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

Alberto Vera
Cornell University

Siddhartha Banerjee
Cornell University

## ABSTRACT

Motivated by the success of using black-box predictive algorithms as subroutines for online decision-making, we develop a new framework for designing online policies given access to an oracle providing statistical information about an offline 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 towards tackling 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 offline benchmark; Second, using this technique, we show that the Bayes Selector has constant expected regret (i.e., independent of the number of arrivals and resource levels) in any online packing and matching problem with a finite type-space. Our results generalize and simplify many existing results for online packing and matching problems, and suggest a promising pathway for obtaining oracle-driven policies for other online decision-making settings.

## KEYWORDS

Stochastic Optimization, Prophet Inequalities, Approximate Dynamic Programming, Revenue Management, Online Packing.

**ACM Reference Format:**
Alberto Vera and Siddhartha Banerjee. 2019. The Bayesian Prophet: A Low-Regret Framework for Online Decision Making . In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '19 Abstracts), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3309697.3331518

## 1 INTRODUCTION

Everyday life is replete with settings where we have to make decisions while facing uncertainty over future outcomes. Some examples include allocating cloud resources, matching an empty car to a ridesharing passenger, displaying online ads, selling airline seats, etc. In many of these instances, 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?*

Our work focuses on two important classes of online decision-making problems: online packing and online matching. These problems are fundamental in MDP theory; they have a rich existing literature and widespread applications in many domains [5]. Nevertheless, our work develops new policies for both these problems which admit performance guarantees that are order-wise better than existing approaches. These policies 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 offline 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* vìs-a-vìs the offline benchmark. Though we focus on online packing and matching problems, we believe our approach provides a new way for designing and analyzing online decision-making policies using predictive models.

**Our Contributions:** We make progress in three aspects

(1) *Technical*: We present a *new stochastic coupling technique*, which we call the *compensated coupling*, for evaluating the regret of online decision-making policies vis-à-vis offline 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 online packing and matching problems, we prove that the Bayes Selector gives regret guarantees that are *independent of the size of the state-space*, i.e., constant with respect to the horizon length and budgets.

## 2 PROBLEM SETTING AND RESULTS

In online packing there are $d$ distinct resources denoted by the set $[d]$. At time $t = T$, we have an initial availability (budget) vector $B \in \mathbb{N}^d$. At every time $t = T, T-1, \ldots, 1$, nature draws an arrival with *type* $\theta^t$ from a finite set of $n$ distinct types $\Theta = [n]$, via some distribution which is known to the algorithm designer (or principal).

An arrival of type $j$ corresponds to a resource request with associated reward $r_j$ and resource requirement $A_j = (a_{ij})_{i\in[d]}$, where $a_{ij}$ denotes the units of resource $i$ required to serve the arrival. At each time, the principal must decide whether to accept the request $\theta^t$ (thereby generating the associated reward while consuming the required resources), or reject it (no reward and no resource consumption). Accepting a request 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.

*Full version available on https://ssrn.com/abstract=3158062


*SIGMETRICS '19 Abstracts, June 24–28, 2019, Phoenix, AZ, USA*

ACM ISBN 978-1-4503-6678-6/19/06.
https://doi.org/10.1145/3309697.3331518

*Online matching* problems are a closely related class of problems, wherein we have the same setup with $d$ resources with initial budget $B$, but now each type comprises of a menu of (requirement, reward) pairs, and is satisfied with *any resource from the given set* (as opposed to all resources from the set).

**Arrival Processes:** To completely define a packing/matching problem, we need to specify the generative model for the type sequence $\theta^T, \theta^{T-1}, \ldots, \theta^1$. An important case is the Multinomial process, where at each time the arrival is of type $j$ with probability $p_j$. More general models allow for non-stationary and/or correlated arrival processes – e.g., non-homogeneous Poisson processes, Markovian models, etc. An important feature of our framework is that it handles a wide variety of such processes in a unified manner.

**The prophet benchmark:** Our performance guarantees are best illustrated by adopting the view that a given problem is simultaneously solved by two '*agents*', Online and Offline, who are primarily differentiated based on their access to information. Online can only take *non-anticipatory* actions, which at each time $t$ can be based on the current state and arrival, past trajectory, and distributional information. On the other hand, Offline at time $t$ is allowed to make decisions with full knowledge of future arrivals $\theta^t, \ldots, \theta^1$. Denoting the total rewards of Offline and Online as $V^{\text{off}}$ and $V^{\text{on}}$ respectively, we define the *regret* of an online policy w.r.t. to an offline benchmark to be the *additive* loss $\textsc{Reg} := V^{\text{off}} - V^{\text{on}}$.

