

Title of dissertation: Optimizing Industrial Processes through Reinforcement Learning

Name of applicant: Abhijeet Pendyala

Abstract

Optimizing complex industrial processes, particularly those involving resource allocation under uncertainty and strict constraints, presents significant challenges for traditional control methods. This thesis addresses the optimization of container management in a real-world, high-throughput plastic sorting facility, a critical stage impacting sustainability. The core task involves scheduling the emptying of multiple containers, which accumulate different materials at stochastic rates, into a limited number of shared Processing Units (PUs). This process is complicated by factors inherent to industrial settings: stochastic material inflow, noisy sensor readings, significant delays between actions and rewards, safety constraints preventing container overflow, and the need to balance competing objectives such as maximizing throughput, minimizing energy consumption, and ensuring operational safety.

Initial investigations revealed that standard Reinforcement Learning (RL) algorithms (e.g., PPO, TRPO, DQN), while promising, struggle to learn effective policies directly due to these complexities. The challenges of sparse rewards, the need for long-term planning under uncertainty, and the difficulty of managing shared resource contention often lead to sub-optimal or unsafe behaviors. To establish a clear basis for research, the problem was first formalized and released as 'ContainerGym', an open-source RL benchmark environment derived directly from the industrial use case, capturing its inherent difficulties.

To overcome the limitations of baseline RL methods, this research proposes and evaluates a progressively sophisticated set of techniques centered around Proximal Policy Optimization (PPO). First, a multi-stage Curriculum Learning (CL) strategy combined with meticulous reward engineering was developed. This approach systematically guides the agent's learning process, starting with simplified environmental dynamics and gradually introducing complexity, stochasticity, resource constraints, and refined reward signals. This CL methodology (PPO-CL) proved effective in enabling the agent to handle delayed rewards, learn basic safety protocols, manage energy considerations, and achieve high performance, significantly outperforming naive PPO training.

However, even with CL, managing "collision" situations—where multiple containers require the scarce PU simultaneously, leading to potential overflows—remained a significant challenge. Purely model-free RL approaches struggle to anticipate and proactively mitigate

these multi-container interactions effectively. To address this, a novel hybrid approach was developed, augmenting the PPO-CL agent with an inference-time planning mechanism. An offline-trained Collision Model (CM), built using Monte Carlo simulations and an XGBoost classifier, predicts the likelihood of imminent collisions. At inference time, if the PPO-CL agent proposes a "do-nothing" action and a high collision risk is detected, the system overrides the agent's inaction and forces a preemptive empty, creating the PPO-CL-CM agent.

Comprehensive experimental evaluations across various container-to-PU ratios (from 7:1 to 12:1) demonstrate the efficacy of this hybrid strategy. The PPO-CL-CM agent significantly reduces the frequency of collision events and safety-limit violations compared to the PPO-CL agent, while maintaining or improving overall throughput. The analysis provides quantitative insights into the scalability of the system, offering practical guidelines for facility design regarding the optimal number of containers that can be managed by a single PU.

This thesis contributes a validated RL benchmark for industrial resource allocation, a robust curriculum learning framework for tackling complex real-world RL problems with delayed rewards and safety constraints, and a novel hybrid RL-planning approach for effective collision avoidance in resource-constrained systems. The findings demonstrate the potential of combining structured learning techniques with domain-specific predictive models to develop safe, efficient, and scalable control solutions for challenging industrial control problems.

Zusammenfassung

Die Optimierung komplexer industrieller Prozesse, insbesondere solcher, die eine Ressourcenallokation unter Unsicherheit und strengen Randbedingungen erfordern, stellt traditionelle Steuerungs- und Regelungsmethoden vor erhebliche Herausforderungen. Diese Dissertation befasst sich mit der Optimierung des Containermanagements in einer realen Hochdurchsatz-Kunststoffsortieranlage, einer kritischen Prozessstufe mit Auswirkungen auf die Nachhaltigkeit. Die Kernaufgabe besteht in der Planung der Entleerung mehrerer Container, in denen sich unterschiedliche Materialien mit stochastischen Raten ansammeln, in eine begrenzte Anzahl gemeinsam genutzter Verarbeitungseinheiten Processing Units (PUs). Dieser Prozess wird durch inhärente Faktoren industrieller Umgebungen erschwert: stochastischer Materialzufluss, verrauschte Sensordaten, erhebliche Verzögerungen zwischen Aktionen und Belohnungen, Sicherheitsbeschränkungen zur Verhinderung von Containerüberläufen und die Notwendigkeit, konkurrierende Ziele wie Durchsatzmaximierung, Energieverbrauchsminimierung und Gewährleistung der Betriebssicherheit auszubalancieren.

