performances du système de production hybride.

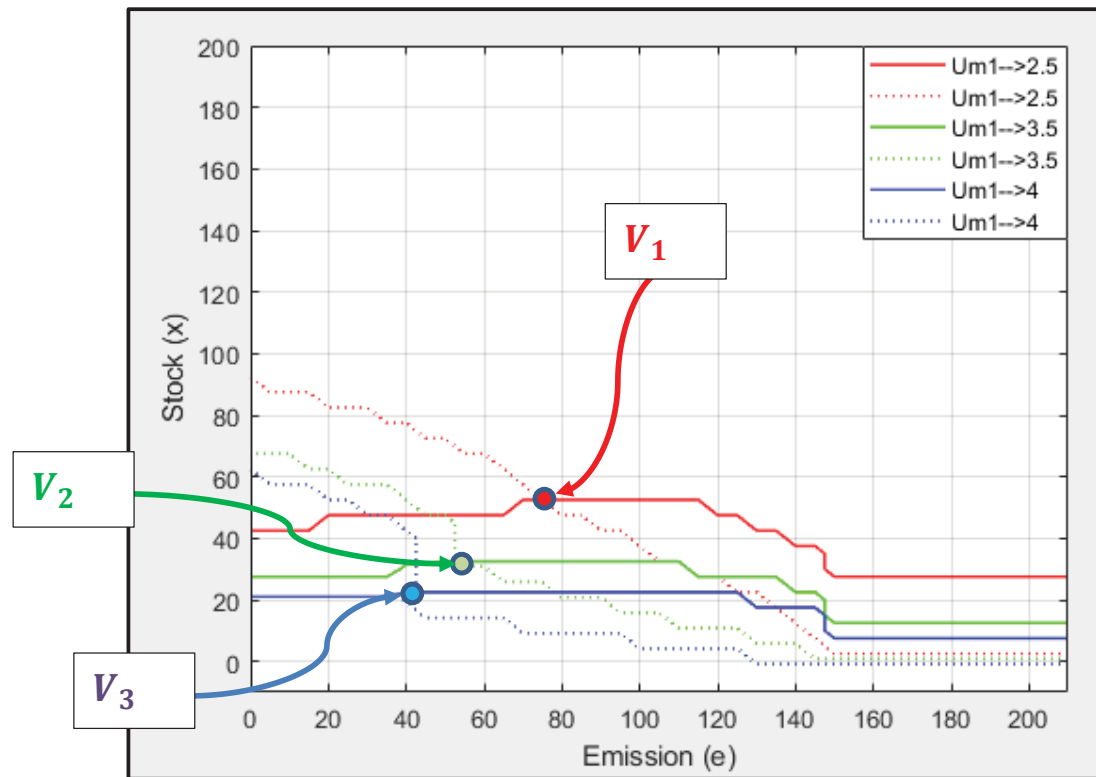


Figure 2.15 Impact de  $U^1_{max}$  sur les machines M1 et M2

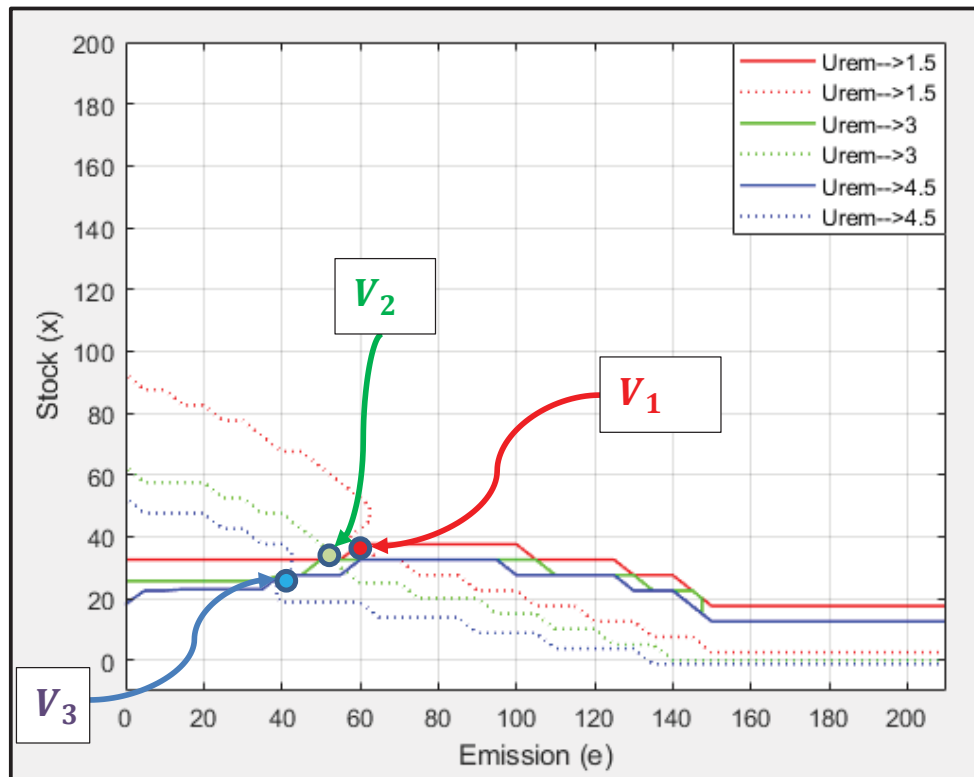


Figure 2.16 Impact de  $U^2_{max}$  sur les machines M1 et M2

Afin de faire le lien entre les conclusions théoriques et les applications pratiques, la section 6 présente une étude comparative visant à évaluer les performances de la politique proposée par rapport aux politiques existantes issues de la littérature. Cette section met en évidence les principales conclusions en matière de gestion et les implications pratiques, en se concentrant sur la manière dont les décideurs peuvent mettre en œuvre efficacement la politique proposée dans des systèmes de production hybrides réels afin d'atteindre un équilibre optimal entre les coûts, les émissions et la fiabilité opérationnelle tout en s'adaptant aux incertitudes du système.

## 2.6 Étude comparative et perspectives managériales

### 2.6.1 Étude comparative

Cette section évalue l'avantage économique de la politique de couverture environnementale dynamique développée en comparant ses performances à celles de deux politiques de référence issues de la littérature. Le coût total encouru, qui comprend les coûts liés aux

stocks, aux commandes en attente, à la production et aux émissions, est utilisé comme principal critère de comparaison. Les politiques comparées sont les suivantes :

- **Politique I (Politique proposée):** La nouvelle politique qui adapte les taux de production selon les niveaux d'émissions, les seuils de stock et la disponibilité des machines pour optimiser le coût total et l'impact environnemental.
- **Politique II** (Afshar-Bakeshloo et al., 2018): Cette politique applique une politique de point de couverture (HPP) pour la fabrication lorsque les émissions sont inférieures à un certain seuil et passe à une stratégie de production basée sur une technologie à faibles émissions (LET) lorsque les émissions dépassent la limite réglementaire. La politique fonctionne avec deux seuils statiques  $Z_1$  et  $Z_2$ , sans adaptation dynamique. La politique, adaptée à notre contexte, est donnée par l'équation suivante :

$$\begin{aligned}
 & \circ \quad e < L: \text{HEM (M2) avec seuil de couverture } Z_1 \\
 & \quad u(2, x, e) = \begin{cases} U^2_{max} & \text{if } x < Z_1 \text{ et } \alpha_2(t) = 1 \\ d - U^1_{max} & \text{if } x = Z_1 \text{ et } \alpha_2(t) = 1 \\ 0 & \text{si } x > Z_1 \text{ ou } \alpha_2(t) = 0 \end{cases} \\
 & \circ \quad e \geq L : \text{LEM (M1) avec seuil de couverture } Z_2
 \end{aligned} \tag{2.16}$$

$$u(1, x, e) = \begin{cases} U^1_{max} & \text{if } x < Z_2 \text{ et } \alpha_2(t) = 1 \\ d - U^2_{max} & \text{if } x = Z_2 \text{ et } \alpha_2(t) = 1 \\ 0 & \text{if } x > Z_2 \text{ ou } \alpha_2(t) = 0 \end{cases}$$

Avec  $Z_1 > Z_2$

- **Politique III** (Entezaminia et al., 2020): Cette approche développe une politique de seuil de couverture (HPP) à double niveau, hiérarchisant l'utilisation des différentes machines de production selon leurs profils d'émissions respectifs. Elle établit un système de basculement déterministe entre les machines à émissions importantes (HEM) et celles à faibles émissions (LEM). Sa formalisation mathématique est détaillée dans l'équation (19) :

$$\begin{aligned}
 & \circ \quad x < Z_2: \\
 & \quad u(1, x, e) = U^1_{max} \quad \text{and} \quad u(2, x, e) = U^2_{max} \\
 & \circ \quad x \geq Z_2 \text{ et } e < L : \text{HPP avec seuil } Z_1 \text{ en utilisant seulement M2} \\
 & \quad u(2, x, e) = \begin{cases} U^2_{max} & \text{if } x < Z_1 \text{ et } \alpha_2(t) = 1 \\ d - U^1_{max} & \text{if } x = Z_1 \text{ et } \alpha_2(t) = 1 \\ 0 & \text{if } x > Z_1 \text{ ou } \alpha_2(t) = 0 \end{cases} \\
 & \circ \quad x \geq Z_2 \text{ et } e \geq L : \text{HPP avec seuil } Z_1 \text{ en utilisant seulement M1}
 \end{aligned} \tag{2.17}$$

$$u(1, x, e) = \begin{cases} U^1_{max} & \text{if } x < Z_1 \text{ et } \alpha_1(t) = 1 \\ d - U^2_{max} & \text{if } x = Z_1 \text{ et } \alpha_1(t) = 1 \\ 0 & \text{if } x > Z_1 \text{ ou } \alpha_1(t) = 0 \end{cases}$$

Avec  $Z_1 > Z_2$

Afin de garantir une comparaison équitable, toutes les politiques ont été testées selon les mêmes paramètres, notamment les coûts de production, de stockage, de pénurie et des émissions. Le coût total encouru a été extrait des résultats de la fonction de valeur en mode 1, comme le montre la figure 2.17, où la fonction de valeur représente le coût à long terme prévu dans le cadre de chaque politique.

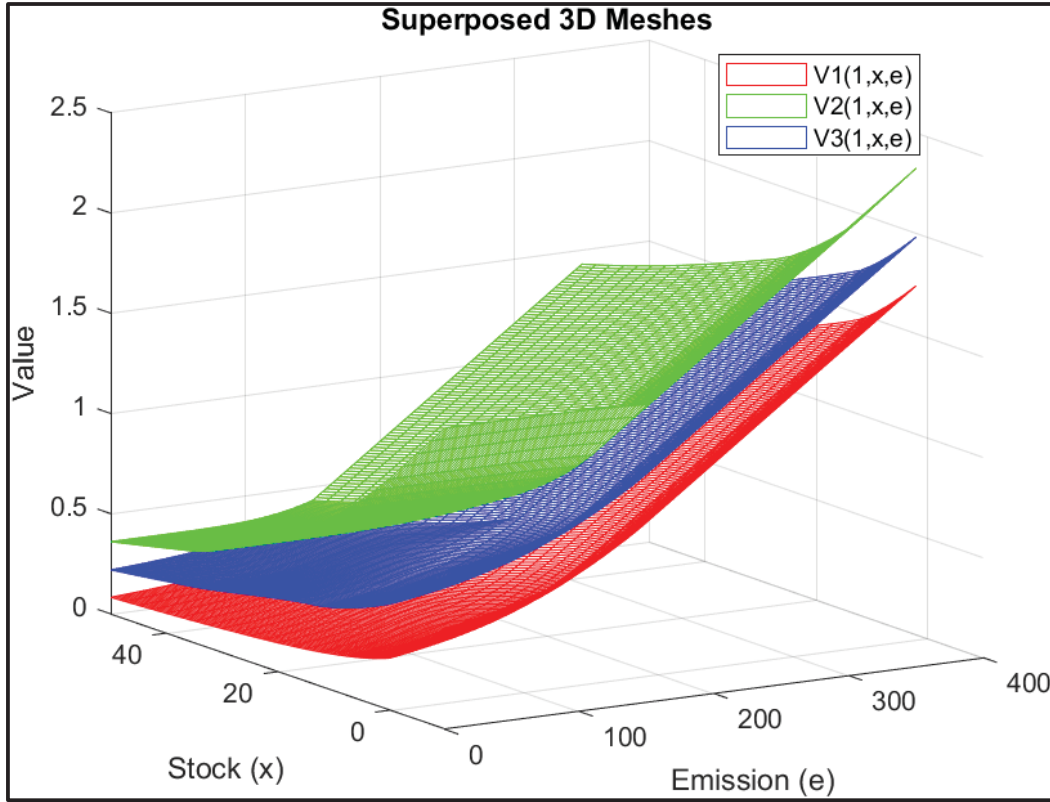


Figure 2.17 Fonction valeur pour les politiques comparées

La Figure 2.17 illustre la superposition des courbes 3D des fonctions de valeur pour l'ensemble des politiques étudiées :

- V1 (**Rouge**) : Politique proposée (Politique I)
- V2 (**Vert**) : (Afshar-Bakeshloo et al., 2018) (Politique II)
- V3 (**Bleu**) : (Entezaminia et al., 2020)(Politique III)

L'analyse de la figure ci-dessus révèle que la politique proposée (V1, surface rouge) génère des coûts inférieurs aux politiques II (V2, surface verte) et III (V3, surface bleue) sur

l'ensemble des niveaux d'inventaire et d'émissions. Cette performance économique améliorée résulte principalement de l'adaptation dynamique des seuils d'inventaire et des décisions de production basées sur les niveaux d'émissions.

La Politique II s'avère la moins efficace des trois approches, comme en témoigne sa fonction de valeur plus élevée sur l'ensemble du domaine. Cette inefficacité découle de son mécanisme de commutation simpliste qui ne profite pas pleinement de la flexibilité du système. La politique repose sur un seuil d'émissions fixe pour alterner entre machines M1 et M2, conduisant à des décisions de production sous-optimales lorsque les émissions demeurent dans des limites acceptables.

La Politique III, bien qu'offrant de meilleures performances que la Politique II, présente encore des coûts totaux supérieurs à la politique proposée en raison de sa structure de seuils rigide qui ne s'ajuste pas dynamiquement aux contraintes d'émissions.

Une limitation critique des deux politiques de référence réside dans l'absence d'une limite volontaire d'émissions  $V$  dans leurs cadres décisionnels. Contrairement à la politique proposée, qui intègre un seuil volontaire pour ajuster proactivement les seuils de production avant d'atteindre la limite réglementaire  $L$ , les Politiques II et III réagissent uniquement lorsque les émissions dépassent un seuil prédéfini. Cette approche réactive entraîne des ajustements de production brusques, générant des inefficacités dans la gestion des stocks et la planification de production.

Les résultats confirment que la politique proposée offre aux décideurs une stratégie plus robuste et économiquement efficace pour optimiser la production dans les systèmes hybrides, particulièrement dans des environnements incertains et contraints par les émissions. L'équilibre optimal entre efficacité économique, contrôle des émissions et fiabilité système constitue l'avantage principal de cette approche par rapport aux méthodes conventionnelles.

## **2.6.2 Analyse comparative sous variations de coût de stockage**

Dans cette sous-section, nous comparons le coût total encouru au niveau des seuils d'inventaire pour les politiques considérées selon différentes valeurs de coût d'inventaire. Le tableau 2.4 présente ces résultats, en soulignant la manière dont chaque politique réagit aux variations des coûts de détention des stocks.

Tableau 2.4 Performance des politiques pour différentes valeurs du coût de stockage

| $c^+$                | Coût total encouru |       |       |       |
|----------------------|--------------------|-------|-------|-------|
|                      | 1                  | 5     | 10    | 20    |
| <i>Politique I</i>   | 63803              | 64419 | 65070 | 67385 |
| <i>Politique II</i>  | 85281              | 86465 | 87710 | 91804 |
| <i>Politique III</i> | 79692              | 80600 | 81550 | 84815 |

Le tableau 2.4 démontre que la Politique I (politique développée) surpasse les politiques de référence sur l'ensemble des valeurs de coût de stockage  $c^+$ , maintenant le coût total le plus bas lorsque  $c^+$  varie de 1 à 20. Les deux politiques de référence s'appuient sur des règles de seuils rigides et ne parviennent pas à ajuster optimalement le niveau d'inventaire selon l'évolution des conditions de coût. Par conséquent, elles tendent à stocker davantage d'inventaires par mesure de précaution, entraînant des coûts de possession significativement plus élevés lorsque  $c^+$  augmente. L'écart de performance s'amplifie à mesure que  $c^+$  s'accroît, confirmant que la Politique I offre un meilleur contrôle des coûts en minimisant l'accumulation d'inventaire non nécessaire. Cette capacité d'optimisation dynamique constitue un avantage déterminant dans la gestion des systèmes de production hybrides soumis à des contraintes économiques et environnementales variables.

### 2.6.3 Analyse comparative sous variations de coût de pénurie

Dans cette sous-section, nous comparons le coût total encouru au niveau des seuils d'inventaire pour les politiques envisagées selon différentes valeurs de coût de pénurie. Le tableau 2.5 présente ces résultats, montrant comment chaque politique se comporte à mesure que le coût de pénurie augmente.

Tableau 2.5 Performance des politiques pour différentes valeurs du coût de pénurie

| $c^-$                | Coût total encouru |       |       |        |
|----------------------|--------------------|-------|-------|--------|
|                      | 200                | 250   | 350   | 450    |
| <i>Politique I</i>   | 63048              | 63803 | 64213 | 64394  |
| <i>Politique II</i>  | 79508              | 85281 | 97450 | 109090 |
| <i>Politique III</i> | 77734              | 79692 | 84580 | 89354  |

Tableau 2.5 présente le coût total engendré par chaque politique à mesure que le coût du retard  $c^-$  augmente de 200 à 450. Comme prévu, le coût total augmente pour toutes les politiques en raison de l'augmentation des pénalités associées à la demande non satisfaite. Cependant, la politique proposée (**Politique I**) maintient le coût total le plus bas pour toutes les valeurs de  $c^-$ , démontrant ainsi sa capacité supérieure à anticiper et à atténuer les risques de pénurie grâce à des ajustements dynamiques des seuils d'inventaire. La politique II affiche la plus forte augmentation des coûts, ce qui indique une réactivité limitée à l'augmentation des pénalités pour pénurie en raison de sa structure fixe et réactive. La politique III obtient de meilleurs résultats que la politique II, mais reste moins performante par rapport à la politique proposée. L'écart de performance croissant, en particulier pour les valeurs élevées de  $c^-$ , confirme que la politique proposée offre une meilleure protection contre les pénuries, minimisant ainsi les scénarios coûteux liés aux retards.

