Dynamic Matching with Post-allocation Service and its Application to Refugee Resettlement

Kirk Bansak
University of California, Berkeley, CA, kbansak@berkeley.edu

Soonbong Lee
Yale School of Management, New Haven, CT, soonbong.lee@yale.edu

Vahideh Manshadi
Yale School of Management, New Haven, CT, vahideh.manshadi@yale.edu

Rad Niazadeh
University of Chicago Booth School of Business, Chicago, IL, rad.niazadeh@chicagobooth.edu

Elisabeth Paulson
Harvard Business School, Boston, MA, epaulson@hbs.edu

Motivated by our collaboration with a major refugee resettlement agency in the U.S., we study a dynamic matching problem where each new arrival (a refugee case) must be matched immediately and irrevocably to one of the static resources (a location with a fixed annual quota). In addition to consuming the static resource, each case requires *post-allocation* services from a server, such as a translator. Given the uncertainty in service time, a server may not be available at a given time, thus we refer to it as a *dynamic resource*. Upon matching, the case will wait to avail service in a first-come-first-serve manner. Bursty matching to a location may result in undesirable congestion at its corresponding server. Consequently, the central planner (the agency) faces a dynamic matching problem with an objective that combines the matching reward (captured by pair-specific employment outcomes) with the cost for congestion for dynamic resources and over-allocation for the static ones. Motivated by the observed fluctuations in the composition of refugee pools across the years, we aim to design algorithms that do not rely on distributional knowledge. We develop learning-based algorithms that are asymptotically optimal in certain regimes, easy to interpret, and computationally fast. Our design is based on learning the dual variables of the underlying optimization problem; however, the main challenge lies in the time-varying nature of the dual variables associated with dynamic resources. Our theoretical development brings together techniques from Lyapunov analysis, adversarial online learning, and stochastic optimization. On the application side, when tested on real data from our partner agency, our method outperforms existing ones making it a viable candidate for replacing the current practice upon experimentation.

Key words: refugee matching, post-allocation service, workload balancing, distribution-free algorithms, online learning, online allocation

Electronic copy available at: https://ssrn.com/abstract=4748762

1. Introduction

According to the United Nations High Commissioner for Refugees (UNHCR), the forcibly displaced population in the world exceeds 100 million as of the end of 2022. This includes over 5 million asylum seekers and over 35 million refugees (UNHCR 2022). The process of providing aid to these individuals often involves making high-stakes decisions in the face of complex operational intricacies. Taking data-driven and algorithmic approaches that capture the unique features of these processes yet provide easy-to-implement solutions can significantly improve these operations and facilitate the well-being of these vulnerable communities. This paper demonstrates such potential in the context of refugee resettlement in the U.S.[1]

This work is motivated by our collaboration with a major refugee resettlement agency in the U.S. Refugee resettlement is a process largely overseen by the UNHCR in which refugees are relocated to and granted long-term/permanent residence in a participating host country. Each year over one hundred thousand refugees are submitted by the UNHCR for resettlement. The U.S. resettles tens of thousands of refugees yearly — the largest number of any host country (UNHCR 2023) — through ten *national resettlement agencies*, one of which is our partner. In the following, we first provide background on the resettlement process, highlighting key features that motivate our modeling framework. Then, we provide an overview of our contributions followed by a discussion on the related work.

1.1. Background on Refugee Resettlement

In the following, we briefly describe the refugee resettlement process in the U.S. Before doing so, we highlight that the refugee resettlement and/or asylum procedures of many host countries (e.g. Switzerland, the Netherlands, Sweden, etc.) share important commonalities with the procedure described below, while the precise details vary by country. After being admitted to the U.S., a refugee case is allocated to one of the national resettlement agencies. The corresponding agency then *assigns* the case to a local service provider (often referred to as an *affiliate*) within its network. Hereafter, for the sake of consistency, we refer to this assignment as a *matching*. Notably, there is evidence that the initial matching has a significant impact on finding employment within 90 days after resettlement (Bansak et al. 2018). This is the key integration outcome in the U.S. that is tracked and reported to Congress. It has also been shown recently that the 90-day employment for any case-affiliate pair can be reasonably predicted using machine learning (ML) models (Bansak et al. 2018, 2024). Equipped with such ML models, a resettlement agency can use these predictions in making matching decisions.

The resettlement agency faces several operational considerations: (i) If the refugee has *U.S. ties* (such as family or close friends) in a particular locality, then the agency will match the case to the affiliate associated with that locality (Bruno 2017); (ii) each affiliate has a target annual *quota* (i.e., the number of refugees to be matched to) approved by the U.S. State Department; (iii) when new cases

[1] There has been prior work in this context that we will discuss thoroughly later in this section.

Electronic copy available at: https://ssrn.com/abstract=4748762

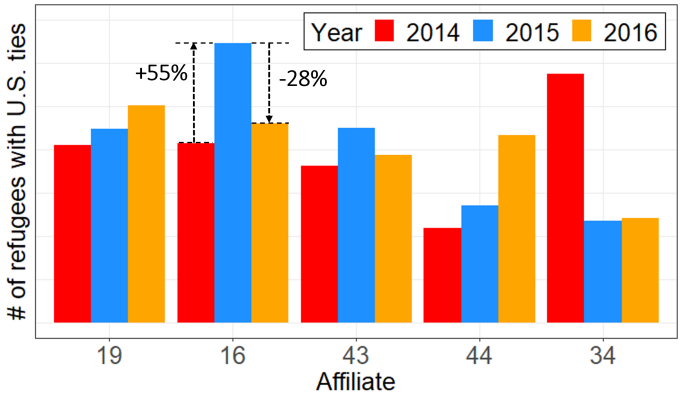


Figure 1 **The number of refugees with U.S. ties at five affiliates normalized by the total number of arrivals.**

arrive at the matched affiliates, case workers and service providers, such as translators, are needed to help resettle the families. Hence, it is not only desirable to respect an affiliate's annual quota but also to ensure smooth matching to avoid creating backlog for such services; and (iv) finally, cases are admitted to the U.S. and then allocated to agencies over time; consequently, the agencies need to make matching decisions *upon arrival of the case* and without foreknowledge on future arrivals.

Lack of information about future arrivals presents a major challenge in optimizing the matching decisions. One approach to address this challenge is to simulate future arrival patterns using data from previous years. As we will discuss later, several recent papers adopt such an approach; see, for example, the work of Bansak and Paulson (2022) and Ahani et al. (2023). Even though natural, the performance of such an approach falters if the pool's composition changes across years, a phenomenon observed in reality. To demonstrate this, in Figure 1, we focus on five selected affiliates of our partner agency. For each affiliate, we plot the (normalized) number of cases with U.S. ties corresponding to that affiliate across three years 2014-2016. We observe a substantial fluctuation in the number of such cases across these years; for example, the portion of cases tied to affiliate 16 increased by 55% in the year 2015 compared to 2014. This can have significant implications for matching decisions: as a toy example, consider affiliate 16 and suppose we are making decisions in the year 2015. If we use data from 2014 to simulate the number of cases with U.S. ties to this affiliate, we are substantially underestimating this number. Given that these cases can only be matched to affiliate 16, we may reach the affiliate's quota well before the end of the year.

The above background and observations motivate our main research question:

How to design a dynamic matching algorithm that does not rely on past years' data yet optimizes for employment outcomes while respecting the affiliates' quotas and desire for a smooth workload?

1.2. Our Contributions

Motivated by the above question, we introduce a framework for dynamic matching with post-allocation services. We then develop two learning-based algorithms that do not rely on any prior distributional knowledge, such as data from past years. Under the assumption that the arrival sequence is independent

Electronic copy available at: https://ssrn.com/abstract=4748762

and identically distributed (i.i.d.) from an unknown distribution, we show that both algorithms are asymptotically optimal in certain regimes. Further, using our partner agency's data, we conduct a comprehensive case study and show that our method outperforms the existing ones.

A Model for Dynamic Matching with Post-allocation Service (Section 2): Following our motivating application, we use terminology from refugee resettlement to introduce our model. We consider a discrete-time, finite-horizon model (e.g., one year) where, at each period, one refugee case arrives. Each case is either "tied" to a target affiliate (to capture the U.S. ties described in Section 1.1) or "free" to be matched with any affiliate. Upon arrival, the agency observes a *reward* (e.g., employment probability predicted from ML models) of matching the case to each affiliate.

We model each affiliate as a *static* resource with a fixed capacity (representing its annual quota). Each case consumes one unit of the static resource.[2] The innovation in our model lies in introducing *post-allocation* services, capturing the need for services *after matching* in applications like refugee resettlement. To model this, we endow each affiliate with a server, also called a *dynamic resource*, whose availability follows an i.i.d. Bernoulli process with a given service rate.[3] Upon the server's availability, a matched case (if any) is served on a first-come-first-served (FCFS) basis. We measure the server's congestion through its *backlog*, representing the number of matched cases (to its corresponding affiliate) awaiting service. The outlined service process resembles a classical queuing system, with a key distinction: the arrival rate to this (hypothetical) queuing system is *endogenously* determined by matching decisions. Despite this distinction, a congested system is still undesirable. In making matching decisions, the agency pursues multiple objectives: maximizing the total matching reward while avoiding excessive congestion for dynamic resources and over-allocation for static ones (i.e., going above an affiliate's annual quota).[4] To that end, we define an objective function consisting of the total reward penalized by *time-averaged* backlog for dynamic resources and total over-allocation for static ones.

As discussed in Section 1.1 (and evidenced in Figure 1), the pool of refugees arriving in a given year may substantially differ from previous years. Consequently, we study the above dynamic matching problem with the assumption of *no knowledge* about the underlying arrival distribution. However, in our theoretical analysis, we assume that the arrival sequence follows an (unknown) i.i.d. process. To evaluate the performance of any algorithm, we compare it to a properly defined benchmark (see Definition 2) and introduce the notion of regret relative to such a benchmark. With this framework, we aim to design algorithms with regret growing sub-linearly with the length of the horizon. En route to this goal, we first establish an impossibility result that implies if the penalty parameter for congestion is "too large," no algorithm can achieve sub-linear regret (see Proposition 1). Excluding this regime, in Sections 3 and 4, we develop two algorithms with a sub-linear regret guarantee.

[2] In practice, a case may consist of a family and therefore consume more than one unit of capacity. For the sake of simplicity, we assume a case size of one. Our results can be straightforwardly generalized to varying case sizes.

[3] As we discuss in Section 2, this model is equivalent to having geometric service time with the same service rate.

[4] Due to the presence of tied cases, over-allocation may be unavoidable (see Section 2).

Electronic copy available at: https://ssrn.com/abstract=4748762

Before providing an overview of our algorithms, we note that, in the absence of dynamic resources, the above problem reduces to online resource allocation (of static resources) — a topic extensively explored in the literature (see Section 1.3 for further discussion). However, the introduction of congestion for dynamic resources changes the nature of the problem (see Proposition 2 and related discussions). Consequently, the design and analysis of algorithms requires new ideas and techniques. We highlight some of the main challenges and new ideas in the following.

Design and Analysis of Algorithms (Sections 3 and 4): The starting point of our algorithm design is to understand the benchmark solution: concretely, had we known the entire sequence of arrivals and server availability upfront, what optimization problem would we solve? In Definition 2, we formulate this optimization problem as a convex program. Notably, to keep track of the (nonlinear) backlog dynamics (see equation (1)), the program includes time-dependent constraints (i.e. one for each period and affiliate). Consequently, in the corresponding dual program, we have the time-varying dual variables for dynamic resources (see Proposition 2), in contrast to the time-invariance of the dual variables for static ones. Nevertheless, had we known the "optimal" dual variables, we could have utilized them to design a simple score-based algorithms. In the absence of such knowledge, we aim to *learn* those dual variables in real-time (i.e., as we make matching decisions).

The dynamic nature of backlog, which also results in the time-varying nature of the problem in the dual space, presents several challenges in the design and analysis of a learning-based algorithm and a major departure from the literature on online resource allocation (see Section 1.3 for more discussion). In our first algorithm (Algorithm 1), we circumvent these challenges by introducing a *surrogate dual* problem (Definition 4) that imposes a time-invariance condition for the dual variables associated with dynamic resources. This algorithm maintains estimates on the dual variables of the surrogate dual problem and makes matching decisions using dual-adjusted scores. The key ingredient of our algorithm is updating (i.e., learning) the dual variables. For those associated with static resources, we use classic techniques from online adversarial learning; however, for updating the dual variables associated with dynamic resources we utilize a crucial structural property that relates a certain learning rule with the backlog dynamics. The resulting adjustment has an interpretable form: we penalize the reward of each affiliate by a scaled version of its current backlog (see equation (9) and related discussions).

As our first technical contribution, we prove that this algorithm achieves a sub-linear regret for all regimes of interest (see Theorem 1 and Corollary 1). To establish the result, we combine techniques from adversarial online learning and the drift-plus-penalty method (Neely 2022), by defining a proper potential function and pseudo-rewards to analyze the regret (see Section 3.3). The novel integration of these two separate frameworks, which is the first of its kind to the best of our knowledge, enables us to obtain a simultaneous guarantee on the algorithm's total matching reward (Lemma 1) and the average backlog in expectation (Lemma 2).

Electronic copy available at: https://ssrn.com/abstract=4748762

We call our first algorithm "*congestion-aware* dual learning" (`CA-DL`) as it requires information of the current backlog. While this is not a practical concern for our partner agency (see Section 5), such information may not be available in other contexts. Motivated by this, we develop a second algorithm that is oblivious to the congestion level. Our second algorithm is based on introducing a *surrogate primal* problem (Definition 5) that disregards the penalty for congestion in its objective. For this problem, we design a dual-adjusted score-based algorithm (Algorithm 2) similar to `CA-DL`, with a few subtle differences in the learning process. We call this second algorithm "*congestion-oblivious* dual-learning" (`CO-DL`), as it operates without knowing the level of current backlog.

Somewhat surprisingly, we show that even though `CO-DL` does not account for the backlog, the algorithm achieves a sub-linear regret under most regimes and some regularity conditions (Theorem 2). The main technical challenge of establishing the result lies in analyzing `CO-DL`'s backlog (Lemma 6), which in turn requires studying the "endogenous" arrival rates induced by `CO-DL`. We overcome this challenge by establishing a high-probability last-iterate convergence of the dual variables constructed by `CO-DL` (Proposition 3). Our proof, while involving many intricacies, relies on (i) establishing a connection between `CO-DL` and online stochastic mirror descent (OSMD) for solving a stochastic optimization problem, and (ii) adapting a recent high-probability last-iterate convergence result on stochastic gradient descent from Harvey et al. (2019) to our variants of OSMD (see Section 4.2.1). However, even though both algorithms achieve sub-linear regrets in many regimes, we show that under certain conditions (i.e., slow service rates and some range of penalty parameters) `CO-DL` cannot achieve a sub-linear regret (Theorem 3) unlike `CA-DL`, implying the inherent advantage of explicitly accounting for the backlog.

Finally, we note that our algorithms are computationally fast and their score-based nature makes them easy to communicate with practitioners (including the managers of our partner agency). Additionally, in the context of our collaboration, as we overview next, our learning-based method outperforms existing ones (that rely on data of past years) when tested on data from our partner agency.

Case Study on Refugee Resettlement Data (Section 5): To demonstrate the practical effectiveness of our learning-based approach, we conduct a case study based on our partner agency's data. In Section 5.1, we detail the construction of primitives and how to adapt our learning-based approach to this specific context. In Section 5.2, we show that had the agency used our proposed algorithm on the actual sequence of arrivals, it would have resulted in substantial improvement in terms of employment outcomes and average backlog. Furthermore, we compare our proposal with another recent proposal based on Bansak and Paulson (2022) (that relies on past year's data to predict future arrivals). Compared to the alternative proposal, we show that ours improves the objective function (that combines the three outcomes of total employment, backlog, and over-allocation) by $10\% - 50\%$ for a broad range of penalty parameters without sacrificing any of the three outcomes (see Table 1 along with Appendix F.3). This promising numerical result combined with the practical advantages of our method (mentioned earlier) has made it a strong contender for experimentation to replace the current system.

Electronic copy available at: https://ssrn.com/abstract=4748762

Moving beyond refugee resettlement, post-allocation services may arise in other assignment problems in contexts, such as foster care and healthcare, where individuals not only consume long-term resources with fixed capacities (such as beds) but also require time-consuming on-boarding processes (such as initial screenings). Similarly, in these contexts, undesirable backlog may arise due to a shortage of dynamic resources. (In the context of emergency departments, see Shi et al. (2016) that discusses post (bed) allocation delay caused by factors such as nurse shortage.) Our framework gives the managers the flexibility to balance the reward of allocation and the cost of (static and dynamic) resources to varying degrees. Further, the computational efficiency of our algorithms enables the policymaker to understand the trade-offs among these objectives by trying different penalty parameters.

1.3. Further Related Work

Our work relates to and contributes to several streams of literature.

Matching Policies for Refugee Resettlement: There has been growing academic interest in designing matching policies for refugee resettlement, leading to the emergence of two main categories. The first category focuses on designing a centralized *preference-based* matching system that can respect preferences of refugees and/or affiliates (see, for example, the work of Nguyen et al. (2021) and Delacrétaz et al. (2023) and references therein). While this approach has promise, most of the current programs, including the U.S. system, lack systematic data on preferences. Instead, they operate as a centralized outcome-based system, focusing on integration outcomes such as employment within 90 days.

As mentioned in Section 1.1, central to the outcome-based approach is the development of ML models that predict employment outcomes for a pair of case and affiliate (Bansak et al. 2018). Equipped with such models, several recent papers optimized matching decisions to improve employment, taking into account some of the operational considerations discussed in Section 1.1 (Gölz and Procaccia 2019, Ahani et al. 2021, Bansak and Paulson 2022, Ahani et al. 2023). Several key factors set our paper apart from the past work: (i) while these impactful papers focused on designing effective heuristics, we take a theoretical approach and propose algorithms with provable guarantees.[5] (ii) the existing works designed algorithms that crucially rely on a pool of past years' data to either estimate the arrival distribution or simulate future arrivals. However, as elaborated in Section 1.1, such an approach may be *non-robust* to discrepancies between the past years' data and the current year's arrivals. In contrast, our algorithms are robust to such discrepancies as they do not rely on data from previous years. (iii) as a result of such robustness and in addition to enjoying theoretical guarantees, our method outperforms its counterpart in practical situations as we discuss next.

Among the existing work, the work of Bansak and Paulson (2022) is the most appropriate point of comparison as they consider an objective that also penalizes congestion. However, tied cases are not

[5] We note that Freund et al. (2024) develops algorithms incorporating a notion of group fairness with a theoretical performance guarantee. However, this paper also assumes full knowledge of arrival distribution.

Electronic copy available at: https://ssrn.com/abstract=4748762

modeled and thus the over-allocation cost is not included in the objective. We modify their proposed sampling-based algorithm to incorporate tied cases and then compare its performance to that of our method. As we show in Section 5, when tested on real data, our method performs substantially better thanks to its robustness.

Online Resource Allocation: Online resource allocation problems have been extensively studied in operations and computer science. Given the broad scope of the literature, we only highlight the stream most closely related to our setting and kindly refer readers to Mehta et al. (2007) and Balseiro et al. (2023a) for an informative survey.[6]

As mentioned in Section 1.2, without dynamic resources, our setting reduces to the online resource allocation problem with an unknown i.i.d. arrival model. Several papers designed and analyzed algorithms for such settings. The primary intuition guiding the design is that, under the stochastic input model, we can *learn* the arrival pattern by observing a small portion of the arrival data. Building on this intuition, earlier studies have explored dual-based (Devanur and Hayes 2009, Feldman et al. 2010, Agrawal et al. 2014, Li and Ye 2022) or primal-based (Kesselheim et al. 2018) methods that periodically solve a linear program using the observed data.

Recognizing the computational costs associated with resolving, recent work has focused on designs that circumvent the need for resolving (Devanur et al. 2011, Agrawal et al. 2014, Gupta and Molinaro 2016, Balseiro et al. 2021, 2023b). These papers use adversarial online learning to efficiently *learn* the dual variables of the associated offline problem. Closest to our work is that of Agrawal et al. (2014) and Balseiro et al. (2021). These studies consider online resource allocation with an assumption that a (non-linear) convex objective can only depend on *total* allocation of resources. Due to this assumption, the offline problem has only a few time-invariant dual variables which can then be learned fast enough to ensure low regret. We leverage similar ideas to the aforementioned papers to learn the dual variables corresponding to static resources. However, such assumption *fails* for the average backlogs in our objective because the associated dual variables are time-varying and thus cannot be learned using similar ideas. This subtle difference led us to go well beyond the machinery of the aforementioned papers and to bring new tools from drift analysis and stochastic first-order methods (notably their high probability last-iterate convergence property) to design and analyze low-regret algorithms.

Related to the above, we also highlight that due to the backlog penalty term, our objective function does not satisfy the "exchangeability" condition imposed in the recent work of Freund and Banerjee (2019). This condition posits that the order and timing of actions do not impact the objective function's value. However, the backlog is inherently dynamic, and thus changing the order of arrivals or actions can significantly change the objective function.

[6] Our dynamic resources are also similar to the reusable resources studied in recent literature (see, for example, Feng et al. (2022) and references therein). However, the models in this stream of work do not allow for waiting for such resources.

Electronic copy available at: https://ssrn.com/abstract=4748762

Dynamic Matching and Control of Queueing Systems: In queuing literature, a stream of work studies matching queues; in these systems, agents arrive over time and wait to be matched according to a deterministic compatibility structure (Aveklouris et al. 2021, Afeche et al. 2022, Gupta 2022, Kerimov et al. 2023, Wei et al. 2023) or a random one (Ashlagi et al. 2013, Anderson et al. 2017, Ashlagi et al. 2019). While most of these papers focus on maximizing total (match-dependent) rewards, the work of Aveklouris et al. (2021) considers an objective similar to ours that penalizes the matching reward by the cost of waiting. However, the crucial distinction of our paper from the aforementioned works is the *reverse order of matching and waiting*. As discussed in Section 1.2, our model can be viewed as a queuing system with an endogenous arrival rate controlled by the matching process. Once again, this subtle difference presents technical challenges in, for example, controlling the size of the backlog.

From a methodological standpoint, the design and analysis of our first algorithm (Algorithm 1) partly rely on the celebrated drift-plus-penalty method (Neely 2022) for stochastic network optimization. We remark that, while in a different context, Wei et al. (2023) also integrates the method with the resolving techniques. Our work complements this approach by combining the drift-plus-penalty method with adversarial online learning. Additionally, our first algorithm implicitly utilizes the scaled backlog as a dual variable, establishing a crucial structural connection to the projected subgradient method (Zinkevich 2003). A similar connection was observed in Ashlagi et al. (2022), albeit in a context different from the setup of our paper.

2. Model and Preliminaries

In this section, we formally introduce our model, specify the objective function, and define the notion of regret for performance evaluation. We conclude by establishing an impossibility result (Section 2.1) that sheds light on the regime of penalty parameters that admit low-regret algorithms.

Notation: Throughout, a "constant" refers to any scalar independent of T. We use $[m]$ to denote a set $\{1, 2, ..., m\}$ for a positive integer m. The positive part of scalar x is denoted by $(x)_+$, and an indicator function is denoted by $\mathbb{1}[\cdot]$. We use bold cases to denote vectors. The non-negative orthant in the m-dimensional Euclidean space is denoted by $\mathbb{R}_+^m$, with $\mathbf{e}_i \in \mathbb{R}_+^m$ denoting the standard basis vector at coordinate i. Finally, we use $\Delta^m := \{\mathbf{z} \in \mathbb{R}_+^m : \sum_{i=1}^m z_i \leq 1\}$ to denote a convex hull of $\{\mathbf{0}, \mathbf{e}_1, ..., \mathbf{e}_m\}$.

