https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *Management Science*

ISSN 0025-1909, EISSN 1526-5501

How Not to Overpackage: AI for Sustainability in HelloFresh's Service Supply Chain

Evelyn Xiao-Yue Gong
Carnegie Mellon University

Michael Johnson
HelloFresh Group

Abstract. Overpackaging is a common challenge faced by meal kit services. For each box of meal kits delivered to customers, meal kit service providers must decide the type of liner materials and the quantity of freezer packs to place inside. These materials ensure the quality of the meal kits delivered; on the other hand, overpackaging leaves a large carbon footprint and imposes psychological burdens on many customers.

As the world's leading meal kit company and integrated food solutions group, HelloFresh is committed to furthering the sustainable nature of its supply chain and decreasing its environmental footprint. One important step is to reduce packaging. While HelloFresh has an in-house thermal engineering team that builds the packaging guidelines with a large number of scientific experiments, artificial intelligence offers numerous possibilities for improvement.

This paper introduces the *contextual packaging* problem, and investigates artificial intelligence solutions to mitigate overpackaging efficiently. We develop a contextual bandit algorithm that takes into account the weather, the conditions of transit, the box contents and other contextual information to adaptively make the packaging decision for each box. Taking advantage of the structural properties of the contextual packaging problem, our algorithm achieves theoretically optimal performance with an $\tilde{O}(\sqrt{T})$ regret bound where T is the horizon length. We test our algorithm on real delivery data from HelloFresh in order to identify potential overpackaging in the current guidelines.

Key words: Artificial intelligence, Sustainability, Reinforcement learning, Overpackaging, Contextual packaging, Meal kit service, Contextual bandits

1. Introduction

Meal kit services have been hot and trending, particularly among the younger generation of consumers. According to a Zion Market Research Report on LinkedIn, the currently 20 billion dollar global meal kit service market is expected to reach 65 billion USD by 2030 (Howard (2024)). Every week, millions of customers select their favorite recipes from the weekly menu and the number of servings they want. After receiving the order, the meal kit service providers send the exact quantities of groceries, condiments and recipe sheets in a box to the customer's door. These boxes also contain packaging materials—freezer packs and liners—to keep the contents temperature-safe.

One of the common challenges faced by meal kit service providers is the risk of overpackaging. Because boxes vary in size, contents and transit conditions (time, weather, refrigeration, etc.), each box requires a nontrivial packaging decision to be made. A New York Times article (Richtel (2016)) chronicles the "guilt" and "embarrassment" that many consumers feel towards the packaging materials that come with their online orders, even when the materials can technically go into the recycling bins. A recent Harvard Business Review (Reichheld et al. (2023)) finds that sustainability promotes trust, particularly among younger generations—one of the major demographics for the meal kit service industry. As long as the service quality is maintained, reducing packaging waste from the start is more appealing than recycling. The potential benefits include cost savings for the providers and customers alike.

Beyond the meal kit service industry, concerns about over-packaging are also ubiquitous in e-commerce and delivery services. Forbes reports that many leading e-commerce retailers such as Amazon and Walmart have taken measures to mitigate this often-criticized aspect of their operations (Bird (2018)). However, the packaging problem in the meal kit service industry is more complex due to the temperature control aspect. The same Forbes article (Bird (2018)) and a Mother Jones article (Butler (2017)) point out that meal kit service provider *Blue Apron* sends out around 8 million meals each month in 2016-2017, and that if each box contains three meals and two six-pound freezer packs, this would generate about 192,000 tons of freezer pack waste per year. To highlight the importance of the problem, these articles describe that the freezer pack waste from this one service provider would have "the weight of nearly 100,000 cars or 2 million adult men" per year (Butler (2017), Bird (2018)). As astonishing as these numbers already sound, the drastic growth of the meal kit industry since 2016, evident in Figure 1, has likely further escalated them.

As the world's leading meal kit company and integrated food solutions group, HelloFresh is committed to furthering the sustainable nature of its service supply chain, of which packaging operations is a cardinal aspect. HelloFresh's in-house thermal engineering team refines the current packaging guidelines over time by constantly conducting chamber experiments. In each experiment, a packaged box of meal kits is placed into a large temperature-controlled scientific chamber for three days, and the temperature change in the box is monitored. Because these experiments are

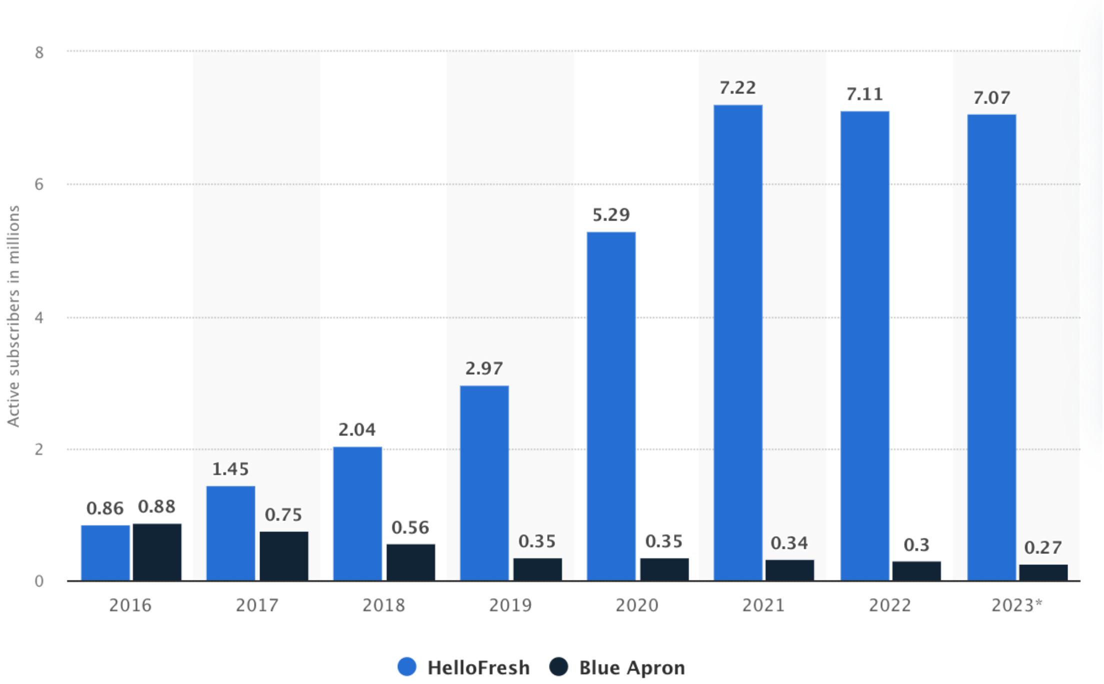


Figure 1 Number of Active Subscribers in Millions by HelloFresh and Blue Apron during 2016-2023. Data collected from Statista (2024).

costly and time-consuming to run, artificial intelligence offers the possibility to improve the current process in terms of speed, cost and optimality.

This paper proposes artificial intelligence solutions to efficiently mitigate overpackaging. We develop a contextual bandit algorithm that takes into account contextual information such as the conditions of transit, the location, the ingredients, the box size, etc., to adaptively make the packaging decision for each box. Taking advantage of the structure of the packaging problem, our algorithm achieves theoretically optimal performance with an $\tilde{O}(\sqrt{T})$ regret bound, where T is the horizon length or, equivalently, the total number of boxes to package. Realistically, to prioritize the safety of the customers and the quality of the meal boxes, we use our algorithm to identify the subset of boxes that are currently likely over-packaged and offer this information to the thermal engineering team to narrow down the set of engineering experiments that should be conducted carefully to refine the packaging rules. Given the enormous scale of operations, this information we learn using our algorithm could potentially cut down an enormous amount of packaging materials and carbon emission, as well as time and resources spent on thermal experiments.

1.1. Motivation

Historically, service providers package meal kit boxes with plentiful materials to ensure the products are still sufficiently cold when they arrive at the customers' doors. The rapid growth of the market size in the past decade (see Figure 1) has heightened the impact of these packaging decisions. Also heightened in recent years are customer expectations on product sustainability, as shown by a recent joint study from McKinsey and NielsenIQ (Am et al. (2023)). Enhanced environmental sustainability is crucial for the sustained future success for the meal kit industry as it promotes trust, particularly for the younger generations who are a major demographic for the meal kit service industry (Reichheld et al. (2023)). Practical sustainability research to cut down on the possible overpackaging in the process in a safe manner is highly desirable. Due to the scale of operations, any small improvement on the packaging guidelines may offer rich rewards, for both cost and environmental impact.

The current guidelines are formed using chamber experiments. As discussed above, each set of chamber experiments requires expensive scientific equipment, too much time, and excessive human resources. The artificial intelligence solution developed in this paper would refine packaging guidelines at a much faster speed and much lower cost than chamber experiments.

Another important benefit of an effective artificial intelligence solution is that it empowers the provider to move freely towards more sustainable packaging options when they are available. New packaging materials that are gradually more environmentally friendly emerge over time in the market. Adopting artificial intelligence would enable the provider to swiftly re-learn and identify the best packaging guidelines for a new set of packaging options with ease. The current practice of conducting chamber experiments, on the other hand, would not allow an easy switch, as the changes in the properties of the materials would render the past chamber experiment results inapplicable to the new materials. Developing new guidelines using chamber experiments again would be a slow and expensive process.

Equally importantly, the industry has developed to a point where artificial intelligence solutions can be introduced with a negligible amount of extra operational difficulty and minimal reconfiguration effort. The meal kit packing process is already highly automated. Currently, at large meal kit packing facilities, human pickers grab packaging materials and groceries and put them into the meal kit box on the conveyor belt by straightforwardly following a sequence of signal lights that indicate what to place into each box next. The signal lights are already automated by a computer program following the current packaging guidelines. An artificial intelligence solution would only make alterations to the computer program that controls the sequences of lights, without causing complications to the on-the-ground operations.

1.2. Our Work

We introduce the *Contextual Packaging* problem, and invest our efforts in an artificial intelligence method to make the packaging decisions adaptively for the tens of millions of boxes shipped out every year within the US market by HelloFresh. Our contributions include the following.

- To the best of our knowledge, we are the first to introduce the *Contextual Packaging* problem in the meal kit service industry.
- Our algorithm leverages the ocean of available past (offline) data to pre-learn structural information that helps accelerate online learning later. In practical applications, online data are often rare and expensive, as they involve some form of real-time experimentation on users, while offline data are abundant and readily available. Yet offline data alone does not suffice, because learning counterfactual information from offline data is difficult. Therefore, it would be tremendously valuable for the providers if we could extract as much as possible structural information from the offline data, and then fill in the missing information with a small amount of online data swiftly. The model formulation and the design of our Alg 1 is in accordance with this philosophy.
- We propose *data injection* solutions for the challenges we encounter in HelloFresh's operational data, which address issues of confounding and sparse signals. We describe the *information* vs. *truth* trade-off in the data injection process in Section 6.
- Our algorithm can be applied in two ways for the packaging problem. The first is to directly make packaging decisions for the shipments in an *online* fashion, and determine the optimal packaging decisions. The second is to identify where potential overpackaging might exist in the current packaging guidelines; once identified, we can pass on the characteristics of these boxes to the thermal engineering team to conduct further scientific experiments.
- Using real data from HelloFresh, our AI method pinpoints the possible overpackaging points and helps the engineering team focus their efforts on refining the packaging rules for these potentially overpackaged boxes. This method offers great benefits to the meal kit service providers in terms of huge savings in time, energy, materials, equipment and human resources.
- This paper provides theoretical results on the complexity of the contextual packaging problem under different modeling assumptions. We identify a natural directional structure property in the problem which enables $\tilde{O}(\sqrt{T})$ regret for finite action sets. In fact the problem exhibits multiple structural properties, and it is subtle to determine which is useful. Indeed, we also identify a

conditional one-sided feedback structure which provably does not decrease the regret, despite appearing superficially similar to structural properties used in related works. These structural properties are explained in Section 4. See Table 1 for a summary of the theoretical results.

	Finite Action Set	Continuous Action Space
Conditional 1-Sided Feedback	$\Theta(T^{2/3})$; Theorems 2 and 3	$\tilde{\Theta}(T^{2/3})$; Theorems 1 and 3
Directional Structure (Property 1)	$\tilde{\Theta}(T^{1/2})$; Theorem 4	$\tilde{\Theta}(T^{2/3})$; Theorems 1 and 3

Table 1 Minimax optimal regret bounds in four problem formulations are shown. With an infinite action set, Theorems 3, 1 and 2 provide positive and negative results. Together, their conclusion is that with an infinite action set or without the directional structure, $\Omega(T^{2/3})$ regret is achievable using a simple context-discretization algorithm and cannot be improved even using conditional 1-sided feedback. Our main positive result, Theorem 4, shows that directional structure (without conditional 1-sided feedback) yields optimal $\tilde{O}(T^{1/2})$ regret in finite action spaces.

1.3. Paper Layout

In Section 2, we discuss further related literature. In Section 3, we formulate the *Contextual Packaging* problem. In Section 4, we introduce the structural properties that are inherent in the contextual packaging problem, including *conditional one-sided feedback structure* and *directional structure*. In Section 4.2, we explain why conditional one-sided feedback is unhelpful for the contextual packaging problem and in fact perilous if used naively. In Section 4.4, we provide negative theoretical results, under some subsets of these structures. For example, achieving less than $\Omega(T^{2/3})$ regret is impossible if the action space is continuous, and moreover there is a simple procedure that obtains $O(T^{2/3})$ regret without requiring this feedback structure.