#### 2.6.4 Analyse comparative sous variations de coût des émissions

Dans cette sous-section, nous comparons le coût total encouru au niveau des seuils d'inventaire pour les politiques envisagées en fonction de différentes valeurs de coût des émissions  $c_e$ . Le tableau 2.6 présente les résultats, en mettant en évidence les performances de chaque politique à mesure que le coût des émissions augmente.

Tableau 2.6 Performance des politiques pour différentes valeurs du coût des émissions

| $c_e$                | Coût total encouru |       |       |        |
|----------------------|--------------------|-------|-------|--------|
|                      | 20                 | 30    | 50    | 70     |
| <i>Politique I</i>   | 30659              | 41918 | 63803 | 84729  |
| <i>Politique II</i>  | 59227              | 68000 | 85281 | 101310 |
| <i>Politique III</i> | 44770              | 56356 | 79692 | 101750 |

Le tableau 2.6 expose l'évolution des coûts totaux pour les trois politiques lorsque le coût des émissions  $c_e$  progresse de 20 à 70. La politique proposée (Politique I) démontre la performance la plus stable, conservant les coûts totaux les plus avantageux sur l'ensemble des valeurs de coût d'émissions. Sa structure dynamique permet des ajustements anticipatifs des taux de production et des seuils d'inventaire, limitant efficacement l'utilisation de la machine à émissions élevées avant que le système n'encoure des pénalités significatives. À l'opposé, les Politiques II et III adoptent des stratégies réactives, ajustant leurs décisions de production

uniquement après que les émissions dépassent une limite prédéterminée. Cette réponse différée génère des émissions accrues et par conséquent des coûts supérieurs. De plus, la structure figée de ces politiques de référence limite leur flexibilité d'adaptation aux contraintes environnementales de plus en plus strictes. Par conséquent, la politique proposée surpasse ses homologues, particulièrement dans des contextes sensibles aux émissions, confirmant son adéquation pour les systèmes hybrides opérant sous réglementations environnementales. Cette performance démontre l'importance cruciale de l'adaptabilité dynamique dans l'optimisation des systèmes de production hybrides.

## 2.7 Perspectives managériales et implémentation

Le défi principal des systèmes hybrides de fabrication/refabrication consiste à synchroniser les opérations afin de minimiser les émissions et les coûts liés à la gestion des stocks. La politique de contrôle proposée fournit un cadre dynamique qui ajuste les taux de production des machines de fabrication et de refabrication en fonction des niveaux de stocks et d'émissions. Cela garantit la rentabilité et la durabilité environnementale du système tout en évitant les ruptures de stock et les coûts de stockage excessifs. Comme l'illustre la figure 2.18, lorsque les émissions restent inférieures au seuil volontaire  $V$  ( $e < 50$ ), le système privilégie la machine de refabrication en raison de ses coûts de production réduits. Si le niveau d'inventaire dépasse le seuil critique  $Z_2(e)$ , les deux machines cessent la production pour éviter la surproduction. Dans le cas contraire, la machine de refabrication fonctionne à son taux maximum tandis que la machine de fabrication demeure inactive. Lorsque l'inventaire atteint le seuil intermédiaire  $Z_1(e)$ , la machine de fabrication s'active pour soutenir la production de la machine de refabrication au taux  $d - U_2max = 2.5$ . Lorsque les niveaux de stock tombent en dessous de  $Z_1(e)$ , les deux machines fonctionnent à leur capacité maximale afin de répondre à la demande et d'éviter les pénuries. Pour reconstituer les niveaux de stock jusqu'à  $Z_2(e)$ , la machine de fabrication fonctionne à sa cadence maximale 3,5 jusqu'à ce que le seuil soit atteint, moment auquel elle est mise en arrêt pour permettre à la machine de refabrication de prendre le relais, car celle-ci est plus rentable en termes de coûts de production. Si les émissions dépassent le seuil volontaire tout en restant sous la limite réglementaire  $L$   $50 \leq e < 150$ , les rôles des machines s'inversent pour privilégier la machine de fabrication, qui constitue l'équipement à faibles émissions. Dans ce scénario, lorsque les niveaux d'inventaire se situent entre les seuils critiques  $Z_1(e)$  et  $Z_2(e)$ , la machine de fabrication fonctionne à son taux maximum tandis que la machine de

refabrication reste inactive. Une fois l'inventaire atteignant le seuil supérieur, la machine de refabrication s'active pour produire au taux  $d - U_{1max} = 1.5$ , soutenant la machine de fabrication.

Lorsque les émissions excèdent la limite absolue  $L$  ( $e \geq 150$ ), la politique de contrôle conserve la même architecture, mais les seuils  $Z_1$  et  $Z_2$  deviennent constants et ne s'ajustent plus dynamiquement aux variations des niveaux d'émissions. Cette configuration garantit que la production demeure dans les contraintes environnementales les plus strictes tout en satisfaisant les exigences d'inventaire et de demande.

La flexibilité de la politique permet aux gestionnaires d'équilibrer efficacement les priorités de production entre les machines de fabrication et de refabrication, de s'adapter dynamiquement aux niveaux variables d'émissions et d'inventaire, et d'assurer la conformité aux réglementations environnementales. En exploitant la surveillance des émissions et de l'inventaire, la politique permet une prise de décision optimale, permettant aux gestionnaires d'adresser simultanément les objectifs économiques et environnementaux.

Ces résultats confirment que la politique proposée offre aux décideurs une stratégie plus robuste et économiquement efficace pour optimiser la production dans les systèmes hybrides, particulièrement dans des environnements incertains et contraints par les émissions.

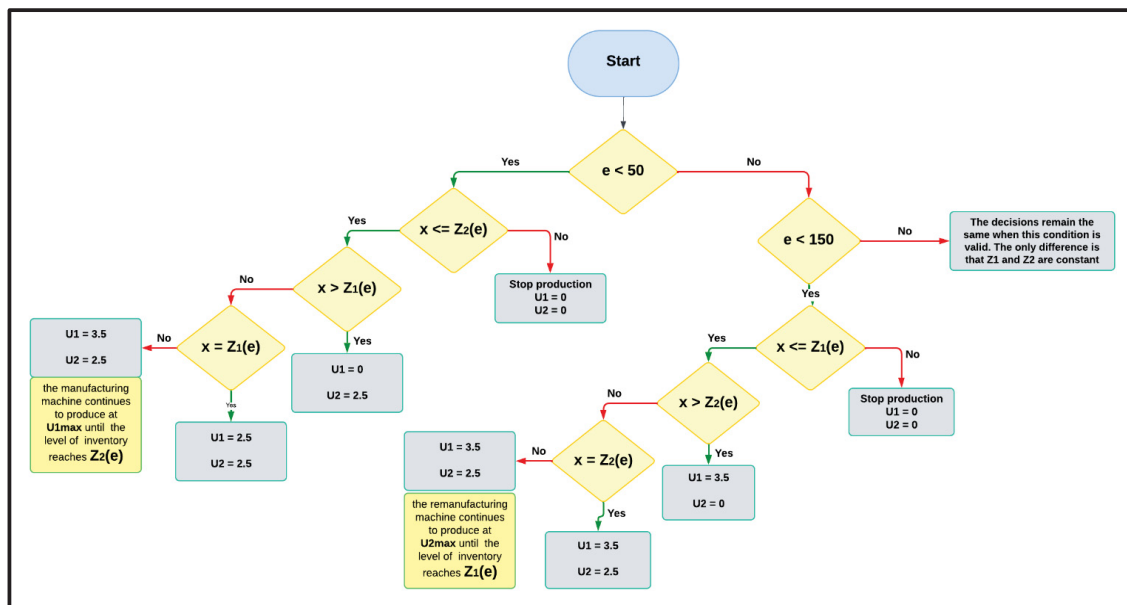


Figure 2.18 Logigramme d'implémentation

Enfin, la section suivante conclut le chapitre en résumant les principales conclusions et contributions.

## **2.8 Conclusion**

Cette étude s'attaque aux défis complexes du contrôle de production dans les systèmes hybrides de fabrication et de refabrication sujets aux pannes et réparations aléatoires et opérant sous contraintes environnementales. Reconnaisant les limitations des Politiques de seuil de couverture traditionnelles (HPP) dans l'intégration des contraintes environnementales, nous avons élaboré une politique environnementale de seuil de couverture dynamique conçue pour minimiser les coûts totaux tout en équilibrant les compromis économiques et environnementaux. En modélisant explicitement les pannes de machines et les émissions de GES, notre approche comble des lacunes critiques dans la littérature, proposant un cadre méthodologique novateur qui hiérarchise et commute dynamiquement entre les opérations de fabrication et de refabrication selon les niveaux d'inventaire et les seuils d'émissions.

La politique proposée exploite les atouts de la théorie du contrôle optimal stochastique pour ajuster dynamiquement les taux de production et les seuils d'inventaire, garantissant la fiabilité opérationnelle et la conformité environnementale. La structure de la politique divise les opérations en zones distinctes, où les priorités de production évoluent selon les niveaux d'inventaire, les seuils d'émissions et la disponibilité des machines. Cette organisation assure un fonctionnement économiquement efficace en optimisant l'utilisation des machines de fabrication et de refabrication, considérant leurs coûts respectifs et leurs impacts environnementaux.

L'approche numérique et l'analyse de sensibilité ont démontré la robustesse de la politique proposée sur un large éventail de paramètres système. Les résultats clés incluent l'impact des coûts, des taux de panne, de réparation et d'émissions, ainsi que des capacités de production maximales sur les seuils d'inventaire et les priorités de production. Ces résultats soulignent l'adaptabilité de la politique aux conditions opérationnelles variables, assurant une synchronisation efficace des machines de fabrication et de refabrication. L'étude comparative confirme la supériorité économique de l'approche développée, avec des améliorations de performance significatives par rapport aux politiques conventionnelles, particulièrement dans

des environnements contraints par les émissions et caractérisés par l'incertitude opérationnelle.



## **CHAPITRE 3**

### **APPRENTISSAGE PAR RENFORCEMENT POUR L'OPTIMISATION DU CONTRÔLE DES SYSTÈMES DE PRODUCTION M1P1**

#### **3.1 Introduction**

Les systèmes de production modernes font face à des défis croissants en raison de l'incertitude imposée par les pannes et réparations aléatoires des machines, ce qui exige le développement de stratégies de production économiquement viables. Le système M1P1 (une machine, un produit) représente un élément fondamental dans la théorie du contrôle de production, où la prise de décision optimale reste cruciale pour minimiser les coûts tout en respectant les contraintes opérationnelles et environnementales auxquelles ce système fait face, particulièrement le respect des réglementations en matière de réduction des émissions de GES.

Ce chapitre présente notre étude sur l'application des méthodes récentes basées sur l'intelligence artificielle (IA), notamment l'apprentissage par renforcement (RL), pour développer des politiques de contrôle optimales des systèmes M1P1 sujets aux pannes et réparations aléatoires dans l'objectif de minimiser le coût total encouru. Nous commençons par développer une approche pour le cas simple avec une seule variable d'état (inventaire), où l'objectif est de minimiser le coût total actualisé composé de coût de stockage et de pénurie. Puis nous étendons notre méthodologie vers un système plus complexe intégrant les contraintes environnementales (inventaire + émissions cumulées), où le coût total actualisé est constitué de coûts d'inventaire et de coûts liés aux émissions dépassant les limites réglementaires imposées par les autorités compétentes.

Notre objectif principal est de développer des stratégies de contrôle de production qui minimisent le coût total actualisé encouru en utilisant l'apprentissage par renforcement et de valider son applicabilité et son efficacité pour ce type de problème de contrôle stochastique. Nous commençons par un cas simple en nous appuyant sur la référence théorique établie par (Bielecki & Kumar, 1988) pour mesurer la performance de notre méthode, puis nous

étendons cette base vers un cas plus complexe afin de démontrer la polyvalence de ce cadre méthodologique pour diverses configurations et niveaux de complexité. Cette recherche démontre l'évolutivité du RL pour des problèmes de complexité croissante, ouvrant la voie à des applications industrielles plus sophistiquées.

## 3.2 Fondements théoriques

L'apprentissage par renforcement (RL) constitue l'une des trois principales branches des techniques d'apprentissage automatique, aux côtés de l'apprentissage supervisé et non supervisé. Contrairement à l'apprentissage supervisé qui utilise des données avec des étiquettes, ou à l'apprentissage non supervisé qui cherche à classer l'information, le RL est la tâche d'apprendre par essai-erreur avec pour objectif d'agir de manière à obtenir la récompense la plus élevée à long terme (Esteso et al., 2023).

L'inspiration du RL provient des systèmes d'apprentissage biologiques, particulièrement des processus d'apprentissage par essai-erreur observés chez les animaux (Zhang et al., 2024). Cette approche élimine le besoin de données pré-étiquetées et distingue le RL des autres méthodes d'apprentissage automatique, démontrant un potentiel prometteur pour la prise de décision séquentielle (Zhang et al., 2024).

### 3.2.1 Composantes fondamentales du RL

Trois éléments principaux sont nécessaires au processus RL, selon (Sutton & Barto, 2018) :

- **La Politique (Policy)** : La politique définit les actions à choisir dans un état donné de l'environnement. Elle constitue la stratégie que l'agent utilise pour déterminer la prochaine action «  $\mathbf{a}$  » basée sur l'état actuel «  $\mathbf{s}$  », représentée comme une distribution de probabilité  $\pi(\mathbf{a}|\mathbf{s})$  qui associe une action «  $\mathbf{a}$  » à un état «  $\mathbf{s}$  » (Esteso et al., 2023).
- **Le Signal de Récompense (Reward)** : Le signal de récompense divise les actions en bonnes ou mauvaises selon le retour immédiat reçu lors de la transition entre états.

L'objectif global de l'agent est de maximiser la récompense totale, ce signal constitue donc la base pour ajuster la politique et la fonction de valeur (Esteso et al., 2023).

- **La Fonction de Valeur (Value Function) :** La fonction de valeur évalue quelles actions ont des effets positifs à long terme en considérant non seulement la récompense immédiate d'un état, mais aussi la récompense susceptible de suivre à long terme. Il existe deux types principaux :
  - Fonction de valeur d'état  $V_{\pi}(s)$  : informe sur le retour total attendu si nous démarrons de cet état en suivant la politique  $\pi$ .
  - Fonction de valeur d'action  $Q_{\pi}(s, a)$  : retour attendu quand l'agent est dans l'état « s » et effectue l'action « a » en suivant la politique  $\pi$ .

Les fonctions état-valeur et action-valeur peuvent être définies récursivement à l'aide de l'équation de Bellman (Sutton & Barto, 2018). Cette équation décompose la fonction de valeur en deux parties : la récompense immédiate et les valeurs futures actualisées.

### 3.2.2 Cadre Mathématique : Processus de décision markovien (MDP)

Les processus RL peuvent être modélisés comme des processus de décision markoviens (MDP) (Sutton & Barto, 2018). Un MDP peut être défini par un tuple de cinq éléments (**S**, **A**, **P**, **R**,  $\gamma$ ) où :

- **S** : L'espace des états contenant tous les états réalisables.
- **A** : L'espace d'action avec un ensemble d'actions possibles.
- **P** : matrice de probabilité de transition  $P(s'/s, a)$  : qui caractérise les dynamiques de l'environnement qui peut être défini comme suit :

$$P(s'/s, a) = \Pr (s_{t+1} = s' / s_t = s \text{ et } a_t = a) \quad (3.1)$$

Avec  $s_t$  est l'état à l'instant t et  $a_t$  l'action à l'instant t.

- **R** : fonction de récompense qui donne la récompense à l'agent en se basant sur l'action prise **a** pour passer de l'état **s** à **s'** (Esteso et al., 2023).

$$R(s'/s, a) = E(r_{t+1} / s_{t+1} = s', s_t = s, a_t = a) \quad (3.2)$$

- $\gamma \in [0,1]$  : facteur d'actualisation utilisé pour actualiser les récompenses futures.

L'une des principales différences du RL par rapport aux autres branches de l'apprentissage automatique est la présence de récompenses retardées. Un agent peut recevoir des récompenses insignifiantes pendant une longue séquence d'actions jusqu'à atteindre un état particulier avec une récompense très significative (par exemple, trouver la sortie d'un labyrinthe). L'agent doit apprendre quelles actions sont favorables sur la base d'une récompense qui peut avoir lieu arbitrairement dans le futur. Ceci contraste avec l'apprentissage supervisé, où les données d'entraînement contiennent déjà la solution désirée (étiquette) (Esteso et al., 2023).

La propriété de Markov garantit que le nouvel état ne dépendra que de l'état actuel plutôt que d'états précédents, faisant du MDP un processus sans mémoire.