To further simplify the presentation of our analytic results, we use the standard asymptotic notation. For functions $f, g : \mathbb{R} \to \mathbb{R}$, we write $f(x) = \mathcal{O}(g(x))$ (resp. $f = \Omega(g(x))$) if $|f(x)|$ is upper-bounded (resp. lower-bounded) by a positive constant multiple of $g(x)$ for all sufficiently large values of x. If $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$, we write it as $f(x) = \Theta(g(x))$. In the similar vein, we write $f(x) = o(g(x))$ if $\frac{f(x)}{g(x)} \to 0$ as $x \to \infty$.

Problem Setup: The problem consists of m affiliates and an arrival sequence of T refugee cases, where case t arrives at the beginning of period t. Upon arrival of case t, the agency observes its type denoted by $\mathbf{A}_t := (\mathbf{w}_t, i_t^\dagger)$, which we assume is an i.i.d. random variable from an *unknown* distribution

Electronic copy available at: https://ssrn.com/abstract=4748762

$\mathcal{F}$. The first component of the type is a vector $\mathbf{w}_t = (w_{t,i})_{i\in[m]}$, where $w_{t,i}$ is a case-affiliate pairwise *reward* from matching case t to affiliate i. Without loss of generality, we assume $w_{t,i} \in [0,1]$. The second component of the type is a *target affiliate* denoted by $i_t^\dagger$, which is a predetermined affiliate that the case must be matched to. If $i_t^\dagger \in [m]$, arrival t is a *tied* case targeted to affiliate $i_t^\dagger$. In contrast, we use $i_t^\dagger = 0$ to denote a *free* case, which does not have any target affiliate.

Upon observing $\mathbf{A}_t$, the agency makes an irrevocable matching decision denoted by $\mathbf{z}_t = (z_{t,i})_{i\in[m]}$. To succinctly represent the feasibility set for both tied and free cases, for a given target affiliate $i^\dagger$, we define a *type-feasibility* set $\mathcal{Z}(i^\dagger)$ as $\mathcal{Z}(i^\dagger) = \Delta^m$ if $i^\dagger = 0$ and $\mathcal{Z}(i^\dagger) = \mathbf{e}_{i^\dagger}$ otherwise. With this notation, the agency chooses the matching decision $\mathbf{z}_t \in \mathcal{Z}(i_t^\dagger)$. Some comments are in order. First, the type-feasibility set allows for fractional allocation, but our algorithms always make a binary decision. Secondly, while inaction (i.e., a zero vector) is not permitted for tied cases, we allow it for free cases for technical reasons.[7] For ease of exposition, we interpret inaction as matching the case to a dummy affiliate with zero reward and unlimited resources.

Resource Constraints: Each affiliate is endowed with two types of resources: static and dynamic. In the following, we elaborate on each type of resource, associated constraints, and cost for the agency.

(i) Static Resource (Capacity): Each affiliate i is endowed with capacity c_i, referred to as a *static* resource, which represents its annual quota. Matching a case to an affiliate consumes one unit of its static resource. We further define $\rho_i = c_i/T$, referred to as the *capacity ratio*. We use $\underline{\rho} = \min_{i\in[m]} \rho_i$ to denote the minimum capacity ratio. We impose two mild assumptions on the capacity ratio: (i) there exists a constant $d > 0$ such that $\rho_i - \mathbb{P}[i_t^\dagger = i] \geq d$ for all $i \in [m]$, meaning that, in expectation, each affiliate has capacity for free cases.[8] (ii) without loss of generality, we assume $\sum_{i=1}^m \rho_i \leq 1$.

The agency aims to respect the endowed capacity as much as possible. Without tied cases, the standard packing constraint ensures that $\sum_{t=1}^T z_{t,i} \leq c_i$, for all $i \in [m]$. However, due to uncertainty in the number of tied cases arriving over the horizon, such constraint is too stringent for our setting. To address this issue, we first introduce a weaker hard constraint: the total number of *free* cases cannot exceed the initial capacity (i.e., $\sum_{t=1}^T \mathbb{1}[i_t^\dagger = 0] z_{t,i} \leq c_i$, for all $i \in [m]$). Because $\mathbf{0} \in \mathcal{Z}(0)$ by our earlier assumption, this constraint can always be satisfied for any arrival sequence. Furthermore, it reduces to the standard packing constraint for affiliates that do not face tied cases. In addition, we introduce a soft constraint where the agency pays $\alpha \geq 0$ for each unit of over-allocation. That is, for each affiliate i, it must pay an over-allocation cost of $\alpha\Big(\sum_{t=1}^T z_{t,i} - c_i\Big)_+$. We will often use (total) *net matching reward* to refer to the total reward penalized by the over-allocation cost, i.e., $\sum_{t=1}^T \sum_{i=1}^m w_{t,i} z_{t,i} - \alpha \sum_{i=1}^m \Big(\sum_{t=1}^T z_{t,i} - c_i\Big)_+$. Here, α is a penalty parameter whose intensity represents

[7] In our case study (Section 5), we relax this assumption and ensure that every refugee is matched with an actual affiliate.

[8] Note that this implies that $\boldsymbol{\rho}$ is also a constant. In practice, the capacities are determined by first creating a proportionality key across affiliates and hence scale with T by construction.

Electronic copy available at: https://ssrn.com/abstract=4748762

the agency's tolerance or physical cost of going over the capacity.[9] Throughout the paper, we assume $\alpha > 1$ to ensure that the cost of matching any case beyond the capacity exceeds the associated reward (because $w_{t,i} \leq 1$). Consequently, had we known the entire arrival sequence upfront and were we maximizing the net matching reward, we would never over-allocate.[10]

(ii) Dynamic Resource (Server): As we motivated earlier, each case further requires *post-allocation* services. To capture this servicing process, we endow each affiliate with a *server*. Upon matching, the case will wait to avail service. Each server is a *dynamic* resource in that its availability changes stochastically over time. In particular, we assume that the service availability of affiliate i follows an i.i.d. Bernoulli process with a success probability of r_i, denoted by $s_{t,i} \sim \texttt{Ber}(r_i)$. We also refer to r_i as a *service rate*. At the end of period t, if the server is available ($s_{t,i} = 1$), it will process one case awaiting (if any) in the FCFS manner. The outlined process is equivalent to a (more conventional) service model with random service time. Specifically, each service time at affiliate i is independently drawn from a geometric distribution with a mean of $1/r_i$. During each period, if the server is available, it initiates serving a new case and remains unavailable until that service is completed.[11]

As motivated in Section 1, congestion of servers is undesirable. To formalize the notion of congestion, we define *backlog*, denoted by $\mathbf{b}_t = (b_{t,i})_{i\in[m]}$, where $b_{t,i}$ is the number of cases waiting for service at affiliate i and the end of period t. Formally, the backlog is recursively given by

$$b_{t,i} = (b_{t-1,i} + z_{t,i} - s_{t,i})_+ \tag{1}$$

with the zero initial backlog ($\mathbf{b}_0 = \mathbf{0}$). The backlog grows large when the affiliate becomes "congested" due to a bursty matching pattern. Hence, we use the time-average backlog $\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{m} b_{t,i}$ as the overall measure of congestion. We further assume that the agency pays a cost of γ for each unit of the (time) average backlog. That is, at the end of the horizon, the agency pays a cost of $\frac{\gamma}{T}\sum_{t=1}^{T}\sum_{i=1}^{m} b_{t,i}$. Similar to the role of α, the congestion penalty γ represents the agency's tolerance for congestion.

In this paper, we assume that the service rate is given by $r_i \geq \rho_i + \epsilon$ for all $i \in [m]$ where $\epsilon \geq 0$ is referred to as *service slack*. Assuming $r_i \geq \rho_i$ is well-motivated both technically and practically. From a technical perspective, the assumption comes from the following "stability sanity check": suppose that we match all cases to actual affiliates with no over-allocation. Then the total number of "arrivals" into affiliate i is $\rho_i T$. To ensure no backlog at the end of the horizon, in expectation, we need at least $\rho_i T$ periods during which the server is available, requiring that $r_i \geq \rho_i$. In practice, each affiliate's quota is (in part) determined based on its service resource availability and hence it is roughly proportional to

[9] For instance, the State Department provides per-capita funding to the resettlement agency in proportion to the approved capacity (U.S. State Department 2023). Consequently, the penalty parameter α carries a straightforward meaning as burden of securing extra funding beyond its initial budget.

[10] Unless the total number of tied cases for an affiliate exceeds its capacity.

[11] If the server is available with no case waiting, it will start serving an "outside" case with the same service time distribution.

Electronic copy available at: https://ssrn.com/abstract=4748762

the amount of service resources (Bansak and Paulson 2022). Regarding the range of ϵ, we distinguish two regimes formally defined as follows when we evaluate our proposed algorithms.

DEFINITION 1 **(Stable vs. Near-Critical Regimes)**. *Given $\epsilon \geq 0$, we categorize the problem into the stable regime if $\epsilon = \Omega(1)$ and the near-critical regime if $\epsilon = \mathcal{O}(1/\sqrt{T})$.*

Information Setting: Before formalizing the agency's objective, we discuss the information available to the agency for decision-making. First, when making a matching decision for case t, the exact realization of server availability $s_{t,i}$ is *not known*. This reflects the uncertainty in service time. However, in regards to fine-grained information about the backlog, we distinguish two cases. In the *congestion-aware* setting (Section 3), at the time of decision-making for case t, the agency is aware of the current backlog $\mathbf{b}_{t-1}$. By contrast, in the *congestion-oblivious* setting (Section 4), the agency does not have any information on the backlog.

Objective: We now formally define the agency's matching process and its objective. The agency employs a dynamic matching algorithm π which, upon arrival of case t, makes an immediate and irrevocable matching decision $\mathbf{z}_t^\pi$. The decision for case t can only depend on the history $\mathcal{H}_{t-1} := \{\mathbf{A}_\tau, \mathbf{s}_\tau, \mathbf{z}_\tau^\pi\}_{\tau=1}^{t-1}$ and the current arrival type $\mathbf{A}_t$. To maintain consistency with the literature, hereafter, we refer to such a dynamic matching algorithm π as an *online* algorithm. For a given penalty parameters α and γ, the agency's objective is given by:

$$\mathsf{ALG}^\pi(\alpha,\gamma) := \sum_{i=1}^{m}\sum_{t=1}^{T} w_{t,i} z_{t,i}^\pi - \alpha \sum_{i=1}^{m} \left(\sum_{t=1}^{T} z_{t,i}^\pi - c_i \right)_+ - \frac{\gamma}{T} \sum_{i=1}^{m}\sum_{t=1}^{T} b_{t,i}^\pi. \tag{2}$$

We refer to $\{\mathbf{z}_1^\pi, ..., \mathbf{z}_T^\pi\}$ (i.e., a sequence of matching decisions across the horizon) as a *matching profile*. As we highlighted earlier, the matching profile must conform to (i) $\mathbf{z}_t^\pi \in \mathcal{Z}(i_t^\dagger)$ for all $t \in [T]$ and (ii) the hard capacity constraint $\sum_{t=1}^{T} \mathbb{1}[i_t^\dagger = 0] z_{t,i}^\pi \leq c_i$ for all $i \in [m]$. We say that a matching profile is *feasible* if it satisfies the two preceding conditions.

Performance Metric: We compare the performance of any online algorithm against an optimal offline benchmark which solves the same optimization problem with foreknowledge of the sequence of arrivals and service availability upfront, formally defined as follows:

DEFINITION 2 **(Optimal Offline Benchmark)**. *Knowing the sample-path $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^T$ of the arrival and service sequences, the optimal offline benchmark solves the following convex problem:*

$$\begin{aligned}
\mathsf{OPT}(\alpha,\gamma) \quad := \quad & \max_{\substack{\mathbf{z}_t \in \mathcal{Z}(i_t^\dagger) \\ \mathbf{b}_t \geq \mathbf{0}}} \sum_{i=1}^{m}\sum_{t=1}^{T} w_{t,i} z_{t,i} - \alpha \sum_{i=1}^{m} \left(\sum_{t=1}^{T} z_{t,i} - c_i \right)_+ - \frac{\gamma}{T} \sum_{i=1}^{m}\sum_{t=1}^{T} b_{t,i} \\
& s.t. \quad \sum_{t=1}^{T} \mathbb{1}[i_t^\dagger = 0] z_{t,i} \leq c_i \qquad \forall i \in [m] \qquad \text{(Packing)} \\
& \qquad b_{t,i} \geq b_{t-1,i} + z_{t,i} - s_{t,i} \qquad \forall t \in [T], i \in [m] \qquad \text{(Backlog Inequality)}
\end{aligned}$$

where we define $b_{0,i} = 0$ for all $i \in [m]$ by convention.

Electronic copy available at: https://ssrn.com/abstract=4748762

It is straightforward to see that, at any optimal solution, at least one of $b_{t,i} \geq 0$ or constraint (Backlog Inequality) is binding for all $t \in [T]$ and $i \in [m]$. Therefore, the optimal solution for $b_{t,i}$ satisfies the backlog dynamic given in equation (1). With a slight abuse of terminology, we use $\mathsf{OPT}(\alpha, \gamma)$ to denote the offline program and its optimal objective value itself.

With the above benchmark, we evaluate an algorithm by its worst-case performance over all instances, where each instance $\mathcal{I}$ consists of (i) the set of affiliates $[m]$ (ii) capacity ratio $\boldsymbol{\rho}$, and (iii) the (unknown) arrival type distribution $\mathcal{F}$. To this end, we define *regret*, measuring the worst-case difference between the expected objective value of the algorithm and the optimal offline:

DEFINITION 3 **(Worst-case Regret).** *The (worst-case) regret of an online algorithm π is given by*

$$\mathsf{Regret}_T^{\pi} := \sup_{\mathcal{I}} \mathbb{E}\left[\mathsf{OPT}(\alpha, \gamma) - \mathsf{ALG}^{\pi}(\alpha, \gamma)\right]. \tag{3}$$

where the expectation is over the arrival distribution $\mathcal{F}$, the Bernoulli service process with a service rate vector $\mathbf{r}$, and (potentially) randomness of the algorithm itself.[12]

2.1. Lower Bound on Achievable Regret

Given our benchmark and notion of regret, our goal is to design algorithms that are asymptotically optimal, meaning that their regret grows sub-linearly with T. However, this goal may not be achievable in certain regimes due to the inherent limitation of making online decisions. The following proposition formalizes this limitation when γ (the congestion penalty parameter) is too large.

PROPOSITION 1 **(Lower Bound on Achievable Regret).** *For any given $\epsilon \geq 0$ and $\gamma = \Omega(T)$, there exists an instance for which the regret (Definition 3) of any online algorithm is $\Omega(T)$.*

We prove Proposition 1 in Appendix A, and provide a proof sketch in the following. For the proof, we construct a one-affiliate instance. For brevity, we omit the subscript for i. Each case t's type is $\mathbf{A}_t = (1, 0)$ — each arrival is a free case with the reward of one. The capacity ratio is $\rho = 0.5$ and the service rate is $r = \rho + \epsilon$ where $\epsilon \in [0, 1 - \rho)$. Note that, because we only have one actual affiliate, the decision is whether to match a case to this affiliate or not. In this instance, the offline can leverage its foreknowledge of service availability to avoid backlogs by matching the case to the actual affiliate only when the server is available. In the full proof, we show that doing so also gives the expected total reward at least $0.5T - \Theta(\sqrt{T})$ with zero average backlog. In contrast, because any online algorithm does not know the current service availability, it must incur an unavoidable backlog without generating more rewards. Based on this intuition, we show that any online algorithm that earns $\Theta(T)$ reward cannot avoid $\Theta(\gamma)$ cost of congestion, which leads to the desired result.

[12] For brevity, we omit the dependence of expectation on the arrival and service distributions.

Electronic copy available at: https://ssrn.com/abstract=4748762

3. Algorithm Design for Congestion-Aware Setting

In this section, we propose our first learning-based algorithm for the congestion-aware setting. In Section 3.1, we first study the dual program of the optimal offline (Definition 2) and introduce our learning-based approach. We then formally introduce the congestion-aware dual-learning algorithm (`CA-DL`) in Section 3.2 and establish the regret guarantee of `CA-DL`. Finally, we provide an overview of our proof technique in Section 3.3.

3.1. Motivation: Surrogate Dual Problem

Our first step is to study the optimal offline benchmark through duality. The following proposition, which we prove in Appendix B, characterizes the dual program of the optimal offline.

PROPOSITION 2 **(Dual of Optimal Offline)**. *For any given sample path* $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^T$*, consider the following dual program:*

$$\begin{aligned}
\mathsf{Dual}(\alpha,\gamma) := \min_{\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\beta}_t \geq \mathbf{0}} \sum_{t=1}^{T} \Big\{ \max_{\mathbf{z}_t \in \mathcal{Z}(i_t^\dagger)} \big(\mathbf{w}_t - \boldsymbol{\theta} - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda} - \boldsymbol{\beta}_t\big) \cdot \mathbf{z}_t + \boldsymbol{\rho} \cdot \boldsymbol{\theta} + \boldsymbol{\rho} \cdot \boldsymbol{\lambda} + \mathbf{s}_t \cdot \boldsymbol{\beta}_t \Big\} \\
s.t. \quad \theta_i \leq \alpha \quad \forall i \in [m] \qquad (4) \\
\beta_{t,i} - \beta_{t+1,i} \leq \frac{\gamma}{T} \quad \forall t \in [T-1], i \in [m], \qquad \beta_{T,i} \leq \frac{\gamma}{T} \quad \forall i \in [m] \qquad (5)
\end{aligned}$$

Then we have the following strong duality: $\mathsf{OPT}(\alpha,\gamma) = \mathsf{Dual}(\alpha,\gamma)$.

In Proposition 2, the dual variables ($\boldsymbol{\theta}$, $\boldsymbol{\lambda}$) are for static resources, respectively corresponding to the over-allocation cost and the hard constraint (Packing). On the other hand, the vector sequence $\{\boldsymbol{\beta}_t\}_{t=1}^T$ is the $m \times T$-dimensional dual variables for dynamic resources, each corresponding to (Backlog Inequality) in Definition 2. Note that each of $\boldsymbol{\beta}_t$ is time-dependent in contrast to $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Such time-varying nature stems from constraints (5), which couples two consecutive $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t+1}$.[13]

Proposition 2 implies a simple method for making optimal matching decisions given the optimal dual variables $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*, \boldsymbol{\beta}_t^*)$: choose decision $\mathbf{z}_t \in \mathcal{Z}(i_t^\dagger)$ that maximizes the *dual-adjusted score* $(\mathbf{w}_t - \boldsymbol{\theta}^* - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda}^* - \boldsymbol{\beta}_t^*) \cdot \mathbf{z}_t$. In light of this observation, we take a *learning-based* approach which iteratively updates the dual variables and chooses the matching decision that maximizes the dual-adjusted score based on the learned values. Because of the time-invariance of $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*)$ and the stationarity of arrivals and services, we can efficiently learn them by employing a first-order method that uses a new gradient of the dual function upon each arrival, similar to Agrawal and Devanur (2014). However, learning the time-varying sequence $\{\boldsymbol{\beta}_t^*\}_{t=1}^T$ with the inter-temporal constraints (5) seems challenging (and possibly infeasible). To overcome this challenge, we introduce a surrogate dual problem as follows.

[13] Indeed, by the complementary slackness, we can show that the optimal dual variable satisfies $\beta_{t,i}^* - \beta_{t+1,i}^* > 0$ if backlog of the optimal offline at period t and affiliate i is nonzero.

Electronic copy available at: https://ssrn.com/abstract=4748762

Algorithm 1 Congestion-Aware Dual-Learning (`CA-DL`) Algorithm

Input: T, $\boldsymbol{\rho}$, η, and ζ.
Initialize $\theta_{1,i} = \exp(-1)$, $\lambda_{1,i} = \exp(-1)$, $c_i^{\mathsf{free}} = \rho_i T$, and $b_{0,i} = 0$ for all $i \in [m]$.
for each arrival $\{\mathbf{A}_t\}_{t=1}^T$ **do**
  if $i_t^\dagger \in [m]$ **then**
    Set $\mathbf{z}_t = \mathbf{e}_{i_t^\dagger}$
  else
    $\mathbf{z}_t = \arg\max_{\mathbf{z} \in \Delta^m} (\mathbf{w}_t - \boldsymbol{\theta}_t - \boldsymbol{\lambda}_t - \zeta \mathbf{b}_{t-1}) \cdot \mathbf{z}$ s.t. $c_i^{\mathsf{free}} - z_{t,i} \geq 0 \quad \forall i \in [m]$
    Update the remaining capacity for the hard constraint: $c_i^{\mathsf{free}} \leftarrow c_i^{\mathsf{free}} - z_{t,i} \quad \forall i \in [m]$
  end if
  Update the backlog: $b_{t,i} = (b_{t-1,i} + z_{t,i} - s_{t,i})_+ \quad \forall i \in [m]$
  Update the dual variables corresponding to the static resource for all $i \in [m]$:

$$\theta_{t+1,i} = \min\{\theta_{t,i} \exp(\eta(z_{t,i} - \rho_i)), \alpha\}, \quad \lambda_{t+1,i} = \min\left\{\lambda_{t,i} \exp(\eta(\mathbb{1}[i_t^\dagger = 0] z_{t,i} - \rho_i)), \frac{1+2\alpha}{\underline{\rho}}\right\} \tag{8}$$

  `// The dual variable for the dynamic resource is implicitly updated as` $\beta_{t+1,i} = (\beta_{t,i} + \zeta(z_{t,i} - s_{t,i}))_+$ `with` $\beta_{0,i} = 0$, `resulting in` $\beta_{t+1,i} = \zeta b_{t,i}$.
end for

Definition 4 **(Surrogate Dual Problem)**. *For any given sample path* $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^T$, *define the per-period surrogate dual function for period* t *as*

$$g_t(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta}) := \max_{\mathbf{z}_t \in \mathcal{Z}(i_t^\dagger)} \left\{(\mathbf{w}_t - \boldsymbol{\theta} - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda} - \boldsymbol{\beta}) \cdot \mathbf{z}_t\right\} + \boldsymbol{\rho} \cdot \boldsymbol{\theta} + \boldsymbol{\rho} \cdot \boldsymbol{\lambda} + \mathbf{s}_t \cdot \boldsymbol{\beta}. \tag{6}$$

The surrogate dual problem is then defined by

$$\mathsf{Dual}^{\mathsf{S}}(\alpha) := \min_{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta} \geq \mathbf{0}} \sum_{t=1}^T g_t(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\beta}) \quad s.t. \quad \theta_i \leq \alpha \quad \forall i \in [m] \tag{7}$$

Compared to the original dual program, the surrogate dual problem addresses the time-varying nature of $\{\boldsymbol{\beta}_t^*\}_{t=1}^T$ by making them *identical* across the periods. As we will show later, the advantage of this surrogate dual problem is that it allows us to uncover a crucial structural connection between the trajectory of online projected subgradient descent for learning the stationary dual solution $\boldsymbol{\beta}^*$ (of the surrogate problem) and the backlog dynamics of each affiliate.

3.2. Algorithm and Analysis

We formally describe out first algorithm, called `CA-DL`, in Algorithm 1. On a high-level, `CA-DL` maintains the dual variables for the surrogate dual problem based on which it makes a primal matching decision. Notably, our algorithm uses the primal decision to obtain gradient information for updating the dual variables via online learning. It then explicitly updates the dual variables $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for static

Electronic copy available at: https://ssrn.com/abstract=4748762

resources using online mirror descent, and implicitly updates the dual variables $\boldsymbol{\beta}$ for dynamic resources (in the surrogate dual problem) using backlog dynamics. As we show later, the scaled backlog plays a natural role as the dual variable for the dynamic resource. To be precise, the algorithm takes the initial dual variables[14] and learning rates $\eta \geq 0$ for static resources and $\zeta \geq 0$ for dynamic resources. For each arrival t, the algorithm is divided into two phases.