**Overview of our 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 offline benchmark and can be interpreted as *probabilities of regretting each particular action in hindsight*. Note that such estimates can easily be obtained, for example, via simulation given knowledge of the arrival process. Moreover, 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, hence we refer to it as the Bayes Selector policy. In particular, when we apply the Bayes Selector to online packing, we obtain the following guarantee, which generalizes and improves on prior and contemporaneous results [1–4, 6].

Theorem 2.1 (Informal). *For any online packing problem with a finite number of resource types and arrival types, for a large class of arrival processes, the Bayes Selector achieves regret which is independent of the horizon $T$ and resource budgets $B$ (both 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 holds under weak assumptions on the arrival process, which admit Multinomial and Poisson arrivals, time-dependent processes, and Markovian arrivals. We also obtain similar results for matching problems.

At the core of our analysis is a *novel stochastic coupling technique* for analyzing online policies based on offline (or *prophet*) benchmarks. In particular, unlike traditional approaches to regret analysis, which are based on showing that an online policy tracks a fixed offline policy, our approach is instead based on *forcing Offline to follow Online's actions*. In the remaining, we briefly describe this approach for general MDPs.

**The Bayes Selector Policy:** Consider a general MDP problem, with available actions $\mathcal{A}$, state space $\mathcal{S}$, and transition function $\mathcal{T}$. For any given arrival sequence, if Offline is at state $s \in \mathcal{S}$, then certain action $A \in \mathcal{A}$ is optimal, while other actions may decrease Offline's to-go reward. Observe that, since Offline uses the knowledge of future arrivals, the optimal action $A$ is a r.v. We say that Offline disagrees with $a \in \mathcal{A}$ if choosing $a$ instead of $A$ does decrease Offline's to-go reward.

Let $q(t, a, s)$ be the *disagreement probability* of action $a \in \mathcal{A}$ at time $t$ in state $s \in \mathcal{S}$, i.e., the probability that Offline disagrees with $a$ if he is in state $s$ at time $t$. Suppose we have an *oracle* that gives us $q(t, a, s)$ for every feasible action $a$; for example, via simulation, learning the probability based on past data, etc.

Given such an oracle, a natural greedy policy is to choose the action $a$ that minimizes the disagreement. Algorithm 1 generalizes this idea for settings where we have estimates $\hat{q}(t, a, s)$ for these disagreement probabilities.

**Algorithm 1** Bayes Selector

**Input:** Access to over-estimates $\hat{q}(t, a, s)$ of the disagreement probabilities, i.e., $\hat{q}(t, a, s) \geq q(t, a, s)$
**Output:** Sequence of decisions for Online.

```
Set S^T as the given initial state
for t = T, ..., 1 do
    Observe the arriving type θ^t
    Choose a ∈ argmin{q̂(t, a, S^t)} (minimize disagreement)
    Update state S^{t-1} ← T(a, S^t, θ^t).
```

Observe that in Algorithm 1 the process $S^t$ corresponds to Online's state. This is critical for ensuring Online is non-anticipatory, since Offline's trajectory could depend on future arrivals; however, it makes it difficult to compare the performance of Online and Offline. The compensated coupling technique we introduce allows us to couple Offline's state to that of Online's, while maintaining the performance of Offline through appropriate 'compensations'.

In our work, we provide a regret bound for Algorithm 1 that holds under complete generality. We then use this to prove Theorem 2.1 by applying Algorithm 1 with $\hat{q}$ obtained from a natural LP relaxation. For details, refer to our full paper.

**Acknowledgements:** AV and SB gratefully acknowledge support from the NSF under grants DMS-1839346 and ECCS-1847393, and the ARL under grant W911NF-17-1-0094.

## REFERENCES

[1] Alessandro Arlotto and Itai Gurvich. Uniformly bounded regret in the multi-secretary problem. *Stocahstic Systems*, 2018. Forthcoming.
[2] Pornpawee Bumpensanti and He Wang. A re-solving heuristic for dynamic resource allocation with uniformly bounded revenue loss. *arXiv preprint arXiv:1802.06192*, 2018.
[3] Stefanus Jasin and Sunil Kumar. A Re-Solving Heuristic with Bounded Revenue Loss for Network Revenue Management with Customer Choice. *Mathematics of Operations Research*, 37(2):313–345, may 2012.
[4] Martin I. Reiman and Qiong Wang. An Asymptotically Optimal Policy for a Quantity-Based Network Revenue Management Problem. *Mathematics of Operations Research*, 33(2):257–282, may 2008.
[5] Kalyan T Talluri and Garrett J Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.
[6] Huasen Wu, R Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems*, pages 433–441, 2015.