Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Bayesian Prophet: A Low-Regret Framework for Online Decision Making

Alberto Vera, Siddhartha Banerjee

To cite this article:
Alberto Vera, Siddhartha Banerjee (2021) The Bayesian Prophet: A Low-Regret Framework for Online Decision Making. Management Science 67(3):1368-1391. https://doi.org/10.1287/mnsc.2020.3624

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

MANAGEMENT SCIENCE
Vol. 67, No. 3, March 2021, pp. 1368–1391
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Bayesian Prophet: A Low-Regret Framework for Online Decision Making

Alberto Vera,[a] **Siddhartha Banerjee**[a]

[a] Cornell University, Ithaca, New York 14853

Contact: aav39@cornell.edu (AV); sbanerjee@cornell.edu, https://orcid.org/0000-0002-8954-4578 (SB)

Received: April 24, 2019
Revised: August 28, 2019; December 6, 2019; January 28, 2020
Accepted: February 21, 2020
Published Online in Articles in Advance: October 5, 2020

https://doi.org/10.1287/mnsc.2020.3624

Abstract. We develop a new framework for designing online policies given access to an oracle providing statistical information about an off-line benchmark. Having access to such prediction oracles enables simple and natural Bayesian selection policies and raises the question as to how these policies perform in different settings. Our work makes two important contributions toward this question: First, we develop a general technique we call *compensated coupling*, which can be used to derive bounds on the expected regret (i.e., additive loss with respect to a benchmark) for any online policy and off-line benchmark. Second, using this technique, we show that a natural greedy policy, which we call the *Bayes selector*, has constant expected regret (i.e., independent of the number of arrivals and resource levels) for a large class of problems we refer to as "online allocation with finite types," which includes widely studied online packing and online matching problems. Our results generalize and simplify several existing results for online packing and online matching and suggest a promising pathway for obtaining oracle-driven policies for other online decision-making settings.

History: Accepted by George Shanthikumar, big data analytics.
Funding: This work was supported by the National Science Foundation Division of Electrical, Communications and Cyber Systems [Grant ECCS-1847393], the National Science Foundation Division of Mathematical Sciences [Grant DMS-1839346], and the Army Research Laboratory [Grant W911NF-17-1-0094].

Keywords: stochastic optimization • prophet inequalities • approximate dynamic programming • revenue management • online allocation • online packing • online matching

Life is the sum of all your choices.—Albert Camus

1. Introduction

Everyday life is replete with settings in which we have to make decisions while facing uncertainty over future outcomes. Some examples include allocating cloud resources, matching an empty car to a ride-sharing passenger, displaying online ads, selling airline seats, etc. In many of these instances, the underlying arrivals arise from some known generative process. Even when the underlying model is unknown, companies can turn to ever-improving machine learning tools to build predictive models based on past data. This raises a fundamental question in online decision-making: how can we use predictive models to make good decisions?

Broadly speaking, an online decision-making problem is defined by a current state and a set of actions, which together determine the next state as well as generate rewards. In Markov decision processes (MDPs), the rewards and state transitions are also affected by some random shock. Optimal policies for such problems are known only in some special cases when the underlying problem is sufficiently simple and knowledge of the generative model sufficiently detailed. For many problems of interest, an MDP approach is infeasible for two reasons: (1) insufficiently detailed models of the generative process of the randomness and (2) the complexity of computing the optimal policy (the so-called "curse of dimensionality"). These shortcomings have inspired a long line of work on approximate dynamic programming (ADP).

We focus on a general class of online resource allocation problems, which we refer to as online allocation (cf. Section 2.1) and which generalize two important classes of online decision-making problems: online packing and online matching. In brief, online allocation problems involve a set of d distinct resources and a principal with some initial budget vector $B \in \mathbb{N}^d$ of these resources, which have to be allocated among T incoming agents. Each agent has a type comprising some specific requirements for resources and associated rewards. The exact type becomes known only when the agent arrives. The principal must make irrevocable accept/reject decisions to try and maximize rewards while obeying the budget constraints.

Online packing and online matching problems are fundamental in MDP theory; they have a rich existing literature and widespread applications in many domains. Nevertheless, our work develops new policies for both these problems that admit performance guarantees that are order-wise better than existing approaches. These policies can be stated in classical ADP terms (for example, see Algorithms 2 and 3) but draw inspiration from ideas in Bayesian learning. In particular, our policies can be derived from a meta-algorithm, the Bayes selector (Algorithm 1), which makes use of a black-box *prediction oracle* to obtain statistical information about a chosen off-line benchmark and then acts on this information to make decisions. Such policies are simple to define and implement in practice, and our work provides new tools for bounding their *regret* vis-à-vis the off-line benchmark. Thus, we believe that, though our theoretical guarantees focus on a particular class of online allocation problems, our approach provides a new way for designing and analyzing much more general online decision-making policies using predictive models.

1.1. Our Contributions

We believe our contributions in this work are threefold:

1. Technical: We present a *new stochastic coupling technique*, which we call *compensated coupling*, for evaluating the regret of online decision-making policies vis-à-vis off-line benchmarks.
2. Methodological: Inspired by ideas from Bayesian learning, we propose a class of policies, expressed as the *Bayes selector*, for general online decision-making problems.
3. Algorithmic: For a wide class of problems that we refer to as online allocation (which includes online packing and online matching problems), we prove that the Bayes selector gives expected regret guarantees that are *independent of the size of the state-space*, that is, constant with respect to the horizon length and budgets.

1.1.1. Organization of the Paper.

The rest of the paper is organized as follows: In Section 2, we introduce a general problem, called *online allocation*, which includes as special cases the multisecretary, online packing, and online matching problems and also more general settings involving agents with complex valuations over bundles. We also define the prophet benchmark and discuss shortcomings of prevailing approaches. In Section 3, we present our main technical tool, compensated coupling, in the general context of finite-state, finite-horizon MDPs; we illustrate its use by applying it to the ski-rental problem. In Section 3.3, we introduce the Bayes selector policy and discuss how compensated coupling provides a generic recipe for obtaining regret bounds for such a policy. In Sections 4 and 5, we use these techniques for the online packing and online matching problems; we analyze them separately to exploit their structure and obtain stronger results. Finally, in Section 6, we analyze the most general problem (online allocation).

In particular, in Section 4, we propose a Bayes selector policy for online packing and demonstrate the following performance guarantee:

Theorem 1 (Informal)**.** *For any online packing problem with a finite number of resource types and arrival types, under mild conditions on the arrival process, the regret of the Bayes selector is independent of the horizon T and budgets B (in expectation and with high probability).*

In more detail, our regret bounds depend on the "resource matrix" A and the distribution of arriving types but are independent of T and B. Moreover, the results hold under weak assumptions on the arrival process, including multinomial and Poisson arrivals, time-dependent processes, and Markovian arrivals. This result generalizes prior and contemporaneous results (Reiman and Wang 2008, Jasin and Kumar 2012, Wu et al. 2015, Arlotto and Gurvich 2019, Bumpensanti and Wang 2019). We show similar results for online matching problems in Section 5.

1.2. Related Work

Our work is related to several active areas of research in MDPs and online algorithms.

1.2.1. Approximate Dynamic Programming.

The complexity of computing optimal MDP solutions can scale with the state space, which often makes it impractical (the so-called curse of dimensionality; Powell 2011). This has inspired a long line of work on ADP (Tsitsiklis and Van Roy 2001, Powell 2011) to develop lower complexity heuristics. Although these methods often work well in practice, they require careful choice of basis functions, and any bounds are usually in terms of quantities that are difficult to interpret. Our work provides an alternate framework, which is simpler and has interpretable guarantees.

1.2.2. Model Predictive Control.

A popular heuristic for ADP and control that is closer to our paradigm is that of *model predictive control* (MPC; Morari et al. 1993, Borrelli 2003, Ciocan and Farias 2012). MPC techniques have also been connected with online convex optimization (OCO; Chen et al. 2015, 2016; Huang 2015) to show how prediction oracles can be used for OCO and applying these policies to problems in power systems and network control. These techniques, however, require continuous controls and do not handle combinatorial constraints.

1.2.3. Information Relaxation.

Parallel to the ADP focus on developing better heuristics, there is a line of work on deriving upper bounds via martingale duality, sometimes referred to as information relaxations (Desai et al. 2012; Brown and Smith 2013, 2014). The main idea in these works is to obtain performance bounds for heuristic policies work by defining more refined outer bounds; in particular, this can be done by adding a suitable martingale term to the current reward in order to penalize "future information." Our off-line benchmarks serve a similar purpose; however, a critical difference is that, instead of using these to analyze a given heuristic, we use the benchmarks directly to derive control policies.

1.2.4. Online Packing and Prophet Inequalities. The majority of work focuses on competitive ratio bounds under worst-case distributions. In particular, there is an extensive literature on the so-called prophet inequalities, starting with the pioneering work of Hill and Kertz (1982) and including more recent extensions and applications to algorithmic economics (Kleinberg and Weinberg 2012, Alaei 2014, Correa et al. 2017, Düetting et al. 2017). We note that any competitive ratio guarantee implies a $O(T)$ expected regret in comparison with our $O(1)$ expected regret guarantees; the cost for this, however, is that our results hold under more restrictive assumptions on the inputs. For example, the policy suggested by Düetting et al. (2017) is static and Arlotto and Gurvich (2019, theorem 1) shows that any static policy has $\Omega(\sqrt{T})$ expected regret; hence, it cannot yield a strong guarantee like ours.

1.2.5. Distribution-Agnostic and Adversarial Models. Though we focus only on the case in which the input is drawn from a stochastic process, we note that there is a long line of work on online packing with adversarial inputs (Buchbinder and Naor 2009a, b; Kesselheim et al. 2018) and also distribution-agnostic approaches (Badanidiyuru et al. 2013, Nikhil et al. 2019). More generally, there is a large body of work on using sublinear expected regret algorithms for solving online linear and convex programs (Agrawal and Devanur 2014, Gupta and Molinaro 2014). The algorithms in these works are incomparable to ours in that, although they cannot get constant expected regret in our setting (stochastic input, finite type space), they provide guarantees under much weaker assumptions.

1.2.6. Regret Bounds in Stochastic Online Packing. For these problems, regret is the most meaningful metric to study, see Zhang et al. (2016) for a discussion, in which approximations to the regret are studied. The first work to prove constant expected regret in a context similar to ours is Arlotto and Gurvich (2019), who prove a similar result for the multisecretary setting with multinomial arrivals; we strengthen their result in Theorem 2. This result is relevant to a long line of work in applied probability. Some influential works are Reiman and Wang (2008), which provides an asymptotically optimal policy under the diffusion scaling, and Jasin and Kumar (2012), who provide a resolving policy with constant expected regret under a certain nondegeneracy condition. In contrast, degeneracy plays no role in the performance of our algorithms. More recently, Bumpensanti and Wang (2019) partially extended the result of Arlotto and Gurvich (2019) for more general packing problems; their guarantee is only valid for independent and identically distributed (i.i.d.) Poisson arrival processes and when the system is scaled linearly, that is, when B is proportional to T (our results and Arlotto and Gurvich (2019) make no such assumption). In Section 7, we demonstrate with a numerical study that the Bayes selector far outperforms all these previous policies.

2. Problem Setting and Overview of Results

As we mentioned in the introduction, our contributions in this work are twofold: (i) we give a technique to analyze the regret of any MDP, and (ii) we apply it to specific problems to obtain constant regret. Our focus in this work is on the subclass of *online packing problems with stochastic inputs*. This is a subclass of the wider class of *finite-horizon online decision-making problems*: given a time horizon $T \in \mathbb{N}$ with discrete time slots $t = T, T-1, \ldots, 1$, we need to make a decision at each time leading to some cumulative reward. Note that, throughout our time, index t indicates the *time to go*. We present the details of our technical approach in this more general context whenever possible, indicating additional assumptions when required.

In what follows, we use $[k]$ to indicate the set $\{1, 2, \ldots, k\}$ and denote the (i, j)th entry of any given matrix A interchangeably by $A_{i,j}$ or $A(i, j)$. We work in an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and the complement of any event $Q \subseteq \Omega$ is denoted $\bar{Q}$. For any optimization problem (P), we use $v(P)$ to indicate its objective value. If S is a finite set, $|S|$ denotes cardinality. The set $\mathbb{N}$ of naturals includes zero.

2.1. The Online Allocation Problem

We now present a generic problem, which we refer to as online allocation, that encompasses both online matching and online packing. The setup is as follows: There are d distinct resource types denoted by the set $[d]$, and at time $t = T$, we have an initial availability (budget) vector $B = (B_1, B_2, \ldots, B_d) \in \mathbb{N}^d$. At every time $t = T, T-1, \ldots, 1$, an arrival with *type* θ^t is drawn from a finite set of n distinct types $\Theta = [n]$ via some distribution, which is known to the algorithm designer (henceforth referred to as the *principal*). We denote $Z(t) = (Z_1(t), Z_2(t), \ldots, Z_n(t)) \in \mathbb{N}^n$ as the cumulative vector of the last t arrivals, where $Z_j(t) := \sum_{\tau \le t} \mathbb{1}_{\{\theta^\tau = j\}}$.

Each arriving agent is associated with a *choice over bundles*. Formally, an arriving agent of type j desires any one among a collection of multisets $S_j \subseteq 2^{[d]}$; the principal can allocate any $s \in S_j$ (referred to as a *bundle*) to the agent, thereby obtaining a reward r_{sj} and consuming one unit of each resource $i \in s$. Observe that we do not assume additive valuations, for example, we do not require $r_{\{1,2\}j} = r_{\{1\}j} + r_{\{2\}j}$. The model can naturally be extended to allow bundles to consume multiple units of any resource.

At each time, the principal must decide whether to allocate a bundle to the request θ^t (thereby generating the associated reward while consuming the required resources) or reject it (no reward and no resource consumption). Allocating a bundle requires that there is sufficient budget of each resource to cover the request. The principal's aim is to make irrevocable decisions so as to maximize overall rewards.

In Section 6, we present results for this general problem. We next describe three particular cases of the general problem, which are each of independent interest. We analyze these three cases separately because they admit improved results over the general case.

2.1.1. Multisecretary. This is a fundamental one-dimensional instance. In this problem, we have $B \in \mathbb{N}$ available positions and want to hire employees with the highest abilities (rewards). There is one resource type ($d = 1$) with budget B; each employee occupies one unit of budget (one position). This is an instance of online allocation with $S_j = \{\{1\}\}$ for all j (all candidates want the same resource) and rewards $r_{\{1\}j} = r_j$.

2.1.2. Online Packing. In this multidimensional problem, each request j is associated to one bundle and one reward if allocated. Specifically, we are given a consumption matrix $A \in \mathbb{N}^{d\times n}$, where a_{ij} denotes the units of resource i required to serve the request j and, for each j, there is a reward r_j if served. This is an instance of online allocation, wherein agents of each type j desire a single bundle: $S_j = \{\{a_{ij}$ units of i for each resource $i \in [d]\}\}$.

2.1.3. Online Matching. There are d resources, but now each type j wants *at most one resource* from among a given set of resources. Formally, a type j agent wants any from A_j instead of all from A_j. The types can be represented by a reward matrix $r \in \mathbb{R}^{d\times n}_{\geq 0}$ and adjacency matrix $A \in \{0,1\}^{d\times n}$; if the arrival is f type $j \in [n]$, we can allocate at most one resource i such that $a_{ij} = 1$, leading to a reward of r_{ij}. This problem can be thought as online bipartite matching, see Section 5 for details. It corresponds to an instance of online allocation with bundles $S_j = \{\{i\}$ for each resource $i \in [d]$ s.t. $a_{ij} = 1\}$ and rewards $r_{\{i\}j} = r_{ij}$.

2.2. Arrival Processes

To specify the generative model for the type sequence $\theta^T, \theta^{T-1}, \dots, \theta^1$, an important subclass is that of *stationary independent* arrivals, which further admits two widely studied cases:

1. The multinomial process is defined by a known distribution $p \in \mathbb{R}^n_{\geq 0}$ over $[n]$; at each time, the arrival is of type j with probability p_j, thus $Z(t) \sim \text{Multinomial}(t, p_1, \dots, p_n)$.
2. The Poisson arrival process is characterized by a known rate vector $\lambda \in \mathbb{R}^n_{\geq 0}$. Arrivals of each class are assumed to be independent such that $Z_j(t) \sim \text{Poisson}(\lambda_j t)$. Note that, although this is a continuous-time process, it can be accommodated in a discrete-time formulation by defining as many periods as arrivals (see Appendix B.1 for details).

We assume without loss of generality (w.l.o.g.) that $p_j > 0$ and $\lambda_j > 0$ for all $j \in [n]$ (if this is not the case for some j, that type never arrives and can be removed from the instance description). More general models allow for nonstationary and/or correlated arrival processes, for example, nonhomogeneous Poisson processes, Markovian models (see Example 6), etc. An important feature of our framework is that it is capable of handling a wide variety of such processes in a unified manner without requiring extensive information regarding the generative model. We discuss the most general assumptions we make on the arrival process in Section 4.4.