Primal Phase (line 4-9): The algorithm selects the decision that maximizes the dual-adjusted score based on the current dual variables $(\boldsymbol{\theta}_t, \boldsymbol{\lambda}_t, \zeta \mathbf{b}_{t-1})$, subject to the hard constraint (Packing) and $\mathbf{z}_t \in \mathcal{Z}(i_t^\dagger)$. Said differently, the algorithm assigns the arriving free case to an affiliate i with capacity for free cases and maximum dual-adjusted score $w_{t,i} - \theta_{t,i} - \lambda_{t,i} - \zeta b_{t-1,i}$ (breaking ties arbitrarily). If there is no affiliate with remaining capacity for free cases and a non-negative adjusted score, the case will be assigned to the dummy affiliate whose dual adjusted score is always zero.

Dual Phase (line 11): Based on the primal decision $\mathbf{z}_t$, we update the dual variables using online learning. To illustrate the connection with online learning, we first focus on the updating rule (8). Note that $\boldsymbol{\rho} - \mathbf{z}_t$ is a subgradient of the per-period surrogate dual function g_t (see line (6) in Definition 4) with respect to (w.r.t.) $\boldsymbol{\theta}$ evaluated at the current dual variables $(\boldsymbol{\theta}_t, \boldsymbol{\lambda}_t, \zeta \mathbf{b}_{t-1})$. We use this gradient to perform the multiplicative weight update (Freund et al. 1997), which can be seen as a special case of online mirror descent. The same idea applies to the updating rule for $\boldsymbol{\lambda}$. In particular, $\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t$ is a subgradient of g_t w.r.t. $\boldsymbol{\lambda}$. The upper bound α in line (8) comes from the domain of $\boldsymbol{\theta}$ in the surrogate dual problem (7). The upper bound on $\boldsymbol{\lambda}$ is solely for a technical purpose.[15]

Our algorithm not only explicitly keeps track of the dual variables $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for the static resource, it also *implicitly* keeps track of the dual variable $\boldsymbol{\beta}$ (in the surrogate dual problem) for dynamic resources through the backlog dynamics. To see this, suppose that we were updating the sequence $\{\boldsymbol{\beta}_t\}_{t=1}^T$ with initial $\boldsymbol{\beta}_1 = \mathbf{0}$ using a slightly different (and simpler) updating rule, which is vanilla gradient descent (Zinkevich 2003) projected onto non-negative orthant $\boldsymbol{\beta}_t \geq \mathbf{0}$.[16] Note that $\mathbf{s}_t - \mathbf{z}_t$ is a subgradient of g_t w.r.t. $\boldsymbol{\beta}$. Using this method with a given learning rate ζ, we obtain the following updating rule:

$$\beta_{t+1,i} = (\beta_{t,i} + \zeta(z_{t,i} - s_{t,i}))_+. \tag{9}$$

We observe that, up to a scalar factor ζ, the updating rule (9) is equivalent to the backlog dynamics (see equation (1) in Section 2). Using induction over period t, we can show that $\beta_{t,i} = \zeta b_{t-1,i}$ for all $t \in [T]$ and $i \in [m]$. In other words, the scaled backlog takes on the role of the dual variable for

[14] Choosing the initial dual as $\exp(-1)$ for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is only for simplicity of explanation. The choice of the initial dual variables does not impact the analytical result regarding the order of regret.

[15] We can show that the dual variables $\boldsymbol{\lambda}_t$ do not exceed $\frac{1+2\alpha}{\underline{\rho}}$ without imposing such upper bound (see Appendix C.4)

[16] This update rule is still a special case of mirror descent when the mirror map $h(\mathbf{x})$ is chosen to be $\|\mathbf{x}\|_2^2$. Note that our multiplicative updating rule in (8) is the result of $h(\mathbf{x})$ being the negative entropy function, i.e., $h(\mathbf{x}) = \sum_i x_i \log(x_i)$.

Electronic copy available at: https://ssrn.com/abstract=4748762

dynamic resources. Intuitively, CO-DL manages congestion by "penalizing" the scaled version of the current backlog and hence avoiding matching arrivals to affiliates with large backlog.

We now formally analyze the regret of CA-DL, first focusing on the stable regime (Definition 1).

THEOREM 1 **(Regret of CA-DL under Stable Regime).** *Let* $\eta = \Theta(1/\sqrt{T})$ *and* $\zeta = \Theta(1/\sqrt{T})$. *Under the stable regime (Definition 1), the regret (Definition 3) of CA-DL is* $\mathcal{O}(\sqrt{T} + \frac{\gamma}{\epsilon})$.

Theorem 1 implies that, under the stable regime, the regret of CA-DL is sublinear in T unless $\gamma = \Omega(T)$. Combined with Proposition 1, Theorem 1 implies that CA-DL achieves a sub-linear regret for all feasible ranges of γ that allow for such a low regret. However, Theorem 1 does not cover the near-critical regime (Definition 1), where we recall $\epsilon = \mathcal{O}(1/\sqrt{T})$. We complement Theorem 1 with the following corollary, which we prove in Appendix C.5, establishing that CA-DL still achieves a sub-linear regret in the near-critical regime (with a difference choice of ζ).

COROLLARY 1 **(Regret of CA-DL under Near-Critical Regime).** *Let* $\eta = \Theta(1/\sqrt{T})$ *and* $\zeta = \Theta(\sqrt{\gamma/T})$. *Under the near-critical regime (Definition 1), the regret (Definition 3) of CA-DL is* $\mathcal{O}(\sqrt{\gamma T})$.

We dedicate the next subsection to outline the proof of Theorem 1. The proof of Corollary 1 uses similar ideas. The main challenge for establishing the regret guarantee lies in the fact that the objective function defined in (2) consists of two terms with fundamentally different natures, the net matching reward and time-average congestion. The innovation of our analysis is to (i) define pseudo-rewards that implicitly contain both parts and (ii) combine techniques from the celebrated drift-plus-penalty method (Neely 2022) and adversarial online learning to bound the pseudo-rewards based on the optimal offline solution (Definition 2) and establish a low regret. Roughly speaking, the two terms in our regret bound respectively correspond to the ones we separately establish for the net matching reward and congestion cost (see Lemmas 1 and 2 in Section 3.3).

3.3. Proof Sketch of Theorem 1

We first introduce some notations. Hereafter, a sample path refers to a sequence of arrivals and services over the horizon (i.e., $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^T$). For a given sample path, let $\{\hat{\mathbf{z}}_t\}_{t=1}^T$ denote an arbitrary feasible matching profile. Furthermore, we use $\{\mathbf{z}_t\}_{t=1}^T$ and $\{\mathbf{z}_t^*\}_{t=1}^T$ to denote the matching profile of CA-DL and the optimal offline benchmark (Definition 2), respectively. For brevity, we also use $\mathsf{NMR}(\cdot;\alpha)$ to denote the net matching reward of a feasible matching profile given α:

$$\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha) = \sum_{t=1}^{T}\sum_{i=1}^{m} w_{t,i}\hat{z}_{t,i} - \alpha \sum_{i=1}^{m}\left(\sum_{t=1}^{T}\hat{z}_{t,i} - c_i\right)_+. \tag{10}$$

Because the congestion cost is non-negative, we have $\mathsf{NMR}(\{\mathbf{z}_t^*\}_{t=1}^T;\alpha) \geq \mathsf{OPT}(\alpha,\gamma)$ for every cost parameter (α,γ) and every sample path. Hence, we have the following decomposition: for any arrival distribution $\mathcal{F}$,

$$\mathbb{E}[\mathsf{OPT}(\alpha,\gamma) - \mathsf{ALG}^{\texttt{CA-DL}}(\alpha,\gamma)] \leq \underbrace{\mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t^*\}_{t=1}^T;\alpha)] - \mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T;\alpha)]}_{\text{(A)=Loss of net matching reward}} + \gamma \cdot \underbrace{\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\Big]}_{\text{(B)=Average backlog}} \tag{11}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

In light of this decomposition, it suffices to upper bound (A) and (B) separately. The following two lemmas are the key results establishing the upper bound on each term.

LEMMA 1 **(Bounding Loss of Net Matching Reward).** *For any arrival distribution $\mathcal{F}$ and $\epsilon \geq 0$, we have*

$$\mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t^*\}_{t=1}^T;\alpha)] - \mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T;\alpha)] \leq \mathcal{O}(\sqrt{T}) \tag{12}$$

LEMMA 2 **(Bounding Average Backlog).** *For any arrival distribution $\mathcal{F}$ and $\epsilon > 0$, we have*

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\Big] \leq \mathcal{O}\Big(\frac{1}{\epsilon}\Big) \tag{13}$$

Combined with the decomposition (11), the two lemmas directly imply Theorem 1. We devote the rest of this section to sketch the proof of Lemmas 1 and 2. The proof consists of three steps. First, we define a proper pseudo-reward—a stochastic process that assists in comparing the objective value of `CA-DL` and that of the optimal offline. We emphasize that the pseudo-reward is defined only for an accounting purpose and is not an actual reward. In the second step, we establish lower and upper bounds on the pseudo-rewards that crucially use the design of `CA-DL`. The final step involves combining the preceding bounds to complete the proof of Lemmas 1 and 2.

Step 1: Defining a Pseudo-reward. Consider the following potential function and its drift.

$$\psi(\mathbf{b}_t) := \frac{1}{2}\|\mathbf{b}_t\|_2^2, \qquad D_t := \psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1}). \tag{14}$$

One simple yet powerful property of this potential function ψ is the following lemma.

LEMMA 3 **(Drift Lemma).** *For any sample path, we have $D_t \leq \mathbf{b}_{t-1}\cdot(\mathbf{z}_t - \mathbf{s}_t) + \mathcal{O}(1)$.*

We prove Lemma 3 in Appendix C.1. Lemma 3 states that the drift of the potential function ψ can be upper bounded by a linear function of the current backlog. With this background, we are ready to define the pseudo-reward at time t, denoted by K_t, as follows:

$$K_t := \mathbf{w}_t\cdot\mathbf{z}_t + \boldsymbol{\theta}_t\cdot(\boldsymbol{\rho} - \mathbf{z}_t) + \boldsymbol{\lambda}_t\cdot(\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t) - \zeta D_t \tag{15}$$

To understand our design of the pseudo reward, we observe that Lemma 3 implies

$$K_t \geq (\mathbf{w}_t - \boldsymbol{\theta}_t - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda}_t - \zeta\mathbf{b}_{t-1})\cdot\mathbf{z}_t + \boldsymbol{\theta}_t\cdot\boldsymbol{\rho} + \boldsymbol{\lambda}_t\cdot\boldsymbol{\rho} + \zeta\mathbf{b}_{t-1}\cdot\mathbf{s}_t - \mathcal{O}(\zeta) \tag{16}$$

for every sample path. Importantly, the first term $(\mathbf{w}_t - \boldsymbol{\theta}_t - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda}_t - \zeta\mathbf{b}_{t-1})\cdot\mathbf{z}_t$ is the adjusted score that `CA-DL` maximizes, which will enable us to utilize the optimality criterion of `CA-DL`.

Step 2: Lower- and Upper-bounding the Pseudo-rewards. The second step involves bounding the expected cumulative pseudo reward up to the last time that `CA-DL` is not constrained by the hard resource constraint. Formally, we define the *stopping time* as follows:

$$T_A := \min\Big\{t \leq T : \sum_{\tau=1}^{t}\mathbb{1}[i_\tau^\dagger = 0]z_{\tau,i} + 1 \geq c_i \text{ for some } i \in [m]\Big\}. \tag{17}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

For all $1 \le t \le T_A$, the primal phase of `CA-DL` (lines 4-9) can be succinctly written as

$$\mathbf{z}_t = \arg\max_{\mathbf{z} \in \mathcal{Z}(i_t^\dagger)} (\mathbf{w}_t - \boldsymbol{\theta}_t - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda}_t - \zeta \mathbf{b}_{t-1}) \cdot \mathbf{z}. \tag{18}$$

The following lemma lower-bounds the total expected pseudo-reward up to the stopping time (in part) via (i) the expected net matching reward of *any feasible matching profile* and (ii) the backlog of `CA-DL`. The fact that this lemma holds for any feasible matching profile plays a crucial role at the end of the proof for Lemmas 1 and 2.

LEMMA 4 **(Lower Bound on Pseudo-Rewards).** *For any feasible matching profile* $\{\hat{\mathbf{z}}_t\}_{t=1}^T$*, we have*

$$\mathbb{E}\Big[\sum_{t=1}^{T_A} K_t\Big] \ge \mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T; \alpha)] + \zeta\epsilon\mathbb{E}\Big[\sum_{t=1}^{T_A - 1} \|\mathbf{b}_t\|_1\Big] - (T - T_A) - \mathcal{O}(\zeta T) \tag{19}$$

We prove Lemma 4 in Appendix C.2. The proof crucially relies on the optimality criterion (18) of `CA-DL`. Specifically, we compare the decision of `CA-DL` at time t with that of a *static* control (see Claim 5 in Appendix C.2), whose value serves as an upper bound on the per-period net matching reward of any feasible matching profile (in expectation).

We now upper bound the total pseudo-rewards (in part) via the net matching reward of `CA-DL` in the following lemma.

LEMMA 5 **(Upper Bound on Pseudo-Rewards).** *For every sample path, we have*

$$\sum_{t=1}^{T_A} K_t \le \mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T; \alpha) - (T - T_A) - \zeta\psi(\mathbf{b}_{T_A}) + \mathcal{O}(\sqrt{T}) \tag{20}$$

The proof of Lemma 5 is presented in Appendix C.3. Similar to Agrawal and Devanur (2014) and Balseiro et al. (2023b), our proof leverages the adversarial learning guarantee, which holds for *every* sample path, of the online mirror descent used to update the dual variables for static resources (see updating rules (8) and related discussion).

Step 3: Putting Everything Together. Combining Lemma 4 and Lemma 5 and taking expectations, we reach to the following crucial inequality: for any feasible matching profile $\{\hat{\mathbf{z}}\}_{t=1}^T$,

$$\underbrace{\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T; \alpha)] - \mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T; \alpha)]}_{\text{(A): Lemma 1}} + \underbrace{\zeta\epsilon\mathbb{E}\Big[\sum_{t=1}^{T_A - 1} \|\mathbf{b}_t\|_1\Big] + \zeta\mathbb{E}[\psi(\mathbf{b}_{T_A})]}_{\text{(B'): Lemma 2}} \le \mathcal{O}(\sqrt{T} + \zeta T) \qquad (\star)$$

To complete the proof of Lemma 1, we set $\hat{\mathbf{z}}_t = \mathbf{z}_t^*$ for all $t \in [T]$ (i.e., the optimal offline allocation) and observe that $(\mathsf{B}') \ge 0$ because the backlog is non-negative. The proof is done by plugging $\zeta = \Theta(1/\sqrt{T})$ as stated in Theorem 1. To prove Lemma 2, we set $\{\hat{\mathbf{z}}_t\}_{t=1}^T$ to be the feasible matching profile that maximizes the net matching reward. By definition of such $\{\hat{\mathbf{z}}_t\}_{t=1}^T$, (A) is nonnegative and hence, (B') is upper bounded by $\mathcal{O}(\sqrt{T} + \zeta T)$. However, note that (B') only involves the backlog *up*

Electronic copy available at: https://ssrn.com/abstract=4748762

to the stopping time T_A. Hence, to translate the upper bound on (B') to Lemma 2, we additionally need to control the backlog accrued *after* the stopping time. In Appendix C.4, we complete this task by showing that (i) the expected backlog at the stopping time is $\mathcal{O}(\sqrt{T})$ and (ii) $\mathbb{E}[(T-T_A)^2] \le \mathcal{O}(T)$. The proof of the latter follows a similar idea from Balseiro et al. (2023b) to study the trajectory of $\boldsymbol{\lambda}_t$ given the fixed learning rate $\eta = \Theta(1/\sqrt{T})$. Together, these claims complete the proof of Lemma 2.

4. Algorithm Design for Congestion-Oblivious Setting

As evident in Algorithm 1, `CA-DL` requires up-to-date information on the backlog. In some contexts, having access to such information may not be practically feasible.[17] Motivated by this, we consider the congestion-oblivious setting (see Section 2) where we design the congestion-oblivious dual-learning algorithm (called `CO-DL`) that does not require any knowledge of backlogs. In Section 4.1, we motivate our algorithm design by introducing a surrogate *primal* problem. In Section 4.2, we formally present `CO-DL`. We then establish that, somewhat surprisingly, `CO-DL` achieves a sub-linear regret under the stable regime and some regularity conditions. In Section 4.3, however, we also show that under the near-critical regime and for a certain range of γ, `CO-DL` cannot achieve a sub-linear regret unlike `CA-DL`, highlighting the inherent benefit of accounting for the backlog.

4.1. Motivation: Surrogate Primal Problem

To further motivate our technical approach, consider a *fluid* approximation of the optimal offline by replacing the random arrivals and services $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^T$ with their expectations. In particular, r_i fraction of arrivals is deterministically processed for every period. Under this approximation, we have a simple *stationary* (fractional) solution $z^*_{t,i} = \rho_i$ that guarantees no backlog. This observation motivates us to investigate a sample path-based primal program that entirely disregards the backlog penalty, which we define as the *surrogate primal* problem.

Definition 5 **(Surrogate Primal Problem).** *The surrgoate primal problem is defined by*

$$\mathsf{Primal}^{\mathsf{S}}(\alpha) := \max_{\mathbf{z_t} \in \mathcal{Z}(i_t^\dagger)} \sum_{t=1}^{T}\sum_{i=1}^{m} w_{t,i} z_{t,i} - \alpha \sum_{i=1}^{m} \left(\sum_{t=1}^{T} z_{t,i} - c_i \right)_+ \quad s.t. \ \sum_{t=1}^{T} \mathbb{1}[i_t^\dagger = 0] z_{t,i} \le c_i \quad \forall i \in [m]$$

In the similar spirit of the surrogate dual problem (Definition 4), the surrogate primal also tackles challenges stemming from the time-varying nature of the optimal dual variables for dynamic resources in the dual of the optimal offline (see Proposition 2). More accurately, by disregarding the backlog penalty in the objective function, a learning-based algorithm for the surrogate primal ignores dual variables for dynamic resources and focuses solely on maintaining the dual variables for static resources.

The idea of using online learning for the surrogate primal has been investigated in the literature (see, for example, Agrawal and Devanur (2014)). However, despite the fluid approximation admitting

[17] One might conjecture that imperfect knowledge of service rates would suffice to "simulate" the backlogs and execute Algorithm 1. As we illustrate in Appendix D.2, such an approach can be sensitive to mis-specification of the service rates to the extent that using our congestion-oblivious algorithm results in better outcomes.

Electronic copy available at: https://ssrn.com/abstract=4748762

no backlog, it is not entirely clear if a learning-based algorithm for the surrogate primal can control backlog under random arrivals and services. In simple terms, to control the backlog, the algorithm's sequence of arrival rates induced by the matching decisions should converge to the stationary solution $\boldsymbol{\rho}$. In particular, a sufficiently strong convergence would be the one that is (i) fast enough, (ii) almost uniform across time, and (iii) happens with high probability. It turns out that, by considering proper time-dependent learning rates, we can achieve such strong convergence property.

4.2. Algorithm and Analysis under Stable Regime

In this section, we study a learning-based algorithm designed based on the surrogate primal problem, called `CO-DL`. The algorithm's architecture mirrors that of `CA-DL`, where decisions in the primal phase are based on current dual variables, followed by updating of dual variables through online learning in the dual phase. The formal description of the algorithm is provided in Appendix D.1 (see Algorithm 2). For brevity, we only highlight the main differences compared to `CA-DL` in the following.

Primal Phase: `CO-DL` chooses the location that maximizes the adjusted score subject to the hard resource constraint where the adjusted score is replaced by $(\mathbf{w}_t - \boldsymbol{\theta}_t - \boldsymbol{\lambda}_t) \cdot \mathbf{z}_t$.

Dual Phase: `CO-DL` performs the multiplicative weight update (Freund et al. 1997) for the dual variables corresponding to static resources in the similar vein of `CA-DL`. However, the main modification is that `CO-DL` uses a *time-varying* step size $\{\eta_t\}_{t=1}^T$ in the dual phase to achieve the strong convergence that was motivated earlier (and will become more clear in Section 4.2.1). More specifically, the updating rules (8) in Algorithm 1 are now replaced by

$$\theta_{t+1,i} = \min\{\theta_{t,i} \exp(\eta_t(z_{t,i} - \rho_i)), \alpha\}, \quad \lambda_{t+1,i} = \min\left\{\lambda_{t,i} \exp(\eta_t(\mathbb{1}[i_t^\dagger = 0] z_{t,i} - \rho_i)), \frac{1+2\alpha}{\underline{\rho}}\right\} \tag{21}$$

where $\eta_t = k/\sqrt{t}$ is the time-varying learning rate with an input parameter $k > 0$.

We now present our main result for `CO-DL` in Theorem 2, first focusing on the stable regime. Our result requires the following assumption on the reward distribution, whose implications will be discussed later in the analysis of the algorithm.

Assumption 1 **(Lipschitz Continuous Reward Distribution)**. *The arrival distribution $\mathcal{F}$ is such that the PDF of the reward vector is L-Lipschitz continuous with respect to ℓ_1-norm.*

Theorem 2 **(Regret of `CO-DL` Under Stable Regime)**. *Let $\eta_t = k/\sqrt{t}$ with a constant $k > 0$. Under Assumption 1 and stable regime (Definition 1), the regret (Definition 3) of `CO-DL` is $\mathcal{O}(\sqrt{T} + \frac{\gamma}{\epsilon})$.*

We outline the proof of Theorem 2 in Section 4.2.2. The main part of the proof is to establish that, somewhat surprisingly, `CO-DL` achieves *a constant average backlog* in expectation under the stable regime, *even though it does not account for the backlog at all*. Before we formally prove this, we use the following example to build some intuition.

Electronic copy available at: https://ssrn.com/abstract=4748762

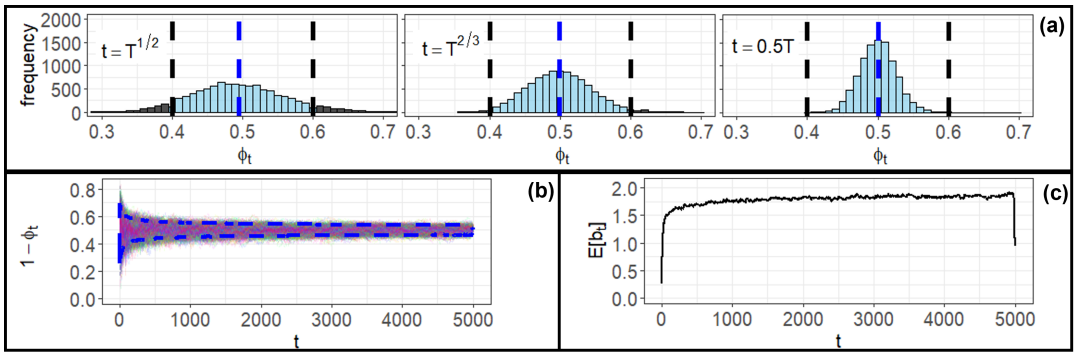


Figure 2 **Numerical illustrations for Example 1 with $\eta_t = 1/\sqrt{t}$, $T = 5000$, and 10,000 sample paths. Panel (a) shows histograms of ϕ_t for three values of t. Panel (b) displays sample paths of the endogenous arrival rates $1 - \phi_t$, with 5% and 95% quantiles (blue dashed lines). Panel (c) depicts the expected backlog for each t.**

EXAMPLE 1. We consider an instance with $T = 5000$, one affiliate, and no tied cases. For brevity, we omit the subscript for $i = 1$. The reward follows a uniform distribution on the interval (0,1) and the capacity ratio is $\rho = 0.5$. The service slack is $\epsilon = 0.1$. For this instance, CO-DL's decision simplifies to $z_t = \mathbb{1}[w_t \geq \theta_t + \lambda_t]$. For ease of exposition, we define $\phi_t = \theta_t + \lambda_t$. The optimal dual variable of the surrogate primal, denoted by $\hat{\phi}$, is the median of $(w_1, .., w_T)$,[18] which is "highly concentrated" around 0.5, the median of the reward distribution. Concretely, over 10000 sample paths, the 5% quantile of $\hat{\phi}$ is 0.5 ± 0.01. Hence, for ease of exposition, we assume $\hat{\phi} = 0.5$.