In Section 5, we break through this regret lower bound by choosing the right subset of structural properties of the contextual packaging problem. We design a provably optimal algorithm with an optimal regret bound of $\tilde{O}(T^{1/2})$ that enjoys polynomial dependence on the number of actions.

In Section 6, we expose some interesting challenges that we encounter in HelloFresh's operational data, and propose corresponding solutions. In Section 7, we test our algorithm on real data from HelloFresh to illustrate how our algorithm identifies over-packaging in the current practice. In Section 8, we conclude and discuss implications of this work.

2. Literature Review

2.1. Contextual Bandits

The contextual bandit problem has been studied for decades, see e.g. Auer et al. (1995), Wang et al. (2005), Pandey et al. (2007), and Li et al. (2010). In such a problem, a decision-maker acts across T

time-steps. In each round t, the learner observes a context x_t for that round and chooses an action a_t from an action set $\mathcal{A}$, which may be continuous or discrete. Then, the learner receives a reward that depends on the arm chosen as well as the context. The learner's goal is to maximize the total reward received over a finite horizon. The contextual bandits problem is a popular model that lies between multi-armed bandit and full-fledged reinforcement learning. It captures a large class of repeated decision problems. In addition, the algorithms developed for the contextual bandits problem have been successfully applied in many domains including ad placement, recommendation systems, and clinical trials.

In this paper, we essentially study a contextual bandit problem with special structural properties. One of the interesting properties is that we are able to observe more feedback than just the reward for the one arm we choose in each round. We define the *rich feedback* and the *directional structure* in more detail in Section 4. Here, we discuss related prior studies on bandits or reinforcement learning with rich feedback.

Zhao and Chen (2019) study bandit learning in problems with *unconditional one-sided feedback*, which means that they can observe rewards for all actions that are lower (or higher) than the chosen action in each round. Gong and Simchi-Levi (2024) develop a reinforcement learning algorithm that uses *unconditional* one-sided feedback efficiently in a Markov Decision Process. In Appendix A, we provide an algorithm that is inspired by their elimination-based algorithm. However, in the packaging problem, what we have is *conditional one-sided feedback*. Therefore, naively applying their techniques would in fact introduce bias to our estimates as discussed in Section 4.2. Therefore, we develop a different approach to utilize the structure of our problem efficiently. Balseiro et al. (2019) consider a contextual bandit problem where in addition to receiving the reward, the learner also learns the values for all other contexts. This differs from the conditional one-sided feedback studied in this paper in that we can observe rewards only for a subset of the actions and contexts.

Of direct relevance for us is the work Agarwal et al. (2011), which addresses the non-contextual problem of minimizing a convex, Lipschitz function over a convex, compact set under a stochastic bandit feedback model. In Section 4.4.2 and Appendix C, we give a $\tilde{O}(T^{2/3})$ regret algorithm for the contextual packaging problem with continuous action space by maintaining $T^{1/3}$ copies of their algorithm. In 3 of the 4 settings shown in Table 1, this simple algorithm cannot be improved, despite not using the special structural properties of the problem. This allows us to precisely pinpoint which structural properties of the contextual packaging problem are useful for efficient online learning.

2.2. Packaging Waste and Food Operations Sustainability

Existing research on packaging waste concentrates on materials engineering or on the after-math waste management. Gavrilescu et al. (2023), Callewaert et al. (2023) and Gritsch and Lederer (2023) study the packaging waste management systems in Romania, Norway and Vienna respectively. Zhang et al. (2023) conclude that food takeaway packaging waste generates the largest amount of plastic packaging waste in China. Dirpan et al. (2023) review current trends in smart packaging for food products including using biodegradable materials. Choi and Burgess (2007), Kucharek et al. (2020), Wang et al. (2020) and others study methods to predict the performance of insulating packages using mathematical models.

Because this paper studies the contextual packaging problem in the meal kit service industry, we also include some literature review on food operations sustainability. Belavina (2021) analyzes the impact of grocery store density on food waste. Belavina et al. (2017) compare per-delivery fees and the subscription model for online grocery shopping from the perspectives of profit, food waste, and emission. Hezarkhani et al. (2023), Kazaz et al. (2023) and Han et al. (2023) study the sustainability benefits of retailers carrying ugly produce. Akkaş et al. (2019) analyze supply chain data to study factors that impact the expiration of packaged food; Akkaş and Honhon (2022) show that the *ship oldest first* policy generates high waste. Kallbekken and Sælen (2013) find that smaller plates in hotel restaurants help reduce food waste; Martins et al. (2016) report that food waste education led to a 33% waste reduction from school lunches; and Schmidt (2016) shows that a related information campaign helped achieve a self-reported 12% reduction in household food-waste. Other works that examine food waste include Akkaş and Honhon (2022), Ata et al. (2019), Lee et al. (2017), and Lee and Tongarlak (2017).

To the best of our knowledge, we are the first to introduce the *Contextual Packaging* problem. Our paper designs artificial intelligence solutions for everyday packaging operations to mitigate unnecessary over-packaging in meal kit services, which can be extended to other thermal packaging applications such as for pharmaceuticals, vaccines, fresh flowers, wine, cheese, sashimi-grade seafood for restaurants, etc. Beyond the contextual packaging problem, artificial intelligence methods are also a promising avenue for future research in sustainable operations.

3. Problem Formulation

The contextual packaging problem is defined as follows. During a time horizon of length T, at every round $1 \leq t \leq T$, a customer's order is due for shipping. As part of the procedure to send out the meal kit box based on the customer's order, the service provider decides the packaging configuration for the box. This configuration entails the grade of the insulation liner material and the quantity of the freezer packs to place inside the box. Before making the decision, the provider receives a vector $\mathbf{x}_t \in \mathcal{X}$, referred to as the "context", for this box. Each context $\mathbf{x}_t$ encodes the relevant information for the meal-box that is to be packaged and sent out at round t, including the contractual transit time, the order contents, the courier, the maximum temperature for the delivery date, and other relevant information. Based on the context, the provider makes a packaging decision $\mathbf{a}_t$ for the meal-box that determines the amount of ice in the box, and the grade (type) of the liner materials. The context vectors are I.I.D. samples from an unknown distribution on the context space $\mathcal{X}$.

Once made, the packaging decision $\mathbf{a}_t$ incurs a per-round Cost, which we denote by $\mathrm{Cost}_t = \mathrm{Cost}_t(\mathbf{x}_t, \mathbf{a}_t, \xi_t)$. It consists of three terms: the *material cost M*, the *environmental footprint E*, and the *failure penalty P*. It is defined by the formula

$$\mathrm{Cost}_t = M(\mathbf{x}_t, \mathbf{a}_t) + E(\mathbf{x}_t, \mathbf{a}_t) + P(\mathbf{x}_t, \mathbf{a}_t, \xi_t) \tag{1}$$

where ξ_t is the realized environmental randomness at round t.

The material cost $M(\mathbf{x}_t, \mathbf{a}_t)$ is a known function that depends on the box size, the chosen liner grade and the chosen freezer packs. $M(\mathbf{x}_t, \mathbf{a}_t)$ is deterministic given the context-action pair $(\mathbf{x}_t, \mathbf{a}_t)$. Similarly, the environmental footprint $E(\mathbf{x}_t, \mathbf{a}_t)$ is modeled deterministically given $(\mathbf{x}_t, \mathbf{a}_t)$. Both $M(\mathbf{x}_t, \mathbf{a}_t)$ and $E(\mathbf{x}_t, \mathbf{a}_t)$ are known and deterministic; therefore they could be combined into one term. Although we do not require it as an assumption, both are monotonically increasing with the quantity and the grade of the materials used. However, the behavior of the failure penalty P is unknown and random, and depends on the I.I.D. environmental randomness ξ_t.

When a box is delivered, the provider may receive a complaint from the customer. The complaint may be temperature-related if for example the courier had an unexpectedly long transit time or the refrigeration malfunctioned during the transit. These scenarios can all be attributed to insufficient packaging materials. When a temperature-related complaint happens, compensation will be made to the customer and a failure penalty $P(\mathbf{x}_t, \mathbf{a}_t, \xi_t)$ is incurred for that box. This penalty includes compensation paid to the customer and any additional penalty imposed internally by the provider. We model P as a known *penalty function Q* multiplied by a stochastic *failure event* which occurs with probability depending on both the context and action. Without loss of generality, we can model the environmental randomness ξ_t as I.I.D. uniform variables in $[0, 1]$. Then with $p(\mathbf{x}_t, \mathbf{a}_t)$ the failure probability, we set:

$$P(\mathbf{x}_t, \mathbf{a}_t, \xi_t) = Q(\mathbf{x}_t, \mathbf{a}_t) \cdot 1_{\xi_t \le p(\mathbf{x}_t, \mathbf{a}_t)}. \tag{2}$$

We note that Q may depend on $\mathbf{a}_t$ as well as $\mathbf{x}_t$; this captures realistic situations where e.g. using more packaging would reduce liability even in the event of failure. See also the discussion following Property 1. We also point out that although it is intuitively helpful to view ξ_t as encoding concrete environmental randomness to model counterfactuals, our results hold as long as the failure probability $p(\mathbf{x}_t, \mathbf{a}_t)$ obeys the corresponding monotonicity properties.

Realistically, the Cost can only be observed after the box is delivered. For modeling simplicity, we assume the Cost is observed before the next round $t+1$. This simplification does not affect the regret analysis of our model and main algorithm, which can tolerate delays as long as $O(\sqrt{T})$; see Remark 1. In our numerical experiments in Section 7, we incorporate delays in the implementation of our main algorithm.

The objective for the provider is to minimize the total expected cost over the horizon $\sum_{t=1}^{T} \text{Cost}_t(\mathbf{x}_t, \mathbf{a}_t)$, or equivalently, to minimize the *regret* of their packaging decisions. The notion of regret is the standard performance measure in online learning, and is defined as the difference between the total cost of a feasible policy and that of a *clairvoyant* optimal policy OPT that knows the true penalty function distributions a priori.

For a learning algorithm ALG, the (expected) regret of ALG over T periods is

$$\begin{aligned} &\text{Regret}_{\text{ALG}}(T) \\ :=&\mathbf{E}\left[\sum_{t=1}^{T} \text{Cost}_t^{\text{ALG}}\right] - \mathbf{E}\left[\sum_{t=1}^{T} \text{Cost}_t^{\text{OPT}}\right]. \end{aligned}$$

4. Structural Properties

In modeling the contextual packaging problem, we would like to leverage the abundantly available past (offline) data to pre-learn structural information that accelerates online learning later. For meal kit delivery services, like many other business operations, online data are rare and expensive, as they involve real-time experimentation on the boxes sent to real users, while offline data are abundant and readily available. On the other hand, online learning is imperative because offline data does not provide us with accurate counterfactual information. It would be tremendously valuable for the providers if we could extract as much as possible structural information from the offline data, and then fill in the missing information with a small amount of online data swiftly.

With this goal in mind, we begin by learning a total ordering ϕ of all of our possible packaging choices $\mathbf{a} \in \mathcal{A}$, as well as a total ordering ψ of all of the potential contexts $x \in \mathcal{X}$. These ordering functions can be learnt from the historical delivery data and the physics knowledge we have on the thermodynamics of a meal kit box. The ordering of a context vector is with respect to the penalty likelihood $p(\mathbf{x}_t, \mathbf{a}_t)$. Since what we ask for is the ordering functions of the actions and the contexts, the mapping does not need to preserve distance relationships between points, and is therefore relatively easy to learn. Despite not knowing the exact scientific value of the expected total heat exerted onto each meal kit box, we know enough to learn an ordering function for the contexts approximately correctly. In Section 6, we discuss some of the challenges in identifying such mappings from the available ocean of offline data and our corresponding solutions.

From here onwards, we assume that we have succeeded in learning the ordering functions $\phi(\cdot)$ and $\psi(\cdot)$. Then we can treat ϕ as a known mapping from the set of all possible packaging actions to a bounded one dimensional set, i.e., actions $\mathbf{a}_t$ live in an action space $\mathcal{A} \subseteq \mathbb{R}$. And we treat ψ as a known mapping from the set of all possible contexts to a bounded one dimensional set, i.e., contexts $\psi(\mathbf{x}_t)$ live in an interval $[0,1] \subseteq \mathbb{R}$. This allows us to, without loss of generality, reduce the dimension of the context space χ to 1 and the dimension of the action space $\mathcal{A}$ to 1. We use $a = \phi(\mathbf{a})$ to denote $\mathbf{a}$ after the mapping, and $x = \psi(\mathbf{x})$ to denote $\mathbf{x}$ after the mapping. This reduction in dimension simplifies our problem and enables faster online learning later on.

In the next three subsections, we identify three structural properties which are present in the contextual packaging problem: finiteness of the action space, conditional 1-sided feedback, directional structure. In Section 4.4, we present results showing that certain combinations of these properties do not enable improved $o(T^{2/3})$ regret. Then in Section 5, we show that for finite action spaces, *directional structure* allows us to break this barrier and achieve optimal $\tilde{O}(\sqrt{T})$ regret.