### 3.2.3 Cycle d'interaction agent - environnement

Le cycle RL fonctionne selon un mécanisme d'interaction continue selon la figure ci-dessous (Sutton & Barto, 2018):

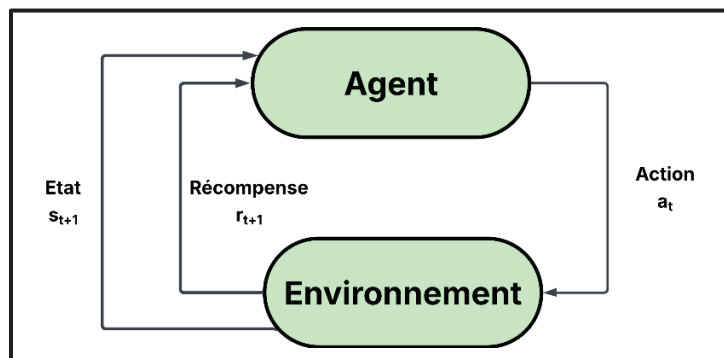


Figure 3.1 Cycle d'interaction Agent-Environnement

L'agent interagit avec l'environnement selon le processus suivant :

- a. L'agent observe l'état actuel de l'environnement.
- b. L'agent sélectionne une action basée sur sa politique.
- c. L'environnement transition vers un nouvel état.
- d. L'agent reçoit une récompense.
- e. Le cycle se répète en permettant l'apprentissage progressif.

Cette interaction permet à l'agent d'identifier par essai-erreur la meilleure action de contrôle à prendre dans chaque état et ainsi de développer progressivement une politique optimale.

Une politique  $\pi: S \rightarrow A$  définit la stratégie de l'agent en associant chaque état à une action. Il détermine à partir de cette politique l'action à prendre dans chaque état. La détermination de la politique optimale  $\pi^*$  peut être réalisée à partir de la fonction valeur. Une fonction valeur d'un état  $s$ ,  $V_\pi(s)$  définit comme la récompense cumulée espérée à partir de cet état en suivant la politique  $\pi$  :

$$V_\pi(s) = E \left[ \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \mid s_t = s \right], \forall s \in S \quad (3.3)$$

Ou à partir de la fonction action-valeur  $Q_\pi(s)$  qui est défini comme la récompense cumulée espérée en prenant l'action  $a$  à l'état  $s$ , puis en suivant la politique  $\pi$  :

$$Q_\pi(s) = E \left[ \sum_{i=0}^{\infty} \gamma^i r_{t+i+1} \mid s_t = s, a_t = s \right], \forall s \in S \text{ et } \forall a \in A \quad (3.4)$$

La politique optimale est la politique qui maximise la récompense cumulée espérée sur tout l'horizon de décision. La fonction valeur optimale  $V^*(s)$ , qui est la fonction valeur suivant la politique optimale  $\pi^*$  peut être estimé à partir de l'équation de Bellman défini comme suit :

$$V^*(s) = \max_a \left\{ r(s, a) + \gamma \sum_{s' \in S} P(s'/s, a) V^*(s') \right\}, \forall s \in S \quad (3.5)$$

Pareil pour  $Q^*(s, a)$  :

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in S} P(s'/s, a) \max_{a'} \{Q^*(s', a')\}, \forall s \in S, \forall a \in A \quad (3.6)$$

Donc la politique optimale peut être déduite à partir de l'une des fonctions précédentes comme suit :

$$\pi^* = \left\{ \begin{array}{l} \operatorname{argmax}_a \left\{ r(s, a) + \gamma \sum_{s' \in S} P(s'/s, a) V^*(s') \right\} \\ \text{ou} \\ \operatorname{argmax}_a Q^*(s, a) \end{array} \right\} \quad (3.7)$$

Pour plus de détails sur l'apprentissage par renforcement et ses différentes composantes, approches et modèles, nous renvoyons le lecteur à l'ouvrage de référence de (Sutton & Barto, 2018).

### 3.3 Formulation de problème

#### 3.3.1 Système de production étudié

Le système étudié est un système de production de base M1P1 (1 machine, 1 produit). Il comprend une machine de fabrication qui produit un seul produit et fait face à des pannes et réparations aléatoires. Ce système constitue un modèle fondamental en théorie du contrôle de production. Il se caractérise par :

- Une machine sujette aux pannes aléatoires.
- Un produit unique à produire.
- Demande déterministe constante  $d$ .
- Capacité de production limitée  $U_{max}$ .

Le système évolue selon un processus de Markov à deux états :

- État opérationnel ( $\alpha = 1$ ) : La machine est en production.
- État de panne ( $\alpha = 0$ ) : La machine est en panne.

Soit  $\varepsilon(t)$  le processus stochastique qui gouverne le système étudié. Le diagramme ci-dessous représente le diagramme de transition d'états :

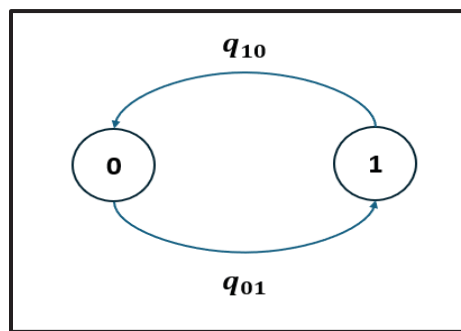


Figure 3.2 Diagramme des états du système étudié.

Avec  $q_{10}$  est le taux de panne et  $q_{01}$  le taux de réparation.

Notre approche consiste à commencer par un système M1P1 de base en appliquant l'apprentissage par renforcement (RL) avec pour objectif d'atteindre la politique optimale connue et validée par approche analytique et numérique. Ensuite, la prochaine étape est d'étendre cette approche vers des systèmes plus complexes. Pour notre cas, nous avons choisi un système M1P1 avec émissions des GES. Nous allons présenter les caractéristiques de ces systèmes ainsi que les différentes composantes du cadre de travail d'apprentissage par renforcement à savoir l'environnement, l'agent, et la fonction de récompense.

La figure 3.3 représente le système de production M1P1 avec émissions des GES :

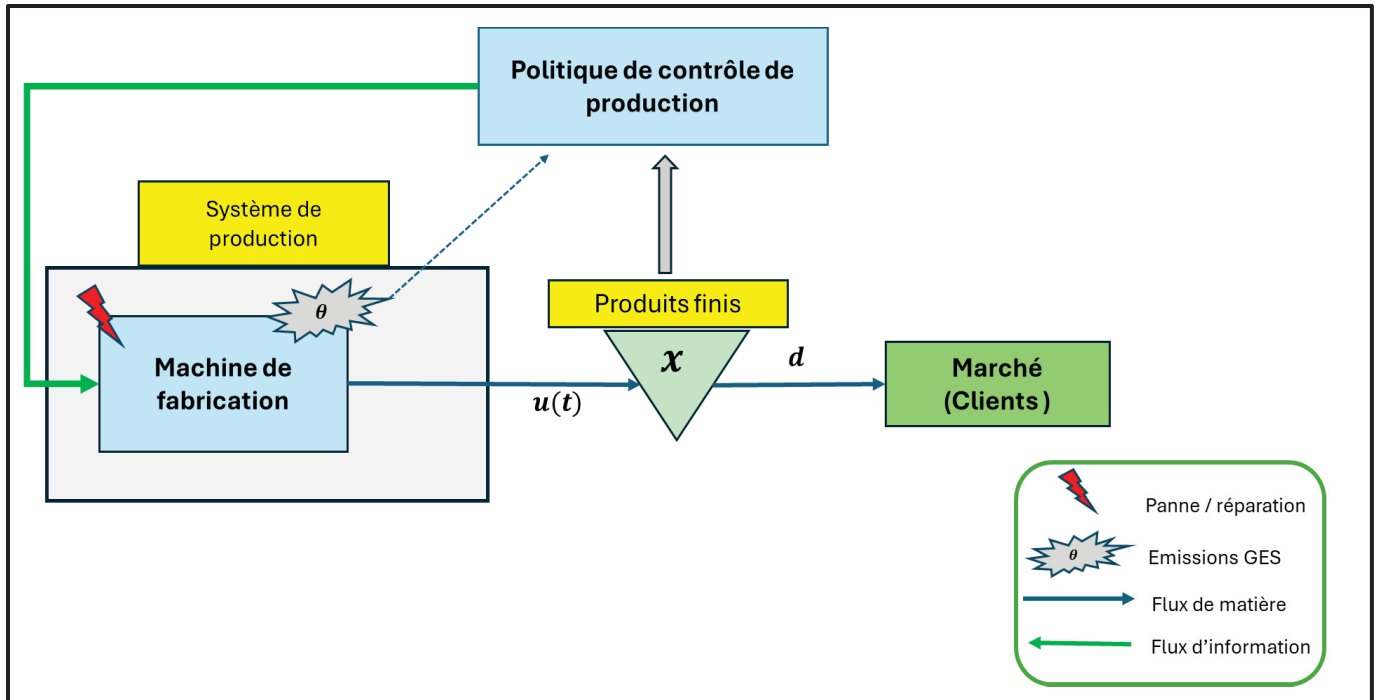


Figure 3.3 Système de production étudié

Pour le reste de ce chapitre, nous considérons les notations suivantes :

- $u(.)$  : *taux de production de la machine de production*
- $x(t)$  : *niveau de stock des produits finis*
- $e(t)$  : *niveau des émissions*
- $\varepsilon(t)$  : *processus stochastique du système de production*
- $x^+$  : *stock positif =  $\max(0, x(t))$  pour  $x(t) > 0$*
- $x^-$  : *stock négatif =  $\max(0, -x(t))$  pour  $x(t) < 0$*
- $d$  : *taux de demande des clients*
- $U_{max}$  : *taux de production maximal de la machine de production*
- $q_{ij}$  : *taux de transition du mode  $i$  vers  $j$  pour le processus  $\varepsilon(t)$  avec  $i, j \in \{0, 1\}$*
- $c^+$  : *coût de stockage des produits*
- $c^-$  : *coût de pénurie des produits*
- $c_e$  : *coût des émissions*
- $\rho$  : *taux d'actulisation*
- $L$  : *limite des émissions*

- $\theta$ : *taux d'émissions pour la machine de production*

### 3.3.2 Dynamiques du Système

Le système étudié suit deux dynamiques principales. La première concerne le stock dans le système. La seconde concerne les émissions générées par la machine de production.

La dynamique de stock est décrite par l'équation suivante :

$$\dot{x}(t) = u(t) - d, \quad x(0) = x_0 \quad (3.8)$$

La dynamique des émissions est décrite par l'équation suivante :

$$\dot{e}(t) = \theta \cdot u(t) \text{ pour } t \in [T_i, T_{i+1}] \text{ et } e(T_i) = 0, i = 0, \dots, \infty \quad (3.9)$$

Les autorités compétentes fixent un plafond des émissions pendant une période  $T_i$ .

### 3.3.3 Espace de commande admissible

L'espace des commandes admissibles est noté  $A(\varepsilon)$ . Il est défini par :

$$A(\varepsilon) = \{ u(.) / 0 \leq u(.) \leq U_{max} \} \quad (3.10)$$

Pour tout  $\varepsilon \in \{0,1\}$ .

### 3.3.4 Fonctions de coût

#### • Fonction de coût instantanée

La fonction de coût instantanée  $g(.)$  est composée de quatre composantes comme suit :

$$g(\varepsilon, x, e) = \text{coût de stockage} + \text{coût de pénurie} + \text{coût des émissions} \quad (3.11)$$

Avec :

- $\text{coût de stockage} = c^+ \cdot x^+$
- $\text{coût de pénurie} = c^- \cdot x^-$
- $\text{coût des émissions} :$

Le coût des émissions à la fin d'une période  $T_i$  est défini comme suit :

$$\text{coût des émissions}(T_i) = c_e * \max(0, e(t_{T_i}) - L) ; i = 0, \dots, \infty \quad (3.12)$$

Avec  $L$  est la limite fixée par les autorités compétentes au-dessus de laquelle des pénalités s'appliquent.

### **3.4 Application au système M1P1 de base**

#### **3.4.1 Conception de l'environnement RL**

##### **3.4.1.1 Architecture générale**

L'implémentation de notre approche d'apprentissage par renforcement (RL) nécessite la création d'un environnement d'entraînement qui simule fidèlement le comportement et les dynamiques du système M1P1. Nous avons développé cet environnement en utilisant le cadre de travail « Framework » Gymnasium, qui constitue le standard pour les environnements d'apprentissage par renforcement et assure une compatibilité directe avec les algorithmes modernes comme PPO. L'environnement encapsule la logique de simulation du système de production et fournit l'interface nécessaire pour l'interaction agent-environnement. Cette architecture permet une séparation entre la modélisation du système physique de production et l'algorithme d'apprentissage.

##### **3.4.1.2 Paramètres du Système**

Les paramètres du système M1P1 sont centralisés dans une configuration flexible qui reflète les caractéristiques réelles d'un système de production :

- **Taux de demande** : 1,0 unité par pas de temps (demande constante)
- **Capacité maximale** : 1,5 unité par pas de temps
- **Taux de panne** :  $1/9 \approx 0,11$  (probabilité par pas de temps)
- **Taux de réparation** : 2,0 (probabilité par pas de temps)
- **Horizon de planification** : 1000 pas de temps par épisode

Nous avons choisi ces paramètres basée sur un exemple numérique développé dans un autre cadre de recherche et nous possédons la solution analytique donnée par la formule développée par (Bielecki & Kumar, 1988).

### 3.4.1.3 Structure économique

Les paramètres économiques reflètent la structure de coût asymétrique typique des systèmes de production :

- **Coût de stockage** : 1,0 unité monétaire par unité stockée
- **Coût de rupture** : 10,0 unités monétaires par unité manquante

## 3.4.2 Représentation d'état

### 3.4.2.1 Espace d'observation

L'état du système est représenté par un vecteur bidimensionnel  $s = [\text{niveau d'inventaire, état de la machine}]$  qui est représenté par les variables  $s = [x, \varepsilon]$  et qui capture l'information complète nécessaire à la prise de décision :

Pour la première variable d'état, le niveau d'inventaire  $x$ , il représente la composante continue du système et dispose des caractéristiques suivantes :

- Domaine de variation :  $\mathbb{R}$  (valeurs réelles)
- Interprétation : valeurs positives = stock disponible, valeurs négatives = pénurie (demande en attente)

La deuxième variable d'état, l'état de la machine, il représente la composante discrète du système et dispose des caractéristiques suivantes :

- Domaine de variation :  $\{0, 1\}$
- Interprétation : 0 = machine en panne, 1 = machine opérationnelle

Cette représentation hybride permet à l'agent de percevoir simultanément les aspects économiques (inventaire) et techniques (disponibilité) du système, information essentielle pour des décisions optimales.

### 3.4.3 Espace d'action et stratégies de production

#### 3.4.3.1 Actions discrètes disponibles

Nous avons opté pour un espace d'action discret avec trois décisions fondamentales :

Tableau 3.1 Espace d'actions admissibles

| Action | Stratégie             | Taux de Production        |
|--------|-----------------------|---------------------------|
| 0      | Arrêt de production   | 0                         |
| 1      | Production équilibrée | 1 (= taux de demande)     |
| 2      | Production maximale   | 1,5 (= capacité maximale) |

Cette discrétisation simplifie l'apprentissage tout en préservant les principales stratégies de contrôle de production. Les trois actions correspondent à une politique de Bang-Bang control qui est déjà démontré être la meilleure politique pour ce type de problème de contrôle.

L'environnement respecte les contraintes physiques du système :

- Production nulle si la machine est en panne.
- Capacité limitée par les spécifications techniques.
- Demande constante indépendante des décisions de production.

### 3.4.4 Dynamiques du Système

Le système implémente les transitions stochastiques selon le processus de Markov défini dans la section précédente :

- **Taux de panne** :  $\lambda = 1/9$  par pas de temps (si machine opérationnelle)
- **Taux de réparation** :  $\mu = 2,0$  par pas de temps (si machine en panne)

L'inventaire évolue selon l'équation suivante :

$$Inventaire(t + 1) = Inventaire(t) + Production(t) - Demande(t) \quad (3.13)$$

Cette équation capture la dynamique fondamentale du système où l'accumulation d'inventaire résulte de l'écart entre production et demande. L'évolution de cette variable d'état détermine les transitions du système d'un état à l'autre.

### 3.4.5 Fonction de Récompense

La fonction de récompense constitue une composante clé du cadre d'apprentissage par renforcement. Sa définition appropriée est critique pour l'apprentissage de l'agent et de la performance globale du modèle. Ce signal de feedback indique à l'agent l'efficacité de ses actions et guide l'amélioration progressive de sa politique de décision.

La fonction de coût instantané reflète la structure économique du système :

$$g(x) = c^+ * x^+ + c^- * x^- \quad (3.14)$$

La récompense d'apprentissage par renforcement est définie comme :

$$R(s, a) = -g(x) \times \text{facteur\_échelle} \quad (3.15)$$

Le facteur d'échelle améliore la stabilité d'entraînement en maintenant les signaux de récompense dans une plage optimale pour la convergence, particulièrement pour les algorithmes basés sur les réseaux de neurones. Sans ce facteur, les gradients peuvent exploser dans le réseau de neurones, compromettant ainsi la convergence vers une politique optimale. Le signe '-' dans la fonction récompense établit une relation inverse entre coût et récompense. Plus le coût encouru augmente, plus la récompense diminue. Cette structure incite naturellement l'agent à minimiser les coûts pour maximiser sa récompense, alignant ainsi l'objectif d'apprentissage avec l'objectif économique du système.

### 3.5 Résultats d'application au premier cas - M1P1 de base

Dans cette section, nous présentons les résultats de l'application de l'apprentissage par renforcement au cas de base M1P1. En premier lieu, nous détaillons les paramètres du système ainsi que le modèle d'apprentissage par renforcement utilisé et ses différentes composantes.

Pour le système étudié, nous avons considéré le système M1P1 de base, c'est-à-dire le système M1P1 présenté dans la figure 3.3 sans la composante des émissions. Cette première application considère un système M1P1 avec une seule variable d'état : le niveau d'inventaire.

