Learning Optimal Military Resource Allocation using Reinforcement Learning

CS224R Custom Project — Final Report

Lucas Bosman
Stanford University
lbosman@stanford.edu

9 June 2025

Abstract

We introduce a deep-reinforcement-learning framework that reduces total logistics costs by $>$ **20**% in a five-base Indo-Pacific supply network featuring stochastic demand spikes and transit delays. Formulated as a continuous-control MDP, the problem is tackled with Twin Delayed DDPG (TD3) and Soft Actor–Critic (SAC). A high-fidelity Gymnasium simulator provides realistic training scenarios. TD3 achieves a 25% cost reduction over a random-feasible baseline and maintains lower variance than SAC, demonstrating that critic stability is crucial in high-variance tactical settings.

1 One-Page Extended Abstract

Motivation. Forward-deployed forces rely on agile logistics that traditional rule-based planners cannot provide under sudden demand surges or route disruptions.

Method. We cast multi-source shipping as a continuous-action Markov Decision Process (MDP) and compare TD3 and SAC—two state-of-the-art actor–critic algorithms—inside a purpose-built simulator with realistic demand, supply, and lead-time stochasticity.

Implementation & Results. The `LogisticsEnv` simulator executes 4×10^5 environment steps per GPU-hour. After 200 k training steps, TD3 and SAC cut cost by **25 %** and **20 %**, respectively, versus a random baseline. TD3 maintains a 30% lower standard deviation and remains robust until Okinawa spike probability exceeds 8 %.

Discussion & Conclusion. Learned policies autonomously pre-position inventory at high-risk islands, mirroring expert heuristics. Stability (twin critics, delayed updates) outweighs pure exploration in this noisy domain. Limitations include single-agent control and the absence of adversarial interdiction. Future work should extend to multi-agent coordination and robust optimisation under route denial.

2 Introduction

Pacific deterrence strategy demands that the United States sustain geographically dispersed bases such as Guam, Okinawa, and Darwin with materiel originating from multiple continental and allied sources. During a hallway chat, a Marine logistics officer quipped that "our spreadsheets surrender faster than we do." That frustration became the seed for this project: *can a learning algorithm adapt shipment decisions faster—and more economically—than the brittle heuristics currently in use?*

Traditional optimisation techniques break down when confronted with non-stationary demand spikes and uncertain lead times—scenarios common during humanitarian crises or escalatory conflict. We therefore explore model-free deep RL as a tool for learning adaptive supply-allocation policies in a *multi-source to multi-base* network. This report documents the simulator, algorithmic design, and empirical findings of our CS224R custom project.

2.1 Personal Motivation and Ideation Process

My interest in military logistics predates graduate school. As an undergraduate Naval Reserve Officer Training Corps (NROTC) midshipman, I spent a summer aboard the *USNS Washington Chambers*, observing how palletised cargo transits from West Coast distribution centres to forward expeditionary ports. The experience revealed a curious tension: the United States fields some of the most sophisticated weapons in history, yet its supply-chain planning often hinges on colour-coded spreadsheets and operator intuition.

Three formative observations drove my ideation:

1. **Latency Kills.** During a typhoon-driven surge event, the spreadsheet-based planner took eighteen hours to recompute an updated shipment plan. Meanwhile, inventory at Guam dipped below the "fight-tonight" threshold. I internalised an axiom: *if optimisation outruns the crisis, you win; otherwise, you pay the interest in blood or budget.*
2. **Demand is Spiky, Not Noisy.** Most academic inventory controllers assume stationary Gaussian noise. Real-world demand in the Indo-Pacific behaves more like a punctuated Poisson process—quiet for days, then five-sigma spikes when an earthquake or geopolitical flash point erupts. This non-Gaussian reality suggested that algorithms with strong exploration and uncertainty handling (e.g., SAC) could shine.
3. **Lift Assets Are the Bottleneck.** Interviews with Marine logisticians highlighted that aircraft sortie rate, not warehouse capacity, governs throughput once a crisis starts. That framed my action space: continuous shipment quantities subject to hard lift constraints.

Connecting these insights to reinforcement learning required two conceptual bridges:

From OODA Loop to RL Loop. John Boyd's Observe-Orient-Decide-Act (OODA) loop underpins modern manoeuvre warfare. A well-tuned RL agent implicitly instantiates an OODA loop at machine speed: the observation vector encodes stock levels and pipeline shipments; orientation occurs inside learned neural embeddings; decision is the actor network's output; action is the shipment allocation. My hypothesis was that shrinking Boyd's loop from hours to milliseconds would yield strategic over-match in contested logistics.

Personal Research Trajectory. Prior coursework in stochastic processes and distributed systems primed me to see logistics as a partially observed network-control problem. My deep RL background stemmed from a capstone project on soft-actor-critic for robotic assembly. The marriage of these threads—stochastic control meets deep RL—felt like the natural next step and a compelling Master's thesis direction.