With this background, we now turn our focus to the backlog process. As mentioned in Section 1.2, our post-allocation service process can be viewed as a queueing system with an arrival rate at period t *endogenously* determined by the CO-DL's matching decision. In this example, such endogenous arrival rate is simply $\mathbb{E}[z_t|\mathcal{H}_{t-1}] = \mathbb{E}[\mathbb{1}[w_t \geq \phi_t]|\mathcal{H}_{t-1}] = 1 - \phi_t$. For a moment, suppose we make a matching decision based on $\hat{z}_t = \mathbb{1}[w_t \geq 0.5]$. In this case, it is straightforward to show that the average backlog is constant (in expectation): its backlog $\hat{b}_t$ is a non-negative jump process with a negative drift for each period t (in the sense that $\mathbb{E}[\psi(\hat{b}_t) - \psi(\hat{b}_{t-1})|\mathcal{H}_{t-1}]$ is upper bounded by a linear function of $\hat{b}_{t-1}$ with a negative slope — see the definition of the potential function ψ in (14) and Lemma 3).

Coming back to CO-DL, even though ϕ_t is not exactly 0.5 and changes across periods, we have designed its learning process (in particular its learning rates) in a way that ϕ_t is highly concentrated around 0.5 at any period (after some initial transient periods). Furthermore, ϕ_t becomes *more* concentrated as t progresses. To illustrate this convergence property, in Figure 2-(a), we display the distribution of ϕ_t for three different values of $t \in \{T^{1/2}, T^{2/3}, 0.5T\}$.[19] The concentration of ϕ_t around 0.5 implies that each endogenous arrival rate $\mathbb{E}[z_t|\mathcal{H}_{t-1}] = 1 - \phi_t$ is concentrated around 0.5. This is further illustrated by Figure 2-(b), from which we observe that the endogenous arrival rates $1 - \phi_t$ remain below the service

[18] The dual function of the surrogate primal problem is $D(\phi) = \sum_{t=1}^{T}\{(w_t - \phi)_+ + 0.5\phi\}$. The derivative of $D(\phi)$ (whenever differentiable) is $-\sum_{t=1}^{T}\mathbb{1}[w_t \geq \phi] + 0.5T$, which is negative (positive) if ϕ is smaller (larger) than the sample median. Hence, the sample median of $(w_1, .., w_T)$ minimizes the dual function.

[19] We set $\alpha = 10$, but the exact value of α does not affect the numerical observation.

Electronic copy available at: https://ssrn.com/abstract=4748762

rate $r = 0.6$ with high probability and gets more concentrated as t progresses. This strong convergence property ensures that, with high probability, the backlog jump process has a negative drift at every period, which is sufficient to show that the average backlog is constant (in expectation) as illustrated in Figure 2-(c).

In summary, the average backlog in expectation of `CO-DL` in Example 1 remains constant because the learning process is designed in a way that the dual variables become more concentrated "with high probability" around the value of 0.5 as time t progresses. Building on these intuitions, to prove Theorem 2, we first formalize a high probability last-iterate convergence of the dual variables constructed by `CO-DL` in Section 4.2.1. We then show that this strong convergence guarantees that the backlog process of `CO-DL` has a negative drift with a high probability (roughly) for every period in Section 4.2.2, which will enable us to prove the constant average backlog of `CO-DL`.

4.2.1. High Probability Last-iterate Convergence of Dual Variables

In this section, we establish a strong convergence for the `CO-DL`'s dual variables, which will be a main building block for proving Theorem 2. To that end, we first define a *static dual problem* as follows.

DEFINITION 6 **(Static Dual Problem & Dual-based Decision)**. *The static dual problem is defined as:*

$$D(\boldsymbol{\nu}) := \mathbb{E}\Big[\max_{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)} (\mathbf{w} - \boldsymbol{\theta} - \mathbb{1}[i^\dagger = 0]\boldsymbol{\lambda})\cdot\mathbf{z}(\mathbf{A}) + \boldsymbol{\rho}\cdot\boldsymbol{\theta} + \boldsymbol{\rho}\cdot\boldsymbol{\lambda}\Big], \quad D(\boldsymbol{\nu}^*) := \min_{\boldsymbol{\nu}\in\mathcal{V}} D(\boldsymbol{\nu})$$

where $\boldsymbol{\nu} := (\boldsymbol{\theta}, \boldsymbol{\lambda})$, $\mathcal{V} := [0,\alpha]^m \times [0, \frac{1+2\alpha}{\underline{\rho}}]^m$, *and the expectation is taken over the arrival* $\mathbf{A} = (\mathbf{w}, i^\dagger) \sim \mathcal{F}$. *Further, for a given dual variable* $\boldsymbol{\nu}\in\mathcal{V}$ *and arrival type* $\mathbf{A}\in\mathcal{A}$, *we define its corresponding (primal) decision as*

$$\mathbf{z}(\boldsymbol{\nu}, \mathbf{A}) := \arg\max_{\mathbf{z}\in\mathcal{Z}(i^\dagger)} \{(\mathbf{w} - \boldsymbol{\theta} - \mathbb{1}[i^\dagger = 0]\boldsymbol{\lambda})\cdot\mathbf{z}\}.$$

The following proposition establishes a high probability last-iterate convergence of $\boldsymbol{\nu}_t$ constructed by `CO-DL` to $\boldsymbol{\nu}^*$ defined above. In simple terms, this ensures that each $\boldsymbol{\nu}_t$ becomes more concentrated around $\boldsymbol{\nu}^*$ as $t \le T_A$ progresses, where T_A is the stopping time defined similarly for `CA-DL` (line (17)).

PROPOSITION 3 **(High-probability Last Iterate Convergence of Dual Variables)**. *Let* T_A *be the stopping time of* `CO-DL`. *For any fixed* $t \le T_A$ *and* $\delta\in(0,1)$, *we have*

$$\mathbb{P}\left[D(\boldsymbol{\nu}_t) - D(\boldsymbol{\nu}^*) \le \mathcal{O}\Big(\frac{\log(t)\log(1/\delta)}{\sqrt{t}}\Big)\right] \ge 1-\delta.$$

The proposition enables us to study the convergence of $\boldsymbol{\nu}_t$ at *any* t (up to the stopping time), which crucially relies on using the time-varying learning rate $\eta_t = \Theta(1/\sqrt{t})$. This is in contrast to setting a fixed learning rate $\eta_t = \Theta(1/\sqrt{T})$ for which we can only establish convergence *at the last period* T (in the absence of the capacity constraint). Because our proof for Theorem 2 involves studying the endogenous arrival rate at any time t, such time-varying learning rate is essential for us.

Electronic copy available at: https://ssrn.com/abstract=4748762

The proof of Proposition 3 is presented in Appendix D.4. The core idea of the proof is the connection of CO-DL and online stochastic mirror descent (OSMD) (Nemirovski et al. 2009) for solving the static dual problem (Definition 6). To see this, note that the gradient of the static dual function $D(\cdot)$ is:

$$\nabla D(\boldsymbol{\nu}) = (\mathbb{E}_{\mathbf{A}}[\boldsymbol{\rho} - \mathbf{z}(\boldsymbol{\nu}, \mathbf{A})], \mathbb{E}_{\mathbf{A}}[\boldsymbol{\rho} - \mathbb{1}[i^\dagger = 0]\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})]).$$

It is straightforward to see that, up to $t \le T_A$, the dual update in (21) can be written as mirror descent:

$$\boldsymbol{\nu}_{t+1} = \arg\min_{\boldsymbol{\nu} \in \mathcal{V}} \hat{\mathbf{g}}_t \cdot \boldsymbol{\nu} + \frac{1}{\eta_t} V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t), \tag{22}$$

where the mirror map is $h(\boldsymbol{\nu}) = \sum_{i=1}^m \theta_i \log(\theta_i) + \sum_{i=1}^m \lambda_i \log(\lambda_i)$ and $V_h(\cdot, \cdot)$ is the Bregman distance with respect to h. Here, the (noisy) gradient is

$$\hat{\mathbf{g}}_t = (\boldsymbol{\rho} - \mathbf{z}(\boldsymbol{\nu}_t, \mathbf{A}_t), \boldsymbol{\rho} - \mathbb{1}[i^\dagger = 0]\mathbf{z}(\boldsymbol{\nu}_t, \mathbf{A}_t)).$$

Importantly, we have $\mathbb{E}[\hat{\mathbf{g}}_t | \mathcal{H}_{t-1}] = \nabla D(\boldsymbol{\nu}_t)$ for all $t \le T_A$, meaning that $\hat{\mathbf{g}}_t$ serves as an unbiased estimator for the true gradient $\nabla D(\boldsymbol{\nu}_t)$. This is because (i) each arrival $\mathbf{A}_t$ is an i.i.d. sample from the distribution $\mathcal{F}$ and (ii) $\boldsymbol{\nu}_t$ is $\mathcal{H}_{t-1}$-measurable. The rest of the proof is to establish the last-iterate high probability convergence result for our variant of OSMD. For this part, we closely follow Harvey et al. (2019), where the same result was established for stochastic gradient descent, a special case of OSMD. Given our main use of the multiplicative updating rule (i.e., exponentiated gradient descent), we need to extend the result to our variants of OSMD. While the proof of Harvey et al. (2019) can be extended relatively straightforwardly, to the best of our knowledge, the exact result we require is not available in the literature. For the sake of completeness, we reproduce the proof of Harvey et al. (2019) in Appendix D.4 and highlight necessary modifications along the way.

4.2.2. Proof Sketch of Theorem 2

Given the aforementioned ingredients, we sketch the proof of Theorem 2. As mentioned earlier, the main part of the proof is to establish a constant upper bound on the average backlog of CO-DL, as stated in the following lemma.

LEMMA 6 **(Bounding Average Backlog of CO-DL).** *For any arrival distribution $\mathcal{F}$ and $\epsilon > 0$, we have*

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} \|\mathbf{b}_t\|_1\Big] \le \mathcal{O}\Big(\frac{1}{\epsilon}\Big)$$

The rest of the proof is to bound the loss of the net matching reward (see Lemma 14 in Appendix D.5.1) in the similar vein of Lemma 1 for CA-DL. The complete proof of Theorem 2 with a more detailed outline can be found in Appendix D.5.

We devote the rest of the section to outline the proof of Lemma 6. To bound the average backlog in expectation over the horizon, in Appendix D.5, we establish (i) $\mathcal{O}(1/\epsilon)$ bound on the average backlog

Electronic copy available at: https://ssrn.com/abstract=4748762

up to the stopping time (Lemmas 15 and 16) and (ii) $\mathcal{O}(1)$ bound on the average backlog *after* the stopping time (Lemma 17). We focus on describing our high-level idea of proving the former. As alluded to in Example 1, the backlog is a non-negative jump process. Hence, it suffices to show that the process has a negative drift with high probability. To that end, we define the following "good event" $\mathcal{B}_T$.

$$\mathcal{B}_T := \left\{ \mathbb{E}_{\mathbf{A}}[z_i(\boldsymbol{\nu}_t, \mathbf{A})] - (\rho_i + \epsilon) \le -\frac{\epsilon}{2} \text{ for all } \sqrt{T} \le t \le T_A \text{ and } i \in [m] \right\} \tag{23}$$

First, let us unpack why the occurrence of the good event $\mathcal{B}_T$ is sufficient for establishing a constant average backlog up to the stopping time. Note that the endogenous arrival rate for period $t \le T_A$ is given by $\mathbb{E}[\mathbf{z}_t|\mathcal{H}_{t-1}] = \mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}_t, \mathbf{A})]$.[20] Furthermore, we recall that the service rate for each affiliate i is $\rho_i + \epsilon$. Hence, $\mathbb{E}_{\mathbf{A}}[z_i(\boldsymbol{\nu}_t, \mathbf{A})] - (\rho_i + \epsilon)$ roughly serves as a drift of the backlog process (for each affiliate i), which is negative under the good event. In the proof of Lemma 16 (Appendix D.5.3), we use a Lyapunov argument with the quadratic potential function ψ (defined in equation (14)) to formally show that, such negative drift leads to $\mathcal{O}(1/\epsilon)$ bound on the average backlog up to the stopping time.

The rest of the proof is to show that the good event occurs with high probability, which we establish in Lemma 15. The proof is fairly intricate and here we only provide a high-level overview. Our proof crucially relies on the convergence result of Proposition 3 and Assumption 1 along with a few other ideas. Under Assumption 1, we can translate the high probability convergence of $D(\boldsymbol{\nu}_t)$ to $D(\boldsymbol{\nu}^*)$ (established in Proposition 1) to that of $\mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}_t, \mathbf{A})]$ to $\mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}^*, \mathbf{A})]$. Further, as we show in Section D.3, the optimal solution $\boldsymbol{\nu}^*$ of the static dual problem (Definition 6) satisfies that $\mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}^*, \mathbf{A})] \le \boldsymbol{\rho}$ with equality whenever $\theta_i^* > 0$. Hence, the endogenous arrival rates $\mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}_t, \mathbf{A})]$, after some transient periods, are concentrated around a value at most $\boldsymbol{\rho}$ for every period $t \le T_A$, which leads to the occurrence of the good event for large T. We refer readers to Appendix D.5.2 for more details of the proof.

4.3. Limitation in Near-critical Regime

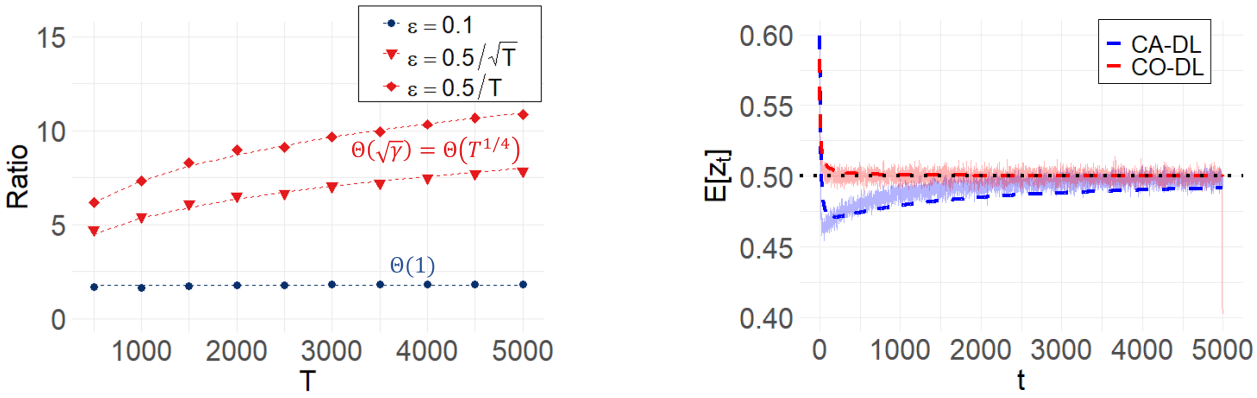
Theorem 2 implies that `CO-DL` is equally powerful to `CA-DL` in the stable regime in the sense of achieving a sub-linear regret (even though it is oblivious to backlogs). However, through the following theorem, we show that `CO-DL` cannot achieve the same order of regret as `CA-DL` under the near-critical regime.

THEOREM 3 **(Lower bound on Regret of `CO-DL` under Near-critical Regime)**. *There exists a constant $q > 0$ and an instance such that for any given $\epsilon \le \frac{q}{\sqrt{T}}$, the regret of `CO-DL` is $\Omega(\gamma\sqrt{T})$.*

Together with Corollary 1 (see Section 3.3), which states that the regret of `CA-DL` is $\mathcal{O}(\sqrt{\gamma T})$ under the near-critical regime, Theorem 3 implies that `CA-DL` outperforms `CO-DL` with a better regret guarantee at least by a factor of $\sqrt{\gamma}$. In particular, `CO-DL` loses its ability to provide a sub-linear guarantee if $\gamma = \Theta(T^\delta)$ with $\delta \in [1/2, 1)$. Intuitively, such difference in the performance stems from the fact that `CA-DL` can adapt its behavior to the backlog by penalizing its scaled version with $\zeta = \Theta(\sqrt{\gamma/T})$, whereas `CO-DL` cannot as it is oblivious to the backlog. Overall, this result highlights the inherent benefit of accounting for the backlog.

[20] We highlight that $\mathbb{E}[\mathbf{z}_t|\mathcal{H}_{t-1}]$ is still a random variable due to the randomness of $\boldsymbol{\nu}_t$.

Electronic copy available at: https://ssrn.com/abstract=4748762



(a) $\mathsf{Diff}^{\texttt{CO-DL}}/\mathsf{Diff}^{\texttt{CA-DL}}$ (b) $\mathbb{E}[z_t]$ for $T = 5000$ and $\epsilon = 0.5/\sqrt{T}$

Figure 3 Numerical Illustrations for Theorem 3 with 1000 sample paths and $\gamma = \sqrt{T}$. We use notation $\mathsf{Diff}^{\pi} := \mathbb{E}[\mathsf{OPT}(\gamma)] - \mathbb{E}[\mathsf{ALG}^{\pi}(\gamma)]$. The learning rate of CA-DL is $\eta = \frac{1}{\sqrt{T}}$, with $\zeta = \frac{10}{\sqrt{T}}$ ($\zeta = \sqrt{\frac{\gamma}{T}}$, resp.) under the stable regime (the near-critical regime, resp.). The learning rate of CO-DL is $\eta_t = \frac{1}{\sqrt{t}}$ under both regimes.

In the proof of Theorem 3, we revisit the one-affiliate instance considered in Example 1 and conduct a delicate analysis of CO-DL's regret under the near-critical regime. We defer the details of the proof to Appendix E. Instead, in the following, we use the same instance to give a high-level intuition for the lower bound of regret in Theorem 3. For each $T \in \{500, 1000, ..., 5000\}$, we use $\gamma = \sqrt{T}$ and $\epsilon \in \{0.1, \frac{0.5}{\sqrt{T}}, \frac{0.5}{T}\}$, where the first value of ϵ belongs to the stable regime and the other two to the near-critical regime. By using the notation of $\mathsf{Diff}^{\pi} := \mathbb{E}[\mathsf{OPT}(\gamma)] - \mathbb{E}[\mathsf{ALG}^{\pi}(\gamma)]$ for an algorithm π (note that α does not play any role in this instance because there are no tied cases), we display $\mathsf{Diff}^{\texttt{CO-DL}}/\mathsf{Diff}^{\texttt{CA-DL}}$ in Figure 3a (as a function of T). Consistent with Theorem 3 and Corollary 1, we note that the ratios (two red curves) diverge in the order of $\sqrt{\gamma} = T^{1/4}$ under the near-critical regimes. In contrast, the ratio under the stable regime (blue curve) remains constant, which is also aligned with Theorems 1 and 2. However, we observe that CA-DL still achieves a lower regret than CO-DL by a constant factor (roughly by 70%).

The main reason behind the performance difference lies in the endogenous arrival rates that each algorithm's decision induces. More specifically, Figure 3b illustrates the empirical average of z_t (solid line) for $T = 5000$ and $\epsilon = 0.5/\sqrt{T}$, which serves as a proxy for the (expected) endogenous arrival rates. The arrival rate of CO-DL quickly converges to $\rho = 0.5$, as evident from the cumulative time-average (dashed red line). In contrast, CA-DL induces an arrival rate strictly less than $\rho = 0.5$ (dashed blue line). Under the stable regime, maintaining an induced arrival rate of ρ ensures a constant average backlog. However, in the near-critical regime, we can show that the approach of CO-DL results in a time-average backlog of $\Omega(T^{1/2})$ for this specific instance (see Lemma 23). To gain some intuition, consider momentarily that the algorithm "exactly" induces the matching rate of ρ at every period. Because $\epsilon = \mathcal{O}(1/\sqrt{T})$, a simple anti-concentration result implies that, with constant probability, there remains $\Omega(\sqrt{t})$ backlog at each period t. By contrast, by penalizing the current backlog level, CA-DL achieves a better order of the backlog (and thus, a better order of regret).

Electronic copy available at: https://ssrn.com/abstract=4748762

5. Case Study on Refugee Resettlement Data

In this section, we utilize real data from our partner agency to conduct a numerical evaluation of our learning-based approach. We first describe the construction of primitives and how to adapt our learning-based approach to our partner agency's context. We then present our numerical result and show the performance improvement of our learning-based approach compared to the benchmarks.

5.1. Data and Benchmarks

Data and Primitives. We use data that includes information about refugees of working age (18 to 64 years old) resettled by our partner agency during the years 2014-2016. In year 2015, our partner agency resettled $T = 3819$ cases to $m = 45$ affiliates. We set the capacity as the actual total number of cases resettled at each affiliate.[21] We use the *actual sequence* of refugee arrivals to construct a sequence of arrival types $\mathbf{A}_t = (\mathbf{w}_t, i_t^\dagger)$. To generate the reward $(\mathbf{w}_t)$ for each case t, we apply the methodology from Bansak et al. (2018), utilizing a supervised machine learning model that predicts the case-affiliate pairwise employment probability based on its demographic characteristics. Our data also includes an indicator for cases with U.S. ties. These cases must be assigned to pre-determined affiliates (as we motivated in Section 1.1). Hence, we set the target affiliate of the cases with U.S. ties to be the affiliate the case was actually placed at. The rest of the refugees are considered as free cases, and we assume that they can be assigned to any affiliate. To measure the backlog, following the approach of Bansak and Paulson (2022), we adopt a deterministic service flow with a service rate $\boldsymbol{\rho}$. This measure of the backlog is currently employed by our partner agency as a measure for congestion. We repeat the aforementioned processes for year 2016 as well to construct primitives for that year. The data of year 2014 is further used for parameter tuning, which we will describe shortly.

Algorithms under Evaluation. Our goal is to compare performance of the following policies:

(i) Robust Online Learning-based Algorithm (`RO-Learing`). The first policy is the learning-based algorithm, which is an adaptation of `CA-DL` (Algorithm 1) to incorporate a few practical considerations that we discuss next.[22] First, we remove the dummy affiliate (i.e., inaction). In practice, once a case is allocated to our partner agency, it must be placed at an actual affiliate. Second, we require that a free case is only matched to affiliates with remaining *aggregate* capacity after accounting for both free *and* tied cases. Said differently, we match the arriving free case to affiliate i with the highest adjusted score (possibly negative) among those with remaining aggregate capacity,[23] followed by the same updating rule in line (8). Although not explicitly considered in the theoretical analysis, this extra constraint mirrors real-world heuristics because the agency avoids matching a free case to an affiliate that has

[21] For institutional reasons, we do not have exact knowledge of the initial capacity for our partner agency.

