LOGEX: Cost-Sensitive Bayesian Experimentation for Adaptive Decision-Making in Supply Chains

Lorenzo Toni
Amazon.com
Supply Chain Analytics
New York, USA
tonilore@amazon.com

ABSTRACT

Running real-world experiments in supply chains is costly, risky, and often limited by operational constraints. Evaluating a new policy—such as a revised inventory heuristic or a routing strategy—requires partial deployment, active monitoring, and foregone opportunity from not exploiting the current best-known alternative. To address this, we propose LOGEX, a Bayesian framework for cost-sensitive experimentation that models uncertain, evolving reward functions using Bayesian Additive Regression Trees (BART). LOGEX quantifies the expected value of experimentation in economic units, enabling practitioners to weigh the benefit of learning against the cost of conducting operational pilots. Unlike conventional black-box optimization, our approach supports partial rollouts, adapts to non-stationary reward landscapes, and maintains interpretability through rule-based posterior estimates. We validate LOGEX in a synthetic supply chain environment and show that it outperforms cost-unaware exploration strategies, achieving higher cumulative reward with fewer, more valuable experiments. The framework offers a practical and theoretically grounded solution for high-stakes experimentation in logistics and operations.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; *Bayesian network models*; • **Applied computing → Operations research**; **Supply chain management**; • **Mathematics of computing →** *Bayesian computation*; *Sequential decision making*.

KEYWORDS

Bayesian optimization, supply chain experimentation, value of experimentation, BART, partial rollout, non-stationary rewards, cost-sensitive learning

ACM Reference Format:
Lorenzo Toni. 2025. LOGEX: Cost-Sensitive Bayesian Experimentation for Adaptive Decision-Making in Supply Chains. In *Proceedings of KDD 2025 Workshop AI4SupplyChain (KDD'25).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Supply chain management increasingly relies on data-driven decision-making to optimize complex operational processes, yet implementing new policies carries substantial operational and financial risks. Unlike digital experimentation platforms where A/B tests can be conducted with minimal cost, supply chain experiments require physical deployment, active monitoring, and acceptance of potential disruptions to established workflows [18]. When organizations evaluate new inventory management heuristics, routing algorithms, or demand forecasting models, they must balance the potential benefits of improved performance against the immediate costs of experimentation and the opportunity cost of not exploiting their current best-known policies [5].

The challenge becomes particularly acute in dynamic supply chain environments where market conditions, supplier relationships, and customer demand patterns evolve continuously. Traditional approaches to policy evaluation often rely on static historical data or simplified simulation models that may not capture the full complexity of real-world operations [11]. Moreover, the high stakes associated with supply chain decisions—where errors can result in stockouts, excess inventory, or service disruptions—demand experimental frameworks that explicitly account for both learning value and implementation costs [6].

Recent advances in Bayesian optimization have demonstrated promising approaches for managing costly experimentation in various domains. The Interpretable Bayesian Optimization for Value Estimation (IBOVE) framework introduced a novel method for translating Bayesian acquisition functions into direct estimates of financial value, enabling stakeholders to make informed decisions about when to experiment versus exploit existing knowledge [2]. However, existing Bayesian optimization approaches face significant limitations when applied to supply chain contexts. First, they typically assume stationary reward functions, which poorly represent the dynamic nature of supply chain performance where external factors continuously influence outcomes [17]. Second, they often rely on Gaussian process models that may struggle to capture the complex, nonlinear relationships common in logistics and operations research [13].

The supply chain literature has long recognized the importance of adaptive experimentation. Early work on supply chain learning focused on demand sensing and forecasting improvement through systematic data collection [7]. More recent research has explored reinforcement learning approaches for inventory management [15] and routing optimization [14], but these methods typically require extensive online learning phases that may be impractical for operational deployment. The concept of "learning while doing" in

KDD'25, August 03, 2025, Toronto, Canada

supply chains has been formalized through various frameworks [3], yet most existing approaches fail to provide explicit guidance on when the cost of continued experimentation exceeds its expected benefits.