2.3. The Off-line Benchmark

Suppose a given problem is simultaneously solved by two "agents," ONLINE and OFF-LINE, who are primarily differentiated based on their access to information. ONLINE can only take *nonanticipatory* actions, that is, use available information only, whereas OFF-LINE is allowed to make decisions with full knowledge of future arrivals. This is known in the literature as a *prophet* or *full-information benchmark*. Denoting the total collected rewards of OFF-LINE and ONLINE as V^{off} and V^{on}, respectively, we define the regret to be the *additive* loss $\text{REG} := V^{\text{off}} - V^{\text{on}}$. Observe that V^{on} depends on the policy used by ONLINE, the underlying policy is always clear from context. Our aim is to design policies with low $\mathbb{E}[\text{REG}]$.

For online packing, the solution to OFF-LINE's problem corresponds to solving an integer programming problem. A looser but more tractable benchmark is given by an LP relaxation of this policy: given arrivals vector $Z(T)$, we assume OFF-LINE solves the following:

$$\begin{aligned} P[Z(T), B]: \quad & \max \ r'x \\ & \text{s.t.} \ Ax \leq B \\ & \qquad x \leq Z(T) \\ & \qquad x \geq 0. \end{aligned} \tag{1}$$

2.3.1. The Unavoidable Regret of the Fluid Benchmark. The most common technique for obtaining policies for online packing is based on the so-called fluid (a.k.a. deterministic or ex ante) LP benchmark $(P[\mathbb{E}[Z(T)], B])$, where (P) is defined in Equation (1). It is easy to see via Jensen's inequality that $v(P[\mathbb{E}[Z(T)], B]) \geq \mathbb{E}[v(P[Z(T), B])]$, and hence, the fluid LP is an upper bound for any online policy. Although the use of this fluid benchmark is the prevalent tool to bound the regret in online packing problems (Talluri and Van Ryzin 2006, Reiman and Wang 2008, Jasin and Kumar 2012, Wu et al. 2015), the following result shows that the approach of using $v(P[\mathbb{E}[Z(T)], B])$ as a benchmark can never lead to a constant expected regret policy as the fluid benchmark can be far off from the optimal solution in hindsight.

Proposition 1. *For any online packing problem, if the arrival process satisfies the central limit theorem and the fluid LP is dual degenerate, that is, the optimal dual variables are not unique, then $v(P[\mathbb{E}[Z(T)], B]) - \mathbb{E}[v(P[Z(T), B])] = \Omega(\sqrt{T})$.*

This gap has been reported in literature, both informally and formally (see Arlotto and Guvich 2019 and Bumpensanti and Wang 2019). For completeness, we provide a proof in Appendix A. Note though that this gap does not pose a barrier to showing constant-factor competitive ratio guarantees, that is, $O(T)$ expected regret; the fluid LP benchmark is widely used for prophet inequalities. In contrast, the gap presents a barrier for obtaining $O(1)$ expected regret bounds. Breaking this barrier, thus, requires a fundamentally new approach.

2.4. Overview of Our Approach and Results

Our approach can be viewed as a meta-algorithm that uses black-box prediction oracles to make decisions. The quantities estimated by the oracles are related to our off-line benchmark and can be interpreted as probabilities of regretting each action in hindsight. A natural Bayesian selection strategy given such estimators is to *adopt the action that is least likely to cause regret in hindsight*. This is precisely what we do in Algorithm 1, and hence, we refer to it as the Bayes selector.

Bayesian selection techniques are often used as heuristics in practice. Our work, however, shows that such policies, in fact, have excellent performance for online allocation; in particular, we show that for matching and packing problems,

1. There are easy-to-compute estimators (in particular, ones that are based on simple adaptive LP relaxations) that, when used for Algorithm 1, give constant expected regret for a wide range of distributions (see Theorems 2–4).
2. Using other types of estimators (for example, Monte Carlo estimates) in Algorithm 1 yields comparable performance guarantees (see Corollaries 4 and 5).

At the core of our analysis is a novel stochastic coupling technique for analyzing online policies based on off-line (or prophet) benchmarks. Unlike traditional approaches for regret analysis that try to show that an online policy tracks a fixed off-line policy, our approach is instead based on *forcing* Off-line *to follow* Online*'s actions*. We describe this in more detail in the next section.

3. Compensated Coupling and the Bayes Selector

We introduce our two main technical ideas: (1) the compensated coupling technique and (2) the Bayes selector heuristic for online decision making. The techniques introduced here are valid for any generic MDP; in subsequent sections, we specialize them to online allocation.

3.1. MDPs and Off-line Benchmarks

The basic MDP setup is as follows: at each time $t = T, T-1, \ldots, 1$ (where t represents the time to go), based on previous decisions, the system *state* is one of a set of possible states $\mathcal{S}$. Next, nature generates an *arrival* $\theta^t \in \Theta$, following which we need to choose from a set of available *actions* $\mathcal{A}$. The state updates and rewards are determined via a transition function $\mathcal{T}: \mathcal{A} \times \mathcal{S} \times \Theta \rightarrow \mathcal{S}$ and a reward function $\mathcal{R}: \mathcal{A} \times \mathcal{S} \times \Theta \rightarrow \mathbb{R}$: for current state $s \in \mathcal{S}$, arrival $j \in \Theta$, and action $a \in \mathcal{A}$, we transition to the state $\mathcal{T}(a, s, j)$ and collect a reward $\mathcal{R}(a, s, j)$. Infeasible actions a for a given state s correspond to $\mathcal{R}(a, s, j) = -\infty$. The sets $\mathcal{A}, \mathcal{S}, \Theta$ as well as the measure over arrival process $\{\theta^t : t \in [T]\}$ are known in advance. Finally, though we focus mainly on maximizing rewards, the formalism naturally ports over to cost minimization.

Recall that we adopt the view that the problem is simultaneously solved by two agents: Online and Off-line. Online can only take nonanticipatory actions, and Off-line makes decisions with knowledge of future arrivals. To keep the notation simple, we restrict ourselves to deterministic policies for Off-line and Online, thereby implying that the only source of randomness is due to the arrival process (our results can be extended to randomized policies).

A sample path $\omega \in \Omega$ encodes the arrival sequence $\{\theta^t : t \in [T]\}$. In other words, there exists a unique sequence of types that is consistent with ω; whenever we fix ω, the type θ^t is uniquely identified, but for notational ease, we do not write $\theta^t[\omega]$. For a given sample path $\omega \in \Omega$ and time t to go, Off-line's value function is specified via the *deterministic* Bellman equations

$$V^{\text{off}}(t, s)[\omega] := \max_{a \in \mathcal{A}} \{\mathcal{R}(a, s, \theta^t) + V^{\text{off}}(t-1, \mathcal{T}(a, s, \theta^t))[\omega]\}, \qquad (2)$$

with boundary condition $V^{\text{off}}(0,s)[\omega] = 0$ for all $s \in \mathcal{S}$. The notation $V^{\text{off}}(t,s)[\omega]$ is used to emphasize that, given sample path ω, Off-line's value function is a deterministic function of t and s.

We require that the DP formulation in Equation (2) is well defined. For simplicity, we enforce this with the following assumption: there are some constants $c_1, c_2 \geq 0$ such that $-c_1 \leq \max_{a\in\mathcal{A}} \mathcal{R}(a,s,j) \leq c_2$ for all $s \in \mathcal{S}, j \in \Theta$. In other words, every state has a feasible action, and the maximum reward is uniformly bounded and attained. The spaces $\mathcal{S}, \Theta, \mathcal{A}$ and functions $\mathcal{T}, \mathcal{R}$ are otherwise arbitrary. We enforce this assumption for clarity of exposition, but we observe that it can be further generalized (c.f. Bertsekas 1995, volume II, appendix A).

On the other hand, Online chooses actions based on *policy* π^{on} defined as follows:

Definition 1 (Online Policy)**.** An online policy π^{on} is any collection of functions $\{\pi^{\text{on}}(t,s,j) : t \in [T], s \in \mathcal{S}, j \in \Theta\}$ such that, if at time t the current state is s and a type j arrives, then Online chooses the action $\pi^{\text{on}}(t,s,j) \in \mathcal{A}$. The function $\pi^{\text{on}}(t,\cdot,\cdot)$ can depend only on $\{\theta^T, \cdots, \theta^t\}$, that is, on the randomness observed at periods $\tau \geq t$ (the history).

Let us denote $\{S^t : t \in [T]\}$ as Online's state over time, that is, the stochastic process $S^t \in \mathcal{S}$ that results from following a given policy π^{on}. We can write Online's accrued value for a given policy π^{on} as

$$V^{\text{on}}(t, S^t)[\omega] := \sum_{\tau \leq t} \mathcal{R}(\pi^{\text{on}}(\tau, S^\tau, \theta^\tau), S^\tau, \theta^\tau)[\omega].$$

For notational ease, we omit explicit indexing of V^{on} on policy π^{on}.

On any sample path ω, we can define the regret of an online policy to be the additive loss incurred by Online using π^{on} with respect to (w.r.t.) Off-line, that is,

$$\text{Reg}[\omega] := V^{\text{off}}(T, S^T)[\omega] - V^{\text{on}}(T, S^T)[\omega].$$

Remark 1 (Regret Is Agnostic of Off-line Algorithm)**.** Our previous definition of Reg depends only on the online policy π^{on}, but it does not depend on the policy (or algorithm) used by Off-line as long as it is optimal. For example, in the case in which there are multiple maximizers in the Bellman Equation (2), different tie-breaking rules for Off-line yield different algorithms, but all of them are optimal and have the same optimal value V^{off}.

3.2. The Compensated Coupling

At a high level, compensated coupling is a sample path-wise charging scheme, wherein we try to couple the trajectory of a given policy to a sequence of off-line policies. Given any nonanticipatory policy (played by Online), the technique works by making Off-line *follow* Online. Formally, we couple the actions of Off-line to those of Online while compensating Off-line to preserve its collected value along every sample path.

Example 1. Consider the multisecretary problem with budget $B = 1$ and three arriving types $\Theta = \{1,2,3\}$ with $r_1 > r_2 > r_3$. The state space in this problem is $\mathcal{S} = \mathbb{N}$, and the action space is $\mathcal{A} = \{\text{accept}, \text{reject}\}$. Suppose for $T = 4$ the arrivals on a given sample path are $(\theta^4, \theta^3, \theta^2, \theta^1) = (1,2,1,3)$. Note that Off-line accepts exactly one arrival of type 1 but is indifferent to which arrival. While analyzing Online, we have the freedom to choose a benchmark by specifying the tie-breaking rule for Off-line; for example, we can compare Online to an Off-line agent who chooses to *front-load* the decision by accepting the arrival at $t = 4$ (i.e., as early in the sequence as possible) or *back-load* it by accepting the arrival at $t = 2$. In conclusion, for this sample path, the following two sequences of actions are optimal for Off-line: (accept, reject, reject, reject) and (reject, reject, accept, reject).

Suppose instead that we choose to reject the first arrival ($t = 4$) and then want Off-line to accept the type 2 arrival at $t = 3$; this would lead to a decrease in Off-line's final reward. The crucial observation is that we can still *incentivize* Off-line to accept arrival type 2 by offering a *compensation* (i.e., additional reward) of $r_1 - r_2$ for doing so. The basic idea behind the compensated coupling is to generalize this argument. We want Off-line to take Online's action; hence, we couple the states of Off-line and Online with the use of compensations.

We start with a general problem: given sample path $\omega \in \Omega$ with arrivals $\{\theta^t[\omega] : t \in [T]\}$, recall $V^{\text{off}}(t,s)[\omega]$ denotes Off-line's value starting from state s with t periods to go. $V^{\text{off}}(t,s)[\omega]$ obeys the Bellman Equation (2). The following definition is about the actions satisfying said Bellman equations.

Definition 2 (Satisfying Action)**.** Fix $\omega \in \Omega$. For any given state s and time t, we say Off-line is *satisfied with an action* a at (s,t) if a is a maximizer in the Bellman equation, that is,

$$a \in \underset{\hat{a}\in\mathcal{A}}{\operatorname{argmax}} \left\{ \mathcal{R}(\hat{a}, s, \theta^t) + V^{\text{off}}\left(t-1, \mathcal{T}(\hat{a}, s, \theta^t)\right)[\omega] \right\}.$$

Observe that a may be satisfying for a sample path ω and not for some other ω'; once the sample path is fixed, satisfying actions are unequivocally identified.

Example 2. Consider the multisecretary problem with $T = 5$, initial budget $B = 2$, types $\Theta = \{1,2,3\}$

with $r_1 > r_2 > r_3$, and a particular sequence of arrivals $(\theta^5, \theta^4, \theta^3, \theta^2, \theta^1) = (2, 3, 1, 2, 3)$. The optimal value of Off-line is $r_1 + r_2$, and this is achieved by accepting the sole type 1 arrival as well as any one out of the two type 2 arrivals. At time $t = 5$, Off-line is satisfied either accepting or rejecting θ^5. Further, at $t = 3$, for any budget $b > 0$, the only satisfying action is to accept.

With the notion of satisfying actions, we can create a coupling as illustrated in Figure 1. Although Off-line may be satisfied with multiple actions (see preceding example and Remark 1), its value remains unchanged under any satisfying action, that is, any tie-breaking rule. We define a valid policy π^{off} for Off-line to be any *anticipatory functional* such that, for every $\omega \in \Omega$, we have a different mapping to actions. Formally, for every $\omega \in \Omega$, $\pi^{\text{off}}[\omega] : [T] \times \mathcal{S} \times \Theta \to \mathcal{A}$ is a function satisfying the optimality principle:

$$V^{\text{off}}(t,s)[\omega] = V^{\text{off}}(t-1, \mathcal{T}(\pi^{\text{off}}(t,s,\theta^t)[\omega], s, \theta^t))[\omega] + \mathcal{R}(\pi^{\text{off}}(t,s,\theta^t)[\omega], s, \theta^t), \quad \forall t \in [T], s \in \mathcal{S}.$$

Next, we quantify by how much we need to compensate Off-line when Online's action is not satisfying, as follows.

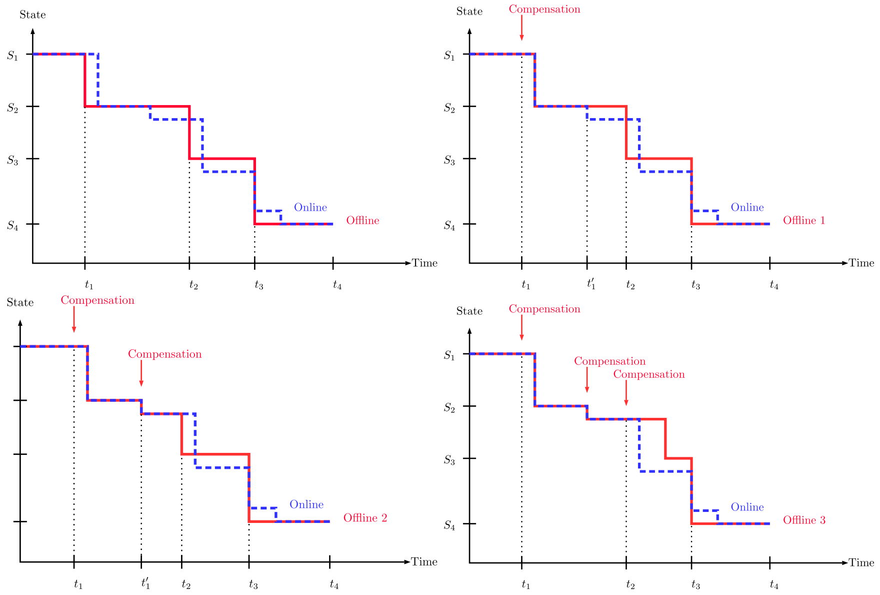
Definition 3 (Marginal Compensation). For action $a \in \mathcal{A}$, time $t \in [T]$, and state $s \in \mathcal{S}$, we denote the random variable ∂R and scalar ∂r:

$$\partial R(t,a,s) := V^{\text{off}}(t,s) - [V^{\text{off}}(t-1, \mathcal{T}(a,s,\theta^t)) + \mathcal{R}(a,s,\theta^t)]$$
$$\partial r(a,j) := \max\{\partial R(t,a,s)[\omega] : t \in [T], s \in \mathcal{S}, \omega \in \Omega \text{ s.t. } \theta^t[\omega] = j\}.$$

The random variable ∂R captures exactly how much we need to compensate Off-line, and $\partial r(a,j)$ provides a uniform (over s,t) bound on the compensation required when Online errs on an arrival of type j by choosing an action a. Though there are several ways of bounding $\partial R(t,a,s)$, we choose $\partial r(a,j)$ as it is clean and expressive and admits good bounds in many problems as the next example shows.