#### 3.5.1 Paramètres du système

Les paramètres utilisés sont présentés dans le tableau suivant :

Tableau 3.2 paramètres du système

| $c^+$ | $c^-$ | $d$ | $U_{max}$ | $q_{10}$ | $q_{01}$ | $x_0$ |
|-------|-------|-----|-----------|----------|----------|-------|
| 1     | 10    | 1   | 1.5       | 1/9      | 2        | 0     |

Pour l'environnement d'apprentissage, l'agent va interagir avec le système M1P1 pendant des épisodes avec les caractéristiques suivantes :

Chaque épisode d'entraînement suit un protocole standardisé conçu pour assurer la reproductibilité et la comparabilité des résultats. L'environnement initialise chaque épisode avec un inventaire nul et une machine en état opérationnel, créant des conditions de départ identiques pour tous les apprentissages. La durée de chaque épisode est fixée à 1000 pas de temps, constituant un horizon de planification suffisant pour observer les effets à long terme des politiques de contrôle. Durant l'exécution, l'environnement enregistre automatiquement l'historique complet des actions prises, des niveaux d'inventaire et des coûts encourus à chaque pas de temps (coût encouru par unité de temps). À la terminaison de chaque épisode, des métriques de performance sont calculées automatiquement : coût total, coût total moyen par pas de temps, inventaire moyen, et fréquence de rupture de stock. Cette architecture

permet non seulement un apprentissage stable et cohérent, mais facilite également l'analyse comparative des différentes politiques apprises et l'évaluation objective de la convergence de l'algorithme d'optimisation. L'évaluation de la qualité de l'apprentissage et de l'efficacité des politiques RL nécessite un ensemble d'indicateurs de performance robustes. Notre environnement calcule automatiquement plusieurs métriques clés à la fin de chaque épisode d'entraînement.

- **Indicateurs économiques principaux**

- **Coût total par épisode** : Somme de tous les coûts instantanés sur l'horizon de planification (longueur d'épisode), constituant l'objectif principal de minimisation.
- **Coût moyen par pas de temps** : c'est le coût moyen par unité de temps.
- **Décomposition des coûts** : Séparation entre coûts de stockage et coûts de rupture pour analyser l'origine des inefficacités.

- **Métriques de contrôle des stocks**

- **Inventaire moyen** : Niveau de stock moyen sur l'épisode.
- **Fréquence de rupture** : Pourcentage de temps passé en situation de rupture de stock (inventaire  $< 0$ ), métrique critique pour l'évaluation du service client.
- **Amplitude des variations** : Écart-type de l'inventaire reflétant la stabilité de la politique de contrôle.

- **Métriques de convergence d'apprentissage**

- **Récompense cumulée** : Évolution de la performance au cours de l'entraînement, indicateur direct de la convergence.
- **Stabilité interépisodes** : Variance des métriques de performance entre épisodes consécutifs.
- **Écart d'optimalité** : Distance relative par rapport à la référence théorique, calculée selon :

$$\text{Écart (\%)} = (\text{Coût\_obtenu} - \text{Coût\_optimal}) / \text{Coût\_optimal} \times 100 \quad (3.16)$$

Ces indicateurs permettent une évaluation multicritère de la performance des politiques RL en facilitant la comparaison avec les méthodes analytiques et numériques existantes, l'analyse de sensibilité aux paramètres du système et de l'algorithme, ainsi que la validation de la robustesse des politiques apprises. Cette approche permet également d'identifier les axes d'amélioration pour l'optimisation continue des performances. L'architecture de métriques adoptée assure une évaluation de la qualité de l'apprentissage par renforcement appliqué au contrôle du système M1P1, offrant une base solide pour la prise de décision et l'amélioration itérative des politiques de contrôle.

### 3.5.2 Modèle d'apprentissage par renforcement : Optimisation de politique proximale (PPO)

#### 3.5.2.1 Justification du choix algorithmique

L'Optimisation de politique proximale (PPO), introduite par (Schulman et al., 2017), représente une méthode de gradient de politique de pointe qui aborde efficacement les défis clés de l'apprentissage par renforcement moderne. Les récentes revues de littérature confirment que PPO fait partie des approches « actor-critic » les plus utilisées pour les problèmes de contrôle de production (Esteso et al., 2023). Cette méthode combine les avantages des approches basées sur les politiques avec la stabilité des méthodes de valeur, ce qui en fait un bon choix pour notre système M1P1. Le choix de PPO se justifie par plusieurs facteurs critiques :

- **Stabilité d'entraînement** : PPO évite les mises à jour destructives de politique grâce à son mécanisme de coupure (clipping).
- **Efficacité d'échantillonnage** : apprentissage efficace même avec des données limitées.
- **Robustesse** : performance consistante dans divers environnements de production.
- **Simplicité d'implémentation** : facilite le déploiement comparé aux méthodes plus complexes.

D'autre part, cette supériorité a été validée empiriquement par une comparaison systématique sur notre système MIP1 : avec une configuration identique d'hyperparamètres et 10 millions de pas d'entraînement, PPO atteint un écart d'optimalité de 17,10 %, soit une performance deux fois meilleure que DQN (33,29%) et sept fois meilleure qu'A2C (119,31%), confirmant ainsi son efficacité pour notre application.

### 3.5.2.2 Architecture « Actor-Critic » du modèle PPO

PPO utilise une architecture «actor-critic » qui combine les éléments des méthodes basées sur les valeurs et sur les politiques (Zhang et al., 2024). Cette architecture se compose de deux composantes principales :

#### ❖ L'Acteur (Actor Network)

- **Fonction** : Estime les actions optimales de contrôle  $\pi(a_t|s_t; \theta)$  pour chaque état.
- **Architecture** : Réseau de neurones paramétré par  $\theta$ .
- **Sortie** : Distribution de probabilité sur l'espace d'action discret  $\{0, 1, 2\}$ .

#### ❖ Le Critique (Critic Network)

- **Fonction** : Évalue la fonction de valeur optimale  $V(s_t; \beta)$  représentant le profit total attendu.
- **Architecture** : Réseau de neurones paramétré par  $\beta$ , partageant la couche d'entrée avec l'acteur.
- **Sortie** : Valeur scalaire estimant la récompense cumulée depuis l'état actuel jusqu'à la fin de l'épisode.

Les entrées des deux réseaux sont identiques et consistent en la représentation d'état du système : [niveau d'inventaire, état de la machine].

### 3.5.2.3 Fonctionnement de l'algorithme PPO

- **Collecte de trajectoires et interaction**

L'agent PPO interagit de manière répétée avec l'environnement pour collecter des trajectoires  $(s_t, a_t, r_t)$  nécessaires à l'entraînement (Zhang et al., 2024). À chaque itération, l'agent collecte un batch d'expériences en implémentant sa politique actuelle, puis met à jour les paramètres  $\theta$  et  $\beta$  des réseaux acteur et critique.

- **Mise à jour de l'acteur : Objectif coupé**

La caractéristique distinctive de PPO réside dans son objectif coupé (clipped surrogate objective) qui contrôle la magnitude des mises à jour de politique. L'acteur est mis à jour en minimisant :

$$L(\theta) = \hat{E}_t[\min\{r_t(\theta)\hat{A}_t, fclip(r_t(\theta), 1 - \delta, 1 + \delta)\hat{A}_t\}] \quad (3.17)$$

Où :

- $r_t(\theta) = \frac{\pi(a_t|s_t; \theta)}{\pi(a_t|s_t; \theta_{old})}$  : rapport de probabilité entre nouvelle et ancienne politique
- $\hat{A}_t$  : estimateur de la fonction d'avantage au pas de temps  $t$ .
- $\delta$ : ratio de coupure limitant le changement maximal autorisé.

La fonction de coupure  $fclip$  garantit la stabilité en limitant les mises à jour excessives de la politique :

$$fclip = \begin{cases} 1 - \delta & r_t(\theta) < 1 - \delta \\ r_t(\theta) & 1 - \delta < r_t(\theta) < 1 + \delta \\ 1 + \delta & r_t(\theta) > 1 + \delta \end{cases} \quad (3.18)$$

Cette formulation empêche l'agent de faire des changements drastiques à sa politique, assurant un apprentissage stable et reproductible.

### ❖ Mise à jour du critique

Le réseau critique est mis à jour en minimisant l'erreur quadratique entre les retours observés et les valeurs prédites :

$$L(\beta) = \hat{\mathbb{E}}_t \left[ \sum_{t' > t} (r_{t'} - v_t(s_{t'}; \beta))^2 \right] \quad (3.19)$$

Cette mise à jour améliore progressivement l'estimation de la fonction de valeur, fournissant des signaux d'apprentissage plus précis pour l'acteur.

## 3.5.3 Implémentation du modèle d'apprentissage pour M1P1 de base

### 3.5.3.1 Architecture réseau du modèle PPO

Notre configuration PPO utilise une architecture compacte, mais efficace, basée sur des réseaux de neurones entièrement connectés. Les réseaux d'acteur et critique partagent une structure similaire avec deux couches cachées. Les figures ci-dessous illustrent l'architecture :

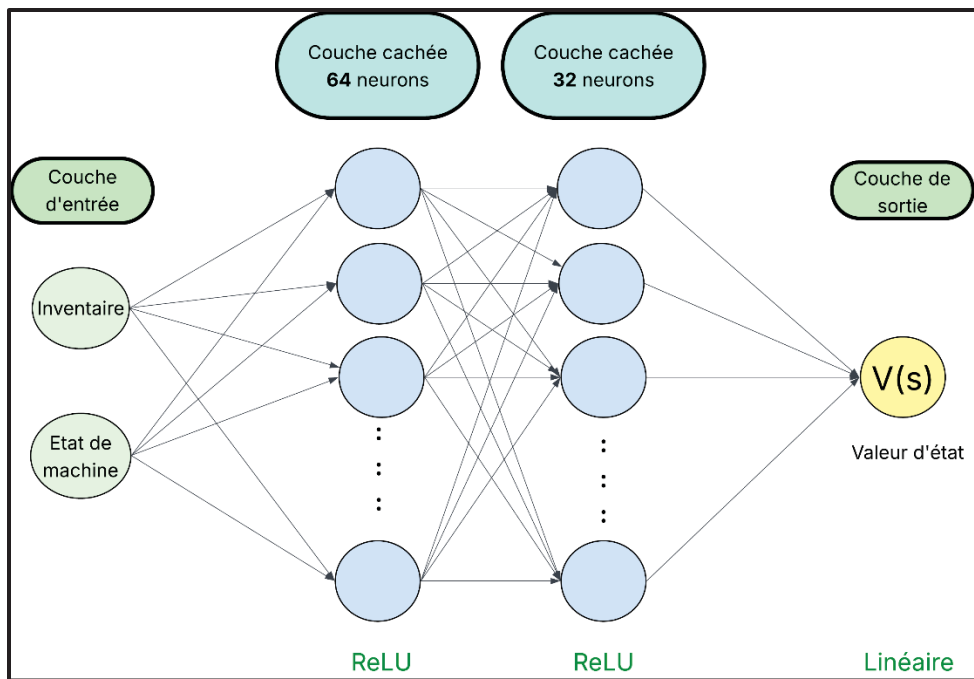


Figure 3.4 Architecture réseau de neurones du critique

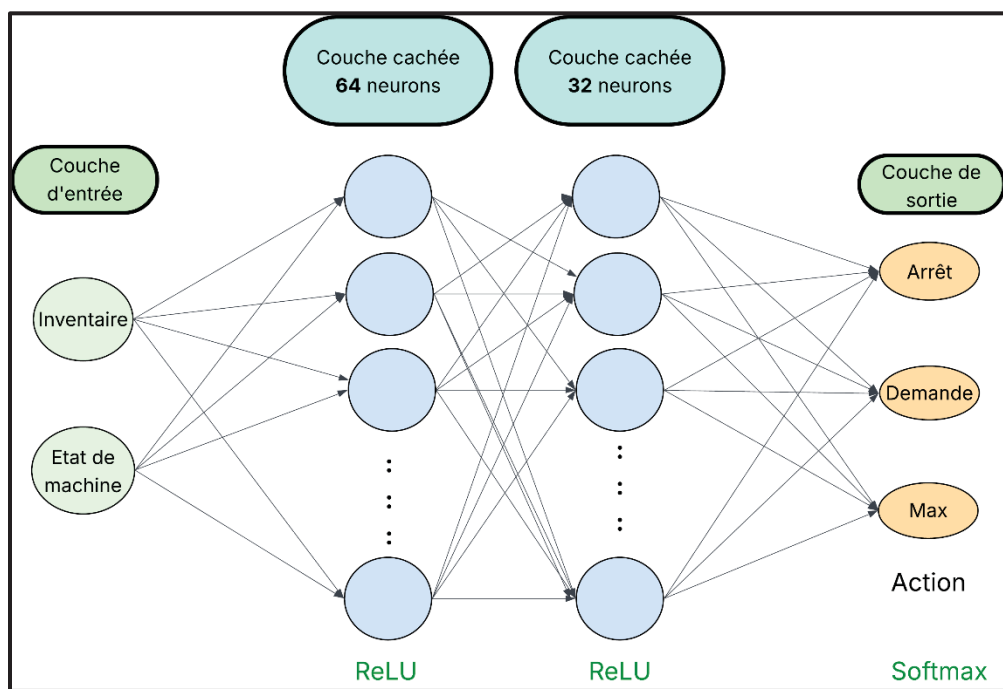


Figure 3.5 Architecture réseau de neurones de l'acteur

Les figures ci-dessus illustrent l'architecture détaillée des réseaux de neurones acteur et critique, tous deux partageant une structure similaire, mais avec des sorties spécialisées. Chaque réseau reçoit en entrée un vecteur bidimensionnel représentant l'état du système M1P1 : le niveau d'inventaire (valeur continue) et l'état de la machine (valeur discrète). Les données traversent ensuite deux couches cachées entièrement connectées de 64 et 32 neurones respectivement, utilisant la fonction d'activation ReLU (Rectified Linear Unit). La différenciation entre les deux réseaux s'opère au niveau de la couche de sortie : le réseau acteur utilise une fonction Softmax pour produire une distribution de probabilité sur les trois actions possibles (arrêt, production à la demande, production maximale), garantissant que la somme des probabilités égale 1 tandis que le réseau critique emploie une couche linéaire pour générer une valeur scalaire  $V(s)$  représentant l'estimation du retour cumulé attendu depuis l'état actuel.

### 3.5.3.2 Hyperparamètres utilisés pour l'apprentissage

Le tableau suivant présente les hyperparamètres utilisés pour entraîner le modèle PPO :

Tableau 3.3 Hyperparamètres du modèle

| <i>Learning rate</i> | <i>Batch size</i> | <i>Clipping Ratio</i> | <i>Discount factor</i> |
|----------------------|-------------------|-----------------------|------------------------|
| $10^{-4}$            | 64                | 0.2                   | 0.99                   |

### 3.5.3.3 Résultats et interprétation

Cette section présente les résultats de l'application du modèle PPO au système M1P1 de base. Après l'entraînement, nous avons analysé la politique apprise par l'agent en fonction du niveau d'inventaire. Le graphique suivant illustre cette politique, avec en abscisse le niveau d'inventaire et en ordonnée les actions sélectionnées : arrêt de production, production à la demande, ou production maximale.

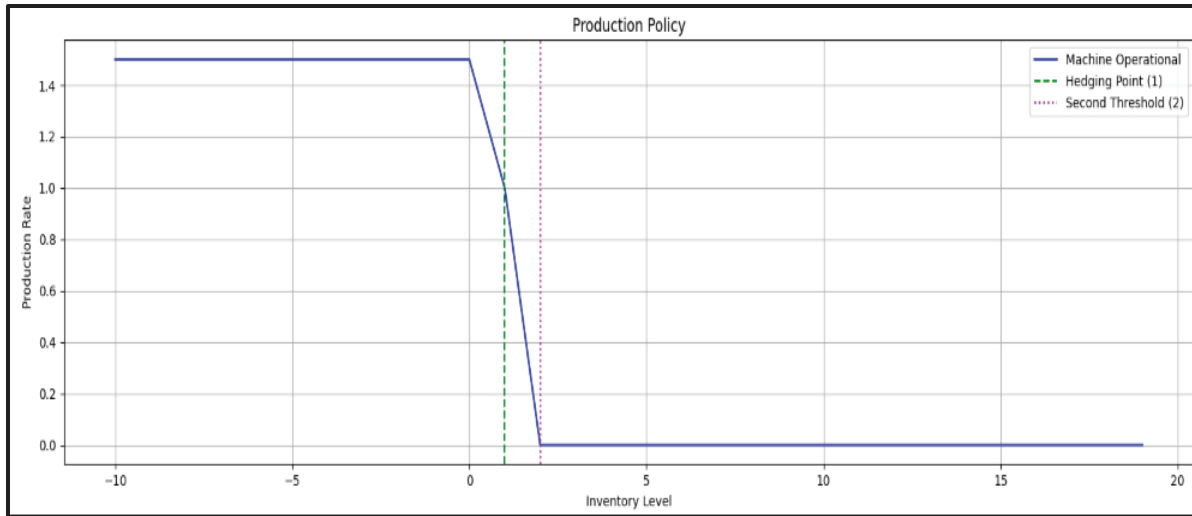


Figure 3.6 Politique de contrôle pour M1P1 appris par l'agent PPO

Nos résultats montrent que l'agent PPO a appris une politique de contrôle très cohérente avec ce qu'on attend théoriquement d'un système de production M1P1. Quand nous analysons le graphique, nous observons une politique de type "hedging point" : l'agent bascule nettement autour d'un niveau d'inventaire spécifique appelé point de couverture ou « hedging point », ce qui est exactement le comportement rationnel recherché. Cette logique démontre que notre modèle a bien compris les enjeux économiques : il réagit immédiatement aux ruptures de stock (coûteuses), mais arrête rapidement la production dès qu'il y a du surstockage. Les seuils de décision que nous avons identifiés révèlent que l'algorithme a saisi les compromis économiques du système M1P1. Le point de basculement principal représente le seuil critique où l'agent passe de la production maximale à la production à la demande puis à l'arrêt complet.

#### 3.5.3.4 Analyse des facteurs d'influence sur la performance d'apprentissage

Bien que la politique apprise démontre une cohérence remarquable avec la théorie du contrôle optimal, la qualité de l'apprentissage et la performance finale dépendent de plusieurs facteurs clés liés à la configuration de l'algorithme PPO. Cette section examine l'impact des hyperparamètres d'entraînement sur l'écart d'optimalité, notre métrique principale de performance, calculée par rapport à la référence théorique de (Bielecki & Kumar, 1988). Le

choix de cette métrique est basé sur le fait qu'elle donne une idée de la qualité de la politique apprise par l'agent de notre modèle.