Bayesian Additive Regression Trees (BART) have emerged as a powerful tool for modeling complex, nonlinear relationships in high-dimensional spaces [4]. Unlike traditional Gaussian processes, BART models provide natural interpretability through their tree-based structure, allowing practitioners to understand which factors drive performance differences [8]. The flexibility of BART in handling non-stationary environments has been demonstrated in various applications, from causal inference [9] to time series forecasting [12]. However, the integration of BART models within cost-sensitive Bayesian optimization frameworks for operational decision-making remains largely unexplored.

In this work, we introduce LOGEX (LOGistics EXperimentation), a novel Bayesian framework specifically designed for cost-sensitive experimentation in supply chain environments. LOGEX addresses three critical limitations of existing approaches. First, it employs BART models to capture complex, evolving reward functions that better represent the nonlinear and non-stationary nature of supply chain performance. Second, it provides explicit economic quantification of experimentation value, enabling practitioners to make principled decisions about resource allocation. Third, it supports partial rollout strategies that allow organizations to test new policies on subsets of their operations while maintaining overall system stability.

Our framework makes several key contributions to the intersection of Bayesian optimization and supply chain management. We develop a theoretically grounded approach for translating BART posterior distributions into expected monetary gains from experimentation. We demonstrate how cost-sensitive acquisition functions can guide sequential decision-making about when to experiment, which policies to test, and how extensively to deploy them. We provide empirical validation through synthetic supply chain scenarios that capture realistic operational constraints and performance dynamics.

The remainder of this paper proceeds as follows. Section 2 formalizes the problem setting and introduces the mathematical framework underlying LOGEX. Section 3 presents the detailed algorithm, including the BART-based reward modeling and cost-sensitive acquisition strategy. Section 4 demonstrates the effectiveness of our approach through comprehensive experimental evaluation using synthetic supply chain scenarios. Section 5 discusses practical implementation considerations and limitations. Section 6 concludes with directions for future research.

2 PROBLEM FORMULATION: COST-SENSITIVE SEQUENTIAL EXPERIMENTATION IN SUPPLY CHAINS

We consider a sequential decision-making setting in which a supply chain planner must evaluate and deploy operational policies under uncertainty, with limited ability to conduct large-scale randomized trials. The planner aims to maximize cumulative long-term value by balancing exploitation of the current best-known policy with selective, cost-aware exploration of alternative configurations. This setting is representative of numerous real-world applications, including inventory control, replenishment timing, fulfillment routing, and labor allocation, where interventions are implemented through resource-intensive pilots.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the continuous space of policy configurations, where each $x \in \mathcal{X}$ encodes an operational decision (e.g., reorder point, routing threshold, forecasting model parameter). Let $f_t : \mathcal{X} \to \mathbb{R}$ denote the unknown reward function at time t, mapping each policy x to an expected outcome—typically expressed in monetary terms, such as margin improvement or cost reduction. Unlike in classical Bayesian optimization, we assume that f_t evolves over time due to latent environmental factors such as seasonal demand shifts, supplier behavior changes, or network reconfigurations.

At each round $t \in \{1, \ldots, T\}$, the planner selects an action $x_t \in \mathcal{X}$ and an allocation ratio $\alpha_t \in [0, 1]$, where α_t denotes the fraction of operational volume or population exposed to the new policy. When $\alpha_t < 1$, the action is treated as an experiment and incurs both a direct cost $C_e(x_t)$ and an opportunity cost due to foregone exploitation. A noisy reward observation is received:

$$y_t = \alpha_t f_t(x_t) + (1 - \alpha_t) f_t(x_t^*) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{m\alpha_t(1 - \alpha_t)}\right) \tag{1}$$

where m denotes the total operational population and the variance expression reflects increased uncertainty under small or imbalanced treatment allocations, consistent with standard randomized trial theory.