Example 3. For online packing problems, define $r_{\max} := \max_{j \in [n]} r_j$ as the maximum reward over all classes. The state space in this problem is $\mathcal{S} = \mathbb{N}^d$, and the actions space is $\mathcal{A} = \{\text{accept}, \text{reject}\}$. Also, for simplicity, we assume that all resource requirements are binary, that is, $a_{ij} \in \{0,1\} \, \forall \, i \in [d], j \in [n]$. For a given sample path $\omega \in \Omega$ and any given budget $b \in \mathbb{N}^d$, if Off-line decides to accept the arrival at t, we can instead

Figure 1. (Color online)



Notes. The top left image shows the traditional approach to regret analysis, wherein one considers a fixed off-line policy (which here corresponds to a fixed trajectory characterized by accept decisions at $t_1, t_2, t_3, \ldots$) and tries to bound the loss resulting from Online (dashed line) oscillating around Off-line (solid line). In contrast, the compensated coupling approach compares Online to an Off-line policy that changes over time. This leads to a sequence of off-line trajectories (top right, bottom left, and bottom right), each "agreeing" more with Online. In particular, Off-line is not satisfied with Online's action at t_1 (leading to divergent trajectories in the top left figure) but is made to follow Online by paying a compensation (top right), resulting in a new Off-line trajectory and a new disagreement at $t_1' \in (t_1, t_2)$. This coupling process is repeated at time t_1' (bottom left) and then at t_2 (bottom right), each time leading to a new future trajectory for Off-line. Coupling the two processes helps simplify the analysis as we now need to study a single trajectory (that of Online) as opposed to all potential Off-line trajectories.

make it reject the arrival while still earning a greater or equal reward by paying a compensation of $r_{\max}$. On the other hand, note that OFF-LINE can at most extract $r_{\max}$ in the future for every resource θ^t uses; hence, on sample paths on which OFF-LINE wants to reject θ^t, it can be made to accept θ^t instead with a compensation of $\|A_{\theta^t}\|_1 r_{\max} \le d r_{\max}$. In conclusion, we have $r_j \le \partial r(a,j) \le d r_{\max}$.

Recall that S^t denotes the random process of ONLINE's state. Additionally, we denote $R^{\text{on}}(t,S^t)[\omega] := \mathcal{R}(\pi^{\text{on}}(t,S^t,\theta^t),S^t,\theta^t)$ as the reward collected by ONLINE at time t, and hence, $V^{\text{on}}(T,S^T)[\omega] = \sum_{t\in[T]} R^{\text{on}}(t,S^t)[\omega]$.

The final step is to fix OFF-LINE's policy to be one that follows ONLINE as closely as possible. For this, given a policy π^{on}, on any sample path ω, we set $\pi^{\text{off}}(t,s,\theta^t)[\omega] = \pi^{\text{on}}(t,s,\theta^t)[\omega]$ if $\pi^{\text{on}}(t,s,\theta^t)[\omega]$ is satisfying and otherwise set $\pi^{\text{off}}(t,s,\theta^t)[\omega]$ to an arbitrary satisfying action. In other words, we start with any valid policy π^{off} and, for every $\omega\in\Omega$, we modify it as described to obtain another valid policy. Abusing notation, we still call π^{off} this modified policy. Recall that this modification does not change the regret guarantees (see Remark 1).

Definition 4 (Disagreement Set)**.** For any state s and time t and any action $a\in\mathcal{A}$, we define the *disagreement set* $Q(t,a,s)$ to be the set of sample paths in which a is not satisfying for OFFLINE, that is,

$$Q(t,a,s) := \{\omega\in\Omega : V^{\text{off}}(t,s)[\omega] > \mathcal{R}(a,s,\theta^t) + V^{\text{off}}(t-1,\mathcal{T}(a,s,\theta^t))[\omega]\}.$$

Finally, let $Q(t,s)\subseteq\Omega$ be the event when OFF-LINE cannot follow ONLINE, that is, $Q(t,s) := Q(t,\pi^{\text{on}}(t,s,\theta^t),s)$. Note that $Q(t,s)$ depends on π^{on}, but we omit the indexing because π^{on} is clear from context. *Only under $Q(t,s)$ do we need to compensate OFF-LINE*; hence, we obtain the following.

Lemma 1 (Compensated Coupling)**.** *For any online decision-making problem, fix any ONLINE policy π^{on} with resulting state process S^t. Then, we have*

$$\text{REG}[\omega] = \sum_{t\in[T]} \partial R(t,\pi^{\text{on}}(t,S^t,\theta^t),S^t)[\omega]\cdot \mathbb{1}_{Q(t,S^t)}[\omega],$$

and thus, $\mathbb{E}[\text{REG}] \le \max_{a\in\mathcal{A},j\in}\{\partial r(a,j)\}\cdot\sum_{t\in[T]}\mathbb{E}[\mathbb{P}[Q\times(t,S^t)|S^t]]$.

Proof. We stress that, throughout, S^t denotes ONLINE's state. We claim that, for every time t,

$$\begin{aligned} V^{\text{off}}(t,S^t)[\omega] - V^{\text{off}}(t-1,S^{t-1})[\omega] \\ = R^{\text{on}}(t,S^t)[\omega] + \partial R(t,\pi^{\text{on}}(t,S^t,\theta^t),S^t)[\omega] \\ \cdot\mathbb{1}_{Q(t,S^t)}[\omega]. \end{aligned} \tag{3}$$

To see this, let $a = \pi^{\text{on}}(t,S^t,\theta^t)$. If OFF-LINE is satisfied taking action a in state S^t, then $V^{\text{off}}(t,S^t)[\omega] - V^{\text{off}}(t-1,S^{t-1})[\omega] = R^{\text{on}}(t,S^t)[\omega]$. On the other hand, if OFF-LINE is not satisfied taking action a, then, by the definition of marginal compensation (Definition 3), we have $V^{\text{off}}(t,S^t)[\omega] - V^{\text{off}}(t-1,\mathcal{T}(a,S^t,\theta^t))[\omega] = \partial R(t,a,S^t)[\omega] + \mathcal{R}(a,S^t,\theta^t)$. Because, by definition, $\mathcal{T}(a,S^t,\theta^t) = S^{t-1}$ and $\mathcal{R}(a,S^t,\theta^t) = R^{\text{on}}(t,S^t)$, we obtain Equation (3). Finally, our first result follows by telescoping the summands and the second by linearity of expectation. □

We list a series of remarks.

- Lemma 1 is a sample path property that makes no reference to the arrival process. Though we use it primarily for analyzing MDPs, it can also be used for adversarial settings. We do not further explore this but believe it is a promising avenue.
- For stochastic arrivals, the regret depends on $\mathbb{E}[\sum_{t\in[T]}\mathbb{P}[Q(t,S^t)]]$; it follows that, if the disagreement probabilities are summable over all t, then the expected regret is constant. In Sections 4 and 5, we show how to bound $\mathbb{P}[Q(t,S^t)]$ for different problems.
- The first part of Lemma 1 provides a distributional characterization of the regret in terms of a weighted sum of Bernoulli variables. This allows us to get high-probability bounds in Section 4.5.
- Lemma 1 gives a tractable way of bounding the regret that does not require either reasoning about the past decisions of ONLINE or the complicated process OFF-LINE may follow. In particular, it suffices to bound $\mathbb{P}[Q(t,S^t)]$, that is, the probability that, given state S^t at time t, OFF-LINE loses optimality in trying to follow ONLINE.
- As mentioned before, Lemma 1 extends to the full generality of MDPs. Indeed, from Van Hentenryck and Bent (2009, chapter 11), it follows that any MDP with random transitions and random rewards can be simulated by the family of MDPs we study here; because the inputs θ^t are allowed to be random, we can define random transitions and rewards based on θ^t, see Van Hentenryck and Bent (2009) for further details.

Lemma 1, thus, gives a generic tool for obtaining regret bounds against the off-line optimum for any online policy. Note also that the compensated coupling argument generalizes to settings in which the transition and reward functions are time dependent, the policies are random, etc. Compensated coupling also suggests a natural greedy policy, which we define next.

3.2.1. Comparison with Traditional Approaches.

As discussed in related work (Section 1.2), there are two main approaches. First is the fluid (or ex ante) benchmark, which can be understood as competing against a fixed value. This is the prevailing technique in competitive analysis (Alaei 2014) and the online packing literature (Reiman and Wang 2008, Jasin and Kumar 2012, Wu et al. 2015). We showed in Proposition 1 that such an approach cannot yield better than $O(\sqrt{T})$ regret bounds, and we prove $O(1)$. Second, the traditional sample path approach, which competes against the random trajectory of Off-line (as illustrated in Figure 1), is based on showing that Online is "close" to a fixed trajectory. This approach is capable of obtaining strong $O(1)$ guarantees (Arlotto and Gurvich 2019), but it is highly involved because it necessitates a complete characterization of Off-line's trajectory. The benefit of our approach is abstracting away from the characterization of Off-line's trajectory and focusing only on Online's (which is the one that the algorithm controls) while yielding strong $O(1)$ guarantees.

The next example illustrates compensated coupling in a different setting. We consider the ski rental problem, which is a well-studied minimum cost covering (not packing) problem.

Example 4 (Ski Rental)**.** Given T days for skiing, each day, we decide whether to buy skis for b dollars or to rent them for one dollar. The snow may melt any day, and we have a distribution over the period we may be able to ski; that is, there is snow during the first $X \in [T]$ periods, and we know the distribution of X. Our aim is to explicitly write the regret of a particular policy (stated later) using compensated coupling.

The optimal off-line solution is trivial: if $X < b$, Off-line rents every day; otherwise, $(X \geq b)$ Off-line buys the first day. In other words, Off-line either buys the first day (which has cost b) or rents every day with cost X. Because Off-line knows X, Off-line picks the minimum.

We map this problem to our framework as follows. The state space is $\mathcal{S} = \{\text{skis}, \text{no-skis}\}$, where "skis" means we own the skis. Arrivals are signals $\theta^t \in \{0,1\}$, where one means there is snow and zero that the season is over. The arrival sequence is always of the form $(\theta^T,\dots,\theta^1) = (1,\dots,1,0,\dots,0)$, where $X = \sum_{t\in[T]} \theta^t$ by definition. Finally, rewards are -1 per day if we rent and $-b$ when we buy.

The compensations are as follows. If the state is skis or if $\theta^t = 0$ (season is over), then no compensation is needed because we know that Off-line does nothing with probability one (either Off-line owns the skis or the problem ended). The only case in which Online and Off-line may disagree is when $\theta^t = 1$ (can ski today) and the state is no skis.

Let us denote X^t as the number of remaining skiing days (including t), and say we observe $\theta^t = 1$ at time t. Off-line is not satisfied renting if $X^t > b$; forcing Off-line to rent in this event requires a compensation of one. On the other hand, Off-line is not satisfied buying if $X^t < b$; forcing Off-line needs a compensation of $b - X^t$. Consider the following policy π^{on}: for fixed $\tau \geq 0$, rent the first τ days and buy on day $\tau + 1$ contingent on seeing snow all these days, that is, contingent on $\theta^t = 1$ for all $t = T, \dots, T-\tau$. Compensated coupling (Lemma 1) allows us to write

$$\text{Reg} = \sum_{t\in[T]} \partial R\big(t, \pi^{\text{on}}(t, S^t, \theta^t), S^t\big)[\omega] \cdot \mathbb{1}_{Q(t,S^t)}[\omega]$$

$$= \sum_{t=T-\tau+1}^{T} \mathbb{1}_{\{X^t > b\}} + \big(b - X^{T-\tau}\big)\mathbb{1}_{\{1 \leq X^{T-\tau} < b\}}.$$

The first term corresponds to the disagreement of the first τ days (pay one dollar each day t such that $X^t > b$), whereas the second is the disagreement of day $\tau + 1$. This is an example in which compensated coupling yields an intuitive way of writing the regret. Furthermore, because the expression is exact, we can take expectations and optimize over τ to get the optimal policy (for example, see Augustine et al. 2004).

3.3. The Bayes Selector Policy

Using the formalism defined in the previous sections, let $q(t,a,s) := \mathbb{P}[Q(t,a,s)]$ be the *disagreement probability* of action a at time t in state s (i.e., the probability that a is not a satisfying action).

For any t, s, θ^t, suppose we have an *oracle* that gives us $q(t,a,s)$ for every feasible action a. Given oracle access to $q(t,a,s)$ (or, more generally, overestimates $\hat{q}$ of q), a natural greedy policy suggested by Lemma 1 is that of choosing action a that minimizes the probability of disagreement. This is similar in spirit to the Bayes selector (i.e., hard thresholding) in statistical learning. Algorithm 1 formalizes the use of this idea in online decision making. The results are essentially agnostic of how we obtain this oracle.

Algorithm 1 (Bayes Selector)

```
Input: Access to overestimates $\hat{q}(t,a,s)$ of the disagreement probabilities, that is, $\hat{q}(t,a,s) \geq q(t,a,s)$.
Output: Sequence of decisions for Online.
    Set $S^T$ as the given initial state.
    for $t = T, \dots, 1$, do
        Observe the arriving type $\theta^t$.
        Take an action minimizing disagreement, that is, $A^t \in \arg\min\{\hat{q}(t,a,S^t) : a \in \mathcal{A}\}$.
        Update state $S^{t-1} \leftarrow \mathcal{T}(A^t, S^t, \theta^t)$.
```

From Lemma 1, we immediately have the following:

Corollary 1 (Regret of Bayes Selector)**.** *Consider Algorithm 1 with overestimates $\hat{q}(t,a,s) \geq \mathbb{P}[Q(t,a,s)]\ \forall (t,a,s)$. If A^t denotes the policy's action at time t, then*

$$\mathbb{E}[\text{Reg}] \leq \max_{a\in\mathcal{A}, j\in\Theta} \partial r(a,j) \cdot \sum_{t\in[T]} \mathbb{E}\big[\hat{q}(t, A^t, S^t)\big].$$

The next result states that, if we can bound the estimation error uniformly over states and actions, then the guarantee of the algorithm increases additively on the error (not multiplicatively as one may suspect). In more detail, our next result is agnostic of the oracle used to obtain the estimators $\hat{q}$. Examples of estimation procedures to obtain $\hat{q}$ include simulation, function approximation, neural networks, etc. Regardless of how $\hat{q}$ is obtained, we can give a regret guarantee based only on the accuracy of the estimators. The following result follows as a special case of Corollary 1; we state it to emphasize that $\hat{q}$ can be estimated with some error.

Corollary 2 (Bayes Selector with Imperfect Estimators)**.** *Assume we have estimators $\hat{q}(t,a,s)$ of the probabilities $q(t,a,s)$ such that $|q(t,a,s)-\hat{q}(t,a,s)| \le \Delta^t$ for all t,a,s. If we run Algorithm 1 with overestimates $\hat{q}(t,a,s)+\Delta^t$ and A^t denotes the policy's action at time t, then*

$$\mathbb{E}[\text{Reg}] \le \max_{a\in\mathcal{A}, j\in\Theta} \partial r(a,j) \cdot \sum_{t\in[T]} \left(\mathbb{E}\left[\hat{q}\left(t, A^t, S^t\right)\right] + \Delta^t\right).$$

Observe that the total error induced as a result of estimation is a constant if, for example, we can guarantee $\Delta^t = 1/t^2$ or $\Delta^t = 1/(T-t)^2$.

It is natural to consider a more sophisticated version of Algorithm 1, wherein we make decisions not only based on disagreement probabilities, but also take into account marginal compensations, that is, the marginal loss of each decision. Although Algorithm 1 is enough to obtain constant regret bounds in the problems we consider, we note that such an extension is possible and can be found in Appendix B.2.

Remark 2. We discussed in Section 3 that compensated coupling extends to the case in which transitions and rewards can be random. By the same argument, the Bayes selector (Algorithm 1) and its guarantees (Corollaries 1 and 2) extend too. Notice that Off-line here is a prophet who has full knowledge of all the randomness (arrivals, transitions, and rewards).

4. Regret Guarantees for Online Packing

We now show that, for the online packing problem, the Bayes selector achieves an expected regret that is *independent of the number of arrivals T and the initial budgets B*; in Section 5, we extend this to online matching problems.

In more detail, we prove that the *dynamic fluid relaxation* (P_t) in Equation (4) provides a good estimator for the disagreement probabilities $q(t,a,s)$ and, moreover, that the Bayes selector based on these statistics reduces to a simple *resolve and threshold* policy.

In this setting, the state space corresponds to resource availability; hence, $\mathcal{S} = \mathbb{N}^d$. There are two possible actions, accept or reject; hence, $|\mathcal{A}| = 2$. Finally, transitions correspond to the natural budget reductions given by the matrix A.

Recall that $Z(t) \in \mathbb{N}^n$ denotes the cumulative arrivals in the last t periods, and $B^t \in \mathbb{N}^d$ denotes Online's budget at time t. Given knowledge of $Z(t)$ and state B^t, we define the ex post relaxation (P_t^*) and fluid relaxation (P_t) as follows.

$$\begin{aligned}
(P_t^*)\ \max\ & r'x \\
\text{s.t.}\ & Ax \le B^t \\
& x \le Z(t) \\
& x \ge 0. \\
(P_t)\ \max\ & r'x \\
\text{s.t.}\ & Ax \le B^t \\
& x \le \mathbb{E}[Z(t)] \\
& x \ge 0.
\end{aligned} \tag{4}$$

Remark 3. Off-line solves (P_t^*) in Equation (4), and Online solves (P_t). Both problems depend on Online's budget at t; this is a crucial technical point and can only be accomplished because of the coupling we have developed.

Let X^t be a solution of (P_t) and $X^{\star t}$ a solution of (P_t^*). Uniqueness of solutions is not required—see Proposition 2—and $X^{\star t}$ is for the analysis only. Our policy is detailed in Algorithms 2.