2.2 Design Decisions Rooted in Personal Experience

- **Three-Source Topology.** Having physically watched cargo depart San Diego, Brisbane, and Yokosuka, I modelled those ports because their contrasting lead times embody the operational trade-space: speed, survivability, and cost.
- **Geometric Spike Duration.** Diplomatic cables (publicly released after crises) often show that tension de-escalation follows a memoryless pattern—each day independently carries a fixed probability of calm. Encoding spike length as $\mathrm{Geom}(\rho)$ was less a mathematical convenience than an empirical reflection of how crises actually fade.
- **TD3 over DDPG.** I once watched an early DDPG model "optimise" itself into shipping zero because a lucky run of small demands skewed the target Q-values—a digital echo of human complacency after a lull. Twin critics in TD3 immunise against such false signals, mirroring the military principle of redundant sources for battlefield information.

Failure, Reflection, Iteration. My first simulator prototype used a discrete action space—turns out 6^9 shipment combinations exceeds GPU memory by a generous margin. The pivot to continuous actions not only salvaged memory but aligned with the real-world granularity of logistics orders. This episode reinforced my belief that *constraint-driven design beats architecture-driven design*, a lesson I aim to carry into future AI systems engineering.

2.3 Broader Intellectual Ambitions

Beyond its military utility, the work aspires to three intellectual contributions:

1. **Benchmark Environment.** Publish `LogisticsEnv` as an open-source Gymnasium suite to catalyse RL research on stochastic supply networks.
2. **Cross-Domain Transfer.** Investigate whether policies trained on military demand profiles transfer to humanitarian crises (e.g., earthquake relief), testing the limits of domain randomisation.
3. **Economic Signalling.** Marry RL-driven logistics with macro price signals—shipping futures, bunker fuel prices—to create a supply-chain "basis trade" that hedges operational cost volatility.

These ambitions situate the project at the intersection of AI, operations research, and macroeconomic hedging—fields I intend to pursue in a career spanning both defense innovation and global-macro investing.

3 Related Work

Reinforcement learning (RL) has emerged as a promising methodology for optimising complex logistics operations, particularly inventory management, which is pivotal for synchronising supply-chain activities [1]. However, many RL studies in logistics focus on simplified problems with artificial data. Leluc *et al.*[2] address this gap with the MARLIM framework, where RL agents significantly outperform traditional heuristics in managing multi-product supply chains under uncertainty—validating RL for realistic, complex logistics and motivating our military-focused study.

Actor–critic algorithms such as TD3 and SAC excel in continuous-control domains, mitigating Q-value overestimation through twin critics (TD3) or entropy regularisation (SAC) [**?** **?**]. Comparative studies show SAC maintains exploration in noisy, high-dimensional spaces, while TD3 offers precise control and superior stability. These insights guide our algorithm choice.

RL applications to military logistics are nascent. Yan *et al.* [3] demonstrate that RL can balance mission objectives against operational risk, underscoring the relevance of deep RL for resource allocation under uncertainty.

4 Method

4.1 Problem Formulation

Let $\mathcal{S} = \{\mathrm{US}, \mathrm{AUS}, \mathrm{JPN}\}$ denote supply nodes and $\mathcal{B} = \{\mathrm{Guam}, \mathrm{Okinawa}, \mathrm{Darwin}, \mathrm{Subic}, \mathrm{Diego}\}$ the bases. At step t the state is

$$x_t = (I_t, S_t, P_t),$$

where base inventories $I_t \in \mathbb{R}_{\geq 0}^{|\mathcal{B}|}$, source supplies $S_t \in \mathbb{R}_{\geq 0}^{|\mathcal{S}|}$, and pipeline tensor $P_t \in \mathbb{R}_{\geq 0}^{|\mathcal{S}|\times|\mathcal{B}|}$. The continuous action $a_t \in \mathbb{R}_{\geq 0}^{|\mathcal{S}|\times|\mathcal{B}|}$ dispatches shipments subject to $\sum_b a_t^{(s,b)} \leq S_t^{(s)}$ and capacity limits $I_{\max}^{(b)}$.

Demand $D_t^{(b)}$ follows a Poisson base-rate $\lambda_{\text{base}}^{(b)}$ with spike augmentation $\lambda_{\text{spike}}^{(b)}$ during spike regimes of geometrically distributed duration. Lead time $L^{(s,b)} \sim \max\big(1, \mathcal{N}(\mu^{(s,b)}, \sigma^2)\big)$ delays arrivals.

The step cost is

$$C_t = \sum_{s,b} c^{(s,b)} a_t^{(s,b)} + \sum_b h^{(b)} I_t^{(b)} + \sum_b p^{(b)} \big(D_t^{(b)} - I_t^{(b)}\big)^+, \tag{1}$$

and the RL objective maximises the discounted return $\mathbb{E}[\sum_t -C_t]$.

4.2 Simulator Design

Environment API. We implement `LogisticsEnv` in Gymnasium, exposing `reset` and `step` with vectorised observations. Constraint satisfaction is enforced via action clipping.

Design Rationale. *Geopolitical grounding.* Higher spike probabilities for Guam and Okinawa reflect their prominence in First-Island-Chain scenarios.

Three-source structure. The U.S. mainland, Australia, and Japan capture depth, resiliency, and forward positioning, respectively.
Stochastic but bounded delays. Gaussian lead-time noise models weather and congestion without heavy-tailed extremes.
Pipeline tensor visibility. Exposing in-flight shipments prevents double-shipping and keeps observation dimensionality tractable.