Erste Untersuchungen zeigten, dass Standardalgorithmen des Verstärkenden Lernens (Reinforcement Learning, RL) (z.B. PPO, TRPO, DQN) aufgrund dieser Komplexität Schwierigkeiten haben, direkt effektive Policies zu erlernen. Die Herausforderungen durch seltene (sparse) Belohnungen, die Notwendigkeit langfristiger Planung unter Unsicherheit und die Schwierigkeit der Bewältigung von Konflikten um gemeinsam genutzte Ressourcen führen oft zu suboptimalem oder unsicherem Verhalten. Um eine klare Forschungsgrundlage zu schaffen, wurde das Problem zunächst formalisiert und als ContainerGym veröffentlicht – eine Open-Source-RL-Benchmark-Umgebung, die direkt aus dem industriellen Anwendungsfall abgeleitet wurde und dessen inhärente Schwierigkeiten abbildet.

Um die Einschränkungen von Basis-RL-Methoden zu überwinden, schlägt diese Arbeit eine Reihe progressiv anspruchsvollerer Techniken vor, die auf der Proximal Policy Optimization (PPO) basieren, und bewertet diese. Zunächst wurde eine mehrstufige Curriculum-Learning-Strategie (CL) in Kombination mit sorgfältigem Reward Engineering (Entwurf von Belohnungsfunktionen) entwickelt. Dieser Ansatz leitet den Lernprozess des Agenten systematisch an, beginnend mit vereinfachten Umgebungsdynamiken und der schrittweisen Einführung von Komplexität, Stochastizität, Ressourcenbeschränkungen und verfeinerten Belohnungssignalen. Diese CL-Methodik (PPO-CL) erwies sich als effektiv, um den Agenten in die Lage zu versetzen, mit verzögerten Belohnungen umzugehen, grundlegende

Sicherheitsprotokolle zu erlernen, Energieaspekte zu berücksichtigen und eine hohe Leistung zu erzielen, die das naive PPO-Training deutlich übertrifft.

Jedoch blieb auch mit CL die Handhabung von "Kollisionssituationen" – in denen mehrere Container gleichzeitig die knappe PU benötigen, was zu potenziellen Überläufen führt – eine wesentliche Herausforderung. Rein modellfreie RL-Ansätze haben Schwierigkeiten, diese Multi-Container-Interaktionen effektiv zu antizipieren und proaktiv zu entschärfen. Um dies zu adressieren, wurde ein neuartiger hybrider Ansatz entwickelt, der den PPO-CL-Agenten um einen Inferenzzeit-Planungsmechanismus erweitert. Ein offline trainiertes Kollisionsmodell (Collision Model, CM), erstellt mittels Monte-Carlo-Simulationen und einem XGBoost-Klassifikator, sagt die Wahrscheinlichkeit bevorstehender Kollisionen voraus. Wenn der PPO-CL-Agent zur Inferenzzeit eine „Nichtstun“-Aktion vorschlägt und ein hohes Kollisionsrisiko erkannt wird, setzt das System die Untätigkeit des Agenten außer Kraft und erzwingt eine präventive Entleerung. Diese Architektur wird als PPO-CL-CM-Agent bezeichnet.

Umfassende experimentelle Evaluationen über verschiedene Container-zu-PU-Verhältnisse (von 7:1 bis 12:1) demonstrieren die Wirksamkeit dieser hybriden Strategie. Der PPO-CL-CM-Agent reduziert die Häufigkeit von Kollisionsereignissen und Sicherheitsgrenzverletzungen im Vergleich zum PPO-CL-Agenten signifikant, während der Gesamtdurchsatz beibehalten oder verbessert wird. Die Analyse liefert quantitative Einblicke in die Skalierbarkeit des Systems und bietet praktische Richtlinien für das Anlagendesign hinsichtlich der optimalen Anzahl von Containern, die von einer einzelnen PU verwaltet werden können.

Diese Dissertation trägt einen validierten RL-Benchmark für industrielle Ressourcenallokation bei, ein robustes Curriculum-Learning-Framework zur Bewältigung komplexer realweltlicher RL-Probleme mit verzögerten Belohnungen und Sicherheitsbeschränkungen, sowie einen neuartigen hybriden RL-Planungsansatz zur effektiven Kollisionsvermeidung in ressourcenbeschränkten Systemen. Die Ergebnisse demonstrieren das Potenzial der Kombination strukturierter Lerntechniken mit domänenspezifischen Vorhersagemodellen zur Entwicklung sicherer, effizienter und skalierbarer Steuerungslösungen für anspruchsvolle industrielle Regelungsprobleme.