Algorithm 2 (Fluid Bayes Selector)

```
Input: Access to solutions $X^t$ of $(P_t)$ and resource matrix $A$.
Output: Sequence of decisions for Online.
Set $B^T$ as the given initial budget levels.
for $t = T, \dots, 1$, do
    Observe arrival $\theta^t = j$ and accept iff $X_j^t \ge \mathbb{E}[Z_j(t)]/2$ and it is feasible, that is, $A_j \le B^t$.
    Update $B^{t-1} \leftarrow B^t - A_j$ if accept and $B^{t-1} \leftarrow B^t$ if reject.
```

Intuitively, we front-load (accept as early as possible) classes j such that $X_j^t \ge \mathbb{E}[Z_j(t)]/2$ and back-load the rest (delay as much as possible). If Off-line is satisfied accepting a front-loaded class (respectively, rejecting a back-loaded class), Off-line does so. Accepting class j is, therefore, an error if Off-line, given the same budget as Online, picks no future arrivals of that class (i.e., $X_j^{\star t} < 1$). On the other hand, rejecting j is an error if $X_j^{\star t} > Z_j(t) - 1$. We summarize this as follows:

1. Incorrect rejection: if $X_j^t < \frac{\mathbb{E}[Z_j(t)]}{2}$ and $X_j^{\star t} > Z_j(t) - 1$.
2. Incorrect acceptance: if $X_j^t \ge \frac{\mathbb{E}[Z_j(t)]}{2}$ and $X_j^{\star t} < 1$.

Observe that compensation is paid only when the fluid solution is far off from the correct stochastic solution. In the proofs of Theorems 2 and 3, we formalize the fact that, because X^t estimates $X^{\star t}$, such an event is highly unlikely; this, along with compensated coupling, provides our desired regret guarantees.

Disagreement Probabilities and the Fluid Bayes Selector: The Bayes selector (Algorithm 1) runs with overestimates $\hat{q}(t, a, b)$ and, at each time, picks the minimum. On the other hand, Algorithms 2 is presented as the "simplified version" in the sense that the decision rule is the one that minimizes suitable overestimates $\hat{q}$. More importantly, we prove the following properties:

$$\operatorname{argmin}\{\hat{q}_j(t, \text{accept}, b), \hat{q}_j(t, \text{reject}, b)\} = \begin{cases} \text{accept} & \text{if } X_j^t \ge \mathbb{E}[Z_j(t)]/2 \\ \text{reject} & \text{if } X_j^t < \mathbb{E}[Z_j(t)]/2 \end{cases}, \quad (5)$$

$$\min\{\hat{q}_j(t, \text{accept}, b), \hat{q}_j(t, \text{reject}, b)\} \le c_1 e^{-c_2 t}. \quad (6)$$

In other words, the property in Equation (5) shows that Algorithms 2 is a Bayes selector, and the property in Equation (6) yields the desired constant regret bound in virtue of compensated coupling and Corollary 1. We give explicit expressions for the values $\hat{q}$ and constants c_1, c_2; see, for example, Equation (B.2).

The Robustness of the Bayes Selector: The probability minimizing disagreement can be uniformly bounded over all budgets $b \in \mathbb{N}^d$; that is, the exponential bound in Equation (6) does not depend on b. This property has the following consequence: because the fluid Bayes selector has strong performance, many other Bayes selector algorithms (using different $\hat{q}$) do too. In other words, the design of algorithms based on the Bayes selector is robust and does not depend on fine-tuning of the parameters $\hat{q}$. We make this precise in Corollaries 3 and 4 and uncover the same phenomenon for matching problems; see Corollary 5.

We need some additional notation before presenting our results. Let $\mathbb{E}_j[\cdot]$ ($\mathbb{P}_j[\cdot]$) be the expectation (probability) conditioned on the arrival at time t being of type j, that is, $\mathbb{P}_j[\cdot] = \mathbb{P}[\cdot | \theta^t = j]$. We denote $r_{\max} := \max_{j \in [n]} r_j$ and $p_{\min} := \min_{j \in [n]} p_j$.

4.1. Special Case: Multisecretary with Multinomial Arrivals

Before we proceed to the general case, we state the result for the multisecretary problem. We present this result separately because, in this one-dimensional problem, we can obtain a better and explicit constant. The proofs of Theorem 2 and Corollary 3 can be found in Appendix B.3.

Theorem 2. *The expected regret of the fluid Bayes selector (Algorithm 2) for the multisecretary problem with multinomial arrivals is at most* $r_{\max} \sum_{j>1} 2/p_j \le 2(n-1) r_{\max}/p_{\min}$.

This recovers the best-known expected regret bound for this problem shown in a recent work (Arlotto and Gurvich 2019). However, although the result in Arlotto and Gurvich (2019) depends on a complex martingale argument, our proof is much more succinct and provides explicit and stronger guarantees; in particular, in Section 4.5, we provide concentration bounds for the regret.

Moreover, Theorem 2, along with Corollary 2, provides a critical intermediate step for characterizing the performance of Algorithm 1 for the multisecretary problem.

Corollary 3. *For the multisecretary problem with multinomial arrivals, the expected regret of the Bayes selector (Algorithm 1) with any imperfect estimators $\hat{q}$ is at most $2r_{\max}(\sum_{j>1} 1/p_j + \sum_{t \in [T]} \Delta^t)$, where Δ^t is the accuracy defined by $|q(t, a, b) - \hat{q}(t, a, b)| \le \Delta^t$ for all $t \in [T], a \in \mathcal{A}, b \in \mathbb{N}$.*

Observe that, if Δ^t is summable, for example, $\Delta^t = 1/t^2$ or $\Delta^t = 1/(T-t)^2$, then Corollary 3 implies constant expected regret for all these types of estimators we can use in Algorithm 1.

4.2. Online Packing with General Arrivals

We consider now the case $d > 1$ and arrival processes other than multinomial. We assume the following condition on the process $Z(t)$, which we refer to as *all-time deviation*.

Definition 5 (All-Time Deviation). Let μ be a given norm in $\mathbb{R}^n$ and $\kappa \in \mathbb{R}^n_{\ge 0}$ a constant parameter. An n-dimensional process $Z(t)$ satisfies the all-time deviation bound w.r.t. μ and κ if, for all $j \in [n]$, there are constants $c_j \ge 0$ and naturals τ_j such that

$$\mathbb{P}\left[\|Z(t) - \mathbb{E}(Z(t))\|_\mu \ge \frac{\mathbb{E}[Z_j(t)]}{2\kappa_j}\right] \le \frac{c_j}{t^2} \quad \forall t > \tau_j. \quad (7)$$

We remark that we do not need exponential tails as it is common to assume, but rather a simple quadratic tail. Additionally, some common tail bounds are valid only for large enough samples; the parameters τ_j capture this technical aspect. In this section, we use the definition with κ_j the same entry for all j, thus denoted simply by $\kappa > 0$. In Section 5, we require the definition with the more general form.

Example 5 (Multinomial and Poisson Tails). In these examples, we actually have the stronger exponential tails, so we do not elaborate on the constant c_j.

For multinomial arrivals, Devroye (1983, lemma 3) guarantees

$$\mathbb{P}[\|Z(t) - \mathbb{E}(Z(t))\|_1 > t\varepsilon] \le e^{-t\varepsilon^2/25}, \forall 0 < \varepsilon < 1, t \ge \frac{\varepsilon^2 n}{20}. \quad (8)$$

By setting $\varepsilon = p_j/2\kappa$, we conclude that Definition 5 is satisfied with constants $\tau_j = (p_j/2\kappa)^2 n/20$.

For Poisson arrivals, from the proof of Devroye (1983, lemma 3), $\mathbb{P}(|X-\lambda| \geq \varepsilon\lambda) \leq 2e^{-\lambda\varepsilon^2/4}$ is valid for $X \sim \text{Poisson}(\lambda)$ and any $\varepsilon > 0$. Using this, we can simply take $\tau_j = 0$.

In the remainder of this section, we generalize our ideas to prove the following.

Theorem 3. *Assume the arrival process* $(Z(t) : t \in [T])$ *satisfies the conditions in Equation* (7). *The expected regret of the fluid Bayes selector (Algorithm* 2*) for online packing is at most* $dr_{\max}M$, *where* M *is independent of* T *and* B. *Specifically, for* $\kappa = \kappa(A)$, *we have*

1. *For multinomial arrivals:* $M \leq 103\kappa^2 \sum_{j\in[n]} 1/p_j$.
2. *For general distributions satisfying Equation* (7): $M \leq \sum_{j\in[n]} p_j(2c_j + \max\{\tau_j, \tilde{\tau}_j\})$, *where* p_j *is an upper bound on* $\mathbb{P}[\theta^t = j]$ *and* $\tilde{\tau}_j$ *is such that* $\mathbb{E}[Z_j(\tilde{\tau}_j)] \geq 2$, *that is, it is large enough.*

The constant $\kappa(A)$ is given by Proposition 2. Just as before, Theorem 3, along with Corollary 2, provides a performance guarantee for Algorithm 1. We state the corollary without proof because it is identical to that of Corollary 3.

Corollary 4. *For the online packing problem, if the arrival process satisfies the conditions in Equation* (7), *the expected regret of the Bayes selector (Algorithm* 1*) with any imperfect estimators* $\hat{q}$ *is at most* $dr_{\max}(M + 2\sum_{t\in[T]} \Delta^t)$, *where* M *is as in Theorem* 3 *and* Δ^t *is the accuracy defined by* $|q(t,a,b) - \hat{q}(t,a,b)| \leq \Delta^t$ *for all* $t \in [T], a \in \mathcal{A}, b \in \mathbb{N}^d$.

To prove Theorem 3, we need to quantify how the change in the right-hand side of an LP impacts optimal solutions. Indeed, as stated in Equation (4), the solutions X^t and $X^{\star t}$ correspond to perturbed right-hand sides ($\mathbb{E}[Z(t)]$ and $Z(t)$, respectively). The following proposition implies that small changes in the arrivals vector do not change the solution by much, and it is based on a more general result from Mangasarian and Shiau (1987, theorem 2.4).

Proposition 2 (LP Lipschitz Property)**.** *Given* $b \in \mathbb{R}^d$ *and any norm* $\|\cdot\|_\mu$ *in* $\mathbb{R}^n$, *consider the following LP:*

$$P(y) \quad \max\{r'x : Ax \leq b, 0 \leq x \leq y, y \in \mathbb{R}^n_{\geq 0}\}.$$

Then, $\exists$ *constant* $\kappa = \kappa_\mu(A)$ *such that, for any* $y, \hat{y} \in \mathbb{R}^n_{\geq 0}$ *and any solution* x *to* $P(y)$, *there exists a solution* $\hat{x}$ *solving* $P(\hat{y})$ *such that* $\|x - \hat{x}\|_\infty \leq \kappa\|y - \hat{y}\|_\mu$.

Proof of Theorem 3. Recall the two conditions derived from our decision rule: (1) Incorrect rejection of j means $X_j^t < \mathbb{E}[Z_j(t)]/2$ and $X_j^{\star t} > Z_j(t) - 1$. (2) Incorrect acceptance of j means $X_j^t \geq \mathbb{E}[Z_j(t)]/2$ and $X_j^{\star t} < 1$. We have to additionally account for feasibility; that is, we can only accept a request j if $B_i^t \geq a_{ij}$ for all $i \in [d]$. In case there are not enough resources, our decision rule is feasible if either $X_j^t < \mathbb{E}[Z_j(t)]/2$ (reject) or $X_j^t \geq 1$ (because X^t is feasible for (P_t)). Only in the case $X_j^t \geq \mathbb{E}[Z_j(t)]/2$ and $X_j^t < 1$ do we need to disregard our decision rule and are forced to reject; under such a condition, we must pay a compensation of r_j. Observe that this condition is never met if $\mathbb{E}[Z_j(t)] \geq 2$; that is, it is vacuous for $t \geq \tilde{\tau}_j$.

The disagreement sets (Definition 4) are, thus, $Q(t,b) = \{\omega \in \Omega : \text{either } (1), (2) \text{ or } t < \tilde{\tau}_j\}$, where (1) and (2) are the previous conditions. Now, we can upper bound the probability of paying a compensation as follows. Call E_j the event $\{\omega \in \Omega : \|Z(t) - \mathbb{E}(Z(t))\|_1 \leq \mathbb{E}[Z_j(t)]/2\kappa\}$. In this event, Proposition 2 implies $|X_j^t - X_j^{\star t}| \leq \mathbb{E}[Z_j(t)]/2$; hence, conditions (1) and (2) do not happen when E_j occurs, that is, $\mathbb{P}_j[Q(t,b)|E_j] \leq_{\{t<\tilde{\tau}_j\}}$. Observe that $\mathbb{P}[\bar{E}_j] \leq f_j(t) +_{\{t<\tau_j\}}$, where $f_j(t) = c_j/t^2$ for general processes satisfying Equation (7) and $f_j(t) = e^{-t(p_j/2\kappa)^2/25}$ for the multinomial process (see Equation (8)). Finally,

$$\begin{aligned} q_j(t,B^t) \leq \mathbb{P}[\bar{E}_j] + \mathbb{P}_j[Q(t,B^t)|E_j] \leq \mathbb{P}[\bar{E}_j] \\ + \mathbb{1}_{\{t<\tilde{\tau}_j\}} \leq f_j(t) + \mathbb{1}_{\{t<\tau_j \text{ or } t<\tilde{\tau}_j\}}. \end{aligned} \tag{9}$$

Summing up over time, we get

$$\sum_{t\in[T]} q(t,B^t) \leq \sum_{j\in[n]} p_j \left(\sum_{t\in[T]} f_j(t) + \max\{\tau_j, \tilde{\tau}_j\} \right).$$

Because $\sum_{t\in[T]} 1/t^2 \leq \pi^2/6 \leq 2$, this finishes the proof for general processes. For the case of multinomial arrivals, we can be more refined. Indeed, $\tilde{\tau}_j$ is defined by $\mathbb{E}[Z_j(\tilde{\tau}_j)] \geq 2$, that is, $\tilde{\tau}_j \geq 2/p_j$ and $\tau_j = (p_j/2\kappa)^2 n/20$ (see Equation (8)). From the previous equation, with the stronger exponential bound $f_j(t) = e^{-t(p_j/2\kappa)^2/25}$, we get

$$\begin{aligned} &\sum_{t\in[T]} q(t,B^t) \\ &\leq \sum_{j\in[n]} p_j \left(\frac{25}{(p_j/2\kappa)^2} + \max\left\{(p_j/2\kappa)^2 n/20, 2/p_j\right\} \right) \\ &\leq 100\kappa^2 \sum_{j\in[n]} \frac{1}{p_j} + 3n. \end{aligned}$$

Because $n \leq \sum_{j\in[n]} \frac{1}{p_j}$, we arrive at the desired bound. The result follows via compensated coupling (Lemma 1) and Corollary 1.

Remark 4. In the multisecretary problem, it is easy to conclude $\kappa(A) = 1$; thus, this analysis recovers the same bound up to absolute constants (namely 103 versus 2). The larger constant comes exclusively from the larger constants in the tail bounds of multinomial compared with binomial random variables (r.v.).

Remark 5. More refined bounds on M can be obtained by not bounding $\mathbb{P}[\theta^t = j] \leq p_j$, but rather by $\mathbb{P}[\theta^t = j] \leq p_j(t)$. For example, a time-varying version of a multinomial process easily fits in our framework, and the proof does not change.

Remark 6. The theorem holds even under Markovian correlations (see Example 6), in which the distribution of $Z(t-1)$ depends on θ^t. It is interesting that, in this case, it is impossible to run the optimal policy for even moderate instance sizes because the state space is huge although the Bayes selector still offers bounded expected regret.

We now give two examples of other arrival processes that satisfy the all-time deviation (Definition 5). The proofs of the bounds are short, but we relegate them to Appendix B.4. We emphasize that Example 7 has quadratic tails (instead of exponential); hence, we term it heavy tailed.