The planner maintains a posterior distribution over f_t based on all prior observed outcomes $\mathfrak{D}_t = \{(x_s, \alpha_s, y_s)\}_{s<t}$, and selects actions to maximize cumulative expected return over a horizon of T rounds:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \left(\alpha_t f_t(x_t) + (1 - \alpha_t) f_t(x_t^*) - C_e(x_t) \cdot \mathbf{1}[\alpha_t < 1]\right)\right] \tag{2}$$

subject to:

- $(x_t, \alpha_t) \sim \pi(\mathfrak{D}_t)$, a policy mapping prior data to allocation decisions;
- $f_t \sim \mathcal{P}(f|\mathfrak{D}_t)$, where $\mathcal{P}$ is a BART posterior;
- f_t allowed to evolve via temporal drift as detailed in Section 4.2

This formulation captures several practical features of real-world experimentation. First, observations are costly: even small-scale pilots require logistics effort, and results are subject to noise. Second, observations are partial: due to ethical, budgetary, or operational concerns, interventions can only be tested on subpopulations. Third, the underlying reward function is non-stationary, requiring the planner to discount outdated evidence and re-explore policies when the environment changes.

To guide decisions, we define the value of experimentation (VoE) as the expected improvement in future deployment value attributable to a new trial, net of its cost. The planner conducts an experiment at round t if and only if the VoE of a candidate action is positive. In the next section, we introduce the LOGEX framework, which uses BART posteriors to estimate this quantity and optimize experimentation choices.

3 THE LOGEX FRAMEWORK: BART-BASED VALUE OF EXPERIMENTATION

In this section, we present LOGEX, a Bayesian framework for cost-sensitive experimentation in supply chain environments. LOGEX uses posterior samples from Bayesian Additive Regression Trees (BART) to estimate the expected value of experimentation (VoE), guiding when and where to deploy pilots that balance potential learning against operational cost and risk.

Given observed data $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $y_i = f(x_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, BART provides posterior draws $\{\hat{f}^{(s)}(x)\}_{s=1}^{S}$ from the posterior distribution over f.

The posterior mean reward for an action x is approximated by:

$$\hat{\mu}(\mathfrak{D}, x) = \frac{1}{S} \sum_{s=1}^{S} \hat{f}^{(s)}(x) \tag{3}$$

The opportunity cost of exploring x is defined as:

$$\hat{O}(\mathfrak{D}, x) = \max_{x' \in \mathcal{X}} \hat{\mu}(\mathfrak{D}, x') - \hat{\mu}(\mathfrak{D}, x) \tag{4}$$

For each posterior draw $\hat{f}^{(s)}$, simulate a new observation $Y^{(s)} \sim \mathcal{N}(\hat{f}^{(s)}(x_{\text{new}}), \sigma^2)$, and update the dataset with $(x_{\text{new}}, Y^{(s)})$ to obtain a new posterior $\mathfrak{D}'_s$ and updated mean:

$$\hat{\mu}'_s = \frac{1}{S'} \sum_{s'=1}^{S'} \hat{f}_s^{(s')}(x^*) \tag{5}$$

The expected value of model improvement is given by:

$$\hat{U}(\mathfrak{D}, x_{\text{new}}) = \frac{1}{S} \sum_{s=1}^{S} \max\left(0, \hat{\mu}'_s - \hat{\mu}(\mathfrak{D}, x^*)\right) \tag{6}$$

The value of experimentation (VoE) is then:

$$V_\gamma(\mathfrak{D}, x_{\text{new}}) = \gamma \cdot \hat{U}(\mathfrak{D}, x_{\text{new}}) - \hat{O}(\mathfrak{D}, x_{\text{new}}) - C_e(x_{\text{new}}) \tag{7}$$

The decision rule is to explore x_{new} if and only if $V_\gamma(\mathfrak{D}, x_{\text{new}}) > 0$. Otherwise, the system exploits the best-known policy:

$$x^* = \arg\max_{x \in \mathcal{X}} \hat{\mu}(\mathfrak{D}, x) \tag{8}$$

In practice, expectations in $\hat{U}(\mathfrak{D}, x)$ are approximated using Monte Carlo simulation over posterior samples from BART.