[22] Under the current practice of measuring the service as a deterministic process, the agency can easily keep track of backlogs without need for excessive communications.

[23] Note that $\sum_{i\in[m]} \rho_i = 1$ because we set the capacity based on the actual number of cases matched at each affiliate. Thus, at any given period, there always exists an affiliate with remaining aggregate capacity.

Electronic copy available at: https://ssrn.com/abstract=4748762

reached full capacity. Finally, recall that `CA-DL` requires two learning rates, η and ζ. We tune them in a data-driven manner to choose values that yield reasonable ranges of the over-allocation and average backlog based on the data of year 2014. Further details on this can be found in Appendix F.1.

(ii) Sampling-based Algorithm (`Sampling`). As motivated in Section 1, the existing proposals in the literature of refugee matching have been relying on a *sampling* approach, which simulates future arrival patterns based on past years' data. We use (an adaptation of) the algorithm from Bansak and Paulson (2022) as our main benchmark relying on such approach. The original algorithm of Bansak and Paulson (2022) is developed for the setting without the tied cases (see discussions in Section 1.3). In Appendix F.2, we describe how we modify the algorithm to account for the tied cases.

`Sampling` was also designed to maximize the same objective function as defined in Section 2. In the similar spirit of `RO-Learning`, `Sampling` also penalizes the current level of congestion, with the degree of penalty increasing in γ. However, a crucial difference from `RO-Learing` is that it requires a *sample trajectory* of future arrivals. More specifically, to make a matching decision for case t, `Sampling` solves an offline problem defined for period t and onward by sampling future arrivals from a *sampling pool*. Consistent with Bansak and Paulson (2022), we use the previous year's arrival as the sampling pool to implement `Sampling`. For example, when tested on the data of year 2015, `Sampling` uses that of year 2014 to simulate future arrivals. Intuitively, the performance of `Sampling` depends on how representative the sampling pool is for the current arrival sequence.

(iii) Actual placement (`Actual`). In the current practice, case officers at our partner agency manually select initial placement decisions for cases. These decisions have been primarily driven by capacity constraints with an eye toward workload balancing, but without a systematic attempt to account for employment outcomes. We use the actual decisions made by the case officers as a benchmark to evaluate how our learning-based approach can improve the current practices of our partner agency.

5.2. Results

In the following, we compare the performance of `RO-Learning` to the benchmarks. We present the results for penalty parameters $\alpha = 3$ and $\gamma = 5$, and defer more extensive results using a broader range of the cost parameters to Appendix F.3.

Main Result: In Table 1, we display the employment rate among free cases,[24] the total over-allocation, and the average backlog. Note that, because we set the capacity based on the actual number of cases matched at each affiliate, `Actual` does not incur any over-allocation. Despite this, we highlight that `RO-Learning` significantly improves upon the actual placement results, both in terms of the employment outcome and average backlog by more than 50% and 30%, respectively.[25]

[24] Because the algorithm cannot affect the tied case's placement, we only report the employment rate among free cases. For both years of 2015 and 2016, roughly 30% of the arrivals were free cases.

[25] That said, we remark that `Actual` may have been influenced by certain unobserved constraints not accounted for in this case study (e.g., unobserved incompatibility between certain free cases and affiliates). Therefore, the improvement over `Actual` should not be interpreted at face value but rather as an indication of the potential benefits that `RO-Learning` could offer when implemented in real-world scenarios.

Electronic copy available at: https://ssrn.com/abstract=4748762

Moving beyond the comparison to the actual placements, we now compare the performance of `RO-Learning` to that of the more sophisticated benchmark, `Sampling`. For $\alpha = 3$ and $\gamma = 5$, compared to `Sampling`, `RO-Learning` improves the objective (that combines the three outcomes of total employment, average backlog, and over-allocation as define in (2)) by 42% for year 2015 (19% for year 2016, resp.). To further unpack how we achieve such an improvement, we compare each of the three outcomes in Table 1. We observe that `RO-Learning` significantly improves both over-allocation and average backlog in both years, with only a small (less than 1%) impact on the employment outcome. For instance, in the year 2015, `RO-Learning` achieves a 25% lower total over-allocation and a 15% lower average backlog compared to `Sampling`. In Appendix F.3, we demonstrate a similar performance improvement, with 10% - 50% improvement of the objective, for a broad range of the penalty parameters.

Performance Improvement through Robustness: To understand the main source of performance improvement, we highlight the advantages of *robustness* inherent in our learning-based approach. Note that `Sampling` heavily relies on a sampling pool of the previous year's arrivals, making it susceptible to discrepancies between the sampling pool and the current year's arrival (as evidenced in 1 in Section 1.1). In contrast, `RO-Learning` remains unaffected by such discrepancies since it *learns* the arrival pattern based on the current year's data. Further, the algorithm uses *adversarial online learning*, which enables it to learn the arrival pattern in a robust manner.

Practical Benefits of `RO-Learning`: We also highlight auxiliary benefits of the adversarial online learning that `RO-Learning` employs. First, it enables the algorithm to be inherently balanced. At the core of robust online learning is its *self-correcting nature*. Specifically, aside from the explicit penalty on the backlog, the way that `RO-Learning` updates the dual variables (line (8) in Algorithm 1) prevents bursty matching to one affiliate because these dual variables naturally increase (decrease) whenever we (do not) match a case to the affiliate. Such self-correcting property naturally "pushes" the algorithm's endogenous arrival rate to $\boldsymbol{\rho}$. In contrast, `Sampling` would maintain such property *only if the sampling pool is exactly the same as the true underlying distribution.* Indeed, as demonstrated in Appendix F.3, `RO-Learning` achieves a reasonably low level of backlog even when $\gamma = 0$, comparable to the level achieved by `Sampling` for higher values of γ. Furthermore, online learning brings significant computational efficiency. In particular, `RO-Learning` does not need to solve any auxiliary optimization problem, which is in contrast to `Sampling`. Such computational efficiency can be valuable because, in practice, a policymaker may need to run multiple iterations of the algorithm for tuning parameters or back-testing the algorithm on various datasets. Finally, we highlight that the `RO-Learning`'s simple score-based rule makes it easy to communicate with our partner agency's practitioners.

6. Conclusion and Future Directions

Motivated by our collaboration with a major refugee resettlement agency in the U.S., and to study the high-stakes matching decisions involved in this application, we introduced a new dynamic matching

Electronic copy available at: https://ssrn.com/abstract=4748762

Table 1 Numerical Performance for year 2015 (2016, resp.) for penalty parameters $\alpha = 3$ and $\gamma = 5$. (For `Sampling`, we take an average over 5 simulations.)

	Employment Rate (%)	Total Over-allocation	Average Backlog
`Actual`	40.8 (44.4)	0 (0)	226.6 (323.5)
`Sampling`	67.7 (70.3)	107.6 (99.8)	180.1 (225.2)
`RO-Learning`	67.3 (70.1)	80 (91)	152.8 (195.7)

problem that captures the following feature: resettling a refugee not only consumes static resources at an affiliate (e.g., funding) but also requires post-matching on-boarding services (e.g., translation) with (potentially uncertain) service time that can be viewed as dynamic resources. As such, the agency's goal is to maximize reward while respecting constraints associated with both static and dynamic resources. While our focus is on refugee resettlement, post-allocation services may naturally arise in other humanitarian or healthcare applications, making our framework applicable beyond this application. Finding other industry partners that can benefit from our work is a direction for future work.

Based on the insights we gained by analyzing historical data from our partner agency, and unlike existing work, we take a prior-independent approach that does not rely on past year's data to make decisions. In particular, we propose learning-based algorithms with provable performance guarantees. As demonstrated in Section 5, our approach not only results in better outcomes due to its robustness but also enjoys other practical advantages. Working with our partner agency to further test our proposed algorithm in practice is an immediate direction that we plan to pursue.

From a theoretical point of view, while our framework shares features of models studied in online (static) resource allocation and matching queue, it does not belong to either group. Consequently, to design and analysis of low-regret online algorithms, we develop new technical ideas that can be of independent interest. Using these novel ideas we establish sub-linear regret with respect to a strong benchmark whenever achieving such guarantees is possible (i.e., when $\gamma = o(T)$). Considering weaker benchmarks that may enable us to design low-regret algorithms for the complement regime ($\gamma = \Omega(T)$) is an interesting theoretical direction for future work.

Acknowledgment

Vahideh Manshadi and Rad Niazadeh gratefully acknowledge the Simons Institute for the Theory of Computing, as this work was done in part while attending the program on Data Driven Decision Processes. The authors thank Global Refuge for access to data and guidance. The data used in this study were provided under a collaboration research agreement with Global Refuge, which requires that these data not be transferred or disclosed. Kirk Bansak and Elisabeth Paulson are Faculty Affiliates of the Immigration Policy Lab (IPL) at Stanford University and ETH Zurich. This work is associated with IPL's GeoMatch project, for which the authors acknowledge funding from the Charles Koch Foundation, Stanford Institute for Human-Centered Artificial Intelligence, Google.org, and Stanford

Electronic copy available at: https://ssrn.com/abstract=4748762

Impact Labs. The authors extend their gratitude to the entire GeoMatch team for helpful feedback. Soonbong Lee and Vahideh Manshadi also extend their gratitude to the Management Science and Engineering Department at Stanford University for hosting them during part of this research. Rad Niazadeh's research is partially supported by an Asness Junior Faculty Fellowship at Chicago Booth School of Business.

References

Philipp Afeche, Rene Caldentey, and Varun Gupta. On the optimal design of a bipartite matching queueing system. *Operations Research*, 70(1):363–401, 2022.

Shipra Agrawal and Nikhil R Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1405–1424. SIAM, 2014.

Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

Narges Ahani, Tommy Andersson, Alessandro Martinello, Alexander Teytelboym, and Andrew C Trapp. Placement optimization in refugee resettlement. *Operations Research*, 69(5):1468–1486, 2021.

Narges Ahani, Paul Gölz, Ariel D Procaccia, Alexander Teytelboym, and Andrew C Trapp. Dynamic placement in refugee resettlement. *Operations Research*, 2023.

Ross Anderson, Itai Ashlagi, David Gamarnik, and Yash Kanoria. Efficient dynamic barter exchange. *Operations Research*, 65(6):1446–1459, 2017.

Itai Ashlagi, Patrick Jaillet, and Vahideh H Manshadi. Kidney exchange in dynamic sparse heterogenous pools. *arXiv preprint arXiv:1301.3509*, 2013.

Itai Ashlagi, Maximilien Burq, Patrick Jaillet, and Vahideh Manshadi. On matching and thickness in heterogeneous dynamic markets. *Operations Research*, 67(4):927–949, 2019.

Itai Ashlagi, Jacob Leshno, Pengyu Qian, and Amin Saberi. Price discovery in waiting lists: A connection to stochastic gradient descent. *Available at SSRN 4192003*, 2022.

Angelos Aveklouris, Levi DeValve, and Amy R Ward. Matching impatient and heterogeneous demand and supply. *arXiv preprint arXiv:2102.02710*, 2021.

Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. Regularized online allocation problems: Fairness and beyond. In *International Conference on Machine Learning*, pages 630–639. PMLR, 2021.

Santiago R Balseiro, Omar Besbes, and Dana Pizarro. Survey of dynamic resource-constrained reward collection problems: Unified model and analysis. *Operations Research*, 2023a.

Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni. The best of many worlds: Dual mirror descent for online allocation problems. *Operations Research*, 71(1):101–119, 2023b.

Kirk Bansak and Elisabeth Paulson. Outcome-driven dynamic refugee assignment with allocation balancing. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1182–1183, 2022.

Kirk Bansak, Jeremy Ferwerda, Jens Hainmueller, Andrea Dillon, Dominik Hangartner, Duncan Lawrence, and Jeremy Weinstein. Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329, 2018.

Electronic copy available at: https://ssrn.com/abstract=4748762

Kirk Bansak, Elisabeth Paulson, and Dominik Rothenhäusler. Learning under random distributional shifts. *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS), to appear*, 2024.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Andorra Bruno. Reception and placement of refugees in the United States. Congressional Research Service (CRS) Report No. R44878, 2017. URL `https://sgp.fas.org/crs/homesec/R44878.pdf`.

Sébastien Bubeck. Introduction to online optimization. *Lecture notes*, 2:1–86, 2011.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

David Delacrétaz, Scott Duke Kominers, and Alexander Teytelboym. Matching mechanisms for refugee resettlement. *American Economic Review*, 113(10):2689–2717, 2023.

Nikhil R Devanur and Thomas P Hayes. The adwords problem: online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 71–78, 2009.

Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 29–38, 2011.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Jon Feldman, Monika Henzinger, Nitish Korula, Vahab S Mirrokni, and Cliff Stein. Online stochastic packing applied to display ad allocation. In *European Symposium on Algorithms*, pages 182–194. Springer, 2010.

Yiding Feng, Rad Niazadeh, and Amin Saberi. Near-optimal bayesian online assortment of reusable resources. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 964–965, 2022.

Daniel Freund and Siddhartha Banerjee. Good prophets know when the end is near. *Available at SSRN 3479189*, 2019.

Daniel Freund, Thodoris Lykouris, Elisabeth Paulson, Bradley Sturt, and Wentao Weng. Group fairness in dynamic refugee assignment, 2024.

Yoav Freund, Robert E Schapire, Yoram Singer, and Manfred K Warmuth. Using and combining predictors that specialize. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 334–343, 1997.

Paul Gölz and Ariel D Procaccia. Migration as submodular optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 549–556, 2019.

Vincent Guigues, Anatoli Juditsky, and Arkadi Nemirovski. Non-asymptotic confidence bounds for the optimal value of a stochastic program. *Optimization Methods and Software*, 32(5):1033–1058, 2017.

Anupam Gupta and Marco Molinaro. How the experts algorithm can help solve lps online. *Mathematics of Operations Research*, 41(4):1404–1431, 2016.

Varun Gupta. Greedy algorithm for multiway matching with bounded regret. *Operations Research*, 2022.

Electronic copy available at: https://ssrn.com/abstract=4748762

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

Süleyman Kerimov, Itai Ashlagi, and Itai Gurvich. Dynamic matching: Characterizing and achieving constant regret. *Management Science*, 2023.

Thomas Kesselheim, Klaus Radke, Andreas Tonnis, and Berthold Vocking. Primal beats dual on online packing lps in the random-order model. *SIAM Journal on Computing*, 47(5):1939–1964, 2018.

Xiaocheng Li and Yinyu Ye. Online linear programming: Dual convergence, new algorithms, and regret bounds. *Operations Research*, 70(5):2948–2966, 2022.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22–es, 2007.

Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. Springer Nature, 2022.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Hai Nguyen, Thành Nguyen, and Alexander Teytelboym. Stability in matching markets with complex constraints. *Management Science*, 67(12):7438–7454, 2021.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Pengyi Shi, Mabel C Chou, Jim G Dai, Ding Ding, and Joe Sim. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science*, 62(1):1–28, 2016.

Maurice Sion. On general minimax theorems. 1958.

Kalyan Talluri and Garrett Van Ryzin. An analysis of bid-price controls for network revenue management. *Management science*, 44(11-part-1):1577–1593, 1998.

UNHCR. Refugee statistics, 2022. URL `https://www.unhcr.org/refugee-statistics/`.

UNHCR. Refugee resettlement facts, 2023. URL `https://www.unhcr.org/us/media/refugee-resettlement-facts`.

U.S. State Department. Fy 2023 notice of funding opportunity for reception and placement program, 2023. URL `https://www.state.gov/fy-2023-notice-of-funding-opportunity-for-reception-and-placement-program/`.

Yehua Wei, Jiaming Xu, and Sophie H Yu. Constant regret primal-dual policy for multi-way dynamic matching. *Available at SSRN 4357216*, 2023.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Electronic copy available at: https://ssrn.com/abstract=4748762

Appendix A: Proof of Proposition 1

We consider an instance with $m = 1$. For brevity, we omit the subscript for i. There are T arrivals with deterministic reward $w_t = 1$ for all $t \in [T]$. There are no tied cases and the capacity is $c = 0.5T$. The service rate is $r = 0.5 + \epsilon$ where $\epsilon \in [0, 0.5)$. Note that α does not play any role here. Hence, we denote the expected value of offline by $\mathsf{OPT}(\gamma)$. In the following, we first obtain a lower bound on $\mathsf{OPT}(\gamma)$ and then an upper bound on the value of any online algorithm.

Step 1: Lower-bounding Value of Optimal Offline.

We first analyze the value of the offline.

Claim 1. *For any $\epsilon \geq 0$,*

$$\mathsf{OPT}(\gamma) \geq 0.5T - \Theta(\sqrt{T})$$

Proof of Claim 1. Consider the following feasible solution of the offline program: set $z_t = \mathbb{1}[s_t = 1]$ while satisfying the capacity constraint. Let S_T^ϵ be a binomial random variable with T trials and success probability $0.5 + \epsilon$. The proposed feasible solution never incurs the backlog, while obtaining $\min(S_T^\epsilon, 0.5T)$ reward. Hence, we have $\mathsf{OPT}(\gamma) \geq \min(S_T^\epsilon, 0.5T)$ for every sample path of arrivals and services. Taking the expectation of the preceding inequality, we have

$$\mathbb{E}[\mathsf{OPT}(\gamma)] \geq \mathbb{E}[\min(S_T^\epsilon, 0.5T)] \geq \mathbb{E}[\min(S_T^0, 0.5T)]$$

where the last inequality is because S_T^ϵ stochastically dominates S_T^0 for all $\epsilon \geq 0$ in the first order.[26] We now lower-bound the right-hand side of the above inequality by $0.5T - \Theta(\sqrt{T})$. Let us define $Z_T^0 := \dfrac{S_T^0 - 0.5T}{\sqrt{\mathrm{Var}[S_T^0]}}$ as the normalized vector. Then we have

$$\mathbb{E}[\min(S_T^0, 0.5T)] = 0.5T - \mathbb{E}[(0.5T - S_T^0)_+] = 0.5T - \sqrt{\mathrm{Var}[S_T^0]}\,\mathbb{E}[(-Z_T^0)_+] = 0.5T - \Theta(\sqrt{T}).$$

where the last equality is because $\mathbb{E}[(-Z_T^0)_+]$ has a constant mean for large T by the central limit theorem. This completes the proof of Claim 1. □

Step 2: Upper-bounding Value of Any Online Algorithm.

We now turn our attention to upper-bounding the value of any online algorithm, denoted by ALG. Fix an online algorithm and let z_t denote the decision at time t. Without loss, we assume that $\mathbb{E}[\sum_{t=1}^T z_t] \geq \frac{T}{4}$ because otherwise, we have $\mathbb{E}[\mathsf{OPT}(\gamma)] - \mathbb{E}[\mathsf{ALG}] \geq 0.25T - \Theta(\sqrt{T}) = \Omega(T)$ from Claim 1.

Claim 2. *The expected value of any online algorithm for which $\mathbb{E}[\sum_{t=1}^T z_t] \geq \frac{T}{4}$ satisfies:*

$$\mathbb{E}[\mathsf{ALG}] \leq 0.5T - 0.25\gamma(0.5 - \epsilon)$$

Proof of Claim 2. We first lower-bound the total cumulative backlog by

$$\mathbb{E}\left[\sum_{t=1}^T b_t\right] \geq \mathbb{E}\left[\sum_{t=1}^T z_t\right](0.5 - \epsilon). \tag{24}$$

[26] Random variable X stochastically dominates (in the first-order) if any only if $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ for all non-decreasing function u. Furthermore, let X and Y be binomial random variables with the respective success probability p_X and p_Y and the common number of trials T. If $p_X \geq p_Y$, X stochastically dominates Y in the first order.

Electronic copy available at: https://ssrn.com/abstract=4748762

To see this, we first observe that

$$b_t \geq \mathbb{1}[s_t = 0] z_t$$

for every sample path. More specifically, if $s_t = 0$, then we have at least z_t amount of backlog at the end of period t. Otherwise, the bound is trivial. Taking the expectation of the preceding inequality conditional on history $\mathcal{H}_{t-1}$, we have

$$\begin{aligned}\mathbb{E}[b_t|\mathcal{H}_{t-1}] &\geq \mathbb{E}[\mathbb{1}[s_t = 0] z_t|\mathcal{H}_{t-1}] \\ &= \mathbb{E}\left[\mathbb{1}[s_t = 0]|\mathcal{H}_{t-1}\right] \mathbb{E}[z_t|\mathcal{H}_{t-1}] \\ &= (0.5 - \epsilon)\, \mathbb{E}[z_t|\mathcal{H}_{t-1}]\end{aligned}$$

For the second line, we used the following fact: since any online algorithm does not know the realization of s_t at the time of making the decision z_t, the random variables z_t and s_t are conditionally independent given $\mathcal{H}_{t-1}$. The last line follows because s_t is an i.i.d. Bernoulli random variable with mean $0.5 + \epsilon$. Taking outer expectation of the preceding inequality, we obtain that

$$\mathbb{E}[b_t] \geq (0.5 - \epsilon)\, \mathbb{E}[z_t]$$

Summing up for all $t \in [T]$ gives the inequality (24). Hence, we conclude that

$$\mathbb{E}[\mathsf{ALG}] = \mathbb{E}\left[\sum_{t=1}^{T} z_t\right] - \frac{\gamma}{T}\mathbb{E}\left[\sum_{t=1}^{T} b_t\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} z_t\right] - \frac{\gamma}{T}(0.5-\epsilon)\mathbb{E}\left[\sum_{t=1}^{T} z_t\right] \leq 0.5T - 0.25\gamma(0.5-\epsilon).$$

In the last inequality, we used that $0.25T \leq \mathbb{E}[\sum_{t=1}^{T} z_t]$ and the fact that capacity is $0.5T$. □

Step 3: Putting Everything Together. From Claim 1 and Claim 2, we conclude that

$$\mathbb{E}[\mathsf{OPT}] - \mathbb{E}[\mathsf{ALG}] \geq 0.25\gamma(0.5-\epsilon) - \Theta(\sqrt{T}).$$

for any online algorithm such that $\mathbb{E}[\sum_{t=1}^{T} z_t] \geq 0.25T$. Hence, whenever $\gamma = \Omega(T)$, the regret must be at least $\Omega(T)$. This completes the proof of Proposition 1.

Appendix B: Proof of Proposition 2

Fix a sample path $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^{T}$. Given the assignment profile $\{\mathbf{z}_t\}_{t=1}^{T}$, let us denote its objective value as

$$f(\{\mathbf{z}_t\}_{t=1}^{T}) = \sum_{t=1}^{T}\sum_{i=1}^{m} w_{t,i} z_{t,i} - \alpha \sum_{i=1}^{m}\left(\sum_{t=1}^{T} z_{t,i} - c_i\right)_+ - \frac{\gamma}{T}\sum_{t=1}^{T}\sum_{i=1}^{m} b_{t,i}$$

where $b_{t,i} = (b_{t,i} + z_{t,i} - s_{t,i})_+$ with $b_{0,i} = 0$ for all $t \in [T]$ and $i \in [m]$.[27] For brevity, we also define a primal and feasible set as:

$$\begin{aligned}\mathcal{P} &:= \left\{\{\mathbf{z}_t\}_{t=1}^{T} \mid \mathbf{z}_t \in \mathcal{Z}(i_t^\dagger), \forall t \in [T]\right\} \\ \mathcal{D} &:= \Big\{(\boldsymbol{\theta}, \boldsymbol{\lambda}, \{\boldsymbol{\beta}_t\}_{t=1}^{T}) \geq \mathbf{0} \mid \ \theta_i \leq \alpha \ \forall i \in [m],\ \beta_{t,i} - \beta_{t+1,i} \leq \frac{\gamma}{T} \ \forall t \in [T-1]\ i \in [m], \\ &\quad \beta_{T,i} \leq \frac{\gamma}{T} \ \forall i \in [m]\Big\}\end{aligned}$$

Consider the following Lagrangian function.

$$L\Big(\{\mathbf{z}_t\}_{t=1}^{T}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \{\boldsymbol{\beta}_t\}_{t=1}^{T}\Big) = \sum_{t=1}^{T} \mathbf{w}_t \cdot \mathbf{z}_t + \underbrace{\sum_{t=1}^{T} \boldsymbol{\theta} \cdot (\boldsymbol{\rho} - \mathbf{z}_t)}_{\mathsf{CS}(\boldsymbol{\theta})} + \underbrace{\sum_{t=1}^{T} \boldsymbol{\lambda} \cdot (\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t)}_{\mathsf{CS}(\boldsymbol{\lambda})} + \underbrace{\sum_{t=1}^{T} \boldsymbol{\beta}_t \cdot (\mathbf{s}_t - \mathbf{z}_t)}_{\mathsf{CS}(\{\boldsymbol{\beta}_t\}_{t=1}^{T})}.$$

[27] Note that this is equivalent to our original objective in Definition 2 by taking the maximum over the decision variables $\{\mathbf{b}_t\}_{t=1}^{T}$ for any given $\{\mathbf{z}_t\}_{t=1}^{T}$ subject to $b_{t,i} \geq 0$ and $b_{t,i} \geq b_{t-1,i} + z_{t,i} - s_{t,i}$ for all $t \in [T]$ and $i \in [m]$.