Example 6 (Markovian Arrival Processes). We consider the case in which θ^t is drawn from an ergodic Markov chain. Let $P \in \mathbb{R}^{n\times n}_{\geq 0}$ be the corresponding matrix of transition probabilities. The process unfolds as follows: at time $t = T$, an arrival $\theta^T \in [n]$ is drawn according to an arbitrary distribution; then, for $t = T, \ldots, 2$, we have $\mathbb{P}[\theta^{t-1} = j|\theta^t] = P_{\theta^t j}$. Let $\nu \in \mathbb{R}^n_{\geq 0}$ be the stationary distribution. *We do not require long-run or other usual stationary assumptions; the process is still over a finite horizon* T. This process satisfies all-time deviation with exponential tails. Specifically, with the norm $\mu = \|\cdot\|_\infty$, for some constants c_j, c' that depend on P only, we have

$$\mathbb{P}\big[\|Z(t) - \nu t\|_\infty \geq \nu_j t/2\kappa_j\big] \leq nc' e^{-c_j t}, \quad \forall t \in [T], j \in [n]. \tag{10}$$

Example 7 (Heavy Tailed Poisson Arrivals). We consider the case in which the arrival process is governed by independent time-varying Poisson processes with arrival rates $\lambda_j(t) > 0$, which we assume for simplicity have finitely many discontinuity points (so that all the expectations are well defined). Under the following conditions, the process satisfies the all-time deviation with *quadratic tails* and norm $\mu = \|\cdot\|_\infty$.

$$\max_{j,k\in[n]} \max_{s\in[0,t]} \frac{\lambda_j(s)}{\lambda_k(s)} \leq g(t) \quad \forall t \geq 0, \tag{11}$$

$$\min_{j\in[n]} \min_{s\in[0,t]} \lambda_j(s) \geq g(t) f(t) \frac{\log(t)}{t}, \quad \text{where} \quad \lim_{t\to\infty} f(t) = \infty. \tag{12}$$

In other words, we require $f(t) = \omega(1)$, and $g(t)$ is any function. Intuitively, Equation (11) guarantees that no type j "overwhelms" all other types; observe that, when the rates are constant, this is trivially satisfied with $g(t)$ constant. On the other hand, Equation (12) controls the minimum arrival rate, which can be as small as $\omega(\log(t)/t)$. Observe that our conditions allow for the intensity to increase closer to the end ($t = 0$); that is, we incorporate the case in which agents are more likely to arrive closer to the deadline.

4.3. High-Probability Regret Bounds

We have proved that $\mathbb{E}[\text{Reg}]$ is constant for packing problems. One may worry that this is not enough because, as it is a random variable, *Reg* may still realize to a large value. We present a bound for the distribution of *Reg* showing that it has light tails.

Proposition 3. *For packing problems, there are constants* τ *and* c_j *for* $j \in [n]$, *depending on* A, p *and the distribution of* Z *only such that*

1. *For multinomial or Poisson arrivals:* $\forall x > \tau$, $\mathbb{P}[\text{Reg} > x] \leq \sum_j p_j e^{-c_j x/r_{\max}}/c_j$.
2. *For general distributions satisfying Equation* (7)*:* $\forall x > \tau$, $\mathbb{P}[\text{Reg} > x] \leq \frac{r_{\max}}{x}\sum_j p_j c_j$.

The proof is based on the following simple lemma. The idea is to first bound the disagreements of our algorithm as defined in Section 3.3. The total number of disagreements is a sum of dependent Bernoulli variables, which we bound next.

Lemma 2. *Let* $\{X^t : t \in [T]\}$ *be a sequence of dependent r.v. such that* $X^t \sim \text{Bernoulli}(p_t)$, *and let* $\{q_t : t \in [T]\}$ *be numbers such that* $q_t \geq p_t$. *If we define* $D := \sum_{t=1}^T X^t$, *then*

$$\mathbb{P}[D \geq d] \leq \sum_{t=d}^T q_t.$$

Proof. Fix $d \in [T]$ and observe that

$$\{\omega \in \Omega : D \geq d\} \subseteq \{\omega \in \Omega : \exists t \geq d, X^t = 1\}.$$

Indeed, if the condition $(\exists t \geq d, X^t = 1)$ fails, then at most $d - 1$ variables X^t can be one.

Finally, a union bound shows $\mathbb{P}[D \geq d] \leq \sum_{t\geq d} \times \mathbb{P}[X^t = 1]$. Because $q_t \geq p_t$, the proof is complete. □

Proof of Proposition 3. As described in the previous sections, we can write $\text{Reg} \leq r_{\max} D$ with D as the number of disagreements. Additionally, D is a sum of T Bernoulli r.v. X^t, each with a parameter bounded by q_t.

In the case of multinomial and Poisson r.v., as described in Section 4.4, we have exponential bounds $q_t \leq \sum_{j\in[n]} p_j e^{-c_j t}$ for $t \geq \tau = \max_{j\in[n]} \tau_j$. We conclude invoking Lemma 2 and upper bounding $\sum_{t=x+1}^T e^{-c_j t} \leq e^{-c_j x}/c_j$.

For general distributions, as described in Section 4, we have the bounds $q_t \leq \sum_{j\in[n]} p_j \frac{c_j}{t^2}$ for $t \geq \tau$. Using Lemma 2 and bounding $\sum_{t=x+1}^T t^{-2} \leq 1/x$ finishes the proof. □

5. Regret Guarantees for Online Matching

We turn to an alternate setting, in which each incoming arrival corresponds to a *unit-demand* buyer. In other words, each arrival wants a unit of a single resource but has different valuations for different resources. This is essentially equivalent to the online bipartite matching problem with edge weights (weights correspond to rewards) in which there can be multiple copies of each node.

As before, we are given a matrix $A \in \{0,1\}^{d\times n}$ characterizing the demand for resources, which can be interpreted as the adjacency matrix in the online matching problem. Define $S_j := \{i \in [d] : a_{ij} = 1\}$. If we allocate any resource $i \in S_j$ to an agent type j, we obtain a reward of r_{ij}, whereas allocating $i \notin S_j$ has no reward. We can allocate at most one item to each agent.

Given resource availability $B \in \mathbb{N}^d$ and total arrivals $Z \in \mathbb{N}^n$, we can formulate Off-line's problem as follows, in which the variable x_{ij} denotes the number of items i allocated to agents of type j.

$$\begin{aligned}(P[Z,B])\quad \max\ & \sum_{i,j} x_{ij} r_{ij} a_{ij}\\ \text{s.t.}\ & \sum_j x_{ij} \le B_i\ \forall i \in [d]\\ & \sum_{i\in[d]} x_{ij} \le Z_j\ \forall j \in [n]\\ & \mathbf{x} \ge 0. \end{aligned} \tag{13}$$

We assume that the process $Z(t)$ satisfies the all-time deviation bound (see Definition 5) w.r.t. the one norm and parameters $\kappa_j = (|S_j|+1)/2$. This condition can be restated as follows. For every $j \in [n]$, there are constants $c_j \ge 0$ and naturals τ_j such that

$$\mathbb{P}\left[\|Z(t) - \mathbb{E}(Z(t))\|_1 \ge \frac{\mathbb{E}[Z_j(t)]}{|S_j|+1}\right] \le \frac{c_j}{t^2} \quad \forall t > \tau_j. \tag{14}$$

We now state the main result of this section, which is based on an instantiation of the Bayes selector. As before, the theorem readily implies performance guarantees for Algorithm 1, which we state without proof because it is identical to that of Corollary 3.

Theorem 4. *For the online matching problem, if the arrival process satisfies the conditions in Equation (14), then the expected regret of the fluid Bayes selector (Algorithm 3) is at most $r_{\max}\sum_{j\in[n]} p_j(c_j+\tau_j)$, where p_j is an upper bound on $\mathbb{P}[\theta^t = j]$.*

Corollary 5. *For the online matching problem, if the arrival process satisfies the conditions in Equation (14), then the expected regret of the Bayes selector (Algorithm 1) with any imperfect estimators $\hat{q}$ is at most $r_{\max}(M + 2\sum_{t\in[T]}\Delta^t)$. The constant $M = \sum_{j\in[n]} p_j(c_j+\tau_j)$ is as in Theorem 4, and Δ^t is the accuracy defined by $|q(t,a,s) - \hat{q}(t,a,s)| \le \Delta^t$.*

5.1. Algorithm and Analysis

We start from the LP in Equation (13) and then add a fictitious item $d+1$ that no agent wants with initial budget $B^T_{d+1} = T$; now, all agents are matched, but if we match an agent to $d+1$, there is no reward. Using compensated coupling, we can write two coupled optimization problems: (P^*_t) for Off-line and (P_t) for Online as follows.

$$\begin{aligned}(P^*_t)\ \max\ & \sum_{i\in[d],j\in[n]} x_{ij} r_{ij} a_{ij}\\ \text{s.t.}\ & \sum_{j\in[n]} x_{ij} \le B^t_i \qquad \forall i \in [d+1]\\ & \sum_{i\in[d+1]} x_{ij} = Z_j(t)\ \ \forall j \in [n]\\ & \mathbf{x} \ge 0.\\ (P_t)\ \max\ & \sum_{i\in[d],j\in[n]} x_{ij} r_{ij} a_{ij}\\ \text{s.t.}\ & \sum_{j\in[n]} x_{ij} \le B^t_i \qquad\qquad \forall i \in [d+1]\\ & \sum_{i\in[d+1]} x_{ij} = \mathbb{E}[Z_j(t)]\quad \forall j \in [n]\\ & \mathbf{x} \ge 0. \end{aligned} \tag{15}$$

Recall that B^t represents Online's budget with t periods to go. We solve (P_t) in Equation (15) and obtain an optimizer X^t. If $\theta^t = j$, let $K \in \operatorname{argmax} \times \{X^t_{i,j} : i \in [d+1]\}$ be the maximal entry, breaking ties arbitrarily and then match j to K. The resulting policy is presented in Algorithm 3. Observe that matching an agent to $K = d+1$ (fictitious resource) is equivalent to rejecting it.

Algorithm 3 (Fluid Bayes Selector for Online Matching)

```
Input: Access to solutions $X^t$ of $(P_t)$ in Equation (15)
Output: Sequence of decisions for Online.
  Set $B^T$ as the given initial budget levels.
  for $t = T, \dots, 1$, do
      Observe arrival $\theta^t = j$ and let
          $K \leftarrow \operatorname{argmax}\{X^t_{ij} : i \in [d+1]\}$, breaking
          ties arbitrarily.
      Match $\theta^t$ to $K$.
      Update $B^{t-1}_i \leftarrow B^t_i$ for $i \ne K$
          and $B^{t-1}_K \leftarrow B^t_K - 1$.
```

5.1.1. Disagreement Sets.

At each time t, matching $\theta^t = j$ to K requires a compensation only if Off-line never matches a type j to K, that is, $X^{\star t}_{K,j} < 1$. On the other hand, Algorithm 3 picks K to be the largest component; hence, we should have $X^t_{K,j} >> 1$ (precisely, $X^t_{K,j} \ge \frac{\mathbb{E}[Z_j(t)]}{d+1}$). More formally, the constraint $\sum_{i\in[d+1]} x_{ij} = \mathbb{E}[Z_j(t)]$ in Equation (15) and the definition of S_j imply $X^t_{K,j} \ge \mathbb{E}[Z_j(t)]/(|S_j|+1)$. We conclude that, if matching to K is not satisfying (see Definition 2), it must be that $\|X^t - X^{\star t}\|_\infty > \mathbb{E}[Z_j(t)]/(|S_j|+1)$. Proposition 4 characterizes exactly this deviation.

Observe that, in Equations (13) and (15), the matrix A appears only on the objective function; this is not the usual LP formulation for this problem, but it allows us to obtain the following result. We remark that not only do we have a Lipschitz property, but the Lipschitz constant is exactly one. We present the proof of Proposition 4 in Appendix B.5.

Proposition 4 (Lipschitz Property for Matching)**.** *Take any z^1, $z^2 \in \mathbb{R}^d_{\geq 0}$ and $b \in \mathbb{R}^d_{\geq 0}$. If $\mathbf{x}^1$ is a solution of $P[z^1, b]$, then there exists $\mathbf{x}^2$ solving $P[z^2, b]$ such that $\|\mathbf{x}^1 - \mathbf{x}^2\|_\infty \leq \|z^1 - z^2\|_1$.*

From here, the proof of Theorem 4 is applying compensated coupling (Lemma 1) and Corollary 1 in the same way as we did in Section 4; hence, we omit it.

5.2. Online Stochastic Matching

A classical problem that fits naturally into this framework is the online bipartite matching problem with stochastic inputs (Manshadi et al. 2012). The reader unfamiliar with the problem can find the details of the setup in Appendix B.6. For this setting, the bound obtained via compensated coupling surprisingly holds with equality:

Lemma 3. *For stochastic online bipartite matching, given an online policy, if U^t denotes the node matched at time t by* Online *and S^t the available nodes, then*

$$V^{\text{off}} - V^{\text{on}} = \sum_{t \in [T]} \mathbb{1}_{Q(t, U^t, S^t)}.$$

Based on this, it is tempting to conjecture that the Bayes selector does, in fact, lead to an optimal policy for this setting. This, however, is not the case although showing this is surprisingly subtle; in Appendix B.6, we discuss this in more detail. Moreover, it is known that this problem cannot admit an expected regret that has better than linear scaling with T (in particular, Manshadi et al. (2012) proves a constant upper bound on the competitive ratio for this setting). That said, the strength of the above bound suggests that the Bayes selector may have strong approximation guarantees; showing this remains an open problem.

6. Regret Guarantees for Online Allocation

We now give the algorithm and analysis for the general online allocation problem defined in Section 2.1. As before, let us introduce a fictitious resource $i = d+1$ with initial capacity $B_{d+1} = T$, zero rewards ($r_{\{d+1\}j} = 0$ for all $j \in [n]$), and such that $\{d+1\} \in S_j$ for all $j \in [n]$. Now we can assume w.l.o.g. that each agent gets assigned a bundle. Finally, for a bundle s, we denote $a_{is} \in \mathbb{N}$ the number of times the resource i appears in s (recall that bundles are multisets).

Given resource availability $B \in \mathbb{N}^{d+1}$ and total arrivals $Z \in \mathbb{N}^n$, we can formulate the coupled problems for Off-line and Online as follows, where the variable x_{sj} denotes the number of times a bundle $s \in S_j$ is allocated to a type j.

$$\begin{aligned}
(P_t^*)\ \max \ & \sum_{j \in [n], s \in S_j} x_{sj} r_{sj} \\
\text{s.t.} \ & \sum_{j \in [n], s \in S_j} a_{is} x_{sj} \leq B_i^t \qquad \forall i \in [d+1] \\
& \sum_{s \in S_j} x_{sj} = Z_j(t) \quad \forall j \in [n] \\
& \mathbf{x} \geq 0. \\
(P_t)\ \max \ & \sum_{j \in [n], s \in S_j} x_{sj} r_{sj} \\
\text{s.t.} \ & \sum_{j \in [n], s \in S_j} a_{is} x_{sj} \leq B_i^t \qquad \forall i \in [d+1] \\
& \sum_{s \in S_j} x_{sj} = \mathbb{E}[Z_j(t)] \quad \forall j \in [n] \\
& \mathbf{x} \geq 0.
\end{aligned} \tag{16}$$

We assume that the process $Z(t)$ satisfies the all-time deviation bound (see Definition 5) w.r.t. some norm μ and parameters $\kappa_j = (d+1)\kappa$, where $\kappa = \kappa_\mu(A)$ depends only on A and μ. This condition can be restated as follows. For every $j \in [n]$, there are constants $c_j \geq 0$ and naturals τ_j such that

$$\mathbb{P}\left[\|Z(t) - \mathbb{E}(Z(t))\|_\mu \geq \frac{\mathbb{E}[Z_j(t)]}{\kappa(|S_j| + 1)}\right] \leq \frac{c_j}{t^2} \quad \forall t > \tau_j. \tag{17}$$

We present the resulting policy in Algorithm 4 with its guarantee in Theorem 5. We remark that the constant κ depends only on the constraint matrix defining the LP in Equation (16); that is, it does depend on the choices of bundles S_j, but it is independent of T and B.

Algorithm 4 (Fluid Bayes Selector for Online Allocation)

```
Input: Access to solutions $X^t$ of $(P_t)$ in Equation (16).
Output: Sequence of decisions for Online.
    Set $B^T$ as the given initial budget levels.
    for $t = T, \ldots, 1$, do
        Observe arrival $\theta^t = j$ and let
            $K \leftarrow \arg\max\{X^t_{sj} : s \in S_j\}$, breaking ties arbitrarily.
        If it is not feasible to assign bundle $K$, then
            reject. Otherwise assign $K$ to $\theta^t$.
        Update $B_i^{t-1} \leftarrow B_i^t$ for $i \notin K$ and $B_i^{t-1} \leftarrow B_i^t - a_{iK}$ for $i \in K$.
```

Theorem 5. *For the online allocation problem, there exists a constant κ that depends on $(S_j : j \in [n])$ only such that, if the arrival process satisfies the conditions in Equation (17), then the expected regret of the fluid Bayes selector (Algorithm 4) is at most $r_{\max} \sum_{j \in [n]} p_j(c_j + \tau_j)$, where p_j is an upper bound on $\mathbb{P}[\theta^t = j]$.*

The proof of Theorem 5 is analogous to that of Theorem 3; hence, we omit it and provide here only the key steps. Recall that, for request j, because we include the fictitious item, there are $|S_j| + 1$ possible bundles. Crucially, incorrect allocation of s to j necessitates $X^t_{sj} \geq \mathbb{E}[Z_j(t)]/(|S_j| + 1)$ (because Algorithm 4 takes the maximum entry) and $X^{\star t}_j < 1$ (OFF-LINE never allocates s to j). By the Lipschitz property of LPs (see Proposition 2), this event requires a large deviation of $Z(t)$ w.r.t. its mean, which has low probability. More formally, the disagreement sets (Definition 4) are $Q(t,b) = \{\omega \in \Omega : X^t_{sj} \geq \mathbb{E}[Z_j(t)]/(|S_j| + 1) \text{ and } X^{\star t}_j < 1\}$. By the Lipschitz property, $Q(t,b) \subseteq \{\omega \in \Omega : \|Z(t) - \mathbb{E}(Z(t))_\mu\| \geq \frac{\mathbb{E}[Z_j(t)]}{\kappa(|S_j|+1)}\}$. The probability of this last event is bounded by Equation (17); hence, compensated coupling concludes the proof.