4.1. Finite Action Space

As just explained, both the context and action spaces can be totally ordered, and thus mapped to $[0,1] \subseteq \mathbb{R}$. The next structural property we identify is an asymmetry between these two spaces. Although the meal kit boxes have innumerable possible contexts, the action space that the service provider can take is finite, with only 50 possible combinations of packaging options. This is because in practice, meal service providers only have two or fewer sizes of freezer packs (e.g. 3lbs and 5lbs) to choose from, and can decide to place up to three freezer packs in each box. There are also just five grades of liner materials to choose from. In order of increasing insulation power, they are: *Linerless* (i.e. no liner), *Winter*, *Spring*, *Summer* and *Super Summer*. There are only 50 combinations of ice packs and liner materials satisfying the above constraints[1]. By contrast, modeling the action space

[1] Given 2 types of ice packs, there are 10 ways to choose at most 3 if order does not matter.

as continuous would be appropriate if the service provider can adjust the packaging decision on a continuous spectrum, by e.g. choosing the exact amount of ice to inject into the freezer packs; however this is not done in practice. We emphasize that the *context* space remains innumerable for the packaging problem: the context for a meal box consists of weather, weight and other relevant continuous information.

4.2. Conditional One-Sided Feedback

Recall that the packaging cost can be decomposed via Equation (1); since M and E are known, the most important term is the penalty function P which takes the form Equation (2). The packaging problem naturally possesses the following counterfactual logic: given actions $\phi(\mathbf{a}) \leq \phi(\mathbf{a}')$, for any context $\mathbf{x}$ and environmental randomness ξ, if failure occurs for $(\mathbf{x}, \mathbf{a}, \xi)$ then it also occurs for $(\mathbf{x}, \mathbf{a}', \xi)$. For example for a meal kit box sent out at time t, if the spring liner and 6 lbs of ice were used and the provider received a temperature complaint, then we can deduce that for the round t, the spring liner and 3 lbs of ice would also have been insufficient packaging. Recalling (2), this is equivalent to the condition that $p(\mathbf{x}, \mathbf{a}) \geq p(\mathbf{x}, \mathbf{a}')$. Similar logic applies if the context is changed: if $\psi(\mathbf{x}) \geq \psi(\mathbf{x}')$, then $p(\mathbf{x}, \mathbf{a}) \geq p(\mathbf{x}', \mathbf{a})$. Of course, these criteria can be combined: if $\phi(\mathbf{a}) \leq \phi(\mathbf{a}')$ and $\psi(\mathbf{x}) \geq \psi(\mathbf{x}')$, then $p(\mathbf{x}, \mathbf{a}) \geq p(\mathbf{x}', \mathbf{a}')$.

This rich feedback structure resembles the (unconditional) *one-sided feedback* definition from Yuan et al. (2021) and Gong and Simchi-Levi (2024). The definition of (unconditional) *one-sided feedback* from Gong and Simchi-Levi (2024) is for Markov Decision Processes. In Appendix A, we adapt their definition to the contextual bandit setting.

Unfortunately, the contextual packaging problem does not actually possess the (unconditional) one-sided feedback. As stated, in round t, if the spring liner and 6 lbs of freezer pack were used and the provider received a complaint, then the spring liner and 3 lbs of freezer pack would have also received a complaint. However, if the provider did not receive a complaint, then we would not know in this case whether the spring liner and 3 lbs of freezer pack would have been enough. Note that we can, in this case, deduce that the spring liner and 9 lbs of freezer pack would also have been enough packaging.

The intricacy is that in the contextual packaging problem, sometimes we can deduce the alternative outcomes for packaging options that are stronger, and sometimes we can deduce for the weaker options. Which side of packaging options we can deduce relies on the realized outcome for the actual action we have taken. This is a “conditional” type of rich feedback that is formerly unstudied. Below we give a formal definition of the *conditional one-sided feedback*.

Definition 1 *A problem has* ***conditional one-sided feedback*** *if there exist ordering functions ψ and ϕ and a criterion κ, such that, when any $(\mathbf{x}_t, \mathbf{a}_t)$ is played, if the reward received satisfies criterion κ, then all rewards (costs) for pairs $(\mathbf{x}, \mathbf{a})$ with $\psi(\mathbf{x}) \leq \psi(\mathbf{x}_t)$ and $\phi(\mathbf{a}) \leq \phi(\mathbf{a}_t)$ are observed in that round t; and if the reward (cost) does not satisfy criterion κ, then all rewards (costs) for pairs $(\mathbf{x}, \mathbf{a})$ with $\psi(\mathbf{x}) \geq \psi(\mathbf{x}_t)$ and $\phi(\mathbf{a}) \geq \phi(\mathbf{a}_t)$ are observed.*

Definition 1 amounts to a form of cross-learning between contexts and actions. We emphasize that the total expected loss has no reason to be monotone, since the material and environmental contributions M, E naturally trade-off against P. We do not even require that $Q(x, a)$ is decreasing in a, although this would likely be the case in realistic settings.

We have seen in the literature that with the formerly studied (unconditional) one-sided feedback, prior works (e.g. Zhao and Chen (2019), Yuan et al. (2021), Gong and Simchi-Levi (2024)) have successfully used this extra feedback to improve the speed and the quality of their reward estimates. Can we follow a similar path with the *conditional* one-sided feedback?

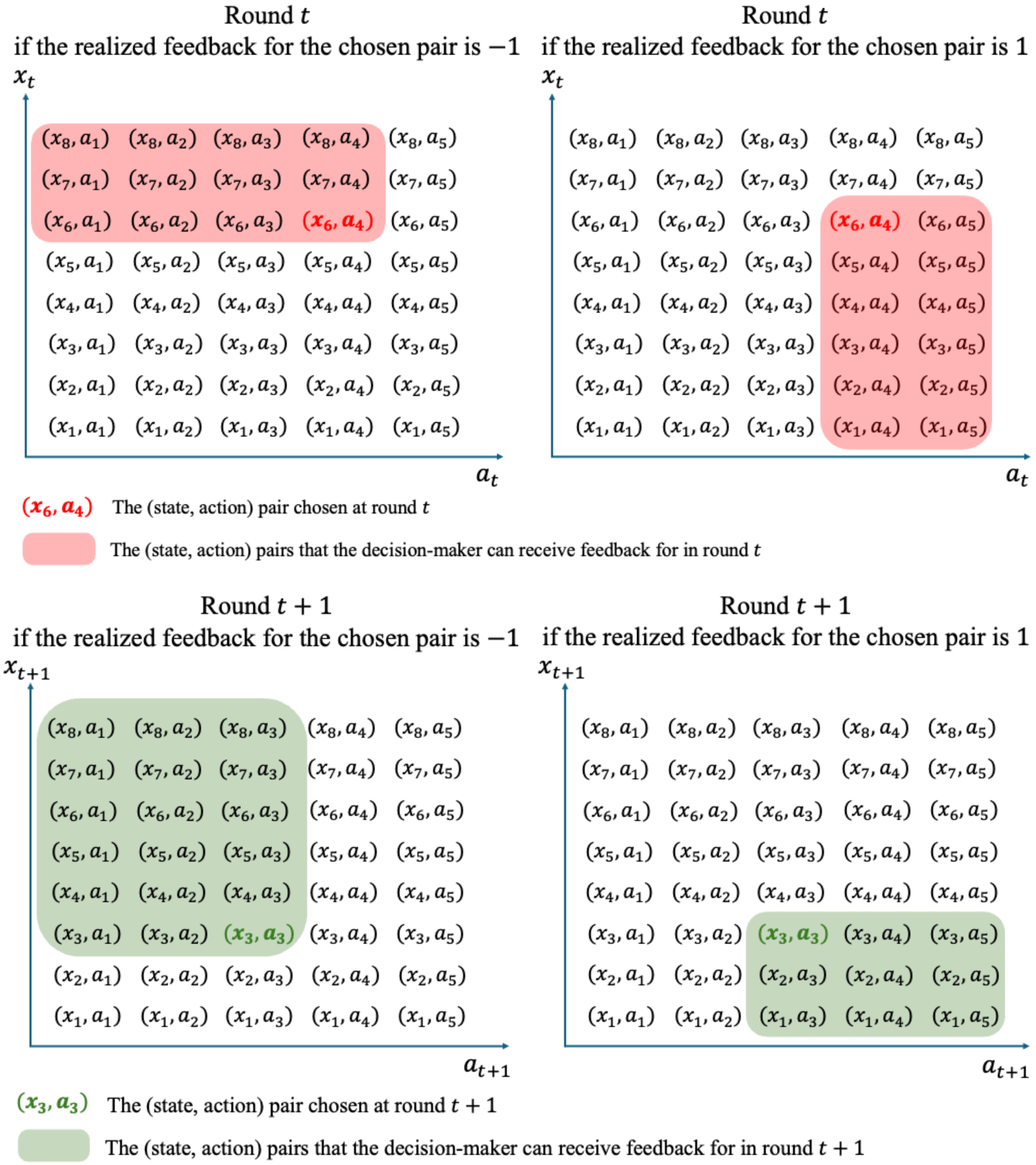
The answer is negative. In what follows, we expose the "danger" of naively using conditional one-sided feedback (to conduct extra updates to the reward estimates) through a toy counterexample.

Toy Counterexample. Suppose in a simple bandit-learning problem, there are three arms to choose from repeatedly over a horizon of $T = 1000$ rounds. The reward received for pulling an arm at time t is either 0 or 1. The true average rewards of arms $1, 2, 3$ are $0.49, 0.5, 0.51$ respectively. We set up the problem so that there exists conditional one-sided feedback in this problem: rewards occur when ξ_t is respectively at most $0.49, 0.5, 0.51$. In particular, if at round t, Arm 2 is pulled and gives reward 0, then if we had pulled Arm 1 at round t, we would have also received reward 0; if at round t, Arm 2 is pulled and gives reward 1, then if we had pulled Arm 3 at round t, we would have also received reward 1. This is a simple toy example of conditional one-sided feedback.

Now if we use this extra feedback to update our reward estimates for the three arms, and we pull Arm 2 for all $T = 1000$ time-steps. Then our reward estimate for Arm 2 would converge correctly to approximately 0.5, while our reward estimates for Arms 1 and 3 would equal 0 and 1 respectively, which would be far from the true expected rewards. This is because we only update our estimate for Arm 1 when the outcome for Arm 2 is bad, and only update for Arm 3 when the outcome for Arm 2 is good. By conducting updating for different arms conditioned on the realized outcomes, we introduce severe biases in our reward estimates for the arms.

The above toy example demonstrates that caution must be taken in order to take advantage of this rich conditional 1-sided feedback. However, it could still be possible to leverage the feedback using a more intricate method to improve the regret; we show later that no such method exists.

Figure 2 Example of Conditional One-Sided Feedback in two rounds t (red) and $t+1$ (green). The (state, action)-pairs chosen by the decision-maker are in bold and colored. In the diagrams at left, the highlighted regions indicate which (state, action)-pairs the decision-maker receives feedback for when the realized feedback is -1. The diagrams at right indicate those pairs the decision-maker receives feedback for when the realized feedback is 1. For ease of presentation, in this example, the contexts and the actions are already indexed by their orderings.


Round t
if the realized feedback for the chosen pair is -1
Round t
if the realized feedback for the chosen pair is 1
x_t
(x_8, a_1) (x_8, a_2) (x_8, a_3) (x_8, a_4) (x_8, a_5)
(x_7, a_1) (x_7, a_2) (x_7, a_3) (x_7, a_4) (x_7, a_5)
(x_6, a_1) (x_6, a_2) (x_6, a_3) (x_6, a_4) (x_6, a_5)
(x_5, a_1) (x_5, a_2) (x_5, a_3) (x_5, a_4) (x_5, a_5)
(x_4, a_1) (x_4, a_2) (x_4, a_3) (x_4, a_4) (x_4, a_5)
(x_3, a_1) (x_3, a_2) (x_3, a_3) (x_3, a_4) (x_3, a_5)
(x_2, a_1) (x_2, a_2) (x_2, a_3) (x_2, a_4) (x_2, a_5)
(x_1, a_1) (x_1, a_2) (x_1, a_3) (x_1, a_4) (x_1, a_5)
a_t
(x_6, a_4) The (state, action) pair chosen at round t
The (state, action) pairs that the decision-maker can receive feedback for in round t
Round $t+1$
if the realized feedback for the chosen pair is -1
Round $t+1$
if the realized feedback for the chosen pair is 1
x_{t+1}
a_{t+1}
(x_3, a_3) The (state, action) pair chosen at round $t+1$
The (state, action) pairs that the decision-maker can receive feedback for in round $t+1$

4.3. Directional Structure

Recall the Cost function in Equation (1); without loss of generality we assume it is normalized to be bounded by $[0, 1]$. We then let ℓ be the expected value of Cost over the environmental randomness

ξ, for consistency with the online learning literature:

$$\ell(x,a) = \mathbb{E}_\xi[\mathrm{Cost}(x,a,\xi)] \in [0,1].$$

Let $a_t^* = a^*(x_t) = \arg\min_a \ell(x_t,a)$ be the true optimal action for x_t. With the same counterfactual logic introduced at the beginning of Section 4.2 for the conditional one-sided feedback property, we state the following *directional structure*:

Property 1 (Directional Structure) *For each pair of contexts $x > x'$, the function $\ell(x,a) - \ell(x',a)$ is decreasing in a.*

Property 1 uses the directional structure in the packaging problem in both the action axis and the context axis. Practically speaking, Property 1 can be comprehended as a "safety" protocol that is widely adopted by providers. It says that if the provider faces two contexts, then they should always use a more "intensive" option on the more "intensive" context. Under no circumstance should the provider decide to "give up" on a meal kit box by intentionally underpackaging, even if the context has so much expected heat exertion that the chance of failure penalty is very high under any action. This property is especially true of the meal kit packaging problem due to safety and customer satisfaction reasons: a more severe failure would be more damaging to the customer relationship.

Property 1 implies the following useful property, which is structurally interesting in its own right. In fact, it will turn out to suffice for our main positive result in Theorem 4.