4 GENERALIZING LOGEX: ALLOCATION RATIOS AND TEMPORAL DRIFT

This section extends the LOGEX framework to support two features essential for real-world supply chain experimentation: (i) partial deployments via allocation ratios, and (ii) evolving reward functions due to environmental drift. These enhancements enable LOGEX to model more realistic conditions under which experimentation is conducted in logistics operations.

4.1 Partial Deployments via Allocation Ratios

Let the extended action space be denoted by $\bar{\mathcal{X}} := \mathcal{X} \times [0, 1]$, where each element $\bar{x} = (x, \alpha)$ consists of an operational policy $x \in \mathcal{X}$ and an allocation ratio $\alpha \in [0, 1]$, representing the fraction of operational volume exposed to the intervention.

The reward from partial deployment scales linearly with α, defined as:

$$\bar{r}(x, \alpha) = \alpha f(x) + (1 - \alpha) f(x^*) \tag{9}$$

The explicit cost of experimentation remains fixed:

$$\bar{C}_e(x, \alpha) = C_e(x) \cdot \mathbf{1}[\alpha < 1] \tag{10}$$

The opportunity cost becomes:

$$\bar{O}(\mathfrak{D}, x, \alpha) = f(x^*) - \bar{r}(x, \alpha) = \alpha(f(x^*) - f(x)) \tag{11}$$

Assuming a total experimental population of size m, the variance of the estimated treatment effect is:

$$\mathrm{Var}[\bar{y}] = \frac{\sigma^2}{m\alpha(1 - \alpha)} \tag{12}$$

Simulated observations under partial allocation are generated via:

$$\bar{Y}^{(s)} \sim \mathcal{N}\left(\hat{f}^{(s)}(x_{\text{new}}, \alpha), \frac{\sigma^2}{m\alpha(1 - \alpha)}\right) \tag{13}$$

The updated expected model improvement is:

$$\bar{U}(\mathfrak{D}, x_{\text{new}}, \alpha) = \frac{1}{S} \sum_{s=1}^{S} \max\left(0, \hat{\mu}'_s - \hat{\mu}(\mathfrak{D}, x^*)\right) \tag{14}$$

And the generalized value of experimentation becomes:

$$\bar{V}_\gamma(\mathfrak{D}, x_{\text{new}}, \alpha) = \gamma \cdot \bar{U}(\mathfrak{D}, x_{\text{new}}, \alpha) - \bar{O}(\mathfrak{D}, x_{\text{new}}, \alpha) - \bar{C}_e(x_{\text{new}}, \alpha) \tag{15}$$

4.2 Modeling Non-Stationarity via Time Decay

To address evolving environments, LOGEX models time decay in its reward estimation. Each past observation (x_i, y_i, t_i) is assigned a weight w_i based on its temporal distance from the current time t:

$$w_i = \exp(-\lambda(t - t_i)) \tag{16}$$

These weights are incorporated into BART's training data, yielding a posterior that discounts older data and increases responsiveness to drift. This mechanism supports re-exploration when uncertainty rises due to environmental change.

5 EXPERIMENTAL EVALUATION

To assess the effectiveness of the LOGEX framework, we conduct a series of experiments in a controlled simulation environment designed to reflect the operational and informational constraints of real-world supply chain experimentation. The simulation captures core characteristics of the domain, including: (i) limited opportunities to experiment due to cost; (ii) the ability to partially deploy policies via allocation ratios; (iii) non-stationary reward functions driven by environmental drift; and (iv) noisy, high-variance feedback typical of real-world logistics operations.