### 3.5.3.5 Facteurs d'influence identifiés

Le choix des hyperparamètres constitue un enjeu critique pour l'efficacité de l'apprentissage par renforcement dans tous ces domaines d'application, y compris le contrôle de production. Nos investigations se concentrent sur les facteurs suivants, identifiés comme potentiellement déterminants pour la performance :

- **Longueur d'entraînement** : nombre total de pas de temps d'interaction agent - environnement.
- **Longueur d'épisode** : horizon de planification par épisode d'apprentissage.
- **Taux d'apprentissage (Learning Rate)** : vitesse de mise à jour des poids du réseau.
- **Facteur d'échelle de récompense (Scale factor)** : normalisation du signal de récompense.

Chaque facteur influence potentiellement la convergence vers la politique optimale et la stabilité d'apprentissage. L'objectif est d'identifier les configurations permettant de minimiser l'écart d'optimalité tout en assurant la reproductibilité des résultats. Notre évaluation s'appuie sur l'écart d'optimalité comme indicateur principal de performance. Cette métrique permet une comparaison objective entre différentes configurations et quantifie la proximité de notre solution RL par rapport à l'optimum théorique. Un écart faible indique une politique de haute qualité, tandis qu'une variance élevée suggère une instabilité d'apprentissage.

Prenons l'exemple du facteur de longueur d'entraînement. Les figures ci-dessous présentent l'évolution de l'écart d'optimalité au fur et à mesure que l'entraînement avance. Pour chaque intervalle de 50k pas de temps, nous avons calculé cet indicateur pour suivre son évolution.

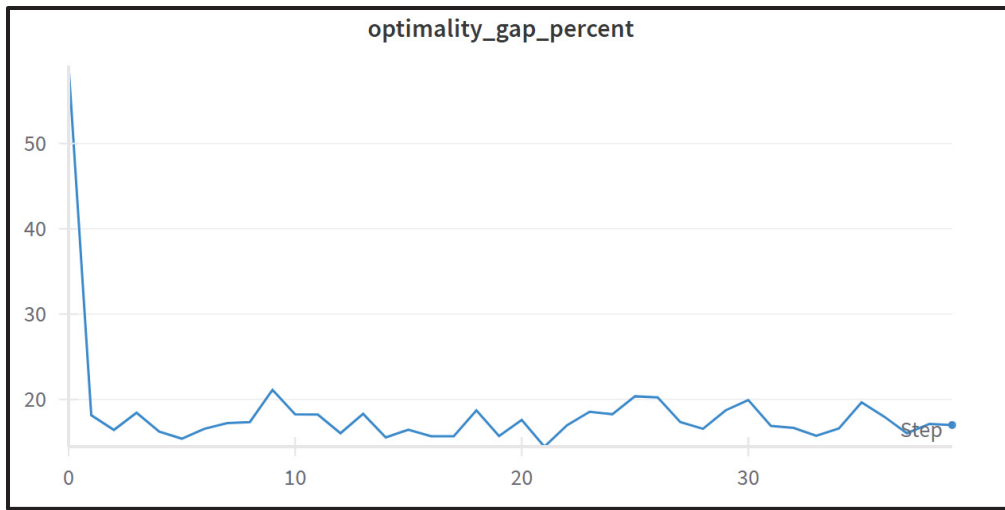


Figure 3.7 Évolution de l'écart d'optimalité pour une longueur d'entraînement de 2M pas de temps

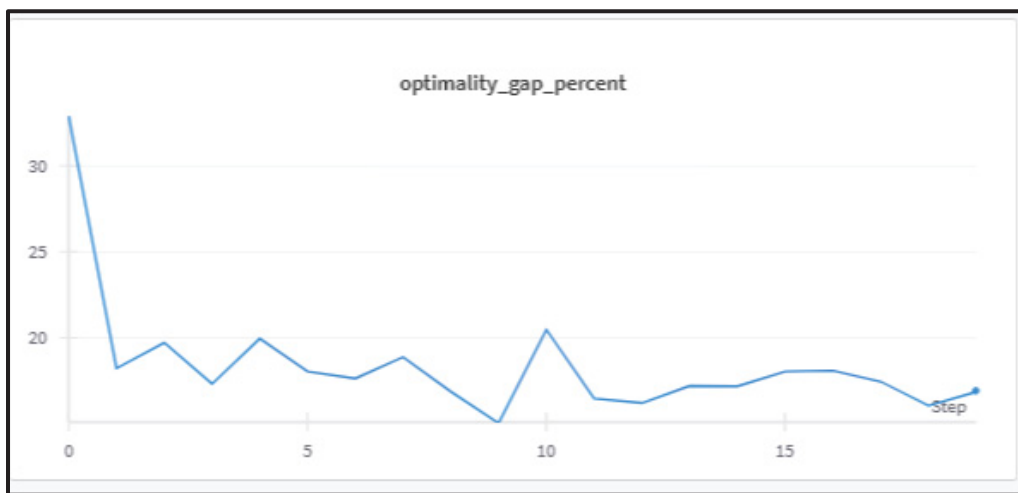


Figure 3.8 Évolution de l'écart d'optimalité pour une longueur d'entraînement de 1M pas de temps

Nous constatons d'après les figures que notre indicateur de performance se stabilise avec l'avancement de l'entraînement. Nous observons aussi que l'écart d'optimalité présente plus de variabilité pour une durée d'entraînement de 1M par rapport à une durée de 2M, ce qui indique que ce facteur impacte la stabilité et la qualité de l'apprentissage de notre modèle. Des analyses similaires ont été menées pour les autres facteurs et nous avons conclu que ces paramètres ont un impact significatif sur la performance et la qualité de l'apprentissage. Face

à ces observations, nous avons mis en place un plan d'expériences pour examiner l'impact de chacun de ces facteurs ainsi que leurs interactions.

### 3.5.3.6 Plan d'expériences

Pour identifier rigoureusement les facteurs critiques, nous avons mis en place un plan d'expériences (DOE) permettant d'évaluer l'impact individuel et les interactions entre ces hyperparamètres. Cette approche assure une exploration efficace de l'espace des configurations possibles.

- **Méthodologie expérimentale**

Pour identifier les facteurs critiques et leurs interactions, nous avons mis en place un plan d'expériences factoriel complet  $2^4$  explorant 16 configurations différentes. Cette approche assure une exploration efficace de l'espace des hyperparamètres tout en quantifiant l'impact statistique de chaque facteur. Les quatre facteurs sélectionnés couvrent les aspects temporels et d'optimisation les plus influents :

- **Longueur d'épisode** : {1000, 10000} pas de temps
- **Longueur d'entraînement** : {1 000 000, 10 000 000} pas de temps total
- **Taux d'apprentissage** :  $\{1 * 10^{-4}, 3 * 10^{-4}\}$
- **Facteur d'échelle de récompense** : {0.001, 0.01}

Le tableau ci-dessous présente les facteurs étudiés :

Tableau 3.4 Facteurs étudiés

| Facteur | Variable                | Niveau bas (-1) | Niveau haut (+1) |
|---------|-------------------------|-----------------|------------------|
| A       | Longueur d'épisode      | 1 000           | 10 000 pas       |
| B       | Longueur d'entraînement | 1000 000        | 10 000 000       |
| C       | Taux d'apprentissage    | 0,0001          | 0,0003           |
| D       | Facteur d'échelle       | 0,001           | 0,01             |

Chaque configuration a été évaluée selon un protocole standardisé assurant la validité statistique des comparaisons :

- Évaluation sur modèle réel : performance des politiques effectivement apprises.
- 500 épisodes d'évaluation par configuration pour assurer la robustesse statistique.

La matrice des expériences est présentée par le tableau suivant :

Tableau 3.5 Matrice des expériences

| <b>Configuration</b> | <b>A</b>  | <b>B</b>  | <b>C</b>  | <b>D</b>  |
|----------------------|-----------|-----------|-----------|-----------|
| <b>DOE_01</b>        | <b>-1</b> | <b>-1</b> | <b>-1</b> | <b>-1</b> |
| <b>DOE_02</b>        | <b>-1</b> | <b>-1</b> | <b>-1</b> | <b>+1</b> |
| <b>DOE_03</b>        | <b>-1</b> | <b>-1</b> | <b>+1</b> | <b>-1</b> |
| <b>DOE_04</b>        | <b>-1</b> | <b>-1</b> | <b>+1</b> | <b>+1</b> |
| <b>DOE_05</b>        | <b>-1</b> | <b>+1</b> | <b>-1</b> | <b>-1</b> |
| <b>DOE_06</b>        | <b>-1</b> | <b>+1</b> | <b>-1</b> | <b>+1</b> |
| <b>DOE_07</b>        | <b>-1</b> | <b>+1</b> | <b>+1</b> | <b>-1</b> |
| <b>DOE_08</b>        | <b>-1</b> | <b>+1</b> | <b>+1</b> | <b>+1</b> |
| <b>DOE_09</b>        | <b>+1</b> | <b>-1</b> | <b>-1</b> | <b>-1</b> |
| <b>DOE_10</b>        | <b>+1</b> | <b>-1</b> | <b>-1</b> | <b>+1</b> |
| <b>DOE_11</b>        | <b>+1</b> | <b>-1</b> | <b>+1</b> | <b>-1</b> |
| <b>DOE_12</b>        | <b>+1</b> | <b>-1</b> | <b>+1</b> | <b>+1</b> |
| <b>DOE_13</b>        | <b>+1</b> | <b>+1</b> | <b>-1</b> | <b>-1</b> |
| <b>DOE_14</b>        | <b>+1</b> | <b>+1</b> | <b>-1</b> | <b>+1</b> |
| <b>DOE_15</b>        | <b>+1</b> | <b>+1</b> | <b>+1</b> | <b>-1</b> |
| <b>DOE_16</b>        | <b>+1</b> | <b>+1</b> | <b>+1</b> | <b>+1</b> |

- **Résultats du plan d'expériences**

Le tableau suivant présente les résultats du plan d'expériences, incluant les coûts moyens par pas de temps et les écarts d'optimalité obtenus après évaluation de chaque modèle entraîné sur 500 épisodes.

Tableau 3.6 Résultats du plan d'expérience

| Configuration | Écart d'optimalité | Coût moyen par pas de temps |
|---------------|--------------------|-----------------------------|
| DOE_01        | 17,36              | 0,92                        |
| DOE_02        | 33,75              | 1,049                       |
| DOE_03        | 33,75              | 1,049                       |
| DOE_04        | 33,75              | 1,049                       |
| DOE_05        | 17,36              | 0,92                        |
| DOE_06        | 33,75              | 1,049                       |
| DOE_07        | 17,36              | 0,92                        |
| DOE_08        | 33,75              | 1,049                       |
| DOE_09        | 33,26              | 1,045                       |
| DOE_10        | 56,49              | 1,227                       |
| DOE_11        | 33,26              | 1,045                       |
| DOE_12        | 33,26              | 1,045                       |
| DOE_13        | 17,04              | 0,918                       |
| DOE_14        | 17,04              | 0,918                       |
| DOE_15        | 6378089            | 50005                       |
| DOE_16        | 33,26              | 1,045                       |

La configuration DOE\_15 a été exclue de l'analyse en raison de ses résultats aberrants (écart d'optimalité  $> 6$  millions %), considérés comme une défaillance d'apprentissage. Quatre configurations atteignent des écarts inférieurs à 18% : DOE\_13 (17,04%), DOE\_14 (17,04%), ainsi que DOE\_01, DOE\_05 et DOE\_07 (17,36% chacune). À l'opposé, la

configuration DOE\_10 présente l'écart le plus élevé avec 56,49%. Les coûts moyens par pas de temps varient de 0,918 (configurations optimales) à 1,227 (configuration DOE\_10), comparés à la référence théorique de 0,784 trouvée en utilisant la formule établie par (Bielecki & Kumar, 1988). Ces résultats constituent la base empirique pour identifier les facteurs déterminants de la performance et établir les principes d'optimisation des hyperparamètres. L'analyse des résultats révèle une meilleure configuration (DOE\_13) qui dépasse les autres combinaisons testées. Les caractéristiques de cette configuration sont :

- **Longueur d'épisode** : 10 000 pas de temps
  - **Longueur d'entraînement** : 10 000 000 pas de temps
  - **Taux d'apprentissage** : 0,0001
  - **Facteur d'échelle** : 0.001
  - **Performance atteinte** : 17,04% d'écart d'optimalité
- 
- **Analyse des effets principaux** :

La figure 3.9 ci-dessous illustre l'analyse des effets principaux des facteurs étudiés :

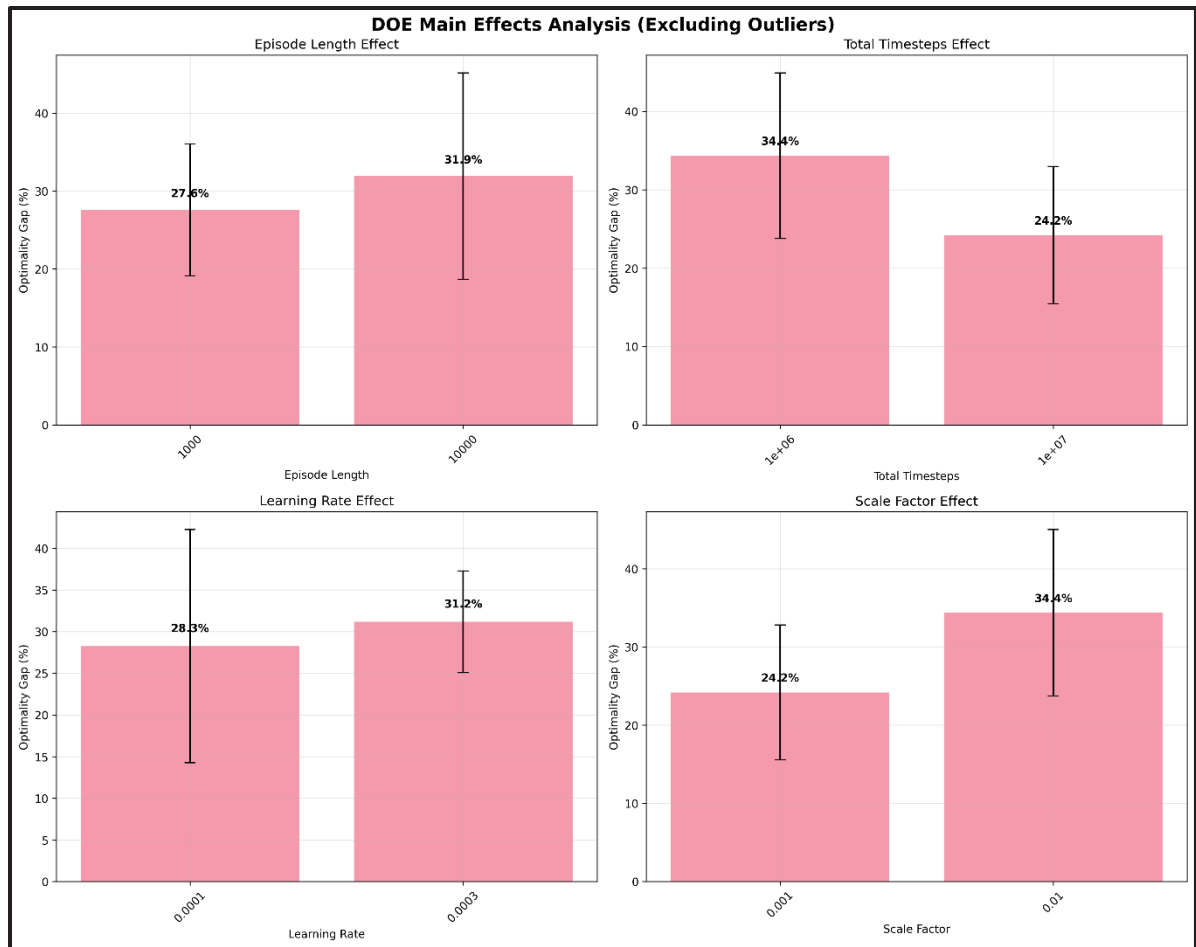


Figure 3.9 Analyse des effets principaux

L'analyse des effets principaux révèle une hiérarchie claire d'importance des facteurs sur l'écart d'optimalité, permettant de prioriser les efforts d'optimisation. Le facteur d'échelle émerge comme le paramètre le plus critique avec un effet de +10,18%, dégradant la performance lorsqu'il passe de 0,001 (24,19% d'écart) à 0,01 (34,38% d'écart). Cette sensibilité confirme l'importance cruciale de la normalisation des signaux de récompense pour la stabilité de l'apprentissage. Le volume d'entraînement présente un effet bénéfique de magnitude comparable (-10,14%), améliorant la performance avec l'extension de 1 000 000 à 10 000 000 pas de temps (34,36% vs 24,22% d'écart). La longueur d'épisode montre un effet modéré (+4,34%) apparemment défavorable aux épisodes longs, tandis que le taux d'apprentissage présente l'impact le plus faible (+2,94%), favorisant légèrement l'approche

conservatrice. Ces résultats quantifient précisément l’impact individuel de chaque hyperparamètre et établissent une base pour l’optimisation.