Property 2 (Weak Directional Structure) *The optimal action $a^*(x)$ is an increasing function of the context x.*

4.4. Negative Results from Table 1

Given the several structural properties presented, it is unclear which are necessary and sufficient for effective online learning. Here we present the negative results shown previously in Table 1. Theorems 1 shows that with a continuous action space, there exist problem instances where regret smaller than $\Omega(T^{2/3})$ is unachievable, even assuming both conditional 1-sided feedback and directional structure (Property 1). Theorem 2 shows that conditional 1-sided feedback is insufficient for $o(T^{2/3})$ regret even for finite action spaces of size 2. As shown in Theorem 3, this bound is often achievable even without relying on any special problem structure. Despite these negative results, we show in Section 5 that the optimal $\tilde{O}(\sqrt{T})$ regret is achievable under the weak directional structure (Property 2) for finite action spaces, even without conditional 1-sided feedback.

4.4.1. Regret Lower Bound of $\Omega(T^{2/3})$

Theorem 1. *For any T there exists a (randomized) problem instance with $\mathcal{X} = \mathcal{A} = [0,1]$ where $\ell(x,a,\xi)$ is Lipschitz in (x,a) such that Property 1 holds and any (possibly randomized) algorithm incurs $\Omega(T^{2/3})$ expected regret, even with conditional 1-sided feedback.*

Theorem 2. *For any T there exists a (randomized) problem instance with $\mathcal{X} = [0,1]$ and $\mathcal{A} = \{0,1\}$ where $\ell(x,a,\xi)$ is Lipschitz in x such that any (possibly randomized) algorithm incurs $\Omega(T^{2/3})$ expected regret, even with conditional 1-sided feedback.*

Theorem 1 is proved below, while Theorem 2 is proved in Appendix B. In both cases, a family of problem instances is carefully chosen to simulate $T^{1/3}$ independent subproblems, while ensuring that conditional 1-sided feedback provides no additional information between different subproblems. However the constructions differ technically because the former also obeys Property 1 while the latter applies to finite action sets in addition to continuous sets. Theorem 4 in Section 5 shows that when both of these additional properties hold, a more efficient online learning algorithm is possible. Therefore the properties must appear separately in the lower bound constructions.

Proof of Theorem 1 We construct an instance of the packaging problem where the unknown failure function ℓ is

$$\ell(x,a) = |a - \gamma(x)|/10$$

for some monotonically increasing function $\gamma : [0,1] \to [0,1]$ to be chosen; Property 1 holds for any such γ. We take realized rewards $L_t = \text{Cost}_t$ as follows. We decompose

$$\begin{aligned} \ell(x,a) &= \left(\frac{a}{10} - \frac{1}{2}\right) + q(x,a) \\ &\equiv \left(\frac{a}{10} - \frac{1}{2}\right) + \begin{cases} \frac{1}{2} - \frac{\gamma(x)}{10}, & a \geq \gamma(x), \\ \frac{1}{2} + \frac{\gamma(x)}{10} - \frac{a}{5}, & a \leq \gamma(x). \end{cases} \end{aligned}$$

Then for $\xi \sim Unif([0,1])$ we take

$$L(x,a,\xi) = \left(\frac{a}{10} - \frac{1}{2}\right) + 1_{\xi \leq q(x,a)}$$

to be a known term plus a Bernoulli variable (note $q(x,a) \in [0,1]$ by construction). It is easy to see that the optimal action for such a construction of ℓ_x is $a^*(x) = \gamma(x)$ which is increasing in

x. Furthermore, this observation model enjoys conditional one-sided feedback: if $\xi \le q(x,a)$ then $\xi \le q(x,a')$ for any $a' \le a$ which determines $L(x,a',\xi)$, and similarly in the other direction when $\xi > q(x,a)$. (Note that although L can be negative, it is uniformly bounded in $[-1,1]$ so one can reparametrize $\tilde{L} = \frac{L+1}{2} \in [0,1]$ to be the loss.)

We suppose the ground truth function satisfies $\gamma(i/T^{1/3}) = i/T^{1/3}$, for each integer $0 \le i \le T^{1/3}$. For each i, at the middle third points of I_i, with equal probability, $\gamma\left(\frac{3i-2}{3T^{1/3}}\right) = \gamma\left(\frac{3i-1}{3T^{1/3}}\right) = \frac{3i-2}{3T^{1/3}}$ or $\gamma\left(\frac{3i-2}{3T^{1/3}}\right) = \gamma\left(\frac{3i-1}{3T^{1/3}}\right) = \frac{3i-1}{3T^{1/3}}$. These choices are made uniformly at random, independently within different I_i. And γ is linear on the three constituent subintervals of I_i. Thus, γ is a concatenation of independently chosen functions $\gamma_i : I_i \longrightarrow I_i$ and knowing the ground truth, one should only play actions $a_t \in I_i$ when presented with $x_t \in I_i$.

Given the above construction of ℓ and γ, it is apparent that each I_i is an independent subproblem with a regret lower bound of $\Omega(T^{1/3})$. (In particular, all possible functions γ are monotone.) As a result, we have constructed a set of $T^{1/3}$ independent learning problems which occur in parallel. In the full problem, each of these subproblems appears $n_i \sim Bin(T, T^{-1/3})$ times, since the contexts $x_t \in [0,1]$ are independent and uniformly random (of course the n_i variables are not independent). By standard tail bounds for Binomial random variables, with high probability $1 - o_T(1)$, we have $n_i \in [T^{2/3}/2, 2T^{2/3}]$ for all i.

By similar reasoning with $Bin(T, T^{-1/3}/3)$ random variables, with probability $1 - o_T(1)$ all of the intervals I_i will have at least $T^{2/3}/10$ contexts landing inside their middle third $J_i = \left[\frac{3i-2}{3T^{1/3}}, \frac{3i-1}{3T^{1/3}}\right]$. It is possible for γ_i to take either the constant value $\frac{3i-2}{3T^{1/3}}$ or $\frac{3i-1}{3T^{1/3}}$ on this interval. For $T^{1/3}/3 \le i \le 2T^{1/3}/3$, it is impossible to distinguish these events with $\ge 2/3$ probability with fewer than $T^{2/3}/C$ samples for some absolute constant C (see e.g. (Slivkins 2019, Chapter 2)). It follows that by taking each γ_i to be independently and uniformly random among two such possibilities, the expected regret from each I_i is $\Omega(T^{1/3})$. Thus the overall expected regret is $\Omega(T^{2/3})$. □

4.4.2. An Intuitive Procedure that Matches the $\Theta(T^{2/3})$ Regret Lower Bound

We design a simple intuitive procedure, and prove that the procedure matches the $\Theta(T^{2/3})$ regret lower-bound. This algorithm uses the mild conditions that ℓ is Lipschitz in (x,a) and convex in a, but does not make use of conditional one-sided feedback.

The main idea of this procedure is to first batch the context space $[0,1]$ into $T^{1/3}$ intervals:

$$I_i = \left[\frac{i-1}{T^{1/3}}, \frac{i}{T^{1/3}}\right], \quad i = 1, 2 \ldots, T^{1/3},$$

allowing us to divide the contextual packaging problem into $T^{1/3}$ problems based on which interval the context falls in. For each of the $T^{1/3}$ parallel copies of problems, we treat each copy as the contextless bandit problem, i.e., we pretend that ℓ is constant on each I_i, and apply a copy of (Agarwal et al. 2011, Algorithm 1), which guarantees $\tilde{O}(\sqrt{n_i}) = \tilde{O}(\sqrt{T^{1/3}})$ regret for the contextless subproblem on I_i, where n_i denotes how many times the context falls in I_i. Therefore, our procedure maintains $T^{1/3}$ parallel copies of (Agarwal et al. 2011, Algorithm 1), one for each interval I_i. Given context x_t, we record the interval I_i containing x_t and call the i-th copy of their algorithm.

Theorem 3. *For the packaging problem with stochastic contexts in $\mathcal{X} = [0,1]$ and continuous or discrete action space $\mathcal{A} = [0,1]$ or $\mathcal{A} = \{0,1\}$, if ℓ is jointly Lipschitz in (x,a) and convex in a, then the above algorithm has expected regret $\tilde{O}(T^{2/3})$.*

The proof for Theorem 3 is provided in Appendix C.

5. Main Algorithm

In what follows, we take advantage of the finite action space to break through the regret lower bound of $\Omega(T^{2/3})$. It is worth noting that conditional one-sided feedback is not needed here: we use only finiteness of the action space $\mathcal{A}$ and the weak directional structure (Property 2).

Let $K = |\mathcal{A}| \in \mathbb{N}^+$ denote the finite cardinality of the action space $\mathcal{A}$. The existence of the ordering function ϕ of the actions now allows us to, without loss of generality, reorder the action set as $0 \le a_1 \le \cdots \le a_K \le 1$. For each $1 \le j \le K$, we use $\ell_j : [0,1] \to [0,1]$ to denote the average cost from playing action a_j given context x. Recall that contexts x_t are I.I.D. from some unknown probability distribution on $[0,1]$, which we denote as μ. Namely, Subsection 4 explains how to use ψ to map the contexts to a 1-dimensional space. The choice of interval $[0,1]$ is then without loss of generality.

We will also tacitly assume contexts are almost surely distinct. This is just for simplicity of presentation, because one can augment a context x_t with an independent uniform "tie-breaking" variable $u_t \sim \text{Unif}([0,1])$. Then if $x_t = x_s$, we say $(x_t, u_t) < (x_s, u_s)$ when $u_t < u_s$; since the u_t variables are almost surely distinct, this breaks all ties. Since our algorithm below only uses that the ordering between contexts is a uniformly random permutation, it directly adapts to this setting with only notational changes.

Our analysis will be based on interval estimates and proceed in epochs $E = 1, 2, \ldots$. Let τ_E denote the beginning of each epoch E. At the first time period τ_E in epoch E, we initiate a collection of feasible closed intervals

$$F_{E,i} \subseteq [0,1],$$

Algorithm 1 Contextual One-Sided Arm Elimination

```
for epoch $E = 1, 2, \ldots$ do
    Reset the definitions of $\hat{\ell}_{t,i}$.
    For a context $x \in [0,1]$, let $S_E(x) = \{i \in [K] \mid x \in F_{E,i}\}$. be the set of feasible intervals containing $x$, and $n_E(x) = |S_E(x)|$ the number of them.
    for round $t = \tau_E, \tau_E + 1, \ldots$ do
        Given context $x_t$, choose $a_t \in S_E(x_t)$ uniformly at random.
        Update $\hat{\ell}$ according to Equation (3). (For efficient implementation, see Remark 3.)
        if there exists $0 \le A \le B \le 1$ and $i, j \in [K]$ such that (recalling (3) )
            $$\hat{\ell}_{t,j}(A,B) - \hat{\ell}_{t,i}(A,B) \ge 10\sqrt{\frac{CK^2 \log(KT)}{t - \tau_E + 1}}.$$
        then
            $E = E + 1$
            $F_{E+1,k} = F_{E,k}$ for $k \in [K] \setminus \{j\}$.
            Compute $M \in [A,B]$ as in (7).
            if $i < j$ then
                $F_{E+1,j} = F_{E,j} \cap [M,1]$.
            end if
            if $i > j$ then
                $F_{E+1,j} = F_{E,j} \cap [0,M]$.
            end if
        end if
    end for
end for
```

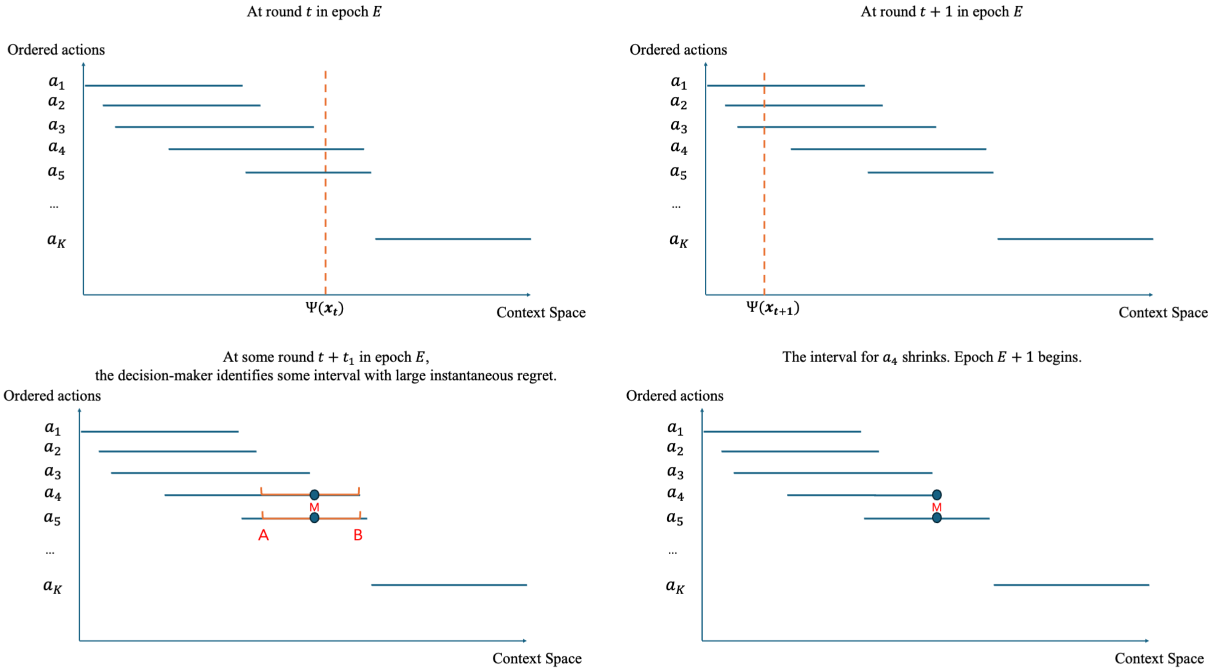
one for each action $i \in [K]$. These $F_{E,i}$'s indicate the range of contexts for which a_i may possibly be the optimal action. In turn, for each context x, let

$$S_E(x) = \{i \in [K] \mid x \in F_{E,i}\}$$

be the set of actions whose interval contains x in epoch E, and let

$$n_E(x) := |S_E(x)|$$

Figure 3 Illustration of Alg 1 Contextual One-Sided Arm Elimination. The y-axis is the ordered action space. The x-axis is the ordered context space. In this example, at round t, the decision-maker receives context x_t, and applies ψ to the context to obtain the ordered context $\psi(x_t)$. The decision-maker observes which actions' intervals intersect with the ordered context, and samples from this set of actions. At round $t+t_1$ in epoch E, the decision-maker identifies large instantaneous regret: Action a_5 is significantly superior to a_4 on some interval $[A,B]$. This indicates that some interval shrinking is about to happen, and a new epoch will start right after.



denote the cardinality of each of these sets.