Electronic copy available at: https://ssrn.com/abstract=4748762

Here, each $\mathsf{CS}(\cdot)$ term can be seen as a complementary slackness term corresponding to each dual variable. We will show that

$$\begin{aligned}\mathsf{Dual}(\alpha,\gamma) &= \min_{(\boldsymbol{\theta},\boldsymbol{\lambda},\{\boldsymbol{\beta}_t\}_{t=1}^T)\in\mathcal{D}} \max_{(\{\mathbf{z}_t\}_{t=1}^T)\in\mathcal{P}} L\Big(\{\mathbf{z}_t\}_{t=1}^T,\boldsymbol{\theta},\boldsymbol{\lambda},\{\boldsymbol{\beta}_t\}_{t=1}^T\Big)\\ &= \max_{(\{\mathbf{z}_t\}_{t=1}^T)\in\mathcal{P}} \min_{(\boldsymbol{\theta},\boldsymbol{\lambda},\{\boldsymbol{\beta}_t\}_{t=1}^T)\in\mathcal{D}} L\Big(\{\mathbf{z}_t\}_{t=1}^T,\boldsymbol{\theta},\boldsymbol{\lambda},\{\boldsymbol{\beta}_t\}_{t=1}^T\Big) \qquad (25)\\ &= \mathsf{OPT}(\alpha,\gamma). \qquad (26)\end{aligned}$$

The first line follows from the definition of the dual function $\mathsf{Dual}(\alpha,\gamma)$ in the main statement. The second line is the consequence of Sion's min-max theorem (Sion 1958).[28] To show the last equality, we show that

$$\min_{(\boldsymbol{\theta},\boldsymbol{\lambda},\boldsymbol{\beta}_t)\in\mathcal{D}} L\Big(\{\mathbf{z}_t\}_{t=1}^T,\boldsymbol{\theta},\boldsymbol{\lambda},\{\boldsymbol{\beta}_t\}_{t=1}^T\Big) = \begin{cases}-\infty & \text{if } \sum_{t=1}^T \mathbb{1}[i_t^\dagger=0]\mathbf{z}_t > \mathbf{c}\\ f(\{\mathbf{z}_t\}_{t=1}^T) & \text{otherwise.}\end{cases} \qquad (27)$$

Note that line (26) directly follows from line (27). Hence, it suffices to show line (27). Recall that $\mathbf{c}=T\boldsymbol{\rho}$ by definition of $\boldsymbol{\rho}$. If $\sum_{t=1}^T \mathbb{1}[i_t^\dagger=0]\mathbf{z}_t > \mathbf{c}$, then the optimal solution $\boldsymbol{\lambda}^* \geq \mathbf{0}$ of the inner minimization problem in line (25) can grow arbitrarily large, hence making $\mathsf{CS}(\boldsymbol{\lambda})$ unbounded below. Otherwise, it is straightforward to see $\mathsf{CS}(\boldsymbol{\lambda}^*)=0$ at the optimal $\boldsymbol{\lambda}^*$ of the inner minimization problem in line (25). We now investigate the minimum value of $\mathsf{CS}(\boldsymbol{\theta})$ and $\mathsf{CS}(\{\boldsymbol{\beta}_t\}_{t=1}^T)$ over the dual feasible set $\mathcal{D}$. We first focus on the former. Note that the inner minimization problem in line (25) is separable over $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_t\}_{t=1}^T$. Hence, we have:

$$\min_{\boldsymbol{\theta}\in[0,\alpha]^m} \mathsf{CS}(\boldsymbol{\theta}) = -\max_{\boldsymbol{\theta}\in[0,\alpha]^m} \sum_{i=1}^m \theta_i\left(\sum_{t=1}^T z_{t,i}-c_i\right) = -\alpha\sum_{i=1}^m\left(\sum_{t=1}^T z_{t,i}-c_i\right)_+.$$

The the last inequality follows from the fact that $\alpha(x)_+ = \max_{\theta\in[0,\alpha]}\theta x$ for any $x\in\mathbb{R}$ (this is also known as Fenchel Duality). In the similar way,

$$\begin{aligned}&\min_{\{\boldsymbol{\beta}\}_{t=1}^T\geq\mathbf{0}} \mathsf{CS}(\{\boldsymbol{\beta}_t\}_{t=1}^T) \quad \text{s.t.} \quad \beta_{t,i}-\beta_{t+1,i}\leq\frac{\gamma}{T}\ \forall t\in[T-1] \text{ and } i\in[m],\ \beta_{T,i}\leq\frac{\gamma}{T}\ \forall i\in[m]\\ &= -\left[\max_{\{\boldsymbol{\beta}_t\}_{t=1}^T\geq\mathbf{0}} \sum_{t=1}^T\sum_{i=1}^m \beta_{t,i}(z_{t,i}-s_{t,i}) \quad \text{s.t} \quad \beta_{t,i}-\beta_{t+1,i}\leq\frac{\gamma}{T}\ \forall t\in[T-1] \text{ and } i\in[m],\ \beta_{T,i}\leq\frac{\gamma}{T}\ \forall i\in[m]\right]\\ &= -\left[\frac{\gamma}{T}\min_{\{\mathbf{b}_t\}_{t=1}^T\geq\mathbf{0}} \sum_{t=1}^T\sum_{i=1}^m b_{t,i} \quad \text{s.t.} \quad b_{t,i}\geq b_{t-1,i}+z_{t,i}-s_{t,i} \quad \forall t\in[T], i\in[m]\right]\\ &= -\frac{\gamma}{T}\sum_{t=1}^T\sum_{i=1}^m (b_{t-1,i}+z_{t,i}-s_{t,i})_+.\end{aligned}$$

In the third line, we used the strong duality of the linear program. Note that, at the optimal solution of the third line, either of $b_{t,i}\geq 0$ or $b_{t,i}\geq b_{t-1,i}+z_{t,i}-s_{t,i}$ must bind. Hence, the last line follows. Combining, we have proved line (27). This completes the proof.

Appendix C: Missing Proofs of Section 3.2

Before going forward, we first present some technical preliminaries that we will frequently use. The following claim is the standard result of online mirror descent.

[28] More specifically, the Lagrangian is a linear function of the primal and dual variables. The primal feasible set $\mathcal{P}$ is convex and compact and the dual feasible set is also convex. Hence, we can switch the minimum and maximum operator by applying Sion's min-max theorem.

Electronic copy available at: https://ssrn.com/abstract=4748762

CLAIM 3 **(Online Mirror Descent (Bubeck 2015))**. *Let $\{f_t : \mathcal{D} \to \mathbb{R}\}$ be a sequence of convex functions and $\boldsymbol{\nu}_t$ be a sequence of iterates such that*

$$\begin{aligned} \nabla h(\tilde{\boldsymbol{\nu}}_{t+1}) &= \nabla h(\boldsymbol{\nu}_t) - \eta_t \hat{\mathbf{g}}_t \\ \boldsymbol{\nu}_{t+1} &= \arg\min_{\mathbf{x}\in\mathcal{D}} V_h(\mathbf{x}, \tilde{\boldsymbol{\nu}}_{t+1}) \end{aligned} \tag{28}$$

where (1) $\mathcal{D}$ is a bounded convex set, (2) $\mathbb{E}[\hat{\mathbf{g}}_t | H_{t-1}] = \nabla f_t(\boldsymbol{\nu}_t)$, and (3) h is a σ-strongly convex function with respect to $\|\cdot\|$ norm on set $\mathcal{D}$,. For any $\boldsymbol{\nu} \in \mathcal{D}$ and positive integers (k, s), we have

$$\sum_{t=k}^{s} (f_t(\boldsymbol{\nu}_t) - f_t(\boldsymbol{\nu})) \le \sum_{t=k}^{s} \frac{\eta_t \|\hat{\mathbf{g}}_t\|_*^2}{2\sigma} + \frac{1}{\eta_k} V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_k) + \sum_{t=k+1}^{s} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) + \sum_{t=k}^{s} \hat{\mathbf{u}}_t \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) \tag{29}$$

where $\hat{\mathbf{u}}_t := \mathbb{E}[\hat{\mathbf{g}}_t | H_{t-1}] - \hat{\mathbf{g}}_t$ (i.e., noise of the gradient) and $V_h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y}) - \nabla h(\mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})$ is the Bregman divergence with respect to h.

The stated theorem is slightly different from the one from Bubeck (2015) because we allow for the step size to be time-varying in Section 4. For the sake of completeness, we reproduce the proof of Claim 3 in Appendix C.6.1.

The following claim shows that we can always "linearize" the total over-allocation cost, which will be crucial throughout the proof.

CLAIM 4 **(Dual Representation of Over-allocation Cost)**. *There exists $\boldsymbol{\theta}^\star \in [0, \alpha]^m$ for which*

$$\alpha \sum_{i=1}^{m} \left(\sum_{t=1}^{T} z_{t,i} - c_i \right)_+ = \sum_{t=1}^{T} \boldsymbol{\theta}^\star \cdot (\mathbf{z}_t - \boldsymbol{\rho}) \tag{30}$$

Proof of Claim 4. The proof simply follows from $\alpha \cdot (x)_+ = \max_{\theta \in [0,\alpha]} \theta x$. □

C.1. Proof of Lemma 3

We first note that $b_{t,i} = (b_{t-1,i} + z_{t,i} - s_{t,i})_+ \le |b_{t-1,i} + z_{t,i} - s_{t,i}|$. Hence, we always have $b_{t,i}^2 \le (b_{t-1,i} + z_{t,i} - s_{t,i})^2$. Rearranging this inequality, we have

$$\frac{b_{t,i}^2 - b_{t-1,i}^2}{2} \le b_{t-1,i}(z_{t,i} - s_{t,i}) + \frac{(z_{t,i} - s_{t,i})^2}{2}.$$

Summing the inequality over $i \in [m]$, we have

$$\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1}) \le \mathbf{b}_{t-1} \cdot (\mathbf{z}_t - \mathbf{s}_t) + \frac{\|\mathbf{z}_t - \mathbf{s}_t\|_2^2}{2}.$$

Finally, the squared norm $\|\mathbf{z}_t - \mathbf{s}_t\|_2^2$ is trivially upper-bounded by $1 + m = \mathcal{O}(1)$. This completes the proof.

C.2. Proof of Lemma 4

Let $\boldsymbol{\beta}_t = \zeta \mathbf{b}_{t-1}$ (see equation (9) in Section 3.2 and related discussions), which is the implicit dual variables for dynamic resources. We recall from line (16) that

$$\begin{aligned} K_t &:= \mathbf{w}_t \cdot \mathbf{z}_t + \boldsymbol{\theta}_t \cdot (\boldsymbol{\rho} - \mathbf{z}_t) + \boldsymbol{\lambda}_t \cdot (\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t) - \zeta D_t \\ &\ge (\mathbf{w}_t - \boldsymbol{\theta}_t - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda}_t - \boldsymbol{\beta}_t) \cdot \mathbf{z}_t + \boldsymbol{\theta}_t \cdot \boldsymbol{\rho} + \boldsymbol{\lambda}_t \cdot \boldsymbol{\rho} + \boldsymbol{\beta}_t \cdot \mathbf{s}_t - \mathcal{O}(\zeta) \end{aligned} \tag{31}$$

where the last inequality follows from Lemma 3. Furthermore, we note that for all $t \le T_A$, the optimality criterion of the algorithm states that

$$\mathbf{z}_t = \arg\max_{\mathbf{z} \in \mathcal{Z}(i_t^\dagger)} (\mathbf{w}_t - \boldsymbol{\theta}_t - \mathbb{1}[i_t^\dagger = 0]\boldsymbol{\lambda}_t - \boldsymbol{\beta}_t) \cdot \mathbf{z} \tag{32}$$

In the similar spirit of Talluri and Van Ryzin (1998), the following claim shows that there always exists a static control whose value upper-bounds the net matching reward of *any feasible matching profile* (part (c)) with a marginal matching rate at most $\boldsymbol{\rho}$ in expectation (part (b)).

Electronic copy available at: https://ssrn.com/abstract=4748762

CLAIM 5 **(Static Control).** *There exists static control* $\bar{\mathbf{z}}:\mathcal{A}\to\Delta^m$ *such that:*

(a) $\bar{\mathbf{z}}(\mathbf{A})\in\mathcal{Z}(i^\dagger)$ *for any* $\mathbf{A}=(\mathbf{w},i^\dagger)$

(b) $\mathbb{E}_{\mathbf{A}}[\bar{\mathbf{z}}(\mathbf{A})]\le\boldsymbol{\rho}$

(c) $\mathbb{E}_{\mathbf{A}}[\mathbf{w}\cdot\bar{\mathbf{z}}(\mathbf{A})]\ge\frac{\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]}{T}$ *for any feasible matching profile* $\{\hat{\mathbf{z}}_t\}_{t=1}^T$ *and* $\alpha\ge1$.

We defer the detailed proof of Claim 5 to Appendix C.6.2. Based on the static control $\bar{\mathbf{z}}(\cdot)$ defined in Claim 5, we define an alternative matching decision as

$$\bar{\mathbf{z}}_t:=\bar{\mathbf{z}}(\mathbf{A}_t). \tag{33}$$

By line (31) and the optimality criterion (32), for $t\le T_A$, we have

$$K_t\ge(\mathbf{w}_t-\boldsymbol{\theta}_t-\mathbb{1}[i_t^\dagger=0]\boldsymbol{\lambda}_t-\boldsymbol{\beta}_t)\cdot\bar{\mathbf{z}}_t+\boldsymbol{\theta}_t\cdot\boldsymbol{\rho}+\boldsymbol{\lambda}_t\cdot\boldsymbol{\rho}+\boldsymbol{\beta}_t\cdot\mathbf{s}_t-\mathcal{O}(\zeta)$$

for every sample path. Taking the expectation conditional on history $\mathcal{H}_{t-1}=\{\mathbf{A}_\tau,\mathbf{s}_\tau\}_{\tau=1}^{t-1}$, we obtain the following: for any feasible matching profile $\{\hat{\mathbf{z}}_t\}_{t=1}^T$,

$$\begin{aligned}
\mathbb{E}[K_t|\mathcal{H}_{t-1}]&\ge\mathbb{E}[(\mathbf{w}_t-\boldsymbol{\theta}_t-\mathbb{1}[i_t^\dagger=0]\boldsymbol{\lambda}_t-\boldsymbol{\beta}_t)\cdot\bar{\mathbf{z}}_t+\boldsymbol{\theta}_t\cdot\boldsymbol{\rho}+\boldsymbol{\lambda}_t\cdot\boldsymbol{\rho}+\boldsymbol{\beta}_t\cdot\mathbf{s}_t-\mathcal{O}(\zeta)|\mathcal{H}_{t-1}]\\
&=\mathbb{E}[\mathbf{w}_t\cdot\bar{\mathbf{z}}_t+\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\bar{\mathbf{z}}_t)+\boldsymbol{\lambda}_t\cdot(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\bar{\mathbf{z}}_t)+\boldsymbol{\beta}_t\cdot(\mathbf{s}_t-\bar{\mathbf{z}}_t)-\mathcal{O}(\zeta)|\mathcal{H}_{t-1}]\\
&=\mathbb{E}[\mathbf{w}_t\cdot\bar{\mathbf{z}}_t|\mathcal{H}_{t-1}]+\boldsymbol{\theta}_t\cdot\mathbb{E}[(\boldsymbol{\rho}-\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]+\boldsymbol{\lambda}_t\cdot\mathbb{E}[(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]+\\
&\quad\boldsymbol{\beta}_t\cdot\mathbb{E}[(\mathbf{s}_t-\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]-\mathcal{O}(\zeta) &(34)\\
&\ge\frac{\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]}{T}+\zeta\epsilon\|\mathbf{b}_{t-1}\|_1-\mathcal{O}(\zeta) &(35)
\end{aligned}$$

Line (34) is because $(\boldsymbol{\theta}_t,\boldsymbol{\lambda}_t,\boldsymbol{\beta}_t)$ are $\mathcal{H}_{t-1}$-measurable (that is, knowing the exact realizations of $\mathcal{H}_{t-1}$ determines the value of the dual variables). In line (35), we used the property of the static control $\bar{\mathbf{z}}(\cdot)$ stated in Claim 5. To be precise, we first explain how we relate $\mathbb{E}[\mathbf{w}_t\cdot\bar{\mathbf{z}}_t|\mathcal{H}_{t-1}]$ of line (34) and $\frac{\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]}{T}$ in line (35). Note that the static control $\bar{\mathbf{z}}(\cdot)$ only depends on the arrival distribution and the current arrival type. Hence, combined with the i.i.d. nature of the arrival process, we have $\mathbb{E}[\mathbf{w}_t\cdot\bar{\mathbf{z}}_t|\mathcal{H}_{t-1}]=\mathbb{E}[\mathbf{w}_t\cdot\bar{\mathbf{z}}_t]$. Furthermore, from Claim 5-(c), we have $\mathbb{E}[\mathbf{w}_t\cdot\bar{\mathbf{z}}_t]\ge\frac{\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]}{T}$. From the same line of reasoning, we can apply Claim 5-(b) to show that $\mathbb{E}[(\boldsymbol{\rho}-\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]\ge\mathbf{0}$. Note that this also implies $\mathbb{E}[(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]\ge\mathbb{E}[(\boldsymbol{\rho}-\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]\ge\mathbf{0}$. Lastly, Claim 5-(b) again implies $\mathbb{E}[(\mathbf{s}_t-\bar{\mathbf{z}}_t)|\mathcal{H}_{t-1}]\ge\epsilon\mathbf{1}$ where $\mathbf{1}$ is the vector of ones. Combining these lower-bounds with $\boldsymbol{\theta}_t\ge\mathbf{0}$, $\boldsymbol{\lambda}_t\ge\mathbf{0}$ and $\boldsymbol{\beta}_t=\zeta\mathbf{b}_{t-1}\ge\mathbf{0}$, we obtain the final line (35).

We now sum line (35) for $t\in[T_A]$ to deduce that, for any feasible matching $\{\hat{\mathbf{z}}_t\}_{t=1}^T$ and filtration $\{\mathcal{H}_t\}_{t=1}^T$:

$$\begin{aligned}
\sum_{t=1}^{T_A}\mathbb{E}[K_t|\mathcal{H}_{t-1}]&\ge\frac{T_A\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]}{T}+\zeta\epsilon\sum_{t=1}^{T_A-1}\|\mathbf{b}_t\|_1-\mathcal{O}(T_A\zeta)\\
&=\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]-\frac{T-T_A}{T}\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]+\zeta\epsilon\sum_{t=1}^{T_A-1}\|\mathbf{b}_t\|_1-\mathcal{O}(T_A\zeta)\\
&\ge\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]-(T-T_A)+\zeta\epsilon\sum_{t=1}^{T_A-1}\|\mathbf{b}_t\|_1-\mathcal{O}(T\zeta).
\end{aligned}$$

In the last line, we used the fact that (i) $\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)\le T$ (because the per-period reward is at most one) and (ii) $T_A\le T$ for every sample path. We finally take the outer expectation of the preceding inequality over the entire history to obtain

$$\mathbb{E}\left[\sum_{t=1}^{T_A}\mathbb{E}[K_t|\mathcal{H}_{t-1}]\right]\ge\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]-\mathbb{E}\left[(T-T_A)\right]+\zeta\epsilon\mathbb{E}\left[\sum_{t=1}^{T_A-1}\|\mathbf{b}_t\|_1\right]-\mathcal{O}(T\zeta). \tag{36}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

It only remains to prove that the left-hand side is equivalent to

$$\mathbb{E}\left[\sum_{t=1}^{T_A}\mathbb{E}[K_t|\mathcal{H}_{t-1}]\right]=\mathbb{E}\left[\sum_{t=1}^{T_A}K_t\right].$$

To prove this, define a stochastic process $Y_t=K_t-\mathbb{E}[K_t|\mathcal{H}_{t-1}]$ and $X_t=\sum_{\tau=1}^{t}Y_t$ with $X_0:=0$. We observe that Y_t is $\mathcal{H}_t$-measurable and $\mathbb{E}[Y_t|\mathcal{H}_{t-1}]=0$. Hence, $\{Y_t\}$ is a Martingale difference sequence with respect to filtration $\{\mathcal{H}_t\}$. Furthermore, we note that T_A is a bounded stopping time with respect to the filtration $\{\mathcal{H}_t\}$. Hence, the optional stopping theorem implies that $\mathbb{E}[X_{T_A}]=X_0=0$, which implies line (C.2). This completes the proof.

C.3. Proof of Lemma 5

By the definition of the pseudo-rewards, we have the following:

$$\begin{aligned}
\sum_{t=1}^{T_A}K_t&=\sum_{t=1}^{T_A}\{\mathbf{w}_t\cdot\mathbf{z}_t+\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\mathbf{z}_t)+\boldsymbol{\lambda}_t\cdot(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\mathbf{z}_t)-\zeta D_t\}\\
&=\sum_{t=1}^{T_A}\mathbf{w}_t\cdot\mathbf{z}_t+\sum_{t=1}^{T_A}\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\mathbf{z}_t)+\sum_{t=1}^{T_A}\boldsymbol{\lambda}_t\cdot(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\mathbf{z}_t)-\zeta\psi(\mathbf{b}_{T_A}) &(37)\\
&=\sum_{t=1}^{T_A}\mathbf{w}_t\cdot\mathbf{z}_t-\alpha\sum_{i=1}^{m}\left(\sum_{t=1}^{T}z_{t,i}-c_i\right)_+ +\sum_{t=1}^{T_A}\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\mathbf{z}_t)-\sum_{t=1}^{T}\boldsymbol{\theta}^\star\cdot(\boldsymbol{\rho}-\mathbf{z}_t)+\\
&\quad\sum_{t=1}^{T_A}\boldsymbol{\lambda}_t\cdot(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\mathbf{z}_t)-\zeta\psi(\mathbf{b}_{T_A}) &(38)\\
&\le\underbrace{\sum_{t=1}^{T}\mathbf{w}_t\cdot\mathbf{z}_t-\alpha\sum_{i=1}^{m}\left(\sum_{t=1}^{T}z_{t,i}-c_i\right)_+}_{\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^{T})}+\underbrace{\sum_{t=1}^{T}\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\mathbf{z}_t)-\sum_{t=1}^{T}\boldsymbol{\theta}^\star\cdot(\boldsymbol{\rho}-\mathbf{z}_t)}_{R_\theta}+\\
&\quad\underbrace{\sum_{t=1}^{T_A}\boldsymbol{\lambda}_t\cdot(\boldsymbol{\rho}-\mathbb{1}[i_t^\dagger=0]\mathbf{z}_t)}_{R_\lambda}+2\alpha(T-T_A)-\zeta\psi(\mathbf{b}_{T_A}) &(39)
\end{aligned}$$

Line (37) is because of the telescoping sum of D_t and $\mathbf{b}_0=\mathbf{0}$. In line (38), we added and subtracted the over-allocation cost and used Claim 4. In line (39), we used the Cauchy-Schwartz inequality to bound $\sum_{t=1}^{T_A}\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\mathbf{z}_t)\le\sum_{t=1}^{T}\boldsymbol{\theta}_t\cdot(\boldsymbol{\rho}-\mathbf{z}_t)+2\alpha(T-T_A)$ by using that (i) $\boldsymbol{\theta}\in[0,\alpha]^m$, (ii) $\|\boldsymbol{\rho}-\mathbf{z}_t\|_1\le\|\boldsymbol{\rho}\|_1+\|\mathbf{z}_t\|_1\le 2$ (by definition of $\mathbf{z}_t$ and our assumption that $\|\boldsymbol{\rho}\|_1\le 1$)[29].