Our goal is to evaluate whether LOGEX can improve cumulative system performance by intelligently selecting which policies to test, how extensively to deploy them, and when to switch from exploration to exploitation.

5.1 Simulation Environment

The simulation models a single-agent sequential decision process over a finite horizon of $T = 30$ rounds. At each round t, the agent selects a policy $x_t \in \mathcal{X} \subset \mathbb{R}$ and an allocation ratio $\alpha_t \in [0, 1]$, representing the fraction of operational capacity or volume exposed to the selected policy.

The true reward function $f_t(x)$ is latent and evolves over time through discrete structural changes. Initially, it is drawn from a smooth nonlinear function perturbed by discontinuities—mimicking typical operational settings where policy performance may abruptly change due to seasonality, capacity shifts, or upstream disruptions. Every 10 rounds, f_t undergoes a drift event that alters its shape. Observed rewards are generated via:

$$y_t = \alpha_t f_t(x_t) + (1 - \alpha_t) f_t(x^*_{t-1}) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{17}$$

where $\sigma^2 = 0.01$ and α_t modulates both the signal and the noise due to treatment sample size. Each experiment incurs a fixed explicit cost $C_e(x) = 0.1$, and no feedback is observed when $\alpha_t = 1$ (i.e., full exploitation). We assess agent performance using cumulative net reward, which combines realized system reward and penalties for experimentation:

$$R_{\text{cum}}(T) = \sum_{t=1}^{T} \left[y_t - C_e(x_t) \cdot \mathbf{1}[\alpha_t < 1] \right] \tag{18}$$

This metric reflects the true operational value captured by the planner and penalizes unnecessary or poorly timed experimentation. We compare LOGEX to a widely used heuristic: Explore-Then-Exploit (ETE). The ETE agent randomly explores 20% of the rounds with uniformly sampled actions from $\mathcal{X}$, each deployed with a fixed allocation $\alpha = 0.3$. After this phase, it exploits the policy with the highest average observed reward for the remaining rounds. This baseline reflects real-world operational test-and-rollout strategies, in which organizations initially run a fixed set of pilots and then deploy the perceived best performer. However, such strategies typically fail to adapt once initial exploration is complete, particularly under dynamic conditions.

5.2 Results

Figure 1 presents the cumulative net reward trajectories of LOGEX and ETE averaged over 20 independent simulation runs. LOGEX consistently outperforms the baseline across the full time horizon, with especially notable gains following reward drift events. Its behavior is characterized by selective and adaptive exploration: the algorithm initiates pilots when uncertainty is high and potential gain exceeds cost, and reverts to exploitation once sufficient evidence is acquired. By contrast, the ETE agent performs comparably to LOGEX during early rounds but degrades significantly following structural shifts in the reward function. Because it lacks any mechanism to detect or respond to environmental change, it remains locked into stale decisions based on early performance—a common failure mode of fixed-horizon experimentation. LOGEX also conducts fewer experiments overall, but with higher average impact, showing its ability to internalize cost-benefit trade-offs at each step. Moreover, LOGEX dynamically adjusts the allocation ratio: it increases when potential gain is modest but uncertainty is high (to limit risk), and decreases when targeting high-potential candidates.