During epoch E, our algorithm will choose actions uniformly from $S_E(x_t)$ when presented with context x_t. Let L_s denote the observed Cost at time s. For each real interval $[A,B] \subseteq [0,1]$ and $t \in E$, we define the importance-weighted unbiased cost estimate

$$\hat{\ell}_{E,t,i}(A,B) = \frac{\sum_{s=\tau_E}^{t} 1_{x_s \in [A,B]} \cdot 1_{a_s=i} \cdot n_E(x_s) L_s}{t - \tau_E + 1} \tag{3}$$

for the true expected cost of a_i from contexts in $[A,B]$:

$$\begin{aligned} \ell_{E,t,i}(A,B) &= \mathbb{E}[\ell_i(x) \cdot \mathbf{1}_{x \in [A,B]}] && (4) \\ &= \int_A^B \ell_i(x) \, \mathrm{d}\mu(x). && (5) \end{aligned}$$

Here we include both endpoints $[A, B]$ in such an integral (in case μ contains atoms).

We maintain confidence bounds for the average reward of each action on each interval $[A, B]$ based on these estimates. Our strategy is to use them to refine the feasible intervals $F_{E,i}$ over time.

To analyze Alg 1, note that Property 2 implies $[0, 1]$ is partitioned into intervals $I_1, \dots, I_K$ where $I_i = [A_i, B_i]$ with $A_1 = 0, A_{i+1} = B_i, B_K = 1$, such that action a_i is optimal on (A_i, B_i). Note that some actions a_j might **never** be optimal for any context, in which case $A_j = B_j$. We say $j \in [K]$ is a *frontier* action if it is optimal on some interval, i.e. $A_j < B_j$.

A key point is that for any frontier action j, we will always have $I_j \subseteq F_{E,j}$. This holds with high probability as we justify below starting with Lemma 1, which upper-bounds the estimation error in Equation (3). This will ensure that only suboptimal actions are eliminated with high probability. We compare to the frontier action on each interval to ensure suboptimal actions are eliminated by confidence bound comparisons sufficiently quickly.

Lemma 1. *There exists an absolute constant C such that the following holds simultaneously over all E, i, t with probability at least $1 - \frac{1}{T^2}$. Suppose that $[A, B] \subseteq F_{E,i}$. Then*

$$\left|\hat{\ell}_{E,t,i}(A, B) - \int_A^B \ell_i(x) \mathrm{d}\mu(x)\right| \le \sqrt{\frac{CK^2 \log(KT)}{t - \tau_E + 1}}.$$

The proof uses a four-dimensional version of the Dvoretzky–Kiefer–Wolfowitz–Massart inequality, and is provided in Appendix D. Having established Lemma 1, we now use it as our basis for action elimination. Each step eliminates action a_j on some interval of contexts, after which we proceed to the next epoch. Epoch E ends at the first time t that there exist A, B, i, j such that $[A, B] \subseteq F_{E,i} \cap F_{E,j}$ and

$$\hat{\ell}_{E,t,j}(A, B) - \hat{\ell}_{E,t,i}(A, B) \ge 10\sqrt{\frac{CK^2 \log(KT)}{t - \tau_E + 1}}. \tag{6}$$

(Here C is a fixed absolute constant.) This event corresponds to deciding that action i is strictly better than action j on average over the interval $[A, B]$, which allows us to eliminate action j on part of the interval. At such a time, defining $\hat{\ell}_{E,t,i,j}(A, M, B) = \left|\hat{\ell}_{E,t,i}(A, M) - \hat{\ell}_{E,t,i}(M, B) - \hat{\ell}_{E,t,j}(A, M) + \hat{\ell}_{E,t,j}(M, B)\right|$, we choose a "weighted midpoint" $M \in [A, B]$ under the criterion

$$|\hat{\ell}_{E,t,i,j}(A, M, B)| \le \frac{5K}{t - \tau_E + 1}, \tag{7}$$

The interpretation is that approximately half of the estimated gap between actions i and j on $[A, B]$ comes from $[A, M]$, and the approximate other half comes from $[M, B]$. Such an M always exists because of the discrete intermediate value theorem. Namely, $\hat{\ell}_{E,t,i}(A,\cdot)$ and $\hat{\ell}_{E,t,i}(\cdot, B)$ have discontinuities of size at most $\frac{K}{t-\tau_E+1}$ assuming all contexts are distinct, so the sum in Equation (3) only gains additional terms 1 at a time (as A and B vary). If contexts are not distinct, then as mentioned before one can introduce random tie-breaking to achieve the same result (nothing changes as long as the contexts are totally ordered with uniformly random permutation).

If $i < j$ (which implies $a_i < a_j$), then we set $F_{E+1,j} = F_{E,j} \cap [M, 1]$, and if $i > j$ we take $F_{E+1,j} = F_{E,j} \cap [0, M]$. We set $F_{E+1,k} = F_{E,k}$ for all other actions k.

We proceed in epochs: when some elimination step happens during epoch E, we restart in epoch $E + 1$ at time τ_{E+1} and forget all existing information except the feasible intervals $F_{E,i}$. This enables us to show in Lemma 4 that progress via eliminations is rapid.

To quantify the regret in our analysis, we define the **total active suboptimality** of action a_i during epoch E to be

$$R_{E,i} = \sum_{j=1}^{K} \int_{x \in I_j \cap F_{E,i}} \ell_i(x) - \ell_j(x) \, \mathrm{d}\mu(x). \tag{8}$$

It is easy to see that the instantaneous regret at time $t \in E$ is upper-bounded by $\sum_{i=1}^{K} R_{E,i}$. Note that the quantity $R_{E,i}$ is not available to the decision-maker and is used only in the analysis. Because the sum in Equation (8) has only K terms and each set $F_{E,i} \cap I_j$ is an interval, Lemma 2 is immediate.

LEMMA 2. *For any epoch E and any i* $\in [K]$*, there exists an interval* $I_j \cap F_{E,i}$ *such that*

$$\int_{x \in F_{E,i} \cap I_j} \ell_i(x) - \ell_j(x) \, \mathrm{d}\mu(x) \geq R_{E,i}/K.$$

Lemma 3 below shows cost estimates are stable in time. The proof is provided in Appendix E.

LEMMA 3. $|\hat{\ell}_{E,t,i} - \hat{\ell}_{E,t-1,i}| \leq \frac{K}{t-\tau_E+1}$

The next lemma ensures that elimination steps occur rapidly.

LEMMA 4. *Suppose there exists an interval* $F_{E,i} \cap I_j$ *and some* $\varepsilon > 0$ *such that*

$$\int_{x \in F_{E,i} \cap I_j} \ell_i(x) - \ell_j(x) \, \mathrm{d}\mu(x) \geq \varepsilon.$$

Then on the event that Lemma 1 applies, the next elimination event involving (A', B', i', j') *satisfies*

$$\int_{A'}^{B'} \ell_{i'}(x) - \ell_{j'}(x) \, \mathrm{d}\mu(x) \geq \varepsilon/3, \tag{9}$$

and occurs within $\tilde{O}(K^2/\varepsilon^2)$ timesteps of the epoch E, i.e.,

$$\tau_{E+1} - \tau_E \le \tilde{O}(K^2/\varepsilon^2). \tag{10}$$

Finally, any midpoint M' satisfies

$$\begin{aligned} &\int_{A'}^{M'} \ell_{i'}(x) - \ell_{j'}(x) \, \mathrm{d}\mu(x) \ge \varepsilon/10, \text{ and} \\ &\int_{M'}^{B'} \ell_{i'}(x) - \ell_{j'}(x) \, \mathrm{d}\mu(x) \ge \varepsilon/10. \end{aligned} \tag{11}$$

The proof for Lemma 4 is provided in Appendix F. Combining the Lemmas 2 and 4, we obtain the following Lemma 5 ensuring that the amount of suboptimality accumulated is small:

Lemma 5. *Suppose that for epoch E, the total amount of suboptimality is*

$$\sum_{i=1}^{K} R_{E,i} \ge \varepsilon. \tag{12}$$

Then on the event that Lemma 1 applies, within $\tilde{O}\left(\frac{K^6}{\varepsilon^2}\right)$ timesteps, an elimiation step will occur and

$$\sum_{i=1}^{K} R_{E+1,i} \le \left(1 - \frac{\Omega(1)}{K^2}\right) \sum_{i=1}^{K} R_{E,i}. \tag{13}$$

Finally, the eliminated action is indeed suboptimal on the interval it was eliminated from.

The proof for Lemma 5 is provided in Appendix G. Now we can deduce Theorem 4.

Theorem 4. *The total regret of Alg 1 is $\tilde{O}\left(K^5\sqrt{T}\right)$.*

Proof Let ε_E be the value of ε in (12) during epoch E. We assume that Lemma 1 applies, as the opposite case incurs $O(1/T)$ total regret as the error probability is at most $1/T^2$. Then for each E the corresponding regret is at most

$$\tilde{O}\left(\min(\varepsilon_E T, K^6 \varepsilon_E^{-1})\right) \le \tilde{O}(K^3\sqrt{T}).$$

By Equation (13), once $E \ge CK^2 \log(KT)$ for some absolute constant C, we have $\sum_{i=1}^{K} R_{E,i} \le K\left(1 - \frac{\Omega(1)}{K^2}\right)^{CK^2 \log(KT)} \le 1/T$. After this point, the future total regret is at most 1. Therefore the total regret is at most $\tilde{O}(K^5\sqrt{T})$ as desired. □

Remark 1. **Delayed Feedback.** The above algorithm works similarly when there is a delay in feedback. Suppose feedback at time t is received at time $t + t_d$, where t_d is deterministic[2]. Then we can perform the above algorithm on the feedback received: whenever an epoch ends, we simply disregard the last t_d units of feedback. Note that the decision-maker uses a fixed policy within each epoch, so performing the same algorithm within an epoch is no issue. As mentioned just above in the proof of Theorem 4, after $O(K^2 \log(T))$ elimination steps, one has $\sum_{i=1}^{K} R_{E,i} \leq 1/T$, after which point the future expected regret is at most 1. Hence the additional regret caused by the feedback delay of t_d is $O(t_d K^2 \log(T))$, which is easily subsumed in the regret bounds because t_d should be a constant $t_d \sim O(1)$. In the extreme case, we can let t_d be as large as $O(\sqrt{T})$ and the delay still would not affect the regret bound.

Remark 2. **Matching $\Omega(\sqrt{T})$ Regret Lower Bound.** It is not difficult to see that our packaging problem with a finite action space has a regret lower bound of $\Omega(\sqrt{T})$. Therefore, Alg 1 obtains the optimal regret bound for contextual packaging with finite action space. Indeed, consider a packaging problem with only one context and two actions. One of the two actions has reward $1/2$, while the other one has reward $1/2 - 1/\sqrt{T}$. The problem becomes the well-studied bandit problem of trying to distinguish which arm has a slightly higher reward for the context. By (Slivkins 2019, Chapter 2), any algorithm for this problem has regret at least $\Omega(\sqrt{T})$.

Remark 3. **Efficient Algorithmic Implementation.** It is natural to wonder about the computational complexity of Alg 1. Although as stated above, A and B range over the continuous context space $[0,1]$, in fact at each time t there are at most t relevant values to check, namely the contexts $x_1, \ldots, x_t$ themselves. This leads to a naive running time of $O(T^3)$ per time-step to check all pairs (A, B) (each of which requires summing up to t values), hence an overall $O(T^4)$ running time. A standard application of the range tree data structure (see e.g. (de Berg et al. 2008, Chapter 5.3)) improves this to $\tilde{O}(T^2)$ per time-step, hence $\tilde{O}(T^3)$ running time. Additionally, instead of checking the elimination condition at every round, the proof would work with almost no modification if the condition is checked only at times of the form $t_i = \tau_E + \lfloor 1.1^k \rfloor$ for $k \in \mathbb{N}_+$. Since the regret analysis still applies if the algorithm performs no eliminations after $\tilde{O}(K^2)$ epochs, this requires checking the elimination condition at only $\tilde{O}(K^2)$ times, hence yields a running time of $\tilde{O}(K^2 T^2)$.

[2] As long as $t_d \leq t_0$ is uniformly bounded, a deterministic delay can be simulated.