La figure 3.10 suivante présente l’analyse des effets d’interaction des facteurs étudiés :

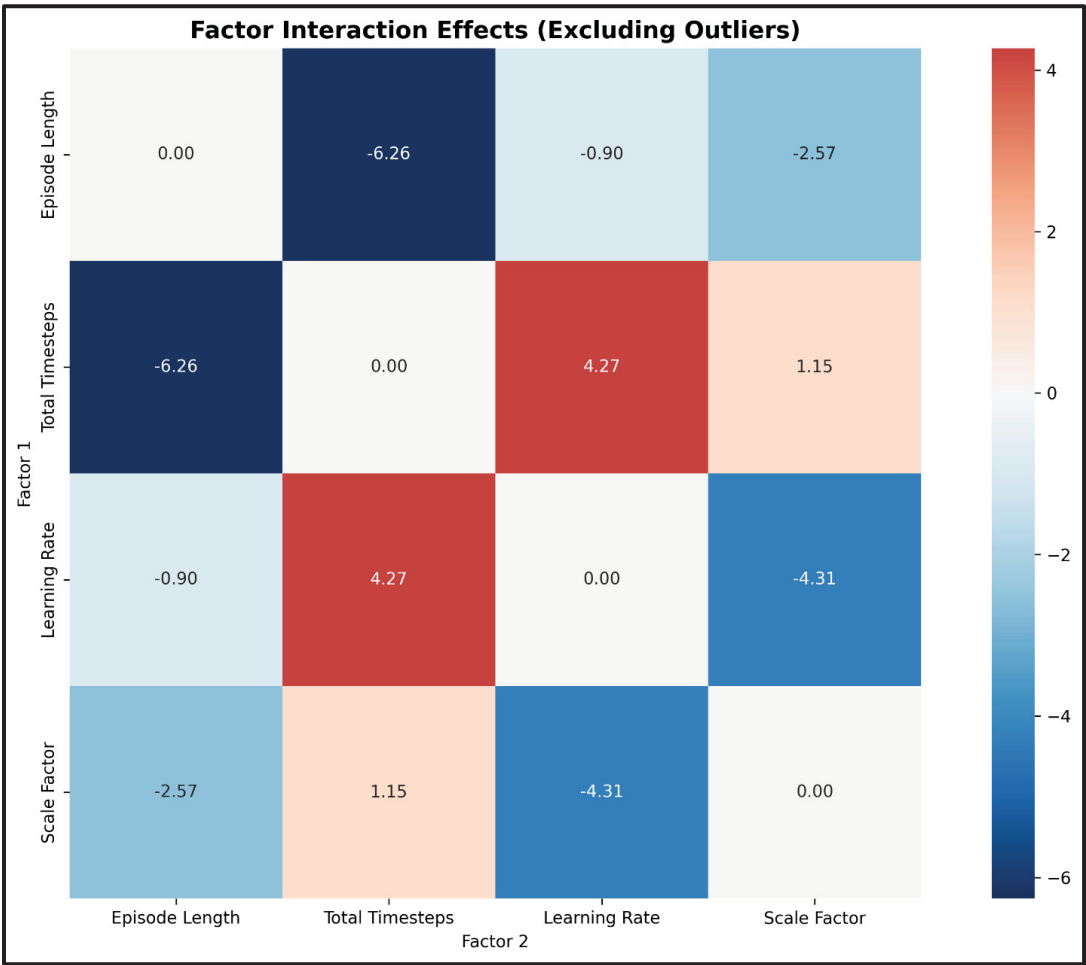


Figure 3.10 Analyse des effets d’interaction

Les interactions entre facteurs révèlent des synergies critiques qui expliquent les résultats contre-intuitifs observés dans les effets principaux. L’interaction la plus forte, longueur d’épisode × Longueur d’entraînement (-6,26%), constitue la clé d’interprétation : les bénéfices des épisodes longs ne se manifestent qu’avec un volume d’entraînement élevé (22,44% d’écart pour les deux à haut niveau vs 29,65% pour les deux à bas niveau). Cette synergie de 7,21 points explique pourquoi la configuration DOE\_13 atteint l’optimum malgré

l'effet principal négatif apparent des épisodes longs. L'interaction Taux d'apprentissage × Facteur d'échelle (-4,31%) révèle un effet de compensation crucial : la combinaison conservatrice (les deux à bas niveau) génère une synergie de stabilité de 12,25 points (21,25% vs 33,50% d'écart). Ces interactions croisées démontrent que l'optimisation des hyperparamètres RL nécessite une approche systémique considérant les effets synergiques plutôt qu'une optimisation isolée de chaque paramètre.

L'analyse DOE confirme l'efficacité de la configuration DOE\_13 qui atteint un écart d'optimalité de 17,04%, représentant une amélioration de 39,46 points par rapport à la configuration la moins performante. Cette performance est obtenue avec 10 000 en longueur d'épisodes, 10 000 000 pas de temps en longueur d'entraînement, un taux d'apprentissage de 0,0001 et un facteur d'échelle de 0,001. L'analyse de 15 configurations (après exclusion de la configuration aberrante DOE\_15) révèle des effets pratiquement significatifs, même si la significativité statistique reste limitée par la taille d'échantillon. Ces résultats valident l'approche d'optimisation systématique et fournissent une configuration de référence robuste pour le système M1P1. La performance obtenue démontre que l'apprentissage par renforcement peut s'approcher de l'optimum théorique avec une optimisation appropriée des hyperparamètres.

Dans cette section, nous avons mis en évidence que l'approche basée sur l'apprentissage par renforcement peut aboutir à des résultats significatifs. Cette méthodologie basée sur le RL a appris une politique de contrôle similaire à celle obtenue au moyen d'approches analytiques, avec des différences au niveau des indicateurs de performance. Ceci est compréhensible et attendu étant donné que l'apprentissage par renforcement est une approche basée sur des estimations et des simplifications des composantes clés de l'apprentissage afin d'obtenir de bons résultats dans un temps raisonnable, ce qui représente une alternative aux méthodes analytiques qui sont souvent limitées ou incapables de fournir des résultats dans des délais raisonnables pour des systèmes complexes. La qualité de l'apprentissage peut être améliorée en agissant sur les hyperparamètres de l'apprentissage. Dans la section suivante, nous allons étendre cette approche vers un système plus complexe que le M1P1 de base. Nous ajoutons une autre variable d'état qui représente le niveau des émissions générées par la machine de

fabrication. Notre objectif est de montrer que l'approche adoptée permet d'apprendre des politiques similaires à des politiques déjà démontrées dans la littérature pour un système de production similaire.

### 3.6 Extension : système M1P1 avec contraintes des émissions

Dans cette section, nous présentons les résultats de l'application de l'approche basée sur l'apprentissage par renforcement sur le système M1P1 avec des émissions des GES. Nous détaillons dans un premier lieu, les caractéristiques du système, les contraintes économiques et environnementales ainsi que l'environnement d'apprentissage de notre modèle.

#### 3.6.1 Paramètres du système

Les paramètres du système M1P1 avec émissions des GES sont présentés dans le tableau suivant

Tableau 3.7 paramètres du système

| $c^+$ | $c^-$ | $c_e$ | $d$ | $U_{max}$ | $q_{10}$ | $q_{01}$ | $\theta$ | $L$ | $x_0$ | max steps |
|-------|-------|-------|-----|-----------|----------|----------|----------|-----|-------|-----------|
| 0.5   | 20    | 12    | 1   | 2         | 0.1      | 2        | 2        | 350 | 0     | 500       |

#### 3.6.2 Représentation d'état

##### 3.6.2.1 Espace d'observation

L'état du système est représenté par un vecteur tridimensionnel  $s = [\text{niveau d'inventaire}, \text{niveau des émissions}, \text{état de la machine}]$  qui est représenté par les variables  $s = [x, e, \varepsilon]$  et qui capture l'information complète et nécessaire à la prise de décision.

Pour la première variable d'état, le niveau d'inventaire  $x$ , il représente la première composante continue du système et dispose des caractéristiques suivantes :

- Domaine de variation :  $\mathbb{R}$  (valeurs réelles)

- Interprétation : valeurs positives = stock disponible, valeurs négatives = pénurie (demande en attente).

La deuxième variable d'état, le niveau des émissions, représente la deuxième composante du système et dispose des caractéristiques suivantes :

- Domaine de variation :  $\mathbb{R}^+$  (valeurs réelles positives).
- Interprétation : niveau des émissions des GES générées par la machine de production.

La troisième variable d'état, l'état de la machine, il représente la composante discrète du système et dispose des caractéristiques suivantes :

- Domaine de variation :  $\{0, 1\}$
- Interprétation : 0 = machine en panne, 1 = machine opérationnelle

### 3.6.3 Dynamiques du Système

Le système implémente les transitions stochastiques selon le processus de Markov défini dans la section précédente :

- **Taux de panne** :  $\lambda = 0.1$  par pas de temps (si machine opérationnelle)
- **Taux de réparation** :  $\mu = 2,0$  par pas de temps (si machine en panne)

L'inventaire évolue selon l'équation suivante :

$$Inventaire(t + 1) = Inventaire(t) + Production(t) - Demande(t) \quad (3.20)$$

Cette équation capture la dynamique du système où l'accumulation d'inventaire résulte de l'écart entre production et demande. L'évolution de cette variable d'état détermine les transitions du système d'un état à l'autre.

Les émissions évoluent selon l'équation suivante :

$$Émissions(t + 1) = Émissions(t) + \theta \times Production(t) \quad (3.21)$$

Où  $\theta$  représente le facteur d'émission par unité produite ( $\theta = 2$  unités CO<sub>2</sub>/unité produite).

Cette équation capture la dynamique d'accumulation des émissions directement liées à l'activité de production. Contrairement à l'inventaire qui peut diminuer selon la demande, les émissions s'accumulent de manière irréversible pendant l'horizon de planification. L'évolution de cette seconde variable d'état introduit une contrainte temporelle supplémentaire dans les décisions de l'agent, créant un compromis nécessaire entre performance économique et respect des limites environnementales.

### 3.6.4 Fonction de Récompense

La fonction de coût instantané pour le système M1P1 avec émissions des GES est donnée par l'équation suivante :

$$g(x, e) = c^+ * x^+ + c^- * x^- + c_e * \max(0, e(t_{T_i}) - L) \quad (3.22)$$

La récompense d'apprentissage par renforcement est définie comme :

$$R(s, a) = -g(x, e) \times \text{facteur\_échelle} \quad (3.23)$$

### 3.6.5 Implémentation du modèle d'apprentissage pour M1P1 avec émissions des GES

#### 3.6.5.1 Architecture du modèle

Nous avons adopté le même modèle d'apprentissage PPO pour ce cas d'extension vue qu'il a déjà démontré son efficacité pour ce type de problème de contrôle.

Notre configuration PPO utilise la même architecture des réseaux de neurones entièrement connectés. Les réseaux d'acteur et critique partagent une structure similaire avec des couches cachées. La différence entre cette architecture et l'architecture précédente réside dans les points suivants :

- La couche d'entrée avec trois valeurs (Vecteur d'entrée tridimensionnel  $[x, e, \varepsilon]$  ).
- Le nombre de couches cachées (2 pour l'acteur et 3 pour le critique).
- Le nombre de neurones dans chaque couche. (Voir figures ci-dessous)
- La fonction d'activation : Tanh.

Les figures ci-dessous illustrent l'architecture :

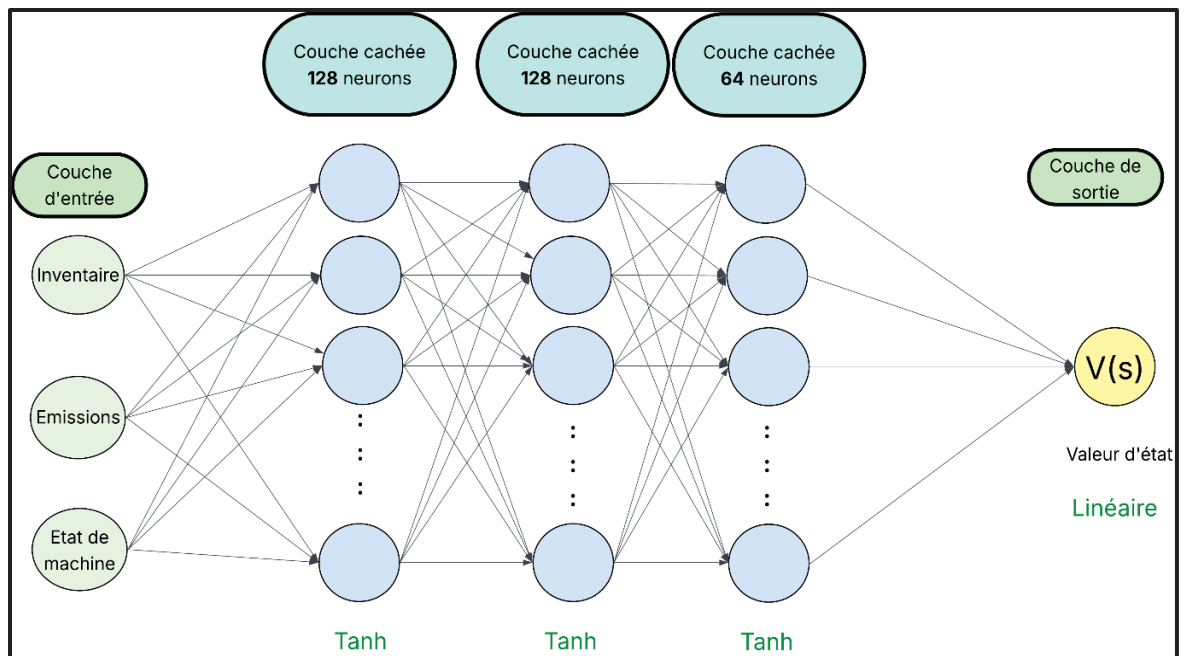


Figure 3.11 Architecture réseau de neurones du critique

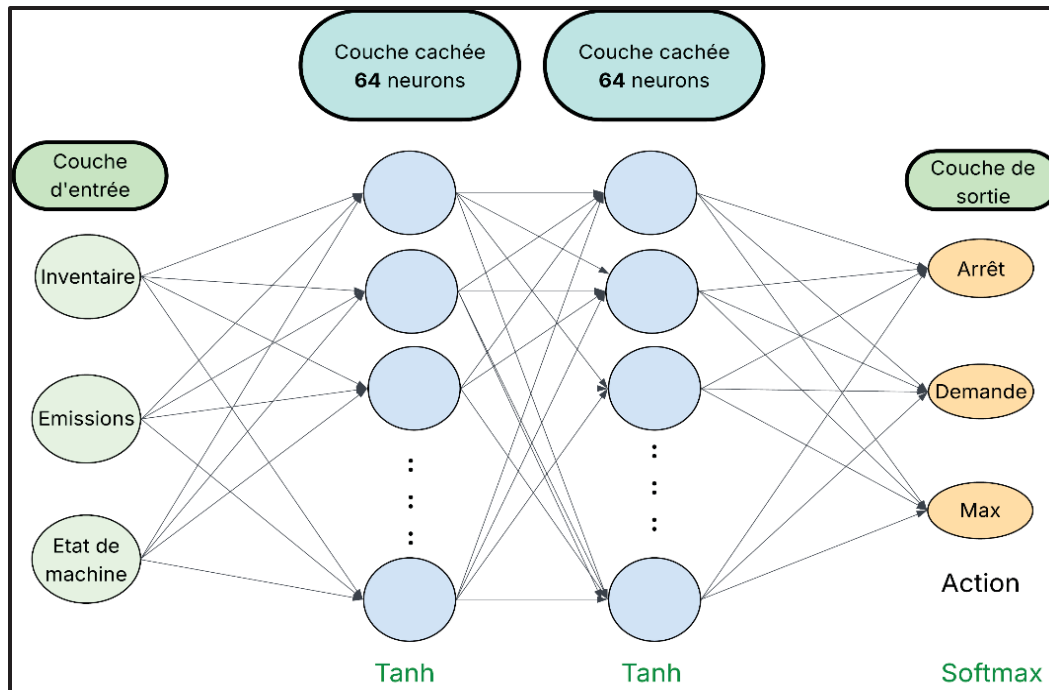


Figure 3.12 Architecture réseau de neurones de l'acteur

Nous avons adapté l'architecture présentée ci-dessus pour gérer la complexité du système étudié. L'acteur utilise une architecture symétrique [64, 64] avec activation Tanh, tandis que le critique emploie une architecture plus profonde [128, 128, 64] pour mieux estimer les valeurs dans l'espace d'état tridimensionnel.

### 3.6.5.2 Hyperparamètres utilisés pour l'apprentissage

Tableau 3.8 Hyperparamètre du modèle

| <i>Learning rate</i> | <i>Batch size</i> | <i>Clipping Ratio</i> | <i>Discount factor</i> |
|----------------------|-------------------|-----------------------|------------------------|
| $10^{-4}$            | 512               | 0.2                   | 0.99                   |

Pour le système étendu, nous adaptons la configuration optimale identifiée (DOE\_13) en conservant les paramètres critiques (learning rate  $1e-4$ , gamma 0.99, clip range 0.2) tout en ajustant les paramètres de volume pour gérer la complexité accrue. L'augmentation du batch size (de 64 à 512) reflète le besoin de plus de données pour l'apprentissage stable dans l'espace d'état tridimensionnel.

### 3.6.6 Résultats et interprétation

Dans cette section, nous présentons les résultats de l'entraînement du modèle d'apprentissage par renforcement PPO appliqué au système M1P1 avec des contraintes des émissions des GES. Ces résultats illustrent la politique de contrôle apprise par l'agent.

La figure ci-dessous présente cette politique sous forme de graphique tridimensionnel où :

- L'axe des abscisses représente le niveau d'inventaire.
- L'axe des ordonnées représente le niveau des émissions.
- L'axe des z indique l'action à prendre.

Cette visualisation 3D permet d'analyser comment l'agent équilibre les contraintes économiques (inventaire) et environnementales (émissions) dans ses décisions de production.