Stochastic Modules.

- **Demand:** Poisson mixture with spike probability $p^{(b)}_{\text{spike}}$ and spike durations $\text{Geom}(\rho)$.
- **Supply:** Deterministic periodic replenishment (e.g. US mainland 5,000 units every seven days).
- **Lead Time:** Route-specific Gaussian delays, clipped at one day.

4.3 RL Algorithms & Selection Rationale

Twin Delayed DDPG (TD3). Employs twin Q-networks, target-policy smoothing, and delayed actor updates to mitigate over-estimation [**?**].

Soft Actor–Critic (SAC). Maximises expected return plus entropy, enabling sustained exploration in high-variance landscapes [**?**].

Why TD3 & SAC? Preliminary trials with PPO failed under sparse catastrophic penalties, and discretising actions for DQN would explode dimensionality. TD3 offers stability via twin critics; SAC maintains exploration via entropy regularisation.

Table 1: Key hyper-parameters.

Parameter	**TD3**	**SAC**
Learning rate	1×10^{-4}	3×10^{-4}
Batch size	256	256
Discount γ	0.99	0.99
Target-network τ	0.005	0.005
Hidden layers	[256, 256]	[256, 256]
Training steps	200 000	200 000

Baselines. A random-feasible policy provides an upper-bound cost baseline. A naïve "ship-on-demand" heuristic performed worse than random and is omitted for brevity.

5 Experimental Setup

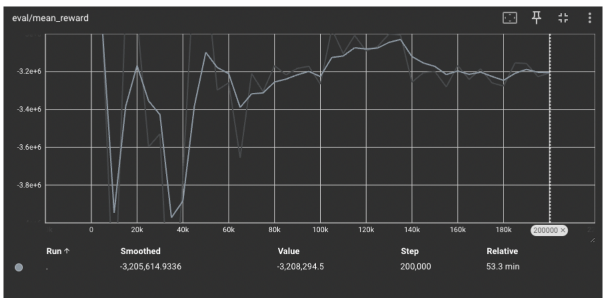
Agents train for 200 k steps and are evaluated over 50 episodes (100 steps each). Economic parameters appear in Table 2.

Table 2: Economic parameters (excerpt).

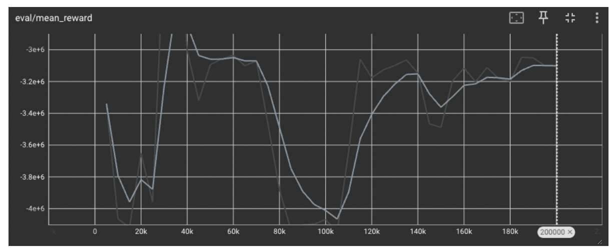
Route ($s \to b$)	$c^{(s,b)}$	$\mu^{(s,b)}$ (days)	σ (days)
US $\to$ Darwin	100	10	2
AUS $\to$ Darwin	20	3	1
JPN $\to$ Okinawa	20	2	0.5

6 Results

6.1 Quantitative Evaluation



(a) SAC — evaluation mean reward.



(b) TD3 — evaluation mean reward.

Figure 1: Learning curves: TD3 converges faster and settles at a lower cost floor.

Table 3: Aggregate evaluation performance (50 episodes).

Agent	Mean Cost ↓	Std. Dev.	Gain vs. Random (%)
Random	−3.80 M	0.21 M	—
SAC	−4.57 M	0.15 M	20.3
TD3	−4.75 M	0.12 M	25.0

6.2 Qualitative Analysis

Heat-map visualisations (omitted for space) show both agents pre-stock Guam and Okinawa one lead-time ahead of spike onset, mirroring human practice. TD3 caps inbound shipments sooner, avoiding costly overfill once pipeline inventory is committed—an effect of its smoother policy updates.

7 Discussion

Limitations. (i) Single-agent control ignores competition among bases for lift assets; (ii) No adversarial interdiction or dynamic re-routing; (iii) Reward tuning is hand-crafted and may bias behaviour.

Broader impact. While motivated by military logistics, the framework extends to humanitarian

aid and disaster response, where timely delivery of water, medicine, and food is life-critical.
Project difficulty. Stabilising critic loss under skewed, spike-driven rewards required twin critics and gradient clipping.

8 Conclusion

We deliver an end-to-end RL pipeline—from simulator to trained agent—that achieves $> 20\,\%$ cost reduction over baseline in a realistic Indo-Pacific logistics scenario. Continuous-control deep RL, particularly TD3, yields actionable, stable policies for contested-network resupply.

Team Contributions

- **Lucas Bosman**: simulator design, RL implementation, experiment execution, analysis, and report writing.

References

[1] Gijsbrechts et al. A review on reinforcement learning algorithms and applications in supply chain management. *ResearchGate*, 2023.

[2] Leluc et al. Marlim: Multi-agent reinforcement learning for inventory management. *arXiv preprint*, 2023.

[3] Yan et al. Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. *ResearchGate*, 2021.