6. Data

This paper formulates the Contextual Packaging problem as an online learning problem. In the classic online learning framework, past data are not available and distributions are unknown a priori. Feedback and data points are collected in real time as actions are taken, and the algorithm learns the best action as time goes on. Nevertheless, the abundance of past HelloFresh delivery data should not go to waste. We would like to extract as much useful structural information as possible from the past (offline) data to assist with our modeling and design of the algorithm, so that online learning can be done faster.

The past delivery dataset contains hundreds of millions of meal-boxes delivered from HelloFresh to the customer around the globe since 2016, documenting the packaging decision for each of the hundreds of millions of meal kit boxes. In addition, several other associated datasets are instrumental to extracting feedback and contextual information for our learning about the corresponding action outcomes. These associated datasets include customer complaint data, weather data, menu data, recipe ingredients data, and other relevant information.

As discussed in Section 3, we take advantage of the abundant past data by learning two functions ϕ and ψ that help us order the actions and the contexts by their intensity/strength. These orderings allow us to reduce the dimensionality and enable efficient online learning for Alg 1. Below we highlight a number of interesting challenges we encounter in learning these ordering functions.

6.1. Data Challenges

1. *Sparsity of Signal*: historically, insufficient thermal packaging failures rarely happen, as the top priority of the providers is to successfully deliver the boxes to the customers, even at the price of overpackaging. The ratio of temperature complaints in the historical dataset is lower than 0.1%.

2. *Noise*: the failure signals for each delivered meal-box are gathered by applying filters to the customer complaint data set to collect any complaint data points that contain relevant keywords such as “warm”, “temperature”, “melted”, etc. These complaints could come from customers submitting a complaint online or from phone calls with customer services. Customers are rationally incentivized to submit a complaint if their box failed, but it cannot be guaranteed that 100% of the customers who encountered a failed box would necessarily submit a complaint. Over the years, HelloFresh has made an effort to make it increasingly easy for customers to submit a complaint. Still, this noise is inevitable. There is reason to believe that the disturbance of this noise should be rather mild, as rational customers would not continue to use the service without complaining if they received an insufficiently packaged box, because of the absence of a compensation.

3. *Confounding*: we would run into serious issues of confounding if we attempt to learn the ordering functions directly from historical delivery and complaint data. For example, a simple linear regression would suggest that higher local temperature leads to lower probability of receiving a complaint $p(\mathbf{x}_t, \mathbf{a}_t)$, contradicting the laws of physics. Upon further investigation, the reason for this particular confounding is that historically, whenever the temperature is higher, the packaging guidelines would increase the freezer packs significantly, resulting in a smaller probability of warm complaint. We must resolve such confounding before we can safely utilize the historical data.

6.2. Our Solution to the Data Challenges

A simple popular solution to the sparse signal issue would be to sample similar numbers of positive and negative data points. However, that would too drastically reduce the size of our dataset. Instead, we develop a solution that tackles the sparse signal issue and the confounding issue at the same time. We balance the dataset manually by injecting synthetic deliveries that result in temperature complaints.

An example *data injection* is as follows: for an observation in the dataset that did not result in a complaint, we hypothetically increase the temperature by 20 degrees Farenheit and reduce the amount of ice in the box by half. We then declare that it must have ended in a thermal complaint from the customer. These injections restore the positive association of high temperature and low amounts of packaging materials with customer complaints by decoupling the correlation between past decisions and past contexts.

Trade-off between information *and* truth. It is important that all the data injections must be "true": we must distort the contextual information and the action chosen far enough for us to confidently claim that this hypothetical data point must have certainly resulted in a failure for this box. If we barely distort the original data point and claim that it ended up in a failed delivery, then we could be adding an *untrue* data point, introducing bias to our learning. On the other hand, if we over-exaggerate too far in distorting the original data point, this data point would certainly be true, but the information that our algorithm can learn from such an injected data point would be minimal. The ideal data points to inject are those just distorted enough to end in failures; those injected data points would be the ones that give the most information to learn about the boundary conditions without introducing bias. However, asking for such perfect injections is unrealistic. When unsure, we prefer to err on the safe side and lean towards over-exaggerating, as the more conservative approach still largely corrects the confounding without introducing error into our learning.

We conducted a number of such injections, largely resolving the confounding factors in the dataset and resolving the sparse signal issue simultaneously.

7. Numerical Experiments

We accessed multiple hundreds of millions of records of orders and packaging decisions made by HelloFresh between 2017 and 2022. We compare the performance of our Algorithm 1 and the performance of the current packaging guidelines in place at HelloFresh.

As we described in Section 1.1, implementing our algorithm in real time in production is technologically attainable due to the high level of automation in the meal kit facilities, as our method only alters the computer program that shows the sequences of light for the human pickers to follow in a straightforward manner. However, to prioritize safety and quality of the products, instead of directly applying our algorithm to make real-time packaging decisions, we use our algorithm to internally serve the thermal engineering team that stipulates the packaging guidelines. As mentioned in Section 1.1, the chamber experiments done by the thermal engineering team are expensive and time-consuming. It is a highly valuable use of our algorithm to identify subsets of contexts where overpackaging seems to exist in the current packaging guidelines. Once we have identified these interest points, the thermal engineering team can direct their efforts in a much more focused way, and refine the current guidelines around these interest points.

We select a recent dataset of over 4 million deliveries from a major distribution center in New Jersey, US between Jan 2022 and May 2023, to create a simulated environment that generates simulated customer feedback for each box our algorithm hypothetically packages and ships out. After the data injections described in Section 6, the augmented sampled dataset has approximately 8 million observations.

transit	max temperature	size	refrigerated	our average	current average
2 days	65-67	all	all	12.4	14.6
2 days	85-90	large	all	17.6	21.4
1 day	90-93	medium	all	23.2	26.7
1 day	30-36	large	all	31.6	35.2
1 day	33-35	all	all	19.8	20.4
1 day	90-92	all	yes	29.3	30.4

Table 2 Example subsets identified by Alg 1 where overpackaging could exist in current guidelines.

Figure 4 Cumulative pseudo-cost for all boxes within $65-67°$ with 2-day transit.

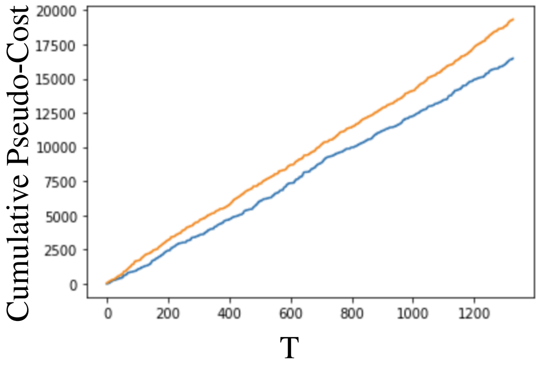


Figure 5 Cumulative pseudo-cost for large boxes within $85-90°$ with 2-day transit.

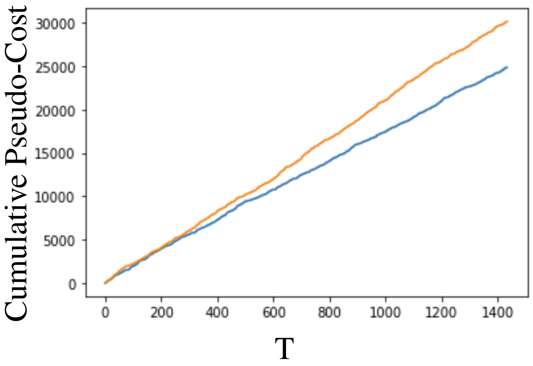


Figure 6 Cumulative pseudo-cost for large boxes within $30-36°$ with 1-day transit.

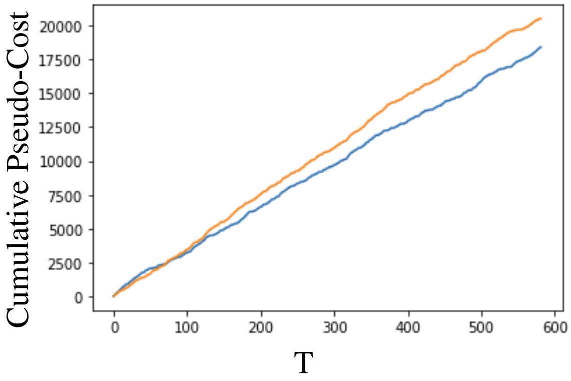


Figure 7 Cumulative pseudo-cost for all boxes within $33-35°$ with 1-day transit.

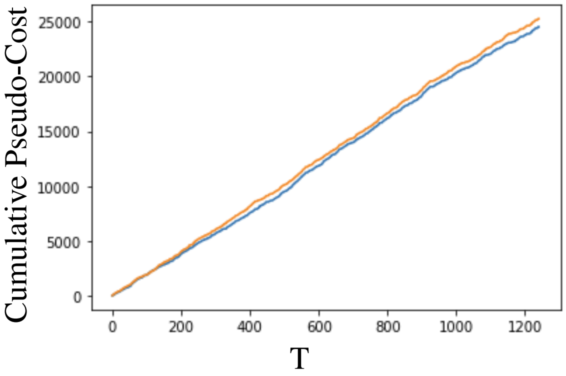
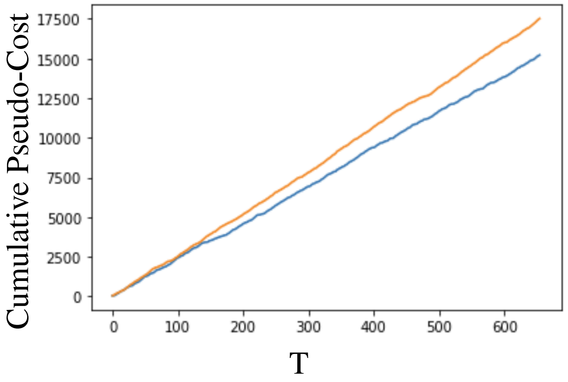
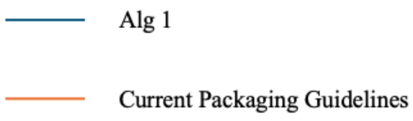


Figure 8 Cumulative pseudo-cost for medium boxes within $90-93°$ with 1-day transit.





Realistically, there is a delay between the time the provider makes the packaging decision and the time that feedback is received for the box sent out. We discussed in Remark 1 how this delay does not affect the regret bounds. Nevertheless, in the numerical experiment, we incorporate the delay into the experiments by letting Alg 1 make decisions in *batches* and only update the algorithm after each batch is finished. In the following experiments, we used a batch size of 1,000 boxes. Alg 1 explores among a small number of actions adjacent to the current guidelines.

To protect proprietary information in the business, we blur the actual costs by multiplying them with some random positive constant. The pseudo-cost includes the material cost, the failure penalty, and any environmental footprint cost that the provider would like to consider for a packaging option. We referred to these costs after scaling as the *pseudo-costs*.

Table 2 shows some examples of subsets that our contextual bandit algorithm identifies as settings under which overpackaging could be present. Figures 4–8 show the cumulative cost difference between our algorithm and the current guidelines in some settings of interest where Alg 1 identified potential overpackaging. In the majority of settings that were examined, the cumulative costs of Alg 1 are indistinguishable from those of the current packaging guidelines, meaning that the current guidelines are approximately optimal for these subsets of contexts. However, for these interest points identified in Table 2 and Figures 4–8, our algorithm was able to find packaging solutions that results in a lower average cost. Having narrowed down to a much smaller space for experimentation, the thermal engineering team is able to conduct more focused chamber experiments to identify overpackaging with higher efficiency and lower cost, to refine the current packaging guidelines.

8. Conclusions

In this work, we develop artificial intelligence solutions to reduce overpackaging in the meal kit service industry. Our algorithm takes advantage of the abundance of offline data to pre-learn structural information that accelerates online learning.

One of the greatest advantages of the artificial intelligence solution proposed compared to the existing practice is that our solution allows for future flexibility to move towards more efficient and environmentally friendly packaging materials with ease. The current practice of conducting long and expensive thermal experiments to decide the packaging guidelines is slow and past results become invalid once the service provider decides to adopt a new set of liner materials, ice pack sizes or box sizes. With the help of our contextual bandit algorithm, if such changes happen, the provider would be able to swiftly generate a new set of packaging guidelines. As more environmentally friendly options for the packaging materials emerge in the market, the flexibility to switch to more favorable materials would be an important competitive edge for any meal kit service provider.

This flexibility in fact encompasses more than just the packaging materials evolution. The online learning nature of the method also allows the service provider to adapt more agilely to other visible or non-visible changes in the environment, such as changes in transit logistics and customer behavioral shifts. For example, when remote work is popular, customers are more likely to be home and open the meal kit box soon after it arrives. As companies ask their employees to return to office, it is possible that more packaging materials might be needed on average. The behavioral aspect of the environmental randomness is something that can be learned and adjusted for intelligently by our contextual bandit algorithm, as compared with the traditional methods of conducting chamber experiments.

Sustainability is the most viable path to the future of services and supply chains, especially when being sustainable aligns with cost-saving. Artificial intelligence solutions have the potential to realize this alignment. There are many future directions of research. How to use artificial intelligence to target pain-points in service operations and supply chains, and to find efficient sustainable solutions much faster than human power is a great question of interest that should bring fruitful results in the next decades to come.

References

Agarwal A, Foster DP, Hsu DJ, Kakade SM, Rakhlin A (2011) Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems* 24.

Akkaş A, Gaur V, Simchi-Levi D (2019) Drivers of Product Expiration in Consumer Packaged Goods Retailing. *Management Science* 65(5):2179–2195.