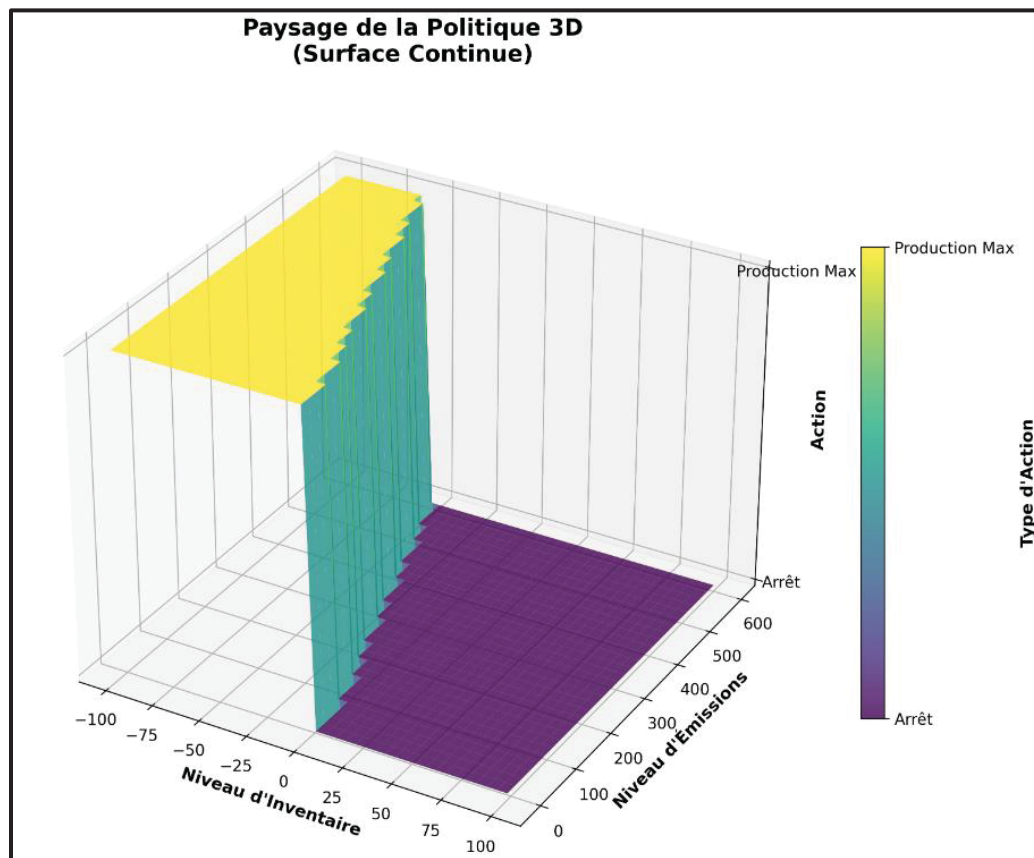


Figure 3.13 Politique de contrôle appris par le modèle PPO

La Figure 3.13 présente la politique de contrôle apprise par l'agent PPO pour le système MIP1 avec des contraintes des émissions des GES. La politique apprise est une politique de forme « Hedging point policy » avec des seuils multiples d'inventaire. Elle présente une segmentation claire de l'espace d'état en trois régions distinctes :

- Zone de production maximale (Jaune) : Correspond aux situations où l'inventaire est faible et les émissions n'ont pas encore atteint des niveaux critiques. L'agent privilégie une production maximale pour reconstituer le stock.
- Zone de production équilibrée (Bleu) : Représente une région de transition où l'agent adapte sa production en fonction de l'équilibre entre besoins de stockage et contrainte d'émissions.
- Zone d'arrêt de production (Violet) : Lorsque les émissions et le niveau d'inventaire atteignent des niveaux élevés, l'agent arrête la production pour éviter les coûts de stockage et des émissions excessifs.

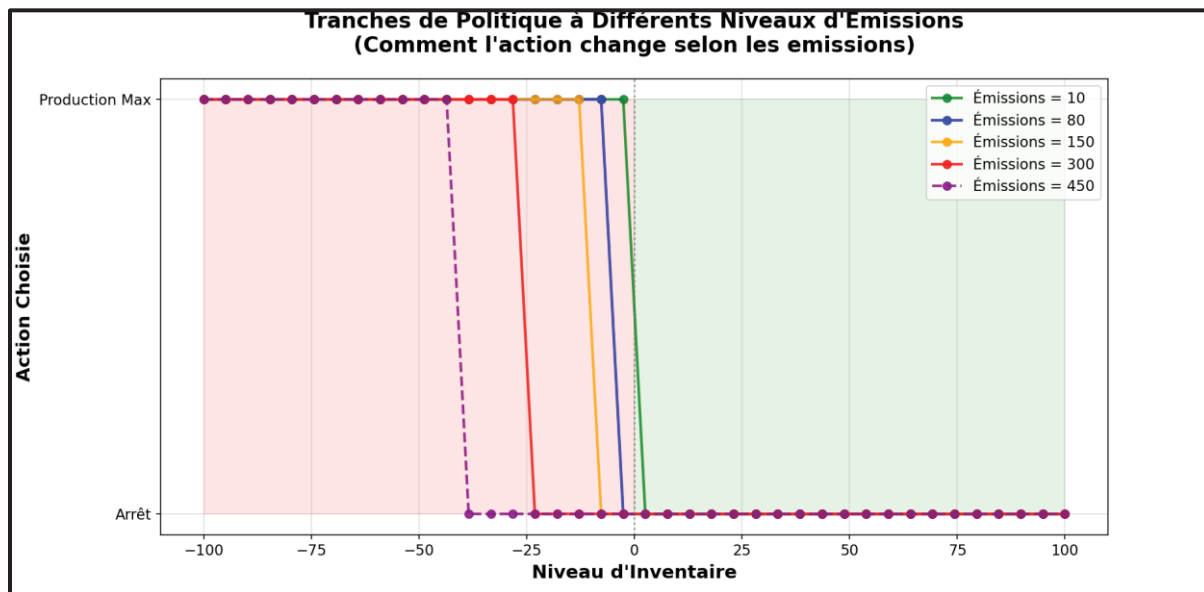


Figure 3.14 Tranches de politique de contrôle à différents niveaux d'émissions

La figure 3.14 représente le seuil d'inventaire à différents niveaux d'émissions (10, 80, 150, 300, 450). On constate que le seuil d'inventaire est dynamique. Il diminue au fur et à mesure

que les émissions augmentent. Le seuil s'ajuste pour réduire progressivement les émissions générées par la machine de production même avant que ce niveau dépasse la limite réglementaire  $L$ .

### 3.6.7 Comparaison qualitative avec les politiques de référence

Notre politique apprise présente des similarités remarquables avec la politique EHPP proposée par (Ben-Salem et al., 2015). L'approche à seuils multiples est clairement visible dans notre surface 3D, où l'agent a développé des frontières de décision qui correspondent conceptuellement aux deux seuils d'inventaire identifiés dans leur travail.

La structure obtenue s'aligne également avec les conclusions de (Kenné et al., 2024) concernant la nécessité d'une politique à seuils multiples. Notre agent PPO a spontanément développé des seuils adaptatifs qui régulent la production en fonction des niveaux d'émissions et d'inventaire, reproduisant qualitativement le comportement quasi optimal décrit dans leur résolution des équations Hamilton-Jacobi-Bellman.

Points de convergence significatifs :

- **Contrôle couplé** : Comme préconisé par les deux références, notre agent a appris que les décisions de production ne peuvent être prises indépendamment des niveaux d'émissions et d'inventaire.
- **Politique de Type Bang-Bang** : La transition nette entre les zones d'action confirme l'apprentissage d'une politique de contrôle discontinue de type bang-bang, caractéristique des solutions optimales dans ce type de problème démontré numériquement dans les travaux de (Kenné et al., 2024) et d'autres travaux pour des problématiques similaires.
- **Contrainte environnementale** : L'arrêt de production dans les zones à fortes émissions démontre que l'agent a intégré l'importance des contraintes environnementales dans sa stratégie.

Nous pouvons donc conclure que notre modèle a appris une politique de contrôle identique à des politiques valides par des approches numériques et par simulation pour des problèmes de contrôle de production similaires en termes de conception et de contraintes.

Nous pouvons donc conclure que notre modèle d'apprentissage par renforcement a appris une politique de contrôle de type Environmental Hedging Point Policy (EHPP) cohérente avec :

- La solution optimale de (Kenné et al., 2024) obtenue par résolution des équations Hamilton-Jacobi-Bellman par approche numérique.
- Les politiques de (Ben-Salem et al., 2015) validées par simulation et méthodologie des surfaces de réponse (RSM).

Cette convergence vers une structure à seuils multiples et dynamiques, sans connaissance préalable de la théorie du contrôle optimal, valide l'efficacité de l'approche d'apprentissage par renforcement pour le contrôle de systèmes de production avec contraintes environnementales.

### **3.7 Conclusion**

La convergence qualitative de notre politique apprise avec les solutions optimales de la littérature pour le cas de M1P1 de base et le cas de M1P1 avec émissions des GES valide l'efficacité de l'apprentissage par renforcement pour le contrôle de systèmes de production dans un environnement stochastique (pannes et réparations aléatoires) avec ou sans contraintes environnementales. Cette cohérence démontre que l'agent PPO a réussi à capturer l'essence des compromis économiques environnementaux, seulement par interaction avec l'environnement et en exploitant le signal de récompense pour optimiser ses décisions, sans connaissance a priori de la théorie du contrôle optimal.

Cette validation conceptuelle ouvre la voie à l'application de l'apprentissage par renforcement à des systèmes de production plus complexes où l'intégration de multiples contraintes rendrait les approches analytiques traditionnelles computationnellement prohibitives.

## CONCLUSION

Cette recherche a atteint ses objectifs en développant avec succès des politiques de contrôle innovantes pour les systèmes de production sous contraintes environnementales selon deux approches complémentaires. L'exploration simultanée des méthodes numériques pour systèmes hybrides et des méthodes d'intelligence artificielle pour systèmes simples a révélé leur potentiel respectif et leur complémentarité pour relever les défis industriels contemporains.

Le développement de la politique environnementale de seuil de couverture dynamique constitue une avancée pour le contrôle des systèmes hybrides fabrication/refabrication. Cette politique fournit un cadre pour faire face aux défis complexes de coordination entre machines en proposant une structure décisionnelle claire qui divise les opérations en zones d'action distinctes. La politique développée intègre simultanément les coûts de stockage, pénurie, production et émissions. L'approche proposée définit des règles de priorisation entre fabrication et refabrication basées sur l'état système et les contraintes environnementales. Les résultats confirment la supériorité de la politique proposée avec des gains de performance substantiels par rapport aux politiques pertinentes de la littérature.

L'application de l'apprentissage par renforcement aux systèmes M1P1 a démontré concrètement l'utilité des techniques d'intelligence artificielle pour le contrôle de production sous contraintes environnementales dans un contexte stochastique. Les politiques obtenues moyennant cette approche sont pareilles à celles obtenues de façon analytique pour le système simple ou numérique pour le cas d'extension, ce qui montre l'efficacité de cette approche pour le contrôle des systèmes de production. L'application réussie de l'apprentissage par renforcement établit concrètement l'utilité des techniques d'intelligence artificielle pour le contrôle de production. La convergence qualitative avec les solutions optimales valide la pertinence de ces méthodes pour les systèmes complexes où les approches traditionnelles atteignent leurs limites.

Les deux approches démontrent leur capacité d'adaptation aux contraintes environnementales. Cette validation croisée renforce la confiance dans les solutions proposées et confirme leur pertinence pour les applications industrielles réelles.

Le chapitre 1 a établi les fondements de notre recherche à travers une analyse critique de la littérature existante. Nous avons défini le contexte de notre recherche, les concepts clés relatifs aux systèmes de production simples et hybrides sous contraintes environnementales et identifié les principales approches méthodologiques. Cette analyse nous a permis de positionner notre contribution scientifique et de démontrer l'originalité de notre démarche. La seconde partie de ce chapitre a formulé la problématique ainsi que les objectifs de recherche attendus de ce travail.

Dans le chapitre 2 nous avons développé une politique de contrôle optimal pour les systèmes hybrides fabrication/refabrication sous contraintes environnementales. Notre contribution consiste en une politique environnementale de seuil de couverture dynamique qui optimise simultanément les aspects économiques et environnementaux tout en coordonnant efficacement les opérations entre machines. Cette politique traite les défis complexes de priorisation selon l'état du système et les contraintes d'émissions. L'analyse de sensibilité et l'analyse comparative révèlent la supériorité de notre approche par rapport aux politiques benchmark dans la littérature.

Dans le chapitre 3, nous avons exploré le potentiel de l'apprentissage par renforcement pour le contrôle des systèmes de production M1P1 non fiables avec des contraintes environnementales. Notre approche repose sur l'application de l'algorithme PPO dont nous avons défini les facteurs influençant la performance de la politique grâce à un plan d'expériences. L'approche a été progressivement validée, d'abord sur une configuration de base puis étendue aux contraintes environnementales. Les politiques obtenues sont similaires qualitativement aux politiques optimales pour ce type de problème de contrôle et confirment ainsi l'utilité des techniques d'intelligence artificielle et établissent une base méthodologique robuste pour des applications futures plus complexes.

Cette étude présente certaines limites qu'il convient de reconnaître. Notre modélisation se concentre sur des configurations simplifiées (M1P1 pour le RL et M1-M2 pour l'approche numérique) qui, bien que représentatives, ne capturent pas toute la complexité des systèmes industriels multiproduits et multimachines. L'écart d'optimalité de 17,10% obtenu avec PPO suggère des opportunités d'amélioration via des techniques d'optimisation plus avancées.

Notre modélisation des contraintes environnementales par pénalités linéaires et l'hypothèse de qualité homogène des retours constituent des simplifications par rapport aux mécanismes réglementaires réels et à la variabilité observée en pratique. Enfin, nos résultats reposent sur des simulations numériques et comparaisons avec des benchmarks théoriques sans validation sur données industrielles réelles, limitant ainsi notre capacité à évaluer la robustesse des politiques face aux perturbations imprévues.

Plusieurs perspectives d'extension émergent de ces travaux. L'application des techniques d'apprentissage par renforcement validées aux systèmes hybrides développés constitue une voie naturelle de convergence entre nos deux axes de recherche. L'extension vers des configurations multimachines et multiproduit représenterait un défi intéressant nécessitant des mécanismes de coordination avancés. D'autres dimensions peuvent enrichir ces modèles : l'intégration de la maintenance corrective et/ou préventive, la prise en compte de la variabilité qualité des retours, ou encore l'adaptation aux fluctuations de demande. L'optimisation des hyperparamètres et l'exploration du transfert d'apprentissage entre systèmes apparentés constituent également des directions de recherche prometteuses.



## ANNEXE I

### CONDITIONS D'OPTIMUM

- **Conditions d'optimum:**

**Mode 1 :**

$$\begin{aligned} \rho v(1, x, e) = \min_{(u_1, u_2) \in A(1)} \{ & g(1, x, e) + \frac{\partial v(1, x, e)}{\partial x_1} \cdot (u_1 + u_2 - d) + \frac{\partial v(1, x, e)}{\partial e} \cdot (\theta_1 \cdot u_1 \\ & + \theta_1 \cdot u_2) + q_{11} \cdot v(1, x, e) + q_{12} \cdot v(2, x, e) + q_{13} \cdot v(3, x, e) \\ & + \sum_{i=1}^{\infty} c_e * \max(0, e(t_{Ti}) - L) \} \end{aligned} \quad (\text{A I-1})$$

**Mode 2 :**

$$\begin{aligned} \rho v(2, x, e) = \min_{(u_1, u_2) \in A(2)} \{ & g(2, x, e) + \frac{\partial v(2, x, e)}{\partial x_1} \cdot (u_1 - d) + \frac{\partial v(2, x, e)}{\partial e} \cdot \theta_1 \cdot u_1 \\ & + q_{21} \cdot v(1, x, e) + q_{22} \cdot v(2, x, e) + q_{24} \cdot v(4, x, e) \\ & + \sum_{i=1}^{\infty} c_e * \max(0, e(t_{Ti}) - L) \} \end{aligned} \quad (\text{A I-2})$$

**Mode 3 :**

$$\begin{aligned} \rho v(3, x, e) = \min_{(u_1, u_2) \in A(3)} \{ & g(3, x, e) + \frac{\partial v(3, x, e)}{\partial x_1} \cdot (u_2 - d) + \frac{\partial v(3, x, e)}{\partial e} \cdot \theta_2 \cdot u_2 \\ & + q_{31} \cdot v(1, x, e) + q_{33} \cdot v(2, x, e) + q_{34} \cdot v(4, x, e) \\ & + \sum_{i=1}^{\infty} c_e * \max(0, e(t_{Ti}) - L) \} \end{aligned} \quad (\text{A I-3})$$

**Mode 4 :**

$$\begin{aligned} \rho v(4, x, e) = \min_{(u_1, u_2) \in A(4)} \{ & g(4, x, e) + \frac{\partial v(4, x, e)}{\partial x_1} \cdot (-d) + q_{42} \cdot v(2, x, e) \\ & + q_{43} \cdot v(3, x, e) + q_{44} \cdot v(4, x, e) + \sum_{i=1}^{\infty} c_e * \max(0, e(t_{Ti}) - L) \} \end{aligned} \quad (\text{A I-4})$$



## ANNEXE II

### APPROCHE DE RÉOLUTION NUMÉRIQUE DES ÉQUATIONS HJB

La résolution analytique des équations HJB est quasi impossible en raison de la complexité du problème. Par conséquent, nous utiliserons une résolution numérique via des approximations des dérivées partielles comme suit :

Nous notons  $v_h(\varepsilon, x, e)$  l'approximation pour la fonction de valeur  $v(\varepsilon, x, e)$  et les pas  $h_x$ ,  $h_e$  de discrétisation pour  $x$  (stock) et  $e$  (émissions) respectivement.

L'approximation des dérivées partielles est donnée par :

$$\frac{\partial v(\varepsilon, x, e)}{\partial x} = \begin{cases} \frac{v_h(\varepsilon, x + h_x, e) - v_h(\varepsilon, x, e)}{h_x} & \text{si } u_1 + u_2 - d \geq 0 \\ \frac{v_h(\varepsilon, x, e) - v_h(\varepsilon, x - h_x, e)}{h_x} & \text{si } u_1 + u_2 - d < 0 \end{cases} \quad (\text{A II-1})$$

$$\frac{\partial v(\varepsilon, x, e)}{\partial e} = \frac{v_h(\varepsilon, x, e + h_e) - v_h(\varepsilon, x, e)}{h_e}$$

La variation du stock peut être soit positive soit négative. D'autre part, le niveau d'émissions est uniquement croissant.