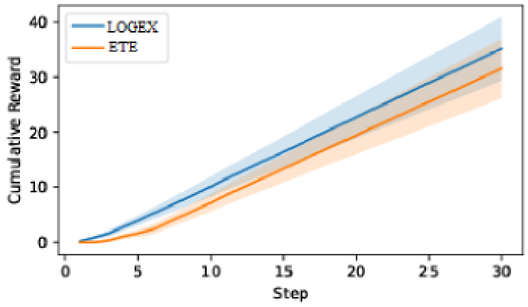

Figure 1: Cumulative Reward of LOGEX and the ETE-baseline over 30 steps

6 CONCLUSION AND FUTURE WORK

This paper introduced LOGEX, a cost-sensitive Bayesian optimization framework for guiding experimentation in complex, high-stakes supply chain environments. By leveraging Bayesian Additive Regression Trees (BART) to model uncertain, non-stationary reward functions and computing value-of-experimentation (VoE) scores in interpretable monetary units, LOGEX enables planners to strategically allocate limited experimentation capacity across time, actions, and subpopulations. The framework extends and operationalizes foundational ideas from value-based experimental design [1, 16], cost-aware Bayesian optimization [10, 19], and nonparametric modeling [4, 9] into the domain of applied supply chain experimentation—an area where these methods have seen limited adoption despite significant relevance.

Our results demonstrate that LOGEX consistently outperforms static heuristics such as explore-then-exploit baselines, achieving greater cumulative operational value with fewer, more targeted experiments. Moreover, the framework adapts dynamically to structural shifts in the environment and supports partial deployments, addressing gaps in the current literature on experimentation in operations research with expensive setups.

Nonetheless, the current implementation has important limitations. LOGEX assumes scalar policy parameters and continuous action spaces, whereas real-world interventions often involve structured, discrete, or combinatorial policies (e.g., routing rules, batching thresholds, multi-knob tuning). The computational cost of repeated posterior simulation under BART may also become prohibitive at scale, particularly when the action space or data size grows. Moreover, our evaluation uses synthetic environments, which, while representative, lack the full complexity of real-world operational frictions such as delayed feedback, interdependent policies, and stakeholder-driven constraints.

Future work will extend LOGEX in three directions. First, we aim to generalize the acquisition framework to accommodate structured and high-dimensional action spaces, including discrete and graph-based decision domains. Second, we plan to integrate causal inference techniques into the reward estimation layer, improving robustness when unobserved confounders or selection effects are present. Third, we will pursue deployment in real-world systems, particularly in supply chain settings where test-and-learn culture is emerging but still underutilized due to cost, scale, and interpretability concerns. We believe LOGEX offers a promising foundation for embedding formal experimentation into the everyday operational logic of data-driven supply chains.

Algorithm 1 LOGEX with BART Posteriors

```
Require: Initial dataset $\mathfrak{D}_0$, horizon $T$, cost function $C_e$, value multiplier $\gamma$
Ensure: Sequence of actions and rewards
for $t = 1$ to $T$ do
    Fit BART model on weighted dataset $\mathfrak{D}_{t-1}$ with time decay weights
    Sample $S$ posterior functions $\{\hat{f}^{(s)}\}_{s=1}^{S}$ from BART
    Compute current best policy: $x^* = \arg\max_x \hat{\mu}(\mathfrak{D}_{t-1}, x)$
    for each candidate $(x, \alpha) \in \bar{\mathcal{X}}$ do
        Compute opportunity cost: $\bar{O}(\mathfrak{D}_{t-1}, x, \alpha)$
        Simulate posterior observations and compute $\bar{U}(\mathfrak{D}_{t-1}, x, \alpha)$
        Compute VoE: $\bar{V}_\gamma(\mathfrak{D}_{t-1}, x, \alpha)$
    end for
    Select action: $(x_t, \alpha_t) = \arg\max_{(x,\alpha)} \bar{V}_\gamma(\mathfrak{D}_{t-1}, x, \alpha)$
    if $\bar{V}_\gamma(\mathfrak{D}_{t-1}, x_t, \alpha_t) > 0$ then
        Execute experiment with allocation $\alpha_t$
        Observe reward $y_t$ and pay cost $C_e(x_t)$
    else
        Exploit: set $(x_t, \alpha_t) = (x^*, 1)$ and observe $y_t = f_t(x^*)$
    end if
    Update dataset: $\mathfrak{D}_t = \mathfrak{D}_{t-1} \cup \{(x_t, \alpha_t, y_t, t)\}$
end for
```

REFERENCES

[1] Eduardo M. Azevedo, Alex Deng, Jose Luis Montiel Olea, Justin Rao, and E. Glen Weyl. 2020. A/B testing with fat tails. *Journal of Political Economy* 128, 12 (Dec. 2020), 4614–4672.