7. Numerical Experiments

The theoretical results we have presented, together with known lower bounds for previous algorithms, show that our approach outperforms existing heuristics for online packing and online matching problems. We now reemphasize these results via simulation with synthetic data, which demonstrates both the suboptimality of existing heuristics (in terms of expected regret, which scales with T) as well as the fact that the Bayes selector has constant expected regret.

We run experiments for both online packing and online matching with multinomial arrivals. For each problem, we consider two instances, that is, two sets of parameters (r, A, p), and then we scale each instance to obtain a family of ever larger systems. For each scaling, we run 100 simulations. In conclusion, we run four sets of parameters (two for packing and two for matching), each scaled to generate many systems. The code for all the algorithms can be found at https://github.com/albvera/bayes_selector.

7.1. Online Packing

We compare the Bayes selector against three policies: (i) Static randomized is the first known policy with regret guarantees; it is based on solving the fluid LP once and using the solution as a randomized acceptance rule (Talluri and Van Ryzin 2006). (ii) Resolve and randomize is based on resolving the fluid LP at each time and using the solution as a randomized acceptance rule (Jasin and Kumar 2012). (iii) Infrequent resolve with thresholding is based on resolving the fluid LP at carefully chosen times, specifically at times $\{T^{(5/6)^u} : u = 0, 1, \ldots, \log\log(T)/\log(6/5)\}$, and then either randomize or threshold depending on the value of the solution (Bumpensanti and Wang 2019).

Our first instance has $d = 2$ resources and $n = 6$ agent types. Types $j \in \{1, 2\}$ require one unit of resource $i = 1$, types $j \in \{3, 4\}$ require one unit of $i = 2$, and types $j \in \{5, 6\}$ require one unit of each resource. All the parameters are presented in Table 1. We consider a base system with capacities $B_1 = B_2 = 40$ and horizon $T = 200$. The base system is chosen such that the problem is near dual degenerate (which is the regime in which heuristics based on the fluid benchmark are known to have poor performance; see Proposition 1). Finally, for a scaling $k \in \mathbb{N}$, the kth system has capacities kB and horizon $(k + k^{0.7})T$. We remark that, traditionally, the horizon is scaled as kT, but we chose this slightly different scaling to emphasize that our result does not depend on the specific way the system is scaled.

The results for the first instance are summarized in Figure 2, in which we also present a log-log plot that allows better appreciation of how the regret grows. Static randomized has the worst performance in our study; indeed, we do not include it in the plot because it is orders of magnitude higher. We note that not only does the Bayes selector outperform previous methods, but the regret is very small (both in average and sample path-wise), especially in comparison with the overall reward, which grows linearly with k, that is, $V^{\text{off}} = \Omega(k)$ (in expectation and with high probability).

The second instance has $n = 15$ agent types and $d = 20$ resources, the specific parameters are presented in Table C.1 and were generated randomly. We take a base system with horizon $T = 50$ and capacities $B_i = 10$ for all $i \in [20]$, and then, the kth system has horizon kT and capacities kB. The performance of different algorithms is presented in Figure 3. We notice that this instance is not degenerate, and we are scaling linearly; hence, all the algorithms except static randomize (which we again omit from the plots) are known to achieve constant regret. Nevertheless, we observe that the Bayes selector has the best performance by a large margin.

7.2. Online Matching

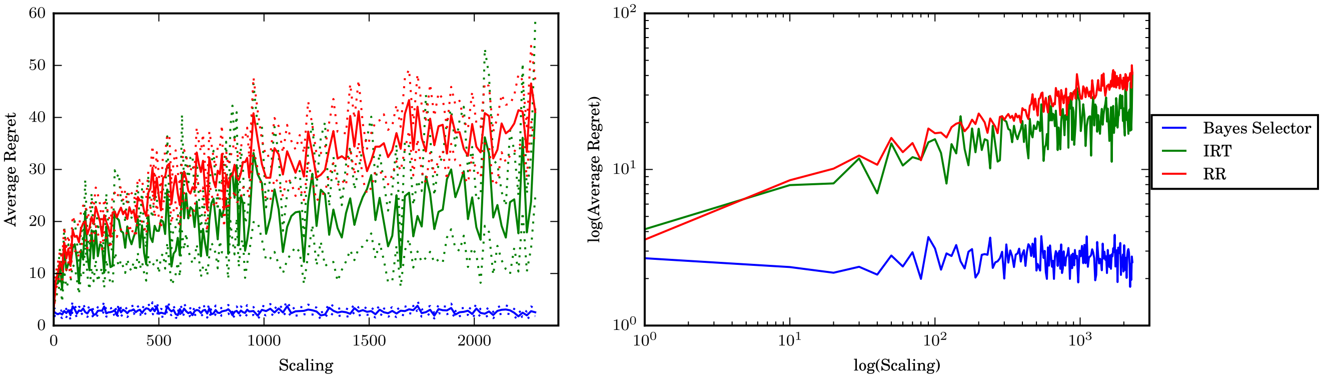
As we mentioned in Section 5, our problem corresponds to stochastic matching with edge weights. There has been previous work studying constant factor approximations for worst-case distributions. In particular, the state of the art is a 0.705 competitive ratio (Brubach et al. 2016), and a previous algorithm

Table 1. Parameters for the First Online Packing Instance

	Type j					
	1	2	3	4	5	6
Resource $i = 1$	1	1	0	0	1	1
Resource $i = 2$	0	0	1	1	1	1
p_j	0.2	0.2	0.2	0.2	0.1	0.1
r_j	10	6	10	5	9	8

Note. Coordinates (i, j) represent the consumption A_{ij}.

Figure 2. (Color online) Average Regret of Different Policies for Online Packing in the First Instance



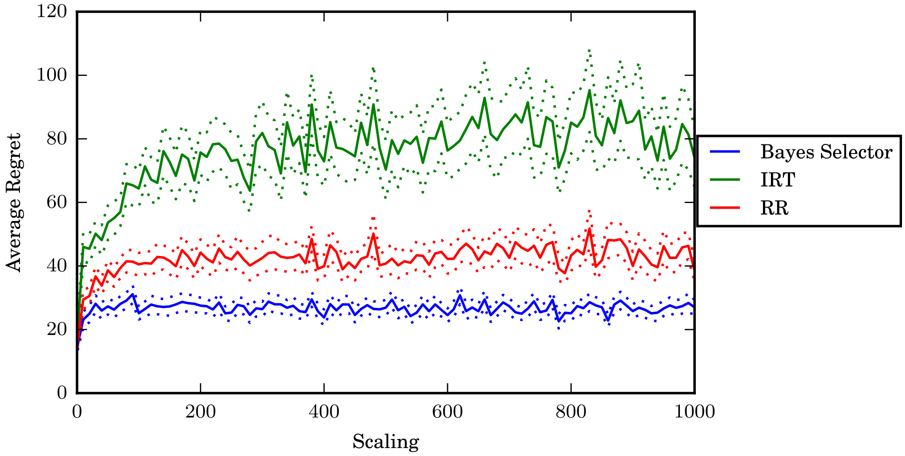
Notes. We present a plot on the left and a log-log plot on the right. We run the Bayes selector, infrequent resolve with thresholding (IRT) (Bumpensanti and Wang 2019), resolve and randomize (RR) (Jasin and Kumar 2012), and static randomized (Talluri and Van Ryzin 2006) (this last one is not reported because its high regret distorts the figures). The plot shows the regret incurred by the policies versus the off-line optimum for different scalings. Dotted lines represent 90% confidence intervals.

achieved a 0.667 competitive ratio (Haeupler et al. 2011). Both algorithms are impractical because they require a sampling procedure over poly($T \cdot \max_{i\in[d]} B_i$) many matchings. To the best of our knowledge, the best guarantee of a practical algorithm is a $1 - 1/e \approx 0.63$ competitive ratio and is achieved by the base algorithm in Haeupler et al. (2011) (that, when built upon, achieves the 0.667 guarantee). We, therefore, benchmark against this algorithm, which we call "competitive."

Competitive is based on solving a big LP once (it has $\Omega(T \cdot \max_{i\in[d]} B_i)$ variables) and using the solution as a probabilistic allocation rule. We also compare against a contemporaneous algorithm, called marginal allocation, that is based on bid prices (Wang et al. 2018). Marginal allocation uses approximate dynamic programming to obtain the marginal benefit of a matching and then uses this marginal value as a bid price so that, if the reward exceeds it, then we match the request. We give further details for both marginal allocation and competitive in Appendix C.

The first instance we consider has $d = 2$ resources and $n = 6$ agent types. The specific parameters are presented in C.1, in which reward $r_{ij} = 0$ implies that type j cannot be matched to that resource i, that is, $A_{ij} = 0$. We consider a base system with horizon $T = 20$ and capacities $B = (4, 5)'$ and then scale it linearly so that the kth system has horizon kT and capacities kB. Our second instance has $d = 6$ resources and $n = 10$ agent types; the specific parameters are presented in Table C.2. We consider a base system with horizon $T = 200$ and capacities $B = (40, 50, 40, 30, 20, 40)'$ and then scale it linearly so that the kth system has horizon kT and capacities kB.

Figure 3. (Color online) Average Regret of Different Policies for Online Packing in the Second Instance.



Notes. We run the Bayes selector, infrequent resolve with thresholding (IRT) (Bumpensanti and Wang 2019), resolve and randomize (RR) (Jasin and Kumar 2012), and static randomized (Talluri and Van Ryzin 2006) (this last one is not reported because its high regret distorts the figures). The plot shows the regret incurred by the policies versus the off-line optimum for different scalings. Dotted lines represent 90% confidence intervals.

Table 2. Parameters Used for the First Online Matching Instance

	Type j					
	1	2	3	4	5	6
Resource 1	10	6	0	0	9	8
Resource 2	0	0	5	10	20	20
p_j	0.2	0.2	0.2	0.2	0.1	0.1

Note. Coordinates (i, j) represent the reward r_{ij}, and $r_{ij} = 0$ implies that it is not possible to match i to j.

The results are presented in Figure 4. We do not include the regret of competitive because it is so high that it distorts the plots (it starts at 80 times the regret of the other algorithms and then grows linearly with k). We can confirm that the Bayes selector has constant regret and, additionally, offers the best performance. Marginal allocation offers a much better performance than competitive, but its regret still grows and seems to scale as $\Omega(\sqrt{T})$.

8. Conclusions

We reiterate that our contributions in this paper are to develop both new online policies that achieve constant regret for a large class of online resource allocation problems and, also, a new technique for analyzing online decision-making heuristics.

Our work herein has developed a new technical tool—compensated coupling—for analyzing online decision-making policies with respect to off-line benchmarks. In short, the main insight is that, through the use of compensations, we can couple Off-line's state to that of Online on every sample path. This simplifies the analysis of online policies because, in contrast to existing approaches, we do not need to track the complicated off-line process.

Next, we presented a general class of problems, which we referred to as online allocation, wherein different agents request different bundles of resources. This problem captures, among others, online packing and online matching. For all of these problems, we present a tractable policy, the Bayes selector, based on resolving an LP, that achieves constant regret.

Our analysis is based on compensated coupling, and thanks to its versatility, we can accommodate a large class of arrival processes, including correlated processes, heavy tailed, and the classical Poisson and multinomial (i.i.d.).

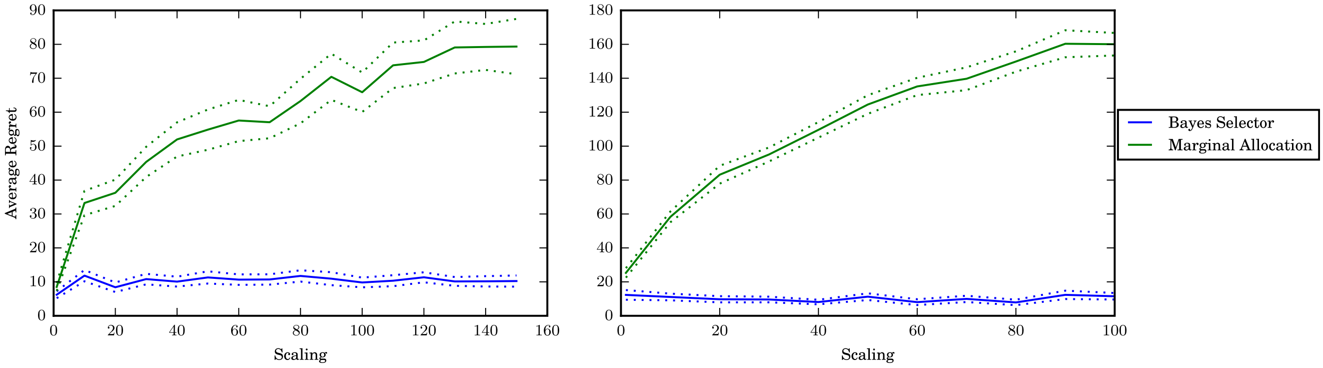
Although we instantiate the Bayes selector for online allocation, we defined it for general MDPs; we hope this policy is useful for other types of problems too. We remark two properties of the Bayes selector: (i) it works on interpretable quantities, namely the estimation of disagreement probabilities $\hat{q}$, and (ii) it is amenable to simulation because $\hat{q}$ can be estimated by running off-line trajectories. We, therefore, think that a promising avenue for further research is to apply this policy to other problems using modern estimation techniques.

The assumption of finite types of agents is well founded in revenue management problems, but there are settings in which the number of types could be very large or even continuous. Based on reported numerical results (Arlotto and Gurvich 2019), the Bayes selector appears to have good performance even in this setting. An interesting problem is to obtain parametric guarantees (not worst case) in the case in which the number of types is very large or continuous.

Acknowledgments

A preliminary version of this work was presented at SIGMETRICS 2019 and accompanied by one-page extended abstracts (Vera and Banerjee 2019). The authors thank the anonymous reviewers for their helpful feedback.

Figure 4. (Color online) Average Regret of Different Policies for Online Matching



Notes. First instance on the left and second on the right. We run the Bayes selector, marginal allocation (Wang et al. 2015), and competitive (Haeupler et al. 2011) (this last one is not reported because its high regret distorts the figures). The plot shows the regret incurred by the policies versus the off-line optimum for different scalings.

Appendix A. The Fluid Benchmark

Proof of Proposition 1. To build intuition, we start with a description of dual degeneracy for the online knapsack problem with budget $B \le T$. We assume w.l.o.g. $r_1 \ge r_2 \ge \ldots \ge r_n$ and denote $Z = Z(T)$. The primal and dual are given by

$$
\begin{aligned}
(P[Z])\ \max\quad & r'x\\
\text{s.t.}\ & \sum_{j\in[n]} x_j \le B\\
& x \le Z\\
& x \ge 0,\\
(D[Z])\ \min\quad & \alpha B + \beta' Z\\
\text{s.t.}\quad & \alpha + \beta_j \ge r_j\ \forall j\\
& \alpha \ge 0\\
& \beta \ge 0.
\end{aligned}
$$

Let us denote $\mu := \mathbb{E}[Z]$. If the fluid $(P[\mu])$ is degenerate, then we have $n+1$ active constraints. It is straightforward to conclude that there must be an index $j^\star$ such that $\sum_{j\le j^\star}\mathbb{E}[Z_j] = B$. The fluid solution is, thus, $x_j = \mathbb{E}[Z_j]$ for $j \le j^\star$ and $x_j = 0$ for $j > j^\star$. We can construct two dual solutions as follows. Let $\alpha^1 = r_{j^\star}$ and $\alpha^2 = r_{j^\star+1}$; these correspond to the shadow prices for alternative budgets $B-\varepsilon$ and $B+\varepsilon$, respectively. The corresponding variables β^1, β^2 are given by $\beta^k_j = (r_j - \alpha^k)_+$ for $k = 1, 2$. Intuitively, the fluid is indifferent between these two dual bases, but given a realization of Z, Off-line prefers one over the other; this causes a discrepancy between the expectations.

Now, we turn to the case of any packing problem. The assumption is that we are given two optimal dual solutions (α^k, β^k) with $\beta^1 \ne \beta^2$. The dual is a minimization problem, and (α^k, β^k) are always dual feasible, thus defining $\beta := \beta^1 - \beta^2$ and $\alpha := \alpha^1 - \alpha^2$,

$$
\begin{aligned}
v(D[Z]) \le \min_{k=1,2}\{B'\alpha^k + Z'\beta^k\} &= \left(B'\alpha^1 + Z'\beta^1\right)\\
&\quad \times \mathbb{1}_{\{B'\alpha + Z'\beta < 0\}} + \left(B'\alpha^2 + Z'\beta^2\right)\mathbb{1}_{\{B'\alpha + Z'\beta \ge 0\}}.
\end{aligned}
$$

The rest of the proof is reasoning that interchanging expectations $\mathbb{E}[\min_{k=1,2}\{B'\alpha^k + Z'\beta^k\}]$ for $\min_{k=1,2}\{B'\alpha^k + \mathbb{E}[Z]'\beta^k\}$ induces a $\Omega(\sqrt{T})$ error.