Ainsi, les équations HJB deviennent :

$$\begin{aligned} v_h(\varepsilon, x, e) = & \min_{(u_1, u_2) \in A(\varepsilon)} \left( \frac{g(\cdot)}{\Omega_h^\varepsilon \cdot \left(1 + \frac{\rho}{\Omega_h^\varepsilon}\right)} + \frac{1}{1 + \frac{\rho}{\Omega_h^\varepsilon}} \cdot (P_x^+(\varepsilon) \cdot v_h(\varepsilon, x + h_x, e) \right. \\ & + P_x^-(\varepsilon) v_h(\varepsilon, x - h_x, e) + P_e(\varepsilon) \cdot v_h(\varepsilon, e + h_e) + \sum_{\beta \neq \alpha} P^\beta(\alpha) \cdot v_h(\beta, x, e)) \\ & \left. + \frac{1}{\Omega_h^\varepsilon \cdot \left(1 + \frac{\rho}{\Omega_h^\varepsilon}\right)} \cdot \sum_{i=1}^{\infty} c_e * \max(0, e(t_{Ti}) - L) \right) \end{aligned} \quad (\text{A II-2})$$

Avec :

$$\Omega_h^\varepsilon = |q_{\varepsilon\varepsilon}| + \frac{|u_1 + u_2 - d_c|}{h_x} + \frac{\theta_1 \cdot u_1 + \theta_1 \cdot u_2}{h_e}$$

$$P_x^+(\varepsilon) = \begin{cases} \frac{u_1 + u_2 - d_c}{h_x \cdot \Omega_h^\varepsilon} & \text{si } u_1 + u_2 - d_c \geq 0 \\ 0 & \text{ailleurs} \end{cases} \text{ and } P_x^-(\varepsilon) = \begin{cases} \frac{d_c - u_1 - u_2}{h_x \cdot \Omega_h^\varepsilon} & \text{si } u_1 + u_2 - d_c < 0 \\ 0 & \text{ailleurs} \end{cases}$$

$$\text{and } P^\beta(\varepsilon) = \frac{q_{\varepsilon\beta}}{\Omega_h^\varepsilon} \quad \text{and} \quad P_e(\varepsilon) = \frac{\theta_1 \cdot u_1 + \theta_1 \cdot u_2}{h_e \cdot \Omega_h^\varepsilon}$$

▪ Pour  $\varepsilon = 1$  :

$$\begin{aligned} v_h(1, x, e) = & \min_{(u_1, u_2) \in A(1)} \left( \frac{g(\cdot)}{\Omega_h^1 \cdot \left(1 + \frac{\rho}{\Omega_h^1}\right)} \right. \\ & + \frac{1}{1 + \frac{\rho}{\Omega_h^1}} \cdot (P_x^+(1) \cdot v_h(1, x + h_x, e) + P_x^-(1) \cdot v_h(1, x - h_x, e)) \\ & + P_e(1) \cdot v_h(1, x, e + h_e) + P^2(1) \cdot v_h(2, x, e) + P^3(1) \cdot v_h(3, x, e)) \\ & \left. + \frac{1}{\Omega_h^1 \cdot \left(1 + \frac{\rho}{\Omega_h^1}\right)} \cdot \sum_{i=1}^{\infty} C^e * \max(0, e(t_{Ti}) - L) \right) \end{aligned} \quad (\text{A II-3})$$

▪ Pour  $\alpha = 2$  ( $u_2 = 0$ ):

$$\begin{aligned} v_h(2, x, e) = & \min_{(u_1) \in A(2)} \left( \frac{g(\cdot)}{\Omega_h^2 \cdot \left(1 + \frac{\rho}{\Omega_h^2}\right)} \right. \\ & + \frac{1}{1 + \frac{\rho}{\Omega_h^2}} \cdot (P_x^-(2) \cdot v_h(1, x - h_x, e) + P_e(2) \cdot v_h(2, x, e + h_e)) \\ & + P^1(2) \cdot v_h(1, x, e) + P^4(2) \cdot v_h(4, x, e)) \\ & \left. + \frac{1}{\Omega_h^2 \cdot \left(1 + \frac{\rho}{\Omega_h^2}\right)} \cdot \sum_{i=1}^{\infty} C^e * \max(0, e(t_{Ti}) - L) \right) \end{aligned} \quad (\text{A II-4})$$

▪ Pour  $\alpha = 3$  ( $u_1 = 0$ ):

$$\begin{aligned}
v_h(3, x, e) = & \min_{(u_2) \in A(3)} \left( \frac{g.}{\Omega_h^3 \cdot \left(1 + \frac{\rho}{\Omega_h^3}\right)} \right. \\
& + \frac{1}{1 + \frac{\rho}{\Omega_h^3}} \cdot (P_x^-(3) \cdot v_h(3, x_2 - h_x, e) + P_e(3) \cdot v_h(3, x_1, x_2, e + h_3) \\
& + P^1(3) \cdot v_h(1, x, e) + P^4(3) \cdot v_h(4, x, e)) \\
& \left. + \frac{1}{\Omega_h^3 \cdot \left(1 + \frac{\rho}{\Omega_h^3}\right)} \cdot \sum_{i=1}^{\infty} C^e * \max(0, e(t_{Ti}) - L) \right)
\end{aligned} \tag{A II-5}$$

▪ Pour  $\alpha = 4$  ( $u_1 = u_2 = 0$ ):

$$\begin{aligned}
v_h(4, x_1, x_2, 0) \\
= & \left( \frac{g(.)}{\Omega_h^4 \cdot \left(1 + \frac{\rho}{\Omega_h^4}\right)} \right. \\
& + \frac{1}{1 + \frac{\rho}{\Omega_h^4}} \cdot (P_x^-(4) v_h(4, x - h_x, e) + P^2(4) \cdot v_h(2, x, e) + P^3(4) \cdot v_h(3, x, e)) \\
& \left. + \frac{1}{\Omega_h^4 \cdot \left(1 + \frac{\rho}{\Omega_h^4}\right)} \cdot \sum_{i=1}^{\infty} C^e * \max(0, e(t_{Ti}) - L) \right)
\end{aligned} \tag{A II-6}$$



## LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

- Afshar-Bakeshloo, M., Bozorgi-Amiri, A., Sajadi, S. M., & Jolai, F. (2018). A multi-objective Environmental Hedging Point Policy with customer satisfaction criteria. *Journal of Cleaner Production*, 179, 478–494. <https://doi.org/10.1016/j.jclepro.2018.01.041>
- Akella, R., & Kumar, P. (1986). Optimal control of production rate in a failure prone manufacturing system. *IEEE Transactions on Automatic Control*, 31(2), 116–126. <https://doi.org/10.1109/TAC.1986.1104206>
- Arora, S., & Cason, T. N. (1996). Why Do Firms Volunteer to Exceed Environmental Regulations? Understanding Participation in EPA's 33/50 Program. *Land Economics*, 72(4), 413. <https://doi.org/10.2307/3146906>
- Barange, S., & Agarwal, S. (2016). *Green Supply Chain Management – A Review*. 3(4).
- BASNET, C., & and MIZE, J. H. (1995). A rule-based, object-oriented framework for operating flexible manufacturing systems. *International Journal of Production Research*, 33(5), 1417–1431. <https://doi.org/10.1080/00207549508930218>
- Ben-Salem, A., Gharbi, A., & Hajji, A. (2015). An Environmental Hedging Point Policy to control production rate and emissions in unreliable manufacturing systems. *International Journal of Production Research*, 53(2), 435–450. <https://doi.org/10.1080/00207543.2014.946161>
- Bielecki, T., & Kumar, P. R. (1988). Optimality of Zero-Inventory Policies for Unreliable Manufacturing Systems. *Operations Research*, 36(4), 532–541. JSTOR.
- Blackman, A., & Boyd, J. (2002). Tailored Regulation: Will Voluntary Site-Specific Environmental Performance Standards Improve Welfare? *Southern Economic Journal*, 69(2), 309–326. JSTOR. <https://doi.org/10.2307/1061674>
- Boukas, E.-K., & Haurie, A. (1990). Manufacturing flow control and preventing maintenance: A stochastic control approach. *IEEE Transactions on Automatic Control*, 35(9), 1024–1031. <https://doi.org/10.1109/9.58530>
- Bouslah, B., Gharbi, A., Pellerin, R., & Hajji, A. (2013). Optimal production control policy in unreliable batch processing manufacturing systems with transportation delay. *International Journal of Production Research*, 51(1), 264–280. <https://doi.org/10.1080/00207543.2012.676217>
- Brouhle, K., Griffiths, C., & Wolverton, A. (2009). Evaluating the role of EPA policy levers: An examination of a voluntary program and regulatory threat in the metal-finishing

- industry. *Journal of Environmental Economics and Management*, 57(2), 166–181. <https://doi.org/10.1016/j.jeem.2008.07.006>
- Carter, C. R., & Rogers, D. S. (2008). A framework of sustainable supply chain management: Moving toward new theory. *International Journal of Physical Distribution & Logistics Management*, 38(5), 360–387. <https://doi.org/10.1108/09600030810882816>
- Corbett, C. J., & Kirsch, D. A. (2001). INTERNATIONAL DIFFUSION OF ISO 14000 CERTIFICATION. *Production and Operations Management*, 10(3), 327–342. <https://doi.org/10.1111/j.1937-5956.2001.tb00378.x>
- Dolgui, A., Ivanov, D., & Sokolov, B. (2020). Reconfigurable supply chain: The X-network. *International Journal of Production Research*, 58(13), 4138–4163. <https://doi.org/10.1080/00207543.2020.1774679>
- Elkington, J. (1998). Partnerships from *cannibals with forks: The triple bottom line of 21st-century business*. *Environmental Quality Management*, 8(1), 37–51. <https://doi.org/10.1002/tqem.3310080106>
- Entezaminia, A., Gharbi, A., & Ouhimmou, M. (2020). Environmental hedging point policies for collaborative unreliable manufacturing systems with variant emitting level technologies. *Journal of Cleaner Production*, 250, 119539. <https://doi.org/10.1016/j.jclepro.2019.119539>
- Entezaminia, A., Gharbi, A., & Ouhimmou, M. (2021). A joint production and carbon trading policy for unreliable manufacturing systems under cap-and-trade regulation. *Journal of Cleaner Production*, 293, 125973. <https://doi.org/10.1016/j.jclepro.2021.125973>
- Esteso, A., Peidro, D., Mula, J., & Díaz-Madroñero, M. (2023). Reinforcement learning applied to production planning and control. *International Journal of Production Research*, 61(16), 5772–5789. <https://doi.org/10.1080/00207543.2022.2104180>
- Fetting, C. (2020). The European Green Deal. *ESDN Report*, December 2020, ESDN Office, Vienna.
- Francie, K. A., Jean-Pierre, K., Pierre, D., Victor, S., & Vladimir, P. (2015). Stochastic models and numerical solutions for manufacturing/remanufacturing systems with applications to the printer cartridge industry. *Journal of Manufacturing Systems*, 37, 662–671. <https://doi.org/10.1016/j.jmsy.2014.11.019>
- Geyer, R., & Jackson, T. (2004). Supply Loops and Their Constraints: The Industrial Ecology of Recycling and Reuse. *California Management Review*, 46(2), 55–73. <https://doi.org/10.2307/41166210>
- Gottinger, H. W. (1995). Regulatory policies under uncertainty, value of information and greenhouse gas emissions. *Energy policy*, 23(1), 51–56.

- Hajej, Z., Rezg, N., & Gharbi, A. (2017a). Ecological optimization for forecasting production and maintenance problem based on carbon tax. *The International Journal of Advanced Manufacturing Technology*, 88(5–8), 1595–1606. <https://doi.org/10.1007/s00170-016-8877-9>
- Hajej, Z., Rezg, N., & Gharbi, A. (2017b). Joint optimization of production and maintenance planning with an environmental impact study. *The International Journal of Advanced Manufacturing Technology*, 93(1–4), 1269–1282. <https://doi.org/10.1007/s00170-017-0580-y>
- Hoen, K. M. R., Tan, T., Fransoo, J. C., & Van Houtum, G. J. (2014). Effect of carbon emission regulations on transport mode selection under stochastic demand. *Flexible Services and Manufacturing Journal*, 26(1–2), 170–195. <https://doi.org/10.1007/s10696-012-9151-6>
- Huisingh, D., Zhang, Z., Moore, J. C., Qiao, Q., & Li, Q. (2015). Recent advances in carbon emissions reduction: Policies, technologies, monitoring, assessment and modeling. *Journal of Cleaner Production*, 103, 1–12. <https://doi.org/10.1016/j.jclepro.2015.04.098>
- Ivanov, D., & Dolgui, A. (2021). OR-methods for coping with the ripple effect in supply chains during COVID-19 pandemic: Managerial insights and research implications. *International Journal of Production Economics*, 232, 107921. <https://doi.org/10.1016/j.ijpe.2020.107921>
- Jorjani, S., Leu, J., & Scott, C. (2004). Model for the allocation of electronics components to reuse options. *International Journal of Production Research*, 42(6), 1131–1145. <https://doi.org/10.1080/00207540310001632466>
- Kang, K., Pu, W., & Ma, Y. (2018). A Dynamic Programming-Based Sustainable Inventory-Allocation Planning Problem with Carbon Emissions and Defective Item Disposal under a Fuzzy Random Environment. *Mathematical Problems in Engineering*, 2018, 1–18. <https://doi.org/10.1155/2018/9376080>
- Kenne, J. P., & Gharbi, A. (2000). Production planning problem in manufacturing systems with general failure and repair time distributions. *Production Planning & Control*, 11(6), 581–588. <https://doi.org/10.1080/095372800414142>
- Kenné, J.-P., Gharbi, A., Takengny, A. L. K., & Assid, M. (2024). Optimal Control Policy of Unreliable Production Systems Generating Greenhouse Gas Emission. *Sustainability*, 16(13), 5760. <https://doi.org/10.3390/su16135760>
- Kimemia, J., & Gershwin, S. B. (1983). An Algorithm for the Computer Control of a Flexible Manufacturing System. *IIE Transactions*, 15(4), 353–362. <https://doi.org/10.1080/05695558308974659>

- Kushner, H. J., & Dupuis, P. (1992). *Numerical Methods for Stochastic Control Problems in Continuous Time* (Second Edition). Springer. <https://doi.org/10.1007/978-1-4613-0007-6>
- Lambert, S., & Riopel, D. (2003). *Logistique inverse: Revue de littérature* (Les Cahiers du GERAD, G-2003-61, p. 3). École Polytechnique de Montréal.
- Lavoie, P., Gharbi, A., & Kenné, J.-P. (2010). A comparative study of pull control mechanisms for unreliable homogenous transfer lines. *International Journal of Production Economics*, 124(1), 241–251. <https://doi.org/10.1016/j.ijpe.2009.11.022>
- Lee, W., & Yik, F. (2004). Regulatory and voluntary approaches for enhancing building energy efficiency. *Progress in Energy and Combustion Science*, 30(5), 477–499. <https://doi.org/10.1016/j.pecs.2004.03.002>
- Melnyk, S. A., & Smith, R. T. (1996). *Green manufacturing*. Computer Automated Systems of the Society of Manufacturing Engineers.
- Olsder, G., & Suri, R. (1980). Time-optimal control of parts-routing in a manufacturing system with failure-prone machines. *1980 19th IEEE Conference on Decision and Control Including the Symposium on Adaptive Processes*, 722–727. <https://doi.org/10.1109/CDC.1980.271895>
- Paris agreement COP21—2015 (2016). United Nations, *Treaty Series*, vol. 3156, p.19. <https://treaties.un.org/doc/Publication/UNTS/Volume%203156/v3156.pdf>
- Rishel, R. (1975). Control of systems with jump Markov disturbances. *IEEE Transactions on Automatic Control*, 20(2), 241–244. <https://doi.org/10.1109/TAC.1975.1100943>
- Schmalensee, R., & Stavins, R. N. (2013). The SO<sub>2</sub> Allowance Trading System: The Ironic History of a Grand Policy Experiment. *Journal of Economic Perspectives*, 27(1), 103–122. <https://doi.org/10.1257/jep.27.1.103>
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms* (No. arXiv:1707.06347). arXiv. <https://doi.org/10.48550/arXiv.1707.06347>
- Sethi, S. P., Qing, Z., & Xun, Y. Z. (1997). *Hierarchical Production Controls in a Stochastic Two-Machine Flowshop with a Finite Internal Buffer*. 13(1).
- SHIN, R. W., & CHEN, Y.-C. (2000). SEIZING GLOBAL OPPORTUNITIES FOR ACCOMPLISHING AGENCIES' MISSIONS: THE CASE OF ISO 14000. *Public Administration Quarterly*, 24(1), 69–94. JSTOR Business & Economics Collection.
- Stevenson, W. J., Benedetti, C., & Youssef, Y. A. (2012). *La gestion des opérations: Produits et services* (3e édition). Chenelière McGraw-Hill; WorldCat.

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (second edition). MIT Press.
- Thesen, A. (1999). Some simple, but efficient, push and pull heuristics for production sequencing for certain flexible manufacturing systems. *International Journal of Production Research*, 37(7), 1525–1539. <https://doi.org/10.1080/002075499191111>
- United Nations General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development, A/RES/70/1. <https://docs.un.org/en/A/RES/70/1>
- Varsei, M., Soosay, C., Fahimnia, B., & Sarkis, J. (2014). Framing sustainability performance of supply chains with multidimensional indicators. *Supply Chain Management: An International Journal*, 19(3), 242–257. <https://doi.org/10.1108/SCM-12-2013-0436>
- Wallace, D. (2017). *Environmental Policy and Industrial Innovation: Strategies in Europe, the USA and Japan* (1st ed.). Routledge. <https://doi.org/10.4324/9781315145358>
- Wang, Y., Wiegerinck, V., Krikke, H., & Zhang, H. (2013). Understanding the purchase intention towards remanufactured product in closed-loop supply chains. *International Journal of Physical Distribution & Logistics Management*, 43(10), 866–888. <https://doi.org/10.1108/IJPDLM-01-2013-0011>
- Zhang, Q., Liu, Y., Xiang, Y., & Xiahou, T. (2024). Reinforcement learning in reliability and maintenance optimization: A tutorial. *Reliability Engineering & System Safety*, 251, 110401. <https://doi.org/10.1016/j.ress.2024.110401>