In the following claims, we show that $R_\theta\le\mathcal{O}(\sqrt{T})$ and $R_\lambda\le\mathcal{O}(\sqrt{T})-(1+2\alpha)(T-T_A)$. Plugging the bound of R_θ and R_λ into inequality (39) completes the proof of Lemma 5.

Claim 6. *$R_\theta\le\mathcal{O}(\sqrt{T})$ for every sample path.*

Proof of Claim 6. We first describe how we can view $\{\boldsymbol{\theta}_t\}_{t=1}^{T}$ as a sequence of the online mirror descent iterates for properly defined primitives stated in Claim 3. Let $f_t(\boldsymbol{\theta}):=\boldsymbol{\theta}\cdot(\boldsymbol{\rho}-\mathbf{z}_t)$, $\mathcal{D}:=[0,\alpha]^m$, $\hat{\mathbf{g}}_t:=\nabla\boldsymbol{\rho}-\mathbf{z}_t$, and $h(\boldsymbol{\theta}):=\sum_{i=1}^{m}\theta_i\log(\theta_i)$. Note that the gradient has no noise (i.e., $\hat{\mathbf{u}}_t=\mathbf{0}$). With these primitives, it is straightforward to see that the updating rule (28) is reduced to that of line (8) (for dual variables $\boldsymbol{\theta}$) in

[29] The assumption that $\|\boldsymbol{\rho}\|_1\le 1$ is without loss, and in general $\|\boldsymbol{\rho}\|_1$ only affects the regret guarantee by a constant factor.

Electronic copy available at: https://ssrn.com/abstract=4748762

Algorithm 1. Furthermore, $\|\hat{\mathbf{g}}\|_1 \le \|\mathbf{z}_t\|_1 + \|\boldsymbol{\rho}\|_1 \le 2$ and $h(\cdot)$ is $(1/\alpha)$-strongly convex with respect to $\|\cdot\|_\infty$ in $\mathcal{D}$. Hence, we invoke Claim 3 with the fixed step size $\eta_t = \eta$:

$$R_\theta = \sum_{t=1}^{T} (f_t(\boldsymbol{\theta}_t) - f_t(\boldsymbol{\theta}^\star)) \le 2\alpha\eta T + \frac{1}{\eta} V_h(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_1) \tag{40}$$

The bregman distance $D(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_1)$ is bounded because $\boldsymbol{\theta}^\star$ and $\boldsymbol{\theta}_1$ are all in $\mathcal{D}$ by definition. Hence, with $\eta = \Theta(1/\sqrt{T})$, the right-hand side of (40) is $\mathcal{O}(\sqrt{T})$. □

Claim 7. *$R_\lambda \le \mathcal{O}(\sqrt{T}) - (1+2\alpha)(T - T_A)$ for every sample path.*

Proof of Claim 7. Let us first denote the upper-bound of λ as $\bar{\lambda} := \frac{1+2\alpha}{\underline{\rho}}$. Similar to the proof of Claim 6, we define $f_t(\boldsymbol{\lambda}) := \boldsymbol{\lambda} \cdot (\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t)$, $\mathcal{D} := [0, \bar{\lambda}]^m$, $\hat{\mathbf{g}}_t := \boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t$, and $h(\boldsymbol{\lambda}) := \sum_{i=1}^m \lambda_i \log(\lambda_i)$. It is straightforward to see that the updating rule (28) is reduced to that of line (8) (for dual variable $\boldsymbol{\lambda}$) in Algorithm 1. Furthermore, $\|\hat{\mathbf{g}}\|_1 \le 2$ and $h(\cdot)$ is $1/\bar{\lambda}$-strongly convex with respect to $\|\cdot\|_\infty$ in $\mathcal{D}$. Hence, we invoke Claim 3 with the fixed step size $\eta_t = \eta$ to obtain that, for any $\boldsymbol{\lambda}^\star \in [0, \bar{\lambda}]^m$,

$$R_\lambda \le \underbrace{2\bar{\lambda}\eta T + \frac{1}{\eta} V_h(\boldsymbol{\lambda}^\star, \boldsymbol{\lambda}_1)}_{\blacklozenge} + \underbrace{\sum_{t=1}^{T_A} \boldsymbol{\lambda}^\star \cdot (\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t)}_{\diamondsuit}$$

Since both $\boldsymbol{\lambda}^\star$ and $\boldsymbol{\lambda}_1$ are bounded, the term $\blacklozenge$ is $\mathcal{O}(\sqrt{T})$ with $\eta = \Theta(1/\sqrt{T})$. We now show that, by properly choosing $\boldsymbol{\lambda}^\star$, the term $\diamondsuit$ is upper bounded by $(1+2\alpha)(T_A - T) + \bar{\lambda}$. We consider two cases. First, if $T_A = T$, then we take $\boldsymbol{\lambda}^\star = \mathbf{0}$ and the bound is trivial. Otherwise, $T_A < T$ and there exists i such that $\sum_{t=1}^{T_A} \mathbb{1}[i_t^\dagger = 0] z_{t,i} + 1 \ge \rho_i T$ (see line (17) in Section 3.3). Take that coordinate i and let $\boldsymbol{\lambda}^\star = \frac{1+2\alpha}{\rho_i}\mathbf{e}_i$. Note that $\boldsymbol{\lambda}^\star \in [0, \bar{\lambda}]^m$ by definition of $\bar{\lambda}$. The term $\diamondsuit$ is then given by

$$\begin{aligned}
\diamondsuit &= \frac{1+2\alpha}{\rho_i} \sum_{t=1}^{T_A} (\rho_i - \mathbb{1}[i_t^\dagger = 0] z_{t,i}) \\
&\le \frac{1+2\alpha}{\rho_i} ((T_A - T)\rho_i + 1) \\
&\le (1+2\alpha)(T_A - T) + \bar{\lambda}
\end{aligned}$$

The second line is by the definition of the coordinate i we have chosen. The last line simply follows from $\underline{\rho} \le \rho_i$. Hence, taking the worst case over $\boldsymbol{\lambda}^* \in \{\mathbf{0}, \frac{1+2\alpha}{\rho_1}\mathbf{e}_1, \frac{1+2\alpha}{\rho_2}\mathbf{e}_2, ..., \frac{1+2\alpha}{\rho_1}\mathbf{e}_m\}$, the term $\diamondsuit$ is always at most $(1+2\alpha)(T_A - T) + \bar{\lambda}$. This completes the proof. □

C.4. Missing Details for Step 3: Finishing the Proof of Lemma 2

In this section, we complete the proof of Lemma 6. We begin by decomposing the average backlog for the entire horizon as

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{b}_t\|_1\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T_A - 1} \|\mathbf{b}_t\|_1\right] + \frac{1}{T}\mathbb{E}\left[\sum_{t=T_A}^{T} \|\mathbf{b}_t\|_1\right] \tag{41}$$

The first (second, resp.) term is the contribution of the backlog until (after, resp.) the stopping time. We first bound the first term. Recall from the inequality ($\star$) that

$$\zeta\epsilon\mathbb{E}\left[\sum_{t=1}^{T_A - 1} \|\mathbf{b}_t\|_1\right] + \zeta\mathbb{E}[\psi(\mathbf{b}_{T_A})] \le \mathcal{O}(\sqrt{T} + \zeta T). \tag{42}$$

Noting that $\psi(\cdot)$ is always non-negative, we straightforwardly obtain

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T_A - 1} \|\mathbf{b}_t\|_1\right] \le \mathcal{O}\Big(\frac{1}{\epsilon}(1 + \frac{1}{\zeta\sqrt{T}})\Big) \le \mathcal{O}\Big(\frac{1}{\epsilon}\Big). \tag{43}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

where the last inequality is because we set $\zeta = \Theta(1/\sqrt{T})$ in the stable regime. We now turn our attention to bounding the contribution of the backlog after $t \geq T_A$. First, observe that, $\|\mathbf{b}_{T_A+\tau}\|_1 \leq \|\mathbf{b}_{T_A-1}\|_1 + \tau + 1$ for all $\tau \geq 0$ and every sample path because the backlog can only increase by one per period. Hence, we have

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=T_A}^{T} \|\mathbf{b}_t\|_1\right] \leq \mathbb{E}\left[\frac{(T-T_A+1)}{T}\|\mathbf{b}_{T_A-1}\|_1 + \frac{(T-T_A+1)(T-T_A+2)}{2T}\right] \tag{44}$$

$$\leq \frac{\sqrt{\mathbb{E}[(T-T_A+1)^2]}\sqrt{\mathbb{E}[\|\mathbf{b}_{T_A-1}\|_1^2]}}{T} + \frac{\mathbb{E}[(T-T_A+1)(T-T_A+2)]}{T} \tag{45}$$

$$\leq \mathcal{O}\left(\frac{\sqrt{\mathbb{E}[(T-T_A)^2]}\sqrt{\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2]} + \mathbb{E}[(T-T_A)^2]}{T}\right) \tag{46}$$

In the second line, we used the Cauchy-Schwartz inequality for expectation. In the last line, for the sake of brevity, we expressed the bound with the asymptotic notations.[30] We will prove that line (46) is $\mathcal{O}(1)$, which will complete the proof. We prove this through two claims: (i) $\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2] \leq \mathcal{O}(T)$ (Claim 8) and (ii) $\mathbb{E}[(T-T_A)^2] \leq \mathcal{O}(T)$. For the latter, we prove a stronger claim by showing that $T - T_A \leq \mathcal{O}(\sqrt{T})$ for every sample path (Claim 9).

Claim 8. $\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2] \leq \mathcal{O}(T)$

Proof of Claim 8. The proof is the direct consequence of line (42). To be precise, because the backlog is non-negative, the line (42) implies that

$$\mathbb{E}[\psi(\mathbf{b}_{T_A})] \leq \mathcal{O}\left(T + \frac{\sqrt{T}}{\zeta}\right) = \mathcal{O}(T)$$

where the last inequality follows since $\zeta = \Theta(1/\sqrt{T})$. We finally recall the definition of $\psi(\mathbf{b}_{T_A}) = \frac{1}{2}\|\mathbf{b}_{T_A}\|_2^2$. By Cauchy-Schwarz inequality, we have $\|\mathbf{b}_{T_A}\|_1^2 \leq m\|\mathbf{b}_{T_A}\|_2^2$ (for every sample path). Hence, $\mathcal{O}(T)$ bound on $\mathbb{E}[\psi(\mathbf{b}_{T_A})]$ is equivalent of showing the same order of bound on $\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2]$.[31] This completes the proof. □

In the following, we obtain a deterministic bound on the stopping time of `CA-DL`. The proof closely follows that of Proposition 2 in Balseiro et al. (2023b). For the sake of completeness, we reproduce and tailor the proof to our setting.

Claim 9 **(Bound on Stopping time for `CA-DL`)**. $T - T_A \leq \mathcal{O}(\sqrt{T})$ *for every sample path.*

Proof of Claim 9. The proof relies on studying the trajectory of the dual variables $\{\boldsymbol{\lambda}_t\}$ up to the stopping time. Recall that the updating rule (8) of Algorithm 1 requires projection of $\lambda_{t,i}$ onto the upper bound $\bar{\lambda} := \frac{1+2\alpha}{\rho}$. We will show that, for all $t \leq T_A$ and for every sample path, the trajectory of $\{\boldsymbol{\lambda}_t\}$ produced by Algorithm 1 satisfies

$$\lambda_{t,i}\exp(\eta(\mathbb{1}[i_t^\dagger = 0]z_{t,i} - \rho_i)) \leq \bar{\lambda} \tag{No Projection}$$

That is, the projection onto the upper-bound $\bar{\lambda}$ never binds. Before proving this, we first finish the proof of Claim 9 building on line (No Projection). Applying induction on t, we deduce from (No Projection) that,

$$\lambda_{t+1,i} = \lambda_{t,i}\exp(\eta(\mathbb{1}[i_t^\dagger = 0]z_{t,i} - \rho_i)) \tag{47}$$

[30] Note that this is without loss of generality since the omitted terms are all constant.

[31] Because of our assumption that the capacity grows linearly with T, the number of affiliates must be $O(1)$.

Electronic copy available at: https://ssrn.com/abstract=4748762

for all $t \le T_A, i \in [m]$ and for every sample path. On the other hand, by the definition of the stopping time T_A (see (17) in Section 3.3), there exists affiliate i such that $\sum_{t=1}^{T_A} \mathbb{1}[i_t^\dagger = 0] z_{t,i} + 1 \ge \rho_i T$, which is equivalent to

$$T - T_A \le \frac{\sum_{t=1}^{T_A} (\mathbb{1}[i_t^\dagger = 0] z_{t,i} - \rho_i)}{\rho_i} + \frac{1}{\rho_i}. \tag{48}$$

For such affiliate i, we note that line (47) is equivalent to

$$\mathbb{1}[i_t^\dagger = 0] z_{t,i} - \rho_i = \frac{\log(\lambda_{t+1,i}) - \log(\lambda_{t,i})}{\eta}$$

for all $t \le T_A$. Plugging this equality into line (48), we have

$$T - T_A \le \frac{\sum_{t=1}^{T_A} \{\log(\lambda_{t+1,i}) - \log(\lambda_{t,i})\}}{\eta \rho_i} + \frac{1}{\rho_i} \tag{49}$$

$$= \frac{\log(\lambda_{T_A+1,i}) + 1}{\eta \rho_i} + \frac{1}{\rho_i} \tag{50}$$

$$\le \mathcal{O}(\sqrt{T}). \tag{51}$$

In the second line, we used the telescoping sum[32] and the fact that $\lambda_{1,i} = \exp(-1)$. The last line follows because $\lambda_{T_A+1,i} \le \bar{\lambda}$ by definition and $\eta = \Theta(1/\sqrt{T})$. Taking the worst case over $i \in [m]$, we conclude that $T - T_A \le \mathcal{O}(\sqrt{T})$ for every sample path.

The remaining step is to prove the line (No Projection). The proof is by induction on $t \le T_A$. First, as we are interested in the asymptotic regime of T, let T be large enough so that $\eta = \Theta(1/\sqrt{T})$ satisfies

$$\eta \le \min_{i \in [m]} \frac{\log(1+\rho_i)}{1-\rho_i}. \tag{52}$$

Base Case. As a base case for the induction, We start from $t = 1$. For any fixed i, recall that the initial dual variable is $\lambda_{1,i} = \exp(-1)$.[33] Hence, we have

$$\begin{aligned} \lambda_{1,i} \exp(\eta(\mathbb{1}[i_1^\dagger = 0] z_{1,i} - \rho_i)) &= \exp(-1) \exp(\eta(\mathbb{1}[i_1^\dagger = 0] z_{1,i} - \rho_i)) \\ &\le \exp(-1 + \log(1+\rho_i)) \\ &= \exp(-1)(1+\rho_i) \le \bar{\lambda}. \end{aligned}$$

The second line is because $\mathbb{1}[i_1^\dagger = 0] z_{1,i} \le 1$ and T is large enough to satisfy (52). The last inequality is because $\rho_i \le 1$ (by our assumption) and $\lambda = \frac{1+2\alpha}{\underline{\rho}} > 1$. Hence, we have proved the base case for $t = 1$.

Inductive Step. Suppose that line (No Projection) holds for all $\tau \le t-1$ for some fixed $2 \le t \le T_A$. By the induction hypothesis and the updating rule (8) (for $\boldsymbol{\lambda}$), we have $\lambda_{\tau,i} \le \bar{\lambda}$ for all $\tau \le t$. Define

$$\tilde{\lambda}_{t+1,i} := \lambda_{t,i} \exp(\eta(\mathbb{1}[i_t^\dagger = 0] z_{t,i} - \rho_i)). \tag{53}$$

To complete the inductive step, we need to show that $\tilde{\lambda}_{t+1,i} \le \bar{\lambda}$. Without loss, assume $\lambda_{t,i} > 0$ since otherwise the result trivially follows. Furthermore, if $i_t^\dagger \ne 0$, the result also follows since $\tilde{\lambda}_{t+1,i} \le \lambda_{t,i} \le \bar{\lambda}$. Hence, assume that $i_t^\dagger = 0$. Recall that the decision for $t \le T_A$ is not constrained by the hard capacity constraint. Hence, it can be written as

$$\mathbf{z}_t = \arg\max_{\mathbf{z} \in \Delta^m} (\mathbf{w}_t - \boldsymbol{\theta}_t - \boldsymbol{\lambda}_t - \boldsymbol{\beta}_t) \cdot \mathbf{z}. \tag{54}$$

[32] Note that, as we highlight in Remark 1, this reasoning breaks when the step-size for $\boldsymbol{\lambda}$ is time-dependent (e.g, $\eta_t = \Theta(1/\sqrt{t})$).

[33] The proof can be straightforwardly extended to a general value of $\boldsymbol{\lambda}_1$ by slightly changing line (52).

Electronic copy available at: https://ssrn.com/abstract=4748762

Because $\mathbf{0} \in \Delta^m$, we always have $(\mathbf{w}_t - \boldsymbol{\theta}_t - \boldsymbol{\lambda}_t - \boldsymbol{\beta}_t) \cdot \mathbf{z}_t \geq 0$. Combining with the fact that the reward is at most one and the dual variables are all non-negative, we have $z_{t,i} \leq \frac{1}{\lambda_{t,i}}$. Plugging this into (53), we have

$$\tilde{\lambda}_{t+1,i} \leq \lambda_{t,i} \exp\left(\eta \min\left\{1, \frac{1}{\lambda_{t,i}}\right\} - \eta\rho_i\right). \tag{55}$$

We now consider two cases. If $\lambda_{t,i} \geq \frac{1}{\rho_i}$, then $\tilde{\lambda}_{t+1,i} \leq \lambda_{t,i} \leq \bar{\lambda}$ and hence the result follows. Otherwise, from line (55) we have

$$\tilde{\lambda}_{t+1,i} \leq \lambda_{t,i} \exp\left(\eta(1-\rho_i)\right) \leq \frac{1}{\rho_i} \exp\left(\eta(1-\rho_i)\right) \leq \frac{1+\rho_i}{\rho_i} \leq \bar{\lambda}. \tag{56}$$

In the second inequality, we used the assumption of the second case that $\lambda_{t,i} \leq \frac{1}{\rho_i}$. In the third inequality, we used the fact that T is large enough that η satisfies (52). The last inequality follows since $\bar{\lambda} = \frac{1+2\alpha}{\underline{\rho}}$ and $\alpha \geq 1$ by our assumption (see Section 2). This completes the inductive step. Hence, we have shown (No Projection) and completed the proof. □

Remark 1. The proof for part (b) fails with the variable step-size $\eta_t = \Theta(1/\sqrt{t})$. To see this, with the variable step-size, line (50) becomes

$$\frac{1}{\rho_i}\left(\underbrace{\sum_{t=1}^{T_A-1} \log(\lambda_{t+1,i})\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}}\right)}_{(\star)} + \frac{1}{\eta_{T_A}}\log(\lambda_{T_A+1,i}) - \frac{1}{\eta_1}\log(\lambda_{1,i})\right) + \frac{1}{\rho_i} \tag{57}$$

Term $(\star)$ cannot be removed or bounded above because the sign of $\log(\lambda_{t+1,i})$ is ambiguous.

C.5. Proof of Corollary 1

Our first step is to show that whenever the backlog is above $1/\zeta$, we have a constant negative drift on the expected backlog.

Claim 10. *Let $t \leq T_A$. Whenever $b_{t-1,i} \geq 1/\zeta$, we have $\mathbb{E}[b_{t,i} - b_{t-1,i} | H_{t-1}] \leq -(\rho_i - \mathbb{P}[i_t^\dagger = i])$*

Proof of Claim 10. Because $1/\zeta = \sqrt{T/\gamma}$ is super-constant (by our assumption that $\gamma = \mathcal{O}(T)$), we can assume that $1/\zeta > 1$ without loss. Hence, $b_{t,i} - b_{t-1,i} = z_{t,i} - s_{t,i}$ whenever $b_{t-1,i} \geq 1/\zeta$. Furthermore, we can write $z_{t,i} = \mathbb{1}[i_t^\dagger = i]$ whenever $b_{t-1,i} \geq 1/\zeta$. To see this, whenever $i_t^\dagger = 0$ (a free case) and $b_{t-1,i} \geq 1/\zeta$, we observe that the adjusted score $w_{t,i} - \theta_{t,i} - \lambda_{t,i} - \zeta b_{t-1,i}$ is always non-positive. Because the adjusted score of the dummy affiliate is always zero (by definition), we have $z_{t,i} = 0$ whenever $b_{t-1,i} > 1/\zeta$. If the arrival t is a tied case to location $j \neq i$, we again have $z_{t,i} = 0$. Combining, we have

$$\mathbb{E}[b_{t,i} - b_{t-1,i} | H_{t-1}] = \mathbb{E}[z_{t,i} - s_{t,i} | H_{t-1}] = \mathbb{P}[i_t^\dagger = i] - \rho_i - \epsilon.$$

The proof is complete because $\epsilon \geq 0$. □

We now prove the following lemma, which establishes $\mathcal{O}\left(\frac{1}{\zeta} + 1\right)$ bound on the average backlog (in expectation) for any $\epsilon \geq 0$.

Lemma 7 **(Upper Bound on Backlog for Algorithm 1).** *For any given $\zeta \geq 0$, the expected average backlog of Algorithm 1 is*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{b}_t\|_1\right] \leq \mathcal{O}\left(\frac{1}{\zeta} + 1\right).$$

Proof of Lemma 7. We apply the following result extracted from Wei et al. (2023) and Gupta (2022).