Because the two dual solutions have the same dual value, $B'\alpha^1 + \mu'\beta^1 = B'\alpha^2 + \mu'\beta^2$, we conclude $B'\alpha = -\mu'\beta$. We can use this condition to rewrite our bound as

$$
\begin{aligned}
v(P[Z]) \le v(D[Z]) &\le \left(B'\alpha^1 + Z'\beta^1\right)\mathbb{1}_{\{(\mu - Z)'\beta > 0\}}\\
&\quad + \left(B'\alpha^2 + Z'\beta^2\right)\mathbb{1}_{\{(\mu - Z)'\beta \le 0\}}.
\end{aligned}
$$

Because $v(P[\mu]) = B'\alpha^k + \mu'\beta^k$ for $k = 1, 2$, we take a random convex combination to obtain

$$
\begin{aligned}
v(P[\mu]) &= \left(B'\alpha^1 + \mu'\beta^1\right)\mathbb{1}_{\{(\mu - Z)'\beta > 0\}} + \left(B'\alpha^2 + \mu'\beta^2\right)\\
&\quad \times \mathbb{1}_{\{(\mu - Z)'\beta \le 0\}}.
\end{aligned}
$$

Now combine the last with our upper bound for $v(P[Z])$ and take expectations to obtain

$$
\begin{aligned}
&v(P[\mu]) - \mathbb{E}[v(P[Z])]\\
&\quad \ge \mathbb{E}\left[(\mu - Z)'\beta^1\mathbb{1}_{\{(\mu - Z)'\beta > 0\}}\right]\\
&\qquad + \mathbb{E}\left[(\mu - Z)'\beta^2\mathbb{1}_{\{(\mu - Z)'\beta \le 0\}}\right]\\
&\quad = \mathbb{E}\left[(\mu - Z)'\beta^1\mathbb{1}_{\{(\mu - Z)'\beta > 0\}}\right]\\
&\qquad + \mathbb{E}\left[(\mu - Z)'\beta^2\left(1 - \mathbb{1}_{\{(\mu - Z)'\beta > 0\}}\right)\right]\\
&\quad = \mathbb{E}\left[(\mu - Z)'\beta\mathbb{1}_{\{(\mu - Z)'\beta > 0\}}\right].
\end{aligned}
$$

Let us define ξ as the normalized vector Z, that is, $\xi := \frac{1}{\sqrt{T}}(\mu - Z)$. We conclude that

$$
v(P[\mu]) - \mathbb{E}[v(P[Z])] \ge \sqrt{T}\mathbb{E}\left[\xi'\beta\mathbb{1}_{\{\xi'\beta > 0\}}\right].
$$

Reducing by the standard deviation and applying the central limit theorem, we arrive at a half normal (also known as folded normal), which has constant expectation. This concludes the desired result. □

Appendix B. Additional Details and Proofs

B.1. Poisson Process in Discrete Periods

We explain how a continuous-time Poisson process can be reduced to our setting. We are given a time horizon T, where time $t \in [0, T]$ still denotes time to go and, according to an exponential clock, arrivals occur at some times $t_1 > t_2 > \ldots > t_N \in [0, T]$, where N is random and corresponds to the total number of arrivals, that is, $N = \sum_{j\in[n]} Z_j(T)$.

Treating times t_k as periods, there is one arrival per period. Observe that Off-line knows N; therefore, Off-line's Bellman equation is well defined. Online acts on these discrete periods; that is, Online is event-driven, thus making at most N decisions. Finally, we note that, at some time t_k, $\mathbb{E}[Z_j(t_k)] = \lambda_j t_k$ if the process is homogeneous or $\mathbb{E}[Z_j(t_k)] = \int_0^{t_k}\lambda_j(t)dt$ if the process is nonhomogeneous In conclusion, Online can compute all the required expectations without knowing N, but rather the knowledge of t_k and $\lambda(\cdot)$ is enough.

B.2. Bayes Selector Based on Marginal Compensations

A somewhat more powerful oracle is one that, for every time t, state s, and action a, returns estimates of the *marginal compensation* $\partial R(t, a, s) \cdot \mathbb{1}_{Q(t,a,s)}$. This suggests a stronger form of the Bayes selector based on marginal compensations as summarized in Algorithm B.1.

The following result follows directly from Lemma 1 and gives a performance guarantee for this algorithm.

Algorithm B.1 (Marginal Compensation Bayes Selector)

```
Input: Access to overestimates $\hat{l}(t,a,s)$ of the expected compensation, that is, $\hat{l}(t,a,s) \geq \mathbb{E}[\partial R(t,a,s)\mathbb{1}_{Q(t,a,s)}]$.
Output: Sequence of decisions for Online.
Set $S^T$ as the given initial state.
for $t = T, \ldots, 1$, do
    Observe arrival $\theta^t$ and take any action that minimizes marginal compensation, that is, $a \in \operatorname{argmin}\{\hat{l}(t, a, S^t) : a \in \mathcal{A}\}$.
    Update state $S^{t-1} \leftarrow \mathcal{T}(a, S^t, \theta^t)$.
```

Corollary B.1 (Regret of Marginal Compensation Bayes Selector)**.** *Consider Algorithm* B.1 *with overestimates* $\hat{l}(t,a,s)$*, that is,* $\hat{l}(t,a,s) \geq \mathbb{E}[\partial R(t,a,s)\mathbb{1}_{Q(t,a,s)}]$*. If* A^t *denotes the policy's action at time* t*, then*

$$\mathbb{E}[\text{Reg}] \leq \sum_t \mathbb{E}\Big[\hat{l}\big(t, A^t, S^t\big)\Big].$$

B.3. Multisecretary Problem

Proof of Theorem 2. Assume w.l.o.g. that $r_1 \geq r_2 \geq \ldots \geq r_n$. This one-dimensional version can be written as follows.

$$\begin{aligned}
(P_t^*) \quad \max \quad & r'x \\
\text{s.t.} \quad & \sum_{j\in[n]} x_j \leq B^t \\
& x_j \leq Z_j(t) \ \forall j \\
& x \geq 0. \\
(P_t) \quad \max \quad & r'x \\
\text{s.t.} \quad & \sum_{j\in[n]} x_j \leq B^t \\
& x_j \leq tp_j \ \forall j \\
& x \geq 0.
\end{aligned}$$

The optimal solution to (P_t^*) is to sort all the arrivals by reward and pick the top ones. The solution to (P_t) is similar except that it can be fractional; we saturate the variable x_1 to tp_1, then x_2 to tp_2, and continue as long as $\sum_{i\leq j} tp_i \leq B^t$ for some j. Define the probability of arrival j or better by $\bar{p}_j := \sum_{i\leq j} p_i$. Observe that we can saturate all variables $1, \ldots, j$ iff $t\bar{p}_j \leq B^t$. The solution to (P_t) is, therefore, to pick the largest j such that $t\bar{p}_j \leq B^t$ and then make $X_i^t = tp_i$ for $i \leq j$ and $X_{j+1}^t = B^t - t\bar{p}_j$. When we round this solution according to Algorithm 2, we arrive at the following policy: First, if $B^t = 0$, end the process. Second (assuming $B^t \geq 1$), always accept class $j = 1$. Third (assuming $B^t \geq 1$), if class $j > 1$ arrives, accept if $B^t/t \geq \bar{p}_j - p_j/2$ and reject if $B^t/t < \bar{p}_j - p_j/2$.

Recall that $q(t,b)$ is the probability that Off-line is not satisfied with Online's action at time t if the budget is b. We denote $q_j(t,b)$ as the probability conditioned on $\theta^t = j$. Our aim in the rest of the section is to show that $q_j(t,b)$ is summable over t.

As we observed before: (1) Off-line is not satisfied rejecting a class j iff Off-line accepts all the future arrivals of type j, that is, $X_j^{\star t} > Z_j(t) - 1$. (2) Off-line is not satisfied accepting class j iff Off-line rejects all future type j arrivals, that is, $X_j^{\star t} < 1$. We use the following standard Chernoff bound in Dubhashi and Panconesi (2009, theorem 1.1). For any $\alpha \in [0,1]$, if $X \sim \text{Bin}(t, \alpha)$,

$$\begin{aligned}
\mathbb{P}[X - \mathbb{E}[X] \leq -t\varepsilon] &\leq e^{-2\varepsilon^2 t}, \\
\mathbb{P}[X - \mathbb{E}[X] \geq t\varepsilon] &\leq e^{-2\varepsilon^2 t}.
\end{aligned} \tag{B.1}$$

We now bound the disagreement probabilities $q_j(t, B^t)$. Take j rejected by Online; that is, it must be that $j > 1$ and $B^t/t < \bar{p}_j - p_j/2$. Because we are rejecting, a compensation is paid only when condition (1) applies; thus, $X_j^{\star t} = Z_j(t)$. By the structure of Off-line's solution, all classes $j' \leq j$ are accepted in the last t rounds; that is, it must be that $X_{j'}^{\star t} = Z(t)_{j'}$ for all $j' \leq j$. We must be in the event $\sum_{j'\leq j} Z(t)_{j'} \leq B^t$. We know that $\sum_{j'\leq j} Z(t)_{j'} \sim \text{Bin}(t, \bar{p}_j)$. Because $B^t/t < \bar{p}_j - p_j/2$, the probability of error is

$$\begin{aligned}
q_j\big(t, B^t\big) &\leq \mathbb{P}\left[\sum_{j'\leq j} Z(t)_{j'} \leq B^t\right] = \mathbb{P}\big[\text{Bin}\big(t, \bar{p}_j\big) \leq B^t\big] \\
&\leq \mathbb{P}\big[\text{Bin}\big(t, \bar{p}_j\big) \leq t\bar{p}_j - tp_j/2\big].
\end{aligned}$$

Using Equation (B.1), it follows that $q_j(t, B^t) \leq e^{-p_j^2 t/2}$.

Now let us consider when j is accepted by Online. A compensation is paid only when $j > 1$ and condition (2) applies; thus, $X_j^{\star t} = 0$. Again, by the structure of $X^{\star t}$, necessarily $X_{j'}^{\star t} = 0$ for $j' \geq j$. Therefore, we must be in the event $\sum_{j'<j} Z(t)_{j'} \geq B^t$. Recall that j is accepted iff $B^t/t \geq \bar{p}_j - p_j/2 = \bar{p}_{j-1} + p_j/2$; thus,

$$\begin{aligned}
q_j\big(t, B^t\big) &\leq \mathbb{P}\left[\sum_{j'<j} Z(t)_{j'} \geq B^t\right] = \mathbb{P}\big[\text{Bin}\big(t, \bar{p}_{j-1}\big) \geq B^t\big] \\
&\leq \mathbb{P}\big[\text{Bin}\big(t, \bar{p}_{j-1}\big) \geq t\bar{p}_{j-1} + tp_j/2\big].
\end{aligned}$$

This event is also exponentially unlikely. Using Equation (B.1), we conclude $q_j(t, B^t) \leq e^{-p_j^2 t/2}$. Overall, we can bound the total compensation as

$$\sum_{t\leq T} q\big(t, B^t\big) \leq \sum_{j>1} p_j \sum_{t\leq T} e^{-p_j^2 t/2} \leq \sum_{j>1} p_j \frac{2}{p_j^2}.$$

Using compensated coupling (Lemma 1), we get our result. □

Proof of Corollary 3. By Corollary 2, if A^t is the action using overestimates $\hat{q}$, then $\mathbb{E}[\text{Reg}] \leq r_{\max} \sum_{t\in[T]} (\mathbb{E}[\hat{q}(t, A^t, B^t)] + \Delta^t)$. Recall that A^t is chosen to minimize disagreement; hence, given the condition $|q(t,a,b) - \hat{q}(t,a,b)| \leq \Delta^t$, we have $\hat{q}(t, A^t, B^t) \leq \min_{a\in\mathcal{A}} q(t, a, B^t) + \Delta^t$. In conclusion,

$$\mathbb{E}[\text{Reg}] \leq r_{\max} \sum_{t\in[T]} \left(\mathbb{E}\left[\min_{a\in\mathcal{A}} q\big(t, a, B^t\big)\right] + 2\Delta^t\right).$$

We prove that $\min_{a\in\mathcal{A}} q_j(t, a, b) \leq e^{-p_j^2 t/2}$ for all $t \in [T]$, $j \in [n]$, $b \in \mathbb{N}$; hence, the corollary follows by summing all the terms.

Let us denote $a = 1$ as the action accept and $a = 0$ reject. In the proof of Theorem 2, we concluded that the following are overestimates of the disagreement probabilities q:

$$\hat{q}_j(t,1,b) = \begin{cases} e^{-p_j^2 t/2} & \text{if } \frac{X_j^t}{tp_j} \geq 1/2 \\ 1 & \text{otherwise.} \end{cases}$$

$$\text{and} \tag{B.2}$$

$$\hat{q}_j(t,0,b) = \begin{cases} e^{-p_j^2 t/2} & \text{if } \frac{X_j^t}{tp_j} < 1/2 \\ 1 & \text{otherwise.} \end{cases}$$

Crucially, observe that the term $e^{-p_j^2 t/2}$ is *independent of the state b*. This proves that $\sup_{b\in\mathbb{N}} \min\{q_j(t,0,b), q_j(t,1,b)\} \leq e^{-p_j^2 t/2}\ \forall\, t \in [T],\ \forall\, j \in \Theta$. The proof is completed. □

B.4. Other Arrival Processes

Proof of Example 6. This follows from an application of Chung et al. (2012, theorem 3.1), which guarantees that, for some constants c', m that depend on P only,

$$\mathbb{P}[|Z_k(t) - \nu_k t| \geq \delta\nu_k t] \leq c' e^{-\frac{\delta^2 \nu_k t}{72m}}, \\ \forall t \in [T], \delta \in [0,1], k \in [n]. \tag{B.3}$$

To obtain Equation (10), we fix $j \in [n]$ and use a union bound taking the worst case in Equation (B.3); we let $\nu_{\min} := \min_{k\in[n]} \nu_k$ and $\nu_{\max} := \max_{k\in[n]} \nu_k$ and set $\delta = \nu_j/2\kappa_j\nu_{\max}$ in Equation (B.3) to obtain the result. The constants are, thus, $c_j = (\nu_j/2\kappa_j\nu_{\max})^2\nu_{\min}/72m$. Finally, we mention that the constants c' and m are related to the spectral gap and mixing time of P; for details see Chung et al (2012). □

Proof of Example 7. To prove the all-time deviation, we use that, from the proof of Devroye (1983, lemma 3), $\mathbb{P}(|X - |\mathbb{E}[X] \geq \varepsilon\mathbb{E}[X]) \leq 2e^{-\mathbb{E}[X]\varepsilon^2/4}$ is valid for any Poisson r.v. X and any $\varepsilon > 0$. Now we proceed as in Example 6: taking $X_k = Z_k(t)$ and $\varepsilon = \frac{\mathbb{E}[Z_j(t)]}{2\kappa_j\mathbb{E}[Z_k(t)]}$, we obtain

$$\mathbb{P}\left[\|Z(t) - \mathbb{E}[Z(t)]\|_\infty \geq \frac{\mathbb{E}[Z_j(t)]}{2\kappa_j}\right] \leq 2\sum_{k\in[n]} e^{-\frac{\mathbb{E}[Z_j(t)]^2}{8\kappa_j^2\mathbb{E}[Z_k(t)]}}.$$

Finally, from Equation (11), we have $\mathbb{E}[Z_k(t)] \leq g(t)\mathbb{E}[Z_j(t)]$, and from Equation (12), we have $\mathbb{E}[Z_j(t)] \geq g(t)f(t)\log(t)$. From these bounds, we conclude $\mathbb{P}[\|Z(t) - \mathbb{E}[Z(t)]\|_\infty] \geq \frac{\mathbb{E}[Z_j(t)]}{2\kappa_j} \leq 2ne^{-f(t)\log(t)/8\kappa_j^2}$ and the existence of constants τ_j, c_j satisfying the all-time deviation follows. □

B.5. Proof of Proposition 4

We denote $\mathbf{x} \in \mathbb{R}^{nd}$ the vector of the form $\mathbf{x} = (x_{11}, x_{21} \ldots, x_{d1}, x_{12}, \ldots)'$; that is, we concatenate the components x_{ij} by j first. We can write the feasible region of $P[z,b]$ as $\{\mathbf{x} : C\mathbf{x} \leq b, D\mathbf{x} \leq z, \mathbf{x} \geq 0$, where $C \in \mathbb{R}^{d\times nd}$ and $D \in \mathbb{R}^{n\times nd}$. It follows from a slight strengthening of Mangasarian and Shiau (1987, theorem 2.4) that $\|\mathbf{x}^1 - \mathbf{x}^2\|_\infty \leq \kappa\|z^1 - z^2\|_1$, where

$$\kappa = \sup\left\{\|v\|_\infty : \|C'u + D'v\|_1 = 1,\ \text{support}\begin{pmatrix} u \\ v \end{pmatrix} \\ \text{corresponds to linearly independent rows}. \\ \text{of } \begin{pmatrix} C \\ D \end{pmatrix}\right\}$$

If we study Equation (13), denoting I_d the d-dimensional identity and $1_d, 0_d$ d-dimensional row vectors of ones and zeros, we can write the matrices C, D as follows. We sketched the multipliers u_i, v_j next to the rows,

$$C = [I_d|I_d|\ldots|I_d] = \begin{array}{cl} \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & \cdots \\ 0 & 1 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots \end{pmatrix} & \begin{matrix} \leftarrow u_1 \\ \leftarrow u_2 \\ \vdots \\ \leftarrow u_d, \end{matrix} \end{array}$$

and similarly,

$$D = \begin{array}{cl} \begin{pmatrix} 1_d & 0_d & \cdots & 0_d \\ 0_d & 1_d & \cdots & 0_d \\ \vdots & \vdots & \ddots & \vdots \\ 0_d & 0_d & \cdots & 1_d \end{pmatrix} & \begin{matrix} \leftarrow v_1 \\ \leftarrow v_2 \\ \vdots \\ \leftarrow v_n. \end{matrix} \end{array}$$

We have two cases: either $u_i = 0$ for some $i \in [d]$ or $u_i \neq 0$ for all $i \in [d]$. On the first case, say w.l.o.g. $u_1 = 0$ and take any $j \in [n]$. Observe that the constraint $\|C'u + D'v\|_1 = 1$ implies (studying all the components involving j) $\sum_{i\in[d]} |u_i + v_j| \leq 1$. Because $u_1 = 0$, this reads as $|v_j| + \sum_{i>1} |u_i + v_j| \leq 1$; thus, $|v_j| \leq 1$ as desired.