Akkaş A, Honhon D (2022) Shipment policies for products with fixed shelf lives: Impact on profits and waste. *Manufacturing & Service Operations Management* 24(3):1611–1629.

Am JB, Doshi V, Malik A, Noble S, Frey S (2023) Consumers care about sustainability—and back it up with their wallets. `https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/consumers-care-about-sustainability-and-back-it-up-with-their-wallets`, accessed: 2024-4-11.

Ata B, Lee D, Sönmez E (2019) Dynamic volunteer staffing in multicrop gleaning operations. *Operations Research* 67(2):295–314.

Auer P, Cesa-Bianchi N, Freund Y, Schapire R (1995) Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science*, 322–331.

Balseiro SR, Golrezaei N, Mahdian M, Mirrokni VS, Schneider J (2019) Contextual bandits with cross-learning. *NeurIPS*.

Belavina E (2021) Grocery store density and food waste. *Manufacturing & Service Operations Management* 23(1):1–18.

Belavina E, Girotra K, Kabra A (2017) Online grocery retail: Revenue models and environmental impact. *Management Science* 63(6):1781–1799.

Bird J (2018) What a waste: Online retail's big packaging problem. URL `https://www.forbes.com/sites/jonbird1/2018/07/29/what-a-waste-online-retails-big-packaging-problem/`.

Butler K (2017) The truth about meal-kit freezer packs. URL `https://www.motherjones.com/environment/2017/06/meal-kit-freezer-packs-blue-apron-hello-fresh/`.

Callewaert P, Raadal HL, Lyng KA (2023) How to achieve ambitious recycling targets for plastic packaging waste? The environmental impact of increased waste separation and sorting in Norway. *Waste Management* 171:218–226.

Choi SJ, Burgess G (2007) Practical mathematical model to predict the performance of insulating packages. *Packaging Technology and Science* 20(6):369–380.

de Berg M, Cheong O, van Kreveld M, Overmars M (2008) *Computational geometry algorithms and applications* (Springer).

Dirpan A, Hidayat SH, Djalal M, Ainani AF, Yolanda DS, Kasmira, Khosuma M, Solon GT, Ismayanti N (2023) Trends over the last 25 years and future research into smart packaging for food: A review. *Future Foods* 8:100252, ISSN 2666-8335.

Gavrilescu D, Seto BC, Teodosiu C (2023) Sustainability analysis of packaging waste management systems: A case study in the Romanian context. *Journal of Cleaner Production* 422:138578.

Gong XY, Simchi-Levi D (2024) Bandits atop Reinforcement Learning: Tackling Online Inventory Models with Cyclic Demands. *Management Science* 70(9):6139–6157.

Gritsch L, Lederer J (2023) A historical-technical analysis of packaging waste flows in Vienna. *Resources, Conservation and Recycling* 194:106975.

Han Z, Hu B, Dawande M (2023) *Seeing Beauty in Ugly Produce: A Food Waste Perspective* (SSRN), URL `https://books.google.com/books?id=9j8A0AEACAAJ`.

Hezarkhani B, Demirel G, Bouchery Y, Dora M (2023) Can "ugly veg" supply chains reduce food loss? *European Journal of Operational Research* 309(1):117–132, ISSN 0377-2217.

Howard M (2024) Meal kit delivery services market size, share, trends, analysis, and future outlook - navigating consumer preferences, market dynamics. URL `https://www.linkedin.com/pulse/meal-kit-delivery-services-market-size-share-trends-analysis-howard-4wm4f`.

Kallbekken S, Sælen H (2013) 'Nudging' hotel guests to reduce food waste as a win–win environmental measure. *Economics Letters* 119(3):325–327.

Kazaz B, Xu F, Yu H (2023) Retailing strategies of imperfect produce and the battle against food waste. *SSRN Electronic Journal* URL `http://dx.doi.org/10.2139/ssrn.4392397`.

Kucharek M, Yang L, Wang K (2020) Assessment of insulating package performance by mathematical modelling. *Packaging Technology and Science* 33(2):65–73.

Lee D, Sönmez E, Gómez MI, Fan X (2017) Combining two wrongs to make two rights: Mitigating food insecurity and food waste through gleaning operations. *Food Policy* 68:40–52.

Lee D, Tongarlak MH (2017) Converting retail food waste into by-product. *European Journal of Operational Research* 257(3):944–956, ISSN 0377-2217.

Li L, Chu W, Langford J, Schapire RE (2010) A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World Wide Web*, 661–670.

Naaman M (2021) On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters* 173:109088.

Pandey S, Agarwal D, Chakrabarti D, Josifovski V (2007) *Bandits for Taxonomies: A Model-based Approach*, 216–227.

Reichheld A, Peto J, Ritthaler C (2023) Research: Consumers' sustainability demands are rising. `https://hbr.org/2023/09/research-consumers-sustainability-demands-are-rising`, accessed: 2024-4-11.

Richtel M (2016) E-commerce: Convenience built on a mountain of cardboard. URL `https://www.nytimes.com/2016/02/16/science/recycling-cardboard-online-shopping-environment.html`.

Schmidt K (2016) Explaining and promoting household food waste-prevention by an environmental psychological based intervention study. *Resources, Conservation and Recycling* 111:53–66.

Slivkins A (2019) Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* 12(1-2):1–286.

Statista (2024) Number of subscribers to meal kit companies worldwide from 2016 to 2022. URL `https://www.statista.com/statistics/947620/meal-kit-companies-number-subscribers-worldwide/`, accessed: 2024-05-05.

Wang CC, Kulkarni SR, Poor HV (2005) Bandit problems with side observations. *IEEE Transactions on Automatic Control* 50(3):338–355.

Wang K, Yang L, Kucharek M (2020) Investigation of the effect of thermal insulation materials on packaging performance. *Packaging Technology and Science* 33(6):227–236.

Yuan H, Luo Q, Shi C (2021) Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science* 67(10):6089–6115.

Zhang L, Liu Y, Zhao Z, Yang G, Ma S, Zhou C (2023) Estimating the quantities and compositions of household plastic packaging waste in China by integrating large-sample questionnaires and lab-test methods. *Resources, Conservation and Recycling* 198:107192, ISSN 0921-3449.

Zhao H, Chen W (2019) Stochastic One-Sided Full-Information Bandit. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'2019).*

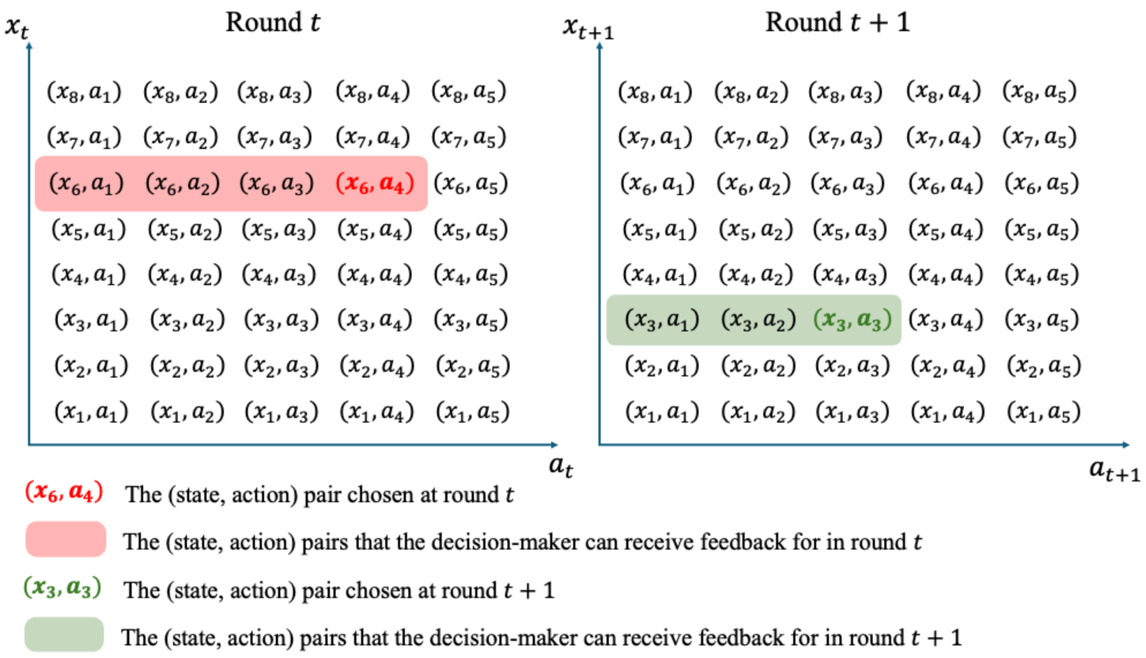
Appendix A: (Unconditional) One-Sided Feedback

We provide a definition of (unconditional) one-sided feedback, adapted to the contextual bandit setting.

Definition 2 *A problem has the lower* one-sided feedback *if there exists an ordering function $\phi : \mathcal{A} \to \mathbb{R}$ such that, for each round t and context $\mathbf{x}_t$ received, if action $\mathbf{a}_t$ is played, then all rewards (costs) for actions $\mathbf{a}$ with $\phi(\mathbf{a}) \le \phi(\mathbf{a}_t)$ are observed in that round t. (In the higher one-sided feedback setting, all rewards (costs) for actions $\mathbf{a}$ with $\phi(\mathbf{a}) \ge \phi(\mathbf{a}_t)$ are observed.)*

See Figure 9 for an example of the lower-side case of one-sided feedback.

Figure 9 An example of One-Sided Feedback (lower side) in two rounds t and $t+1$. For ease of presentation, in this example, the contexts and the actions are already indexed by their orderings.


Round t
Round $t+1$
x_t
a_t
x_{t+1}
a_{t+1}
(x_8, a_1) (x_8, a_2) (x_8, a_3) (x_8, a_4) (x_8, a_5)
(x_7, a_1) (x_7, a_2) (x_7, a_3) (x_7, a_4) (x_7, a_5)
(x_6, a_1) (x_6, a_2) (x_6, a_3) (x_6, a_4) (x_6, a_5)
(x_5, a_1) (x_5, a_2) (x_5, a_3) (x_5, a_4) (x_5, a_5)
(x_4, a_1) (x_4, a_2) (x_4, a_3) (x_4, a_4) (x_4, a_5)
(x_3, a_1) (x_3, a_2) (x_3, a_3) (x_3, a_4) (x_3, a_5)
(x_2, a_1) (x_2, a_2) (x_2, a_3) (x_2, a_4) (x_2, a_5)
(x_1, a_1) (x_1, a_2) (x_1, a_3) (x_1, a_4) (x_1, a_5)

(x_6, a_4) The (state, action) pair chosen at round t

🟥 The (state, action) pairs that the decision-maker can receive feedback for in round t

(x_3, a_3) The (state, action) pair chosen at round $t+1$

🟩 The (state, action) pairs that the decision-maker can receive feedback for in round $t+1$

Appendix B: Proof of Theorem 2

Proof of Theorem 2 We construct another example in which there are essentially $T^{1/3}$ separate bandit problems. The construction again ensures that conditional 1-sided feedback does not provide any new usable information. It depends on a uniformly random sequence $\sigma_1, \dots, \sigma_{T^{1/3}} \overset{I.I.D.}{\sim} \text{Unif}(\{-1, 1\})$ of signs. We define it for the continuous action space setting of $[0,1]$; with two actions, restricting to $\{0,1\}$ will give the same hardness result.

For $x \in [0,1]$, let $|x|_*$ be the distance from x to the nearest multiple of $T^{-1/3}$. Given $1 \le j \le T^{1/3}$ and $x \in I_j \equiv \left[\frac{j-1}{T^{1/3}}, \frac{j}{T^{1/3}}\right]$, the average loss function is

$$\ell(x,a) = (1-a)\sigma_j |x|_*.$$

We can crucially write $\ell(x,a) = \ell_0(x) + \ell_1(x,a)$ as the sum of two terms: a known function of just x and a known function which is increasing in x and decreasing in a. Namely defining

$$\tilde{\sigma}(x) = \begin{cases} \sigma_j, & x \in \left[\frac{2j-2}{2T^{1/3}}, \frac{2j-1}{2T^{1/3}}\right] \\ -\sigma_j, & x \in \left[\frac{2j-1}{2T^{1/3}}, \frac{2j}{2T^{1/3}}\right] \end{cases},$$

we set:

$$\ell_0(x) = x;$$
$$\ell_1(x,a) = \int_0^x (1-a)\tilde{\sigma}(u) - 1 \, \mathrm{d}u.$$

This is a valid decomposition of ℓ because one can easily check that $\int_0^x \tilde{\sigma}(u)\mathrm{d}u = \sigma_j |x|_*$.

We take the realized rewards $L(x,a,\xi)$ to be $\ell_0 + 1_{\xi \le \ell_1(x,a)}$ where $\xi \sim \mathrm{Unif}[0,1]$. (Technically L is valued in $[0,2]$ rather than $[0,1]$ but this is irrelevant by scaling.) Hence the first term is known, while the second is a Bernoulli reward. The contexts $x_1,\ldots,x_T$ are I.I.D. uniform on $[0,1]$.

Since ℓ_1 is increasing in x and decreasing in a, the Bernoulli structure of the realized rewards L reveals no additional information from conditional 1-sided feedback. As a result, we have constructed a set of $T^{1/3}$ independent learning problems which occur in parallel, one for each σ_j. In the full problem, each of these subproblems appears $n_j \sim Bin(T, T^{-1/3})$ times, since the contexts $x_t \in [0,1]$ are independent and uniformly random (of course the n_j variables are not independent). By standard tail bounds for Binomial random variables, with high probability $1 - o_T(1)$, we have $n_j \in [T^{2/3}/2, 2T^{2/3}]$ for all i.