[2] Anonymous Author et al. 2020. Interpretable bayesian optimization for value estimation. *Proceedings of Machine Learning Research* 108 (2020), 1–15.

[3] Omar Besbes and Assaf Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* 57, 6 (Nov. 2009), 1407–1420. https://doi.org/10.1287/opre.1080.0640

[4] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2010. BART: Bayesian additive regression trees. *Annals of Applied Statistics* 4, 1 (Mar. 2010), 266–298. https://doi.org/10.1214/09-AOAS285

[5] Sunil Chopra and Peter Meindl. 2019. *Supply Chain Management: Strategy, Planning, and Operation* (7th ed.). Pearson, Boston, MA.

[6] Martin Christopher. 2016. *Logistics & Supply Chain Management* (5th ed.). Pearson, Harlow, UK.

[7] Marshall L. Fisher, Janice H. Hammond, Walter R. Obermeyer, and Ananth Raman. 1994. Making supply meet demand in an uncertain world. *Harvard Business Review* 72, 3 (May-Jun. 1994), 83–93.

[8] Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (Jan. 2011), 217–240. https://doi.org/10.1198/jcgs.2010.08162

[9] Jennifer Hill, Antonio Linero, and Jared Murray. 2020. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and its Application* 7 (Mar. 2020), 251–278. https://doi.org/10.1146/annurev-statistics-031219-041110

[10] Remi Lam, Karen Willcox, and David H. Wolpert. 2015. Bayesian optimization with a finite budget: An approximate dynamic programming approach. In *Advances in Neural Information Processing Systems* 28. 883–891.

[11] Hau L. Lee. 2004. The triple-A supply chain. *Harvard Business Review* 82, 10 (Oct. 2004), 102–112.

[12] Antonio R. Linero and Yun Yang. 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80, 5 (Nov. 2018), 1087–1110. https://doi.org/10.1111/rssb.12293

[13] John T. Mentzer, William DeWitt, James S. Keebler, Soonhong Min, Nancy W. Nix, Carlo D. Smith, and Zach G. Zacharia. 2001. Defining supply chain management. *Journal of Business Logistics* 22, 2 (Sep. 2001), 1–25. https://doi.org/10.1002/j.2158-1592.2001.tb00001.x

[14] Mohammadreza Nazari, Afshin Oroojlooyjadid, Lawrence Snyder, and Martin Takáč. 2018. Reinforcement learning for solving the vehicle routing problem. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (NeurIPS '18). Curran Associates Inc., Red Hook, NY, 9861–9871.

[15] Afshin Oroojlooyjadid, Mohammadreza Nazari, Lawrence Snyder, and Martin Takáč. 2020. A deep Q-network for the beer game: Deep reinforcement learning for inventory optimization. *Manufacturing & Service Operations Management* 22, 6 (Nov.-Dec. 2020), 1196–1215. https://doi.org/10.1287/msom.2019.0840

[16] Ilya O. Ryzhov, Warren B. Powell, and Peter I. Frazier. 2012. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research* 60, 1 (Jan.-Feb. 2012), 180–195. https://doi.org/10.1287/opre.1110.0999

[17] Edward A. Silver, David F. Pyke, and Douglas J. Thomas. 2016. *Inventory and Production Management in Supply Chains* (4th ed.). CRC Press, Boca Raton, FL.

[18] David Simchi-Levi, Xin Chen, and Julien Bramel. 2014. *The Logic of Logistics: Theory, Algorithms, and Applications for Logistics Management* (3rd ed.). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-9149-1

[19] Jian Wu, Matthias Poloczek, Andrew Gordon Wilson, and Peter Frazier. 2019. Bayesian optimization with gradients. In *Advances in Neural Information Processing Systems* 32. 5267–5278.