For the other case, we assume $u_i \neq 0$ for all i; hence, $v_j = 0$ for some j because, otherwise, we would violate the l.i. restriction on the support. Assume w.l.o.g. $v_1 = 0$ and let us study some v_j. The constraint $\|C'u + D'v\|_1 = 1$ implies (looking at the first n components and the components involving j) $\sum_{i\in[d]} |u_i| + \sum_{i\in[d]} |v_j + u_i| \leq 1$. By triangle inequality,

$$d|v_j| = \left|\sum_{i\in[d]} (u_i + v_j) - \sum_{i\in[d]} u_i\right| \\ \leq \sum_{i\in[d]} |v_j + u_i| + \sum_{i\in[d]} |u_i| \leq 1.$$

This shows $|v_j| \leq 1$, and the proof is complete. □

B.6. Additional Details for Online Stochastic Matching

The stochastic bipartite matching is defined by a set of static nodes U, $|U| = d$, and a random set of nodes arriving sequentially. At each time, a node θ^t is chosen from a set V, $|V| = n$, and we are given its set of neighbors in U. We identify the online bipartite matching problem in our framework as follows. The state S^t encodes the available nodes from U; an action corresponds to matching the arrival $\theta^t \in V$ to a neighbor $u \in U$ of θ^t or to discard the arrival. In the latter case, we say that it is matched to $u = \emptyset$.

For a graph G, we denote the size of its maximum matching as $M(G) \in \mathbb{N} \cup \{0\}$ and $G - (u,v)$ as the usual removal of nodes; in the case $u = \emptyset$, $G - (u,v) = G - v$. Recall that $Q(t,a,s)$ is the event when Off-line is not satisfied with action a and $q(t,a,s) = \mathbb{P}[Q(t,a,s)]$. Let us fix an Online policy and define $G_t = (L,R)$ as the bipartite graph with nodes $L = S^t$ and $R = Z(t)$, that is, the realization of future arrivals and current state. With the convention $\mathbb{1}_\emptyset = 0$ and $\mathbb{1}_u = 1$ for $u \in U$,

$$\bar{Q}(t,u,s) = \{\omega \in \Omega : M(G_t) = \mathbb{1}_u + M(G_t - (u,\theta^t))\}.$$

In words, Off-line is satisfied matching θ^t to u if the size of the maximum matching with and without that edge differs

Table C.1. Parameters Used for the Second Online Packing Instance

		Type j														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Resource i	1	0	1	0	1	0	1	0	0	0	0	1	0	1	0	0
	2	1	1	1	0	0	1	1	1	1	0	0	0	1	1	1
	3	0	1	1	0	1	0	0	1	0	0	1	1	0	1	0
	4	0	1	0	0	0	1	1	0	1	1	1	1	1	0	0
	5	1	1	1	0	0	0	1	0	1	0	0	0	0	1	1
	6	0	1	1	0	1	0	1	0	1	0	1	1	1	1	0
	7	1	0	1	1	1	1	1	0	1	1	1	0	0	1	1
	8	1	1	1	0	0	1	0	0	1	0	0	0	1	0	0
	9	0	1	1	1	1	0	0	1	0	0	0	1	0	0	0
	10	1	1	0	1	1	0	0	0	1	0	1	1	0	0	0
	11	1	1	0	1	0	0	1	0	1	0	0	0	0	0	1
	12	0	1	1	1	0	1	1	1	1	1	0	1	0	0	1
	13	0	0	1	0	0	1	0	1	1	0	1	1	0	1	1
	14	1	0	0	0	0	1	0	1	0	0	1	1	1	0	0
	15	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0
	16	1	0	1	0	0	1	0	1	0	1	1	0	0	0	1
	17	0	1	0	1	1	0	0	0	0	1	1	1	0	0	1
	18	0	0	1	1	0	1	1	0	1	1	0	0	0	0	1
	19	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0
	20	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1
	p_j	0.075	0.075	0.125	0.025	0.05	0.062	0.062	0.1	0.1	0.05	0.125	0.012	0.075	0.062	0.002
	r_j	7	5	16	1	1	20	10	18	7	14	17	19	14	1	2

Note. Coordinates (i, j) represent the consumption A_{ij}.

by exactly one. With this observation, a straightforward application of the compensated coupling Lemma 1 yields Lemma 3.

Finally, we provide an example for a negative result. Despite the fact that the regret is exactly the number of disagreements and the Bayes selector minimizes each term, it is not an optimal policy.

Proposition B.1. *The Bayes selector is suboptimal for stochastic online bipartite matching.*

Proof. Consider an instance with static nodes $U = \{a, b, c\}$ and four types of online nodes $V = [4]$. Type 1 matches to a only, 2 to a and b, 3 to c only, and 4 to b and c. Observe that the only types inducing error are 2 and 4.

Assume the arrival at $t = 3$ is $\theta^3 = 2$. Matching it to a is an error if arrivals are $\{1, 1\}, \{1, 3\}, \{1, 4\}$, so the disagreement is $p_1^2 + 2p_1p_3 + 2p_1p_4$. Matching it to b is an error if arrivals are $\{4, 4\}, \{3, 4\}$ with disagreement $p_4^2 + 2p_4p_3$. Now assume $p_4^2 + 2p_4p_3 = p_1^2 + 2p_1p_3 + 2p_1p_4$, so the Bayes selector is indifferent and, thus, say it matches to a.

At $t = 2$, there is only an error if $\theta^2 = 4$, in which case matching it to b has disagreement p_2 and matching it to c disagreement p_3. In conclusion, the Bayes selector pays $p_1^2 + 2p_1p_3 + 2p_1p_4$ in the first stage plus $\min\{p_2, p_3\}$ in the second with probability p_4.

The strategy that matches at $t = 3$ type 2 to b has disagreement $p_4^2 + 2p_4p_3 = p_1^2 + 2p_1p_3 + 2p_1p_4$, thus lower than the Bayes selector. To see this, note that, if we match to b, there is no error at $t = 2$.

Finally, the equation $p_4^2 + 2p_4p_3 = p_1^2 + 2p_1p_3 + 2p_1p_4$ is satisfied, for example, with $p_1 = p_4/2$ and $p_3 = p_4/4$. □

Appendix C. Additional Details from Numerical Experiments

Competitive is described as follows. For a given horizon T, let $K_j := \lceil p_jT \rceil$. We create a bipartite graph $G = (U, V, E)$, where U is the static side and V the online side. The static side contains B_i copies of each resource i; hence, $|U| = \sum_{i \in [d]} B_i$. The online side contains K_j copies of each type j; hence, $|V| = \sum_{j \in [n]} K_j$. The edge set E is the natural construction

Table C.2. Parameters Used for the Second Online Matching Instance

		Type j									
		1	2	3	4	5	6	7	8	9	10
Resource i	1	10	6	0	0	9	8	2	0	0	1
	2	1	0	0	0	0	0	2	0	0	8
	3	0	0	0	0	0	0	2	0	0	6
	4	0	26	0	0	1	0	3	0	0	11
	5	1	4	0	0	0	0	0	0	0	13
	6	7	4	12	11	10	12	18	2	0	0
	p_j	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Note. Coordinates (i, j) represent the reward r_{ij}, and $r_{ij} = 0$ implies that it is not possible to match i to j.

in which each copy of $j \in [n]$ has edges to all copies of $i \in [d]$ according to the adjacency matrix A. The weight w_e on edge $e = (u, v)$ is r_{ij} if u is a copy of i and v is a copy of j. Finally, define the following matching LP on the graph G, where λ_{iljk} stands for the lth copy of i and kth copy of j.

$$
\begin{aligned}
(P)\ \max & \sum_{i\in[d]} \sum_{l\in[B_i]} \sum_{j\in[n]:A_{ij}=1} \sum_{k\in[K_j]} r_{ij}\lambda_{iljk} \\
\text{s.t.} & \sum_{j\in[n]} \sum_{k\in[K_j]} \lambda_{iljk} \le 1 \quad \forall i \in [d], l \in [B_i] \\
& \sum_{i\in[d]} \sum_{l\in[B_i]} \lambda_{iljk} \le 1 \quad \forall j \in [n], k \in [K_j] \\
& \lambda \ge 0,
\end{aligned}
$$

and let λ^* be a solution to this LP. Whenever a type j arrives, competitive draws $k \in [K_j]$ uniformly at random and then takes a vertex $u = il$ incident to $v = jk$ with probability λ^*_{iljk}, and if $u = il$ is not taken, it matches v to u. We note that the process of copying nodes is not superfluous because the analysis of competitive heavily relies on the fact that the LP is in this form.

Marginal allocation is described as follows. Let x be a solution of (P_T) in Equation (15), that is, of the fluid LP, and let $f_i : [T] \times \{0, \ldots, B_i\} \to \mathbb{R}_{\ge 0}$ be some functions specified later. When a type j arrives at t and the current budgets are $B^t \in \mathbb{N}^d$, marginal allocation uses $f_i(t, B_i^t) - f_i(t, B_i^t - 1)$ as the bid price for each resource $i \in [d]$: the type is rejected if $r_{ij} < f_i(t, B_i^t) - f_i(t, B_i^t - 1)$ for all $i \in [d]$ such that $B_i^t > 0$, and otherwise, it is matched to $\operatorname{argmax}\{r_{ij} - f_i(t, B_i^t) + f_i(t, B_i^t - 1) : i \in [d], B_i^t > 0$. Finally, the functions f are obtained with the following recursion

$$
\begin{aligned}
f_i(t+1, b) = f_i(t, b) + \frac{1}{T} \sum_{j\in[n]} x_{ij} \big(r_{ij} - f_i(t, b) \\
+ f_i(t, b-1)\big)^+, \qquad f_i(1, \cdot) = 0, f_i(\cdot, 0) = 0.
\end{aligned}
$$

References

Agrawal S, Devanur NR (2014) Fast algorithms for online stochastic convex programming. *Proc. 26th Annual ACM-SIAM Sympos. Discrete Algorithms*, 1405–1424.

Alaei S (2014) Bayesian combinatorial auctions: Expanding single buyer mechanisms to many buyers. *SIAM J. Comput.* 43(2):930–972.

Arlotto A, Gurvich I (2019) Uniformly bounded regret in the multisecretary problem. *Stochastic Systems* 9(3):231–260.

Augustine J, Irani S, Swamy C (2004) Optimal power-down strategies. *45th Annual IEEE Sympos. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 530–539.

Badanidiyuru A, Kleinberg R, Slivkins A (2013) Bandits with knapsacks. *2013 IEEE 54th Annual Sympos. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 207–216.

Bertsekas DP (1995) *Dynamic Programming and Optimal Control* (Athena Scientific, Belmont, MA).

Borrelli F (2003) *Constrained Optimal Control of Linear and Hybrid Systems*, vol. 290 (Springer, Berlin, Heidelberg).

Brown DB, Smith JE (2013) Optimal sequential exploration: Bandits, clairvoyants, and wildcats. *Oper. Res.* 61(3):644–665.

Brown DB, Smith JE (2014) Information relaxations, duality, and convex stochastic dynamic programs. *Oper. Res.* 62(6):1394–1415.

Brubach B, Sankararaman KA, Srinivasan A, Xu P (2016) Online stochastic matching: New algorithms and bounds.

Buchbinder N, Naor JS (2009a) The design of competitive online algorithms via a primal–dual approach. *Foundations Trends Theoretical Comput. Sci.* 3(2–3):93–263.

Buchbinder N, Naor J (2009b) Online primal-dual algorithms for covering and packing. *Math. Oper. Res.* 34(2):270–286.

Bumpensanti P, Wang H (2019) A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Sci.* Forthcoming.

Chen N, Agarwal A, Wierman A, Barman S, Andrew LLH. (2015) Online convex optimization using predictions. *ACM SIGMETRICS Performance Evaluation Rev.* 43(1):191–204.

Chen N, Comden J, Liu Z, Gandhi A, Wierman A (2016) Using predictions in online optimization: Looking forward with an eye on the past. *Performance Evaluation Rev.* 44(1):193–206.

Chung K-M, Lam H, Liu Z, Mitzenmacher M (2012) Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified. *Sympos. Theoretical Aspects Comput. Sci.*, vol. 14, 124–135.

Ciocan DF, Farias V (2012) Model predictive control for dynamic resource allocation. *Math. Oper. Res.* 37(3):501–525.

Correa J, Foncea P, Hoeksma R, Oosterwijk T, Vredeveld T (2017) Posted price mechanisms for a random stream of customers. *Conf. Econom. Comput.*, (ACM), 169–186.

Desai VV, Farias VF, Moallemi CC (2012) Pathwise optimization for optimal stopping problems. *Management Sci.* 58(12):2292–2308.

Devanur NR, Jain K, Sivan B, Wilkens CA (2019) Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *J. ACM* 66(1).

Devroye L (1983) The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *Ann. Statist.* 11(3):896–904.

Dubhashi DP, Panconesi A (2009) *Concentration of Measure for the Analysis of Randomized Algorithms* (Cambridge University Press).

Düetting P, Feldman M, Kesselheim T, Lucier B (2017) Prophet inequalities made easy: Stochastic optimization by pricing non-stochastic inputs. *Foundations of Computer Science* (IEEE), 540–551.

Gupta A, Molinaro M (2014) How experts can solve LPs online. *Eur. Sympos. Algorithms* (Springer), 517–529.

Haeupler B, Mirrokni VS, Zadimoghaddam M (2011) *Online Stochastic Weighted Matching: Improved Approximation Algorithms* (Internet and Network Economics).

Hill TP, Kertz RP (1982) Comparisons of stop rule and supremum expectations of iid random variables. *Ann. Probab.* 10(2):336–345.

Huang L (2015) Receding learning-aided control in stochastic networks. *Performance Evaluation* 91(C):150–169.

Jasin S, Kumar S (2012) A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Math. Oper. Res.* 37(2):313–345.

Kesselheim T, Radke K, Tönnis A, Vöcking B (2018) Primal beats dual on online packing LPs in the random-order model. *SIAM J. Comput.* 47(5):1939–1964.

Kleinberg R, Weinberg SM (2012) Matroid prophet inequalities. *ACM Sympos. Theory Comput.* (ACM), 123–136.

Mangasarian OL, Shiau TH (1987) Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems. *SIAM J. Control Optim.* 25(3):583–595.

Manshadi VH, Gharan SO, Saberi A (2012) Online stochastic matching: Online actions based on offline statistics. *Math. Oper. Res.* 37(4):559–573.

Morari M, Garcia CE, Lee JH, Prett DM (1993) *Model Predictive Control* (Prentice Hall, Englewood Cliffs, NJ).

Powell WB (2011) *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, vol. 842 (John Wiley & Sons).

Reiman MI, Wang Q (2008) An asymptotically optimal policy for a quantity-based network revenue management problem. *Math. Oper. Res.* 33(2):257–282.

Talluri KT, Van Ryzin GJ (2006) *The Theory and Practice of Revenue Management*, vol. 68 (Springer Science & Business Media, Boston).

Tsitsiklis JN, Van Roy B (2001) Regression methods for pricing complex American-style options. *IEEE Trans. Neural Networks* 12(4):694–703.

Van Hentenryck P, Bent R (2009) *Online Stochastic Combinatorial Optimization* (MIT Press, Boston).

Vera A, Banerjee S (2019) The Bayesian prophet: A low-regret framework for online decision making. *Performance Evaluation Rev.* 47(1):81–82.

Wang X, Truong VA, Bank D (2018) Online advance admission scheduling for services with customer preferences.

Wu H, Srikant R, Liu X, Jiang C (2015) Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *Advances in Neural Information Processing Systems*, 433–441.

Zhang H, Shi C, Qin C, Hua C (2016) Stochastic regret minimization for revenue management problems with nonstationary demands. *Naval Res. Logist.* 63(6):433–448.