By similar reasoning with $Bin(T, T^{-1/3}/3)$ random variables, with probability $1 - o_T(1)$ all of the intervals I_j will have at least $T^{2/3}/10$ contexts landing inside their middle third $J_i = \left[\frac{3j-2}{3T^{1/3}}, \frac{3j-1}{3T^{1/3}}\right]$. For each j, it is impossible to determine the value of σ_j with $\ge 2/3$ probability while $n_j \le T^{2/3}/C$ samples for some absolute constant C (see e.g. (Slivkins 2019, Chapter 2)). Since each σ_j is an independent uniform sign, the expected regret from each I_j is thus $\Omega(T^{1/3})$, so the overall expected regret is $\Omega(T^{2/3})$. □

Appendix C: Proof of Theorem 3

Proof of Theorem 3 For each parallel problem indexed by i, let $\mathrm{OPT}_i := \min_{a \in \mathcal{A}} \mathbb{E}_{x \sim D}[\ell(x,a) | x \in I_i]$ denote the optimal benchmark developed in Agarwal et al. (2011) for the contextless subproblem on interval I_i. (Here we focus on $\mathcal{A} = [0,1]$; if $\mathcal{A} = \{0,1\}$, then in place of Agarwal et al. (2011) we just use the basic $O(\sqrt{T})$ regret bound for two-armed stochastic bandits.) Let ALG_i denote the total cost incurred by our algorithm on I_i. We let n_i be the total number of contexts $x_t \in I_i$ (as $t \le T$ varies). Using result of Agarwal et al. (2011) and Jensen's inequality, our algorithm incurs

$$\sum_{i=1,\ldots,T^{1/3}} \mathbb{E}[\mathrm{ALG}_i - n_i \mathrm{OPT}_i] \tag{14}$$
$$= \mathbb{E} \sum_{i=1}^{T^{1/3}} \tilde{O}(\sqrt{n_i}) \tag{15}$$
$$\le \tilde{O}(T^{2/3}) \tag{16}$$

expected regret against the contextless optimal benchmark for the interval-averaged failure

$$\ell_{I_i}(a) = \mathbb{E}^{x \in I_i}[\ell(x,a)],$$

which is convex in a because ℓ is convex by assumption. In other words, with the total cost incurred $\mathrm{ALG} = \sum_i \mathrm{ALG}_i$, and with $x_t \in I_{i_t}$,

$$\mathbb{E}\Big[\mathrm{ALG} - \sum_{t=1}^{T} \ell_{I_{i_t}}(a_t)\Big] \le \tilde{O}(T^{2/3}) \tag{17}$$

The remaining suboptimality is from treating each subproblem as contextless (or having one context) and choosing only one action a_i for each interval I_i in OPT_i. Since ℓ is Lipschitz in x by assumption, if x and $\tilde{x} \in I_i$,

$$|\ell(x,a) - \ell(\tilde{x},a)| \le O(T^{-1/3}), \qquad \forall a.$$

Substituting this estimate into each term of the sum in Equation (17) completes the proof. □

Appendix D: Proof of Lemma 1

Proof of Lemma 1 Let $n_s = |S_E(x_s)|$. We note that under any policy, the quadruples (x_s, a_s, n_s, L_s) are I.I.D. (within a fixed epoch E), and the policy used is constant within epoch E. We will apply the multi-dimensional Dvoretzky–Kiefer–Wolfowitz–Massart inequality, see e.g. Naaman (2021). This inequality states that the multi-dimensional cumulative distribution function:

$$F(x,a,n,\ell) = \mathbb{P}[x_t \le x, a_t \le a, n_t \le n_t, L_t \le L]$$

is uniformly approximated by the within-epoch-E empirical estimate

$$\hat{F}(x,a,n,L) = \frac{1}{t-\tau_E+1} \sum_{s=\tau_E}^{t} 1_{x_s \le x, a_s \le a, n_s \le n, L_s \le L}$$

in the sense that at time t, for any $\delta > 0$:

$$\begin{aligned} &\mathbb{P}[\sup_{x,a,n,\ell} |\hat{F}(x,a,n,\ell) - F(x,a,n,\ell)| \ge \delta] \\ &\le 4(t-\tau_E+1)e^{-2(t-\tau_E+1)\delta^2} \\ &\le 4Te^{-2(t-\tau_E+1)\delta^2}. \end{aligned} \tag{18}$$

(Note $t-\tau_E+1$ is the number of samples obtained after time t in epoch E, i.e. the number of IID samples in the multi-dimensional DKWM inequality.) As in the inclusion-exclusion principle, a similar bound holds for any product of intervals $I^{(4)} = [x^{(1)}, x^{(2)}] \times [a^{(1)}, a^{(2)}] \times [n^{(1)}, n^{(2)}] \times [L^{(1)}, L^{(2)}]$. Namely, define the "bounded" version of F:

$$\begin{aligned} &F_{bdd}(x^{(1)}, x^{(2)}, a^{(1)}, a^{(2)}, n^{(1)}, n^{(2)}, L^{(1)}, L^{(2)}) \\ &= \frac{1}{t-\tau_E+1} \sum_{s=\tau_E}^{t} 1_{(x_s, a_s, n_s, L_s) \in I^{(4)}}. \end{aligned}$$

Analogously, define the empirical estimate $\hat{F}_{bdd}(x^{(1)}, x^{(2)}, a^{(1)}, a^{(2)}, n^{(1)}, n^{(2)}, L^{(1)}, L^{(2)})$. Then with $\vec{i} = (i_x, i_a, i_n, i_L)$, the alternating sum of 16 terms

$$\sum_{\vec{i} \in \{1,2\}^4} (-1)^{i_x+i_a+i_n+i_L} F(x^{(i_x)}, a^{(i_a)}, n^{(i_n)}, L^{(i_L)})$$

gives the corresponding value for F_{bdd} corresponding to the half-open intervals $(x^{(1)}, x^{(2)}] \times (a^{(1)}, a^{(2)}] \times (n^{(1)}, n^{(2)}] \times (L^{(1)}, L^{(2)}]$, and similarly for $\hat{F}$. Applying the triangle inequality to (18) then shows the same bound for the product

$(x^{(1)},x^{(2)}]\times(a^{(1)},a^{(2)}]\times(n^{(1)},n^{(2)}]\times(L^{(1)},L^{(2)}]$, with error term $64(T+1)e^{-2(t-\tau_E+1)\delta^2}$. If the context distribution is free of atoms, this finishes the proof. In general, since this bound is uniform in the interval endpoints, we get the same bound for $[x^{(1)},x^{(2)}]\times[a^{(1)},a^{(2)}]\times[n^{(1)},n^{(2)}]\times[L^{(1)},L^{(2)}]$. This is since for any probability measure μ on $\mathbb{R}^4$, $\lim_{x^1\uparrow x^{(1)},a^1\uparrow a^{(1)},n^1\uparrow n^{(1)},\ell^1\uparrow \ell^{(1)}} \mu\Big((x^1,x^{(2)}]\times(a^1,a^{(2)}]\times(n^1,n^{(2)}]\times(\ell^1,\ell^{(2)}]\Big)=\mu\Big([x^{(1)},x^{(2)}]\times[a^{(1)},a^{(2)}]\times[n^{(1)},n^{(2)}]\times[L^{(1)},L^{(2)}]\Big)$.

We apply this bound for each $[x^{(1)},x^{(2)}]=[A,B]$ by fixing $a^{(1)}=a^{(2)}=i$ for action a_i, and with $L^{(1)}=0$. Define $I_{A,B,s,i,n,L}$ to be 1 if $\{x_s\in[A,B];a_s=a_i;n_s\geq n;L_s\leq L\}$, and 0 otherwise. The conclusion is that with probability $O(Te^{-2(t-\tau_E+1)\delta^2})$, simultaneously over $[A,B]$ and n and L:

$$\Big|\sum_{s=\tau_E}^{t} I_{A,B,s,i,n,L}-\mathbb{P}[\mathbf{1^4}(s)]\Big|\leq\delta(t-\tau_E+1).$$

where

$$\mathbf{1^4}(s):=1_{x_s\in[A,B]}\cdot 1_{a_s=a_i}\cdot 1_{n_s\geq n}\cdot 1_{L_s\in[L,1]}.$$

Integrating over $L\in[0,1]$ replaces the indicator $1_{L_s\in[L,1]}$ by the value L_s in both terms, with the same error bound. Thus

$$\Big|\sum_{s=\tau_E}^{t} I_{A,B,s,i,n}\,L_s-\mathbb{E}[\ell(x_s)\,\mathbf{1^3}(s)]\Big|\leq\delta(t-\tau_E+1) \tag{19}$$

where $I_{A,B,s,i,n}=I_{A,B,s,i,n,1}$ and $\mathbf{1^3}(s):=1_{x_s\in[A,B]}\cdot 1_{a_s=a_i}\cdot 1_{n_s\geq n}$.

Then summing (19) over $n=1,2,\ldots,K$ gives an upper bound of $K\delta(t-\tau_E+1)$ on the error between the true expectation and the importance weighted estimate. Indeed, $\sum_{n=1}^{K}\sum_{s=\tau_E}^{t} I_{A,B,s,i,n}\cdot L_t$ exactly gives the importance weighted estimate $\hat{\ell}_{[A,B]}$ while $\sum_{n=1}^{K}\mathbb{E}[\ell(x_s)1_{x_s\in[A,B]}\cdot 1_{a_s=a_i}\cdot 1_{n_s\geq n}]$ gives the expected true cost on a_i from contexts in $[A,B]$. This is because the event $n_s=j$ is counted j times, balancing the $1/j$ probability to choose a_i. Taking $\delta=\frac{10\sqrt{\log(KT)}}{\sqrt{t-\tau_E+1}}$ finishes the proof. □

Appendix E: Proof of Lemma 3

Proof of Lemma 3 Note that if the time t term

$$1_{x_t\in[A,B]}\cdot 1_{a_t=i}\cdot n_E(x_t)L_t$$

equals 0, then $\hat{\ell}_{E,t-1,i}\leq\hat{\ell}_{E,t,i}$. Meanwhile if that term equals K, then $\hat{\ell}_{E,t-1,i}\geq\hat{\ell}_{E,t,i}$. On the other hand, this change affects $\hat{\ell}_{E,t,i}$ by $\frac{K}{t-\tau_E+1}$. This completes the proof. □

Appendix F: Proof of Lemma 4

Proof of Lemma 4 On the event of Lemma 1, we have $[A,B]\subseteq F_{t,j}$ as explained previously. By definition, $t-\tau_E+1$ is the number of extra time steps since the beginning of the epoch. Additionally by the triangle inequality the *first* occurrence of (6) must happen at a time $t=\tau_{E+1}-1$ satisfying

$$\begin{aligned}&\int_{F_{E,i}\cap I_j}\ell_i(x)-\ell_j(x)\,\mathrm{d}\mu(x)\\ \leq&12\sqrt{\frac{CK^2\log(KT)}{t-\tau_E+1}}+\frac{2K}{t-\tau_E+1}\\ \leq&14\sqrt{\frac{CK^2\log(KT)}{t-\tau_E+1}}.\end{aligned}$$

(Here we use Lemma 3 and that Equation (6) does not hold at time $t-1$.) Similarly, if

$$\int_{A'}^{B'} \ell_{i'}(x) - \ell_{j'}(x)\, \mathrm{d}\mu(x) \geq 6\sqrt{\frac{CK^2 \log(KT)}{t-\tau_E+1}}$$

holds for any other (A', B', i', j') then Equation (6) must have already been satisfied by this time, sandwiched by these inequalities. In other words, the minimum possible value ε_* of ε over all $F_{E,i}, I_j$ must satisfy $\varepsilon_* \in \left[6\sqrt{\frac{CK^2\log(KT)}{t-\tau_E+1}}, 14\sqrt{\frac{CK^2\log(KT)}{t-\tau_E+1}}\right]$. This proves Equation (9) since $14/6 < 3$. The next assertion of the lemma follows by rearranging the inequality $\varepsilon \leq 14\sqrt{\frac{CK^2\log(KT)}{t-\tau_E+1}}$ we just obtained; recall that $t-\tau_E$ is the number of extra time steps since the beginning of the epoch. Finally (11) follows by similar reasoning using the definition (7); the error bound in (7) is of lower order than ε_*, so the triangle inequality shows that (11) holds with the value ε_*. □

Appendix G: Proof of Lemma 5

Proof of Lemma 5 Assuming Equation (12), there must exist $i \in [K]$ such that $R_{E,i} \geq \varepsilon/K$. Therefore by Lemma 2, there exists $j \in [K]$ and an interval $F_{E,i} \cap I_j$ such that

$$\int_{x \in F_{E,i} \cap I_j} \ell_j(x) - \ell_i(x)\, \mathrm{d}\mu(x) \geq \varepsilon/K^2.$$

The claims now follow by applying Lemma 4 with ε replaced by ε/K^2; note that $K^2/(\varepsilon/K^2)^2 = K^6/\varepsilon^2$. It follows from Lemma 4 and Equation (11) (via Equation (7)) that the next elimination phase removes $\Omega(\varepsilon/K^2)$ of the total active suboptimality (as defined in Equation (8)). With ε chosen so equality holds in (12), this implies the desired result (since we can assume without loss of generality that Inequality (12) holds with an equal sign). The final assertion follows from (11) and Property 1 (since $\varepsilon/10 \geq 0$). □