Electronic copy available at: https://ssrn.com/abstract=4748762

CLAIM 11 **(Lemma 5 of Wei et al. (2023))**. *Let $\Psi(t)$ be an $\{\mathcal{H}_t\}$-adapted stochastic process satisfying:*

- *Bounded variation:* $|\Psi(t+1)-\Psi(t)| \le K$
- *Negative drift:* $\mathbb{E}[\Psi(t+1)-\Psi(t)|\mathcal{H}_t] \le -d$ *whenever* $\Psi(t) \ge D$
- $\Psi(0) \le K+D$

Then, we have

$$\mathbb{E}[\Psi(t)] \le K\left(1+\lceil \frac{D}{K} \rceil\right) + K\left(\frac{K-d}{2d}\right).$$

To apply Claim 11 to our setting, let $\Psi(t) := b_{t,i}$ let $d_i = \rho_i - \mathbb{P}[i_t^\dagger = i]$. We recall that $d_i > 0$ is a positive constant by our assumption (see Section 2). Furthermore, we have $b_{0,i} = 0$ and $|b_{t,i} - b_{t-1,i}| \le 1$. From Claim 10, we have $\mathbb{E}[b_{t,i} - b_{t-1,i}|H_t] \le -d_i$ whenever $b_{t-1,i} \ge 1/\zeta$. Applying Claim 11 with, $K=1$, and $D = 1/\zeta$, we have

$$\mathbb{E}[b_{t,i}] \le 1 + \lceil \frac{1}{\zeta} \rceil + \frac{1-d_i}{2d_i} \le \frac{1}{\zeta} + \frac{1+3d_i}{2d_i} \tag{58}$$

Hence, we have $\mathbb{E}[\|\mathbf{b}_t\|_1] \le \mathcal{O}\left(\frac{1}{\zeta}+1\right)$ for all $t \le T_A$. Finally, the average backlog is bounded by

$$\begin{aligned}
\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\right] &= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T_A}\|\mathbf{b}_t\|_1\right] + \frac{1}{T}\mathbb{E}\left[\sum_{t=T_A+1}^{T}\|\mathbf{b}_t\|_1\right] \\
&\le \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T_A}\|\mathbf{b}_t\|_1\right] + \mathbb{E}\left[\frac{(T-T_A)}{T}\|\mathbf{b}_{T_A}\|_1 + \frac{(T-T_A)(T-T_A+1)}{2T}\right] \\
&\le \mathcal{O}\left(\frac{1}{\zeta}+1\right)
\end{aligned}$$

The last inequality is because we have $\mathbb{E}[\|\mathbf{b}_t\|_1] \le \mathcal{O}\left(\frac{1}{\zeta}+1\right)$ for all $t \le T_A$ from inequality (58) and $T - T_A \le \mathcal{O}(\sqrt{T})$ almost surely from Claim 9 (Appendix C.4). This completes the proof of Lemma 7. □

Proof of Corollary 1. We now complete the proof of Corollary 1. For ease of reference, we recall the inequality ($\star$) (see Section 3.3) in the following: for *any* feasible matching profile $\{\hat{\mathbf{z}}_t\}_{t=1}^T$,

$$\underbrace{\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)] - \mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T;\alpha)]}_{(\mathsf{A})} + \zeta\epsilon\mathbb{E}\left[\sum_{t=1}^{T_A-1}\|\mathbf{b}_t\|_1\right] + \zeta\mathbb{E}[\psi(\mathbf{b}_{T_A})] \le \mathcal{O}(\sqrt{T}+\zeta T)$$

In fact, we did not rely on the fact that ϵ is a strictly positive constant to derive the line ($\star$). Hence, the above inequality holds for *any* $\epsilon \ge 0$ under the near-critical regime. This implies that the term (A) is still $\mathcal{O}(\sqrt{T}+\zeta T)$. Let $\{\mathbf{z}_t^*\}_{t=1}^T$ denote the the optimal offline solution. By utilizing the same decomposition as line (11), we bound the regret by

$$\begin{aligned}
\mathbb{E}[\mathsf{OPT}(\alpha,\gamma)] - \mathbb{E}[\mathsf{ALG}(\alpha,\gamma)] &\le \underbrace{\mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t^*\}_{t=1}^T;\alpha)] - \mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T;\alpha)]}_{(\mathsf{A})} + \underbrace{\frac{\gamma}{T}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\right]}_{\text{Lemma 7}} \\
&\le \mathcal{O}(\zeta T + \sqrt{T}) + \mathcal{O}\left(\frac{\gamma}{\zeta}+\gamma\right) \\
&\le \mathcal{O}(\sqrt{\gamma T})
\end{aligned}$$

where the last line follows from $\zeta = \Theta(\sqrt{\frac{\gamma}{T}})$ under the near-critical regime.

Electronic copy available at: https://ssrn.com/abstract=4748762

C.6. Proof of Supporting Claims

C.6.1. Proof of Claim 3 For any given $\boldsymbol{\nu} \in \mathcal{D}$ and $\mathbf{g}_t = \nabla f_t(\boldsymbol{\nu}_t)$,

$$
\begin{aligned}
f_t(\boldsymbol{\nu}_t) - f_t(\boldsymbol{\nu}) &\le \mathbf{g}_t \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) \\
&= \hat{\mathbf{g}}_t \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) + (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) \\
&= \frac{1}{\eta_t}(\nabla h(\boldsymbol{\nu}_t) - \nabla h(\tilde{\boldsymbol{\nu}}_{t+1})) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) + (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) && (59) \\
&= \frac{1}{\eta_t}(V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) + V_h(\boldsymbol{\nu}_t, \tilde{\boldsymbol{\nu}}_{t+1}) - V_h(\boldsymbol{\nu}, \tilde{\boldsymbol{\nu}}_{t+1})) + (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) && (60) \\
&\le \frac{1}{\eta_t}(V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) + V_h(\boldsymbol{\nu}_t, \tilde{\boldsymbol{\nu}}_{t+1}) - V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_{t+1}) - V_h(\boldsymbol{\nu}_{t+1}, \tilde{\boldsymbol{\nu}}_{t+1})) + (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) && (61) \\
&= \frac{1}{\eta_t}(V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) - V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_{t+1})) + \frac{1}{\eta_t} \underbrace{(V_h(\boldsymbol{\nu}_t, \tilde{\boldsymbol{\nu}}_{t+1}) - V_h(\boldsymbol{\nu}_{t+1}, \tilde{\boldsymbol{\nu}}_{t+1}))}_{(\star)} + (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) && (62)
\end{aligned}
$$

The first line is because of the definition of the subgradient. Line (59) is from the updating rule (28). Line (60) is from the three-point equality property of V_h (Lemma 5.2 of Bubeck (2011)). Line (61) is due to Generalized Pythagorean Theorem (Lemma 5.3 of Bubeck (2011)). We now bound the term $(\star)$ as

$$
\begin{aligned}
(\star) &= h(\boldsymbol{\nu}_t) - h(\boldsymbol{\nu}_{t+1}) - \nabla h(\tilde{\boldsymbol{\nu}}_{t+1}) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}) && (63) \\
&\le (\nabla h(\boldsymbol{\nu}_t) - \nabla h(\tilde{\boldsymbol{\nu}}_{t+1})) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}) - \frac{\sigma}{2}\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}\|^2 && (64) \\
&= \eta_t \hat{\mathbf{g}}_t \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}) - \frac{\sigma}{2}\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}\|^2 && (65) \\
&\le \eta_t \|\hat{\mathbf{g}}_t\|_* \|\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}\| - \frac{\sigma}{2}\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_{t+1}\|^2 && (66) \\
&\le \frac{\eta_t^2 \|\hat{\mathbf{g}}_t\|_*^2}{2\sigma} && (67)
\end{aligned}
$$

Line (63) is by the definition of V_h. Line (64) is because $h(\cdot)$ is σ-strongly convex with respect to $\|\cdot\|$. Line (65) is again due to the updating rule (28). In the final two lines, we used the generalized Cauchy-Schartz inequality and the fact that $az - bz^2 \le \frac{a^2}{4b}$ for any $a > 0, b > 0$ and $z \in \mathbb{R}$.

Plugging the bound of $(\star)$ in (67) into (62), we have

$$
f_t(\boldsymbol{\nu}_t) - f_t(\boldsymbol{\nu}) \le \frac{1}{\eta_t}(V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) - V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_{t+1})) + \frac{\eta_t \|\hat{\mathbf{g}}_t\|_*^2}{2\sigma} + (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) \tag{68}
$$

We sum up the preceding inequality for $k \le t \le s$ to obtain

$$
\begin{aligned}
\sum_{t=k}^{s}(f_t(\boldsymbol{\nu}_t) - f_t(\boldsymbol{\nu})) &\le \sum_{t=k}^{s} \frac{1}{\eta_t}(V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) - V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_{t+1})) + \sum_{t=k}^{s} \frac{\eta_t \|\hat{\mathbf{g}}_t\|_*^2}{2\sigma} + \sum_{t=k}^{s} (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) \\
&= \sum_{t=k+1}^{s} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) + \frac{1}{\eta_k} V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_k) - \frac{1}{\eta_s} V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_{s+1}) + \sum_{t=k}^{s} \frac{\eta_t \|\hat{\mathbf{g}}_t\|_*^2}{2\sigma} + \\
&\quad \sum_{t=k}^{s} (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) \\
&\le \sum_{t=k+1}^{s} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t) + \frac{1}{\eta_k} V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_k) + \sum_{t=k}^{s} \frac{\eta_t \|\hat{\mathbf{g}}_t\|_*^2}{2\sigma} + \sum_{t=k}^{s} (\mathbf{g}_t - \hat{\mathbf{g}}_t) \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu})
\end{aligned}
$$

where the last inequality follows from the non-negativity of V_h. This completes the proof.

Electronic copy available at: https://ssrn.com/abstract=4748762

C.6.2. Proof of Claim 5 Consider the following optimization problem.

$$\mathsf{P}(\mathbf{z}) := \mathbb{E}\left[\mathbf{w}\cdot\mathbf{z}(\mathbf{A})\right] - \alpha\sum_{i=1}^{m}(\mathbb{E}[z_i(\mathbf{A})]-\rho_i)_+ \quad \text{s.t. } \mathbb{E}\left[\mathbb{1}[i^\dagger=0]\cdot\mathbf{z}(\mathbf{A})\right]\le\boldsymbol{\rho}$$
$$\mathsf{P}^\star(\alpha) := \max_{\substack{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)\\ \forall\mathbf{A}\in\mathcal{A}}} \mathsf{P}(\mathbf{z})$$

where the expectation is with respect to $\mathbf{A}=(\mathbf{w},i^\dagger)\sim\mathcal{F}$. This optimization problem can be viewed as the static problem maximizing the expected per-period net matching reward. The decision is to choose a mapping $\mathbf{z}:\mathcal{A}\to\Delta^m$ subject to $\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)$. That is, for each arrival type $\mathbf{A}=(\mathbf{w},i^\dagger)$, we choose the matching rate $\mathbf{z}(\mathbf{A})\in\Delta^m$ if $i^\dagger=0$ (free case) or we set it as $\mathbf{z}(\mathbf{A})=\mathbf{e}_{i^\dagger}$ otherwise. The objective is to maximize the functional $\mathsf{P}(\mathbf{z})$, which is the expected per-period net matching reward given the control $\mathbf{z}(\cdot)$, subject to the single-period version of the hard constraint on the matching rate of free cases.

Define $\bar{\mathbf{z}}(\cdot)$ as the optimal mapping that solves the single-period primal problem in (C.6.2). In the following, we show that $\bar{\mathbf{z}}(\cdot)$ satisfies the condition (a)-(c) stated in Claim 5.

Proof of Claim 5-(a). This directly follows from the feasibility condition of the mapping $\bar{\mathbf{z}}(\cdot)$. □

Proof of Claim 5-(b). Suppose for contradiction that $\mathbb{E}[\bar{z}_i(\mathbf{A})]>\rho_i$ for some $i\in[m]$. For affiliate i being over-allocated, we define alternative control $\tilde{\mathbf{z}}$ defined by

$$\tilde{z}_i(\mathbf{A}) = \bar{z}_i(\mathbf{A}) - \delta\mathbb{1}[i^\dagger=0] \quad \forall\mathbf{A}\in\mathcal{A}$$
$$\tilde{z}_j(\mathbf{A}) = \bar{z}_j(\mathbf{A}) \quad \forall j\neq i, \forall\mathbf{A}\in\mathcal{A}$$

for infinitesimally small $\delta>0$. To be precise, we set δ small enough to maintain $\mathbb{E}[\tilde{z}_i(\mathbf{A})]\ge\rho_i$. The alternative $\tilde{\mathbf{z}}$ is identical to $\bar{\mathbf{z}}$ except that the matching rate of free cases to affiliate i is decreased by δ.[34] Note that $\tilde{\mathbf{z}}$ is feasible for the program $\mathsf{P}^\star(\alpha)$. We show that $\tilde{\mathbf{z}}$ can strictly improve the objective value (C.6.2) upon the value achieved by $\bar{\mathbf{z}}(\cdot)$, which is a contradiction of the optimality of $\bar{\mathbf{z}}(\cdot)$. To see this,

$$\begin{aligned}\mathsf{P}(\tilde{\mathbf{z}})-\mathsf{P}(\bar{\mathbf{z}}) &= \alpha\sum_{j=1}^{m}\{(\mathbb{E}[\bar{z}_j(\mathbf{A})]-\rho_i)_+ - (\mathbb{E}[\tilde{z}_j(\mathbf{A})]-\rho_i)_+\} + \mathbb{E}\left[\mathbf{w}\cdot\tilde{\mathbf{z}}(\mathbf{A})-\mathbf{w}\cdot\bar{\mathbf{z}}(\mathbf{A})\right]\\ &= \alpha\delta\mathbb{E}[\mathbb{1}[i^\dagger=0]] - \mathbb{E}[w_i\delta\mathbb{1}[i^\dagger=0]]\\ &\ge \delta\Pr[i^\dagger=0](\alpha-1)\end{aligned}$$

The first line is from the definition of the functional $\mathsf{P}(\cdot)$. In the second line, we used the definition of $\tilde{\mathbf{z}}$ in line (C.6.2). In the last line, we used the fact the $w_i\le 1$. By our assumption, we have $\alpha>1$ and $\Pr[i^\dagger=0]>0$. Hence, the change in the objective the objective is always strictly positive. This completes the proof of Claim 5-(b). □

Proof of Claim 5-(c). Let $\{\hat{\mathbf{z}}_t\}_{t=1}^T$ be an arbitrary feasible matching profile. From part (b), we have $\mathsf{P}^\star(\alpha)=\mathbb{E}[\mathbf{w}\cdot\bar{\mathbf{z}}(\mathbf{A})]$. Hence, it suffices to prove $\mathsf{P}^\star(\alpha)\ge\mathbb{E}\left[\frac{\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)}{T}\right]$. We prove this via two steps. First, we define a dual program of $\mathsf{P}^\star(\alpha)$ and show the strong-duality. Second, we show that the value of the dual program is at least $\mathbb{E}\left[\frac{\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)}{T}\right]$. Define the dual problem of $\mathsf{P}^\star(\alpha)$ as:

$$\mathsf{D}(\boldsymbol{\theta},\boldsymbol{\lambda}) := \mathbb{E}\left[\max_{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)}(\mathbf{w}-\boldsymbol{\theta}-\mathbb{1}[i^\dagger=0]\boldsymbol{\lambda})\cdot\mathbf{z}(\mathbf{A})+\boldsymbol{\rho}\cdot\boldsymbol{\theta}+\boldsymbol{\rho}\cdot\boldsymbol{\lambda}\right]$$
$$\mathsf{D}^\star(\alpha) := \min_{\boldsymbol{\theta}\in[0,\alpha]^m,\boldsymbol{\lambda}\ge\mathbf{0}}\mathsf{D}(\boldsymbol{\theta},\boldsymbol{\lambda}).$$

[34] Note that this is well-defined because of our assumption $\mathbb{P}[i^\dagger=i]<\rho_i$.

Electronic copy available at: https://ssrn.com/abstract=4748762

We first show that $\mathsf{P}^\star(\alpha) = \mathsf{D}^\star(\alpha)$.

$$\begin{aligned}
\mathsf{P}^\star(\alpha) &= \max_{\substack{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)\\ \forall \mathbf{A}\in\mathcal{A}}} \min_{\boldsymbol{\lambda}\geq \mathbf{0}} \mathbb{E}\left[\mathbf{w}\cdot\mathbf{z}(\mathbf{A})\right] - \alpha\sum_{i=1}^{m}(\mathbb{E}[z_i(\mathbf{A})]-\rho_i)_+ - \boldsymbol{\lambda}\cdot\left(\mathbb{E}\left[\mathbb{1}[i^\dagger=0]\cdot\mathbf{z}(\mathbf{A})\right]-\boldsymbol{\rho}\right)\\
&= \max_{\substack{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)\\ \forall \mathbf{A}\in\mathcal{A}}} \min_{\boldsymbol{\lambda}\geq \mathbf{0}} \min_{\boldsymbol{\theta}\in[0,\alpha]^m} \mathbb{E}\left[\mathbf{w}\cdot\mathbf{z}(\mathbf{A})\right] - \boldsymbol{\theta}\cdot(\mathbb{E}[\mathbf{z}(\mathbf{A})]-\boldsymbol{\rho}) - \boldsymbol{\lambda}\cdot\left(\mathbb{E}\left[\mathbb{1}[i^\dagger=0]\cdot\mathbf{z}(\mathbf{A})\right]-\boldsymbol{\rho}\right) \qquad (69)\\
&= \max_{\substack{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)\\ \forall \mathbf{A}\in\mathcal{A}}} \min_{\boldsymbol{\lambda}\geq \mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m} \mathbb{E}[(\mathbf{w}-\boldsymbol{\theta}-\mathbb{1}[i^\dagger=0]\boldsymbol{\lambda})\cdot\mathbf{z}(\mathbf{A}) + \boldsymbol{\rho}\cdot(\boldsymbol{\theta}+\boldsymbol{\lambda})]\\
&= \min_{\boldsymbol{\lambda}\geq \mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m} \max_{\substack{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)\\ \forall \mathbf{A}\in\mathcal{A}}} \mathbb{E}[(\mathbf{w}-\boldsymbol{\theta}-\mathbb{1}[i^\dagger=0]\boldsymbol{\lambda})\cdot\mathbf{z}(\mathbf{A}) + \boldsymbol{\rho}\cdot(\boldsymbol{\theta}+\boldsymbol{\lambda})] \qquad (70)\\
&= \min_{\boldsymbol{\lambda}\geq \mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m} \mathbb{E}\left[\max_{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)} (\mathbf{w}-\boldsymbol{\theta}-\mathbb{1}[i^\dagger=0]\boldsymbol{\lambda})\cdot\mathbf{z}(\mathbf{A}) + \boldsymbol{\rho}\cdot\boldsymbol{\theta} + \boldsymbol{\rho}\cdot\boldsymbol{\lambda}\right]\\
&= \mathsf{D}^\star(\alpha)
\end{aligned}$$

Line (69) follows from Fenchel duality $\alpha(x)_+ = \max_{\theta\in[0,\alpha]}\alpha x$. In line (70), we used Sion's min-max theorem (Sion 1958).[35]

Finally, we show, for any feasible matching profile $\{\hat{\mathbf{z}}_t\}_{t=1}^T$, that $\mathsf{D}^\star(\alpha)\geq\mathbb{E}[\frac{\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]}{T}$.

$$\begin{aligned}
&\mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)]\\
&\leq \mathbb{E}\left[\max_{\mathbf{z}_t\in\mathcal{Z}(i_t^\dagger),t\in[T]} \sum_{t=1}^T \mathbf{w}_t\cdot\mathbf{z}_t - \alpha\sum_{i=1}^m\left(\sum_{t=1}^T z_{t,i}-c_i\right)_+ \quad \text{s.t } \sum_{t=1}^T \mathbb{1}[i_t^\dagger=0]z_{t,i}\leq c_i\right] \qquad (71)\\
&\leq \mathbb{E}\left[\max_{\mathbf{z}_t\in\mathcal{Z}(i_t^\dagger),t\in[T]} \sum_{t=1}^T \mathbf{w}_t\cdot\mathbf{z}_t - \alpha\sum_{i=1}^m\left(\sum_{t=1}^T z_{t,i}-c_i\right)_+ - \boldsymbol{\lambda}\cdot\left(\sum_{t=1}^T\mathbb{1}[i_t^\dagger=0]\mathbf{z}_t-\mathbf{c}\right)\right] \quad \forall\boldsymbol{\lambda}\geq\mathbf{0}\\
&\leq \mathbb{E}\left[\max_{\mathbf{z}_t\in\mathcal{Z}(i_t^\dagger),t\in[T]} \sum_{t=1}^T \mathbf{w}_t\cdot\mathbf{z}_t - \sum_{i=1}^m\theta_i\left(\sum_{t=1}^T z_{t,i}-c_i\right) - \boldsymbol{\lambda}\cdot\left(\sum_{t=1}^T\mathbb{1}[i_t^\dagger=0]\mathbf{z}_t-\mathbf{c}\right)\right] \quad \forall\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m \qquad (72)\\
&= \sum_{t=1}^T \mathbb{E}\left[\max_{\mathbf{z}_t\in\mathcal{Z}(i_t^\dagger)} (\mathbf{w}_t-\boldsymbol{\theta}-\mathbb{1}[i_t^\dagger=0]\boldsymbol{\lambda})\cdot\mathbf{z}_t + \boldsymbol{\rho}\cdot\boldsymbol{\theta}+\boldsymbol{\rho}\cdot\boldsymbol{\lambda}\right] \quad \forall\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m \qquad (73)\\
&= T\mathbb{E}\left[\max_{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)} (\mathbf{w}-\boldsymbol{\theta}-\mathbb{1}[i^\dagger=0]\boldsymbol{\lambda})\cdot\mathbf{z}(\mathbf{A}) + \boldsymbol{\rho}\cdot\boldsymbol{\theta}+\boldsymbol{\rho}\cdot\boldsymbol{\lambda}\right] \quad \forall\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m \qquad (74)\\
&= TD(\boldsymbol{\theta},\boldsymbol{\lambda}) \quad \forall\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m.
\end{aligned}$$

Line (71) follows from upper-bounding $\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T;\alpha)$ by the maximum total net matching reward (subject to the same feasibility condition and on the same sample path). In line (72), we used the fact that $\alpha(x)_+\geq\theta x$ for any $\theta\in[0,\alpha]$ and $x\in\mathbb{R}$ (see Claim 4). Line (73) follows from the fact that the feasible set $\mathcal{Z}(i_t^\dagger)$ only depends on the type of arrival t and hence we can move the maximum operator inside the summation. Line (74) is because the arrival type $\mathbf{A}_t=(\mathbf{w}_t,i_t^\dagger)$ is i.i.d. random variables. The proof is complete by taking the minimum over $\boldsymbol{\lambda}\geq\mathbf{0},\boldsymbol{\theta}\in[0,\alpha]^m$. □

Appendix D: Missing Details and Proofs of Section 4

D.1. Formal description of CO-DL

We provide the formal description of the congestion-oblivious algorithm in Algorithm 2.

[35] This is justified because each type-dependent feasible set $\mathcal{Z}(\cdot)$ is convex and compact.

Electronic copy available at: https://ssrn.com/abstract=4748762

Algorithm 2 Congestion-Oblivious Dual-Learning (`CO-DL`) Algorithm

Input: T, $\boldsymbol{\rho}$, constant $k>0$, and $\eta_t = \frac{k}{\sqrt{t}}$ for all t.
Initialize $\theta_{1,i} = \exp(-1)$, $\lambda_{1,i} = \exp(-1)$, and $c_i^{\mathsf{free}} = \rho_i T$ for all $i \in [m]$.
for each arrival $\mathbf{A}_t$ ($t \in [T]$) **do**
  if $i_t^\dagger \in [m]$ **then**
    Set $\mathbf{z}_t = \mathbf{e}_{i_t^\dagger}$
  else
    $\mathbf{z}_t = \arg\max_{\mathbf{z}\in\Delta^m} (\mathbf{w}_t - \boldsymbol{\theta}_t - \boldsymbol{\lambda}_t)\cdot \mathbf{z}$ s.t. $c_i^{\mathsf{free}} - z_{t,i} \geq 0 \quad \forall i \in [m]$
    Update remaining capacity for hard constraint: $c_i^{\mathsf{free}} \leftarrow c_i^{\mathsf{free}} - z_{t,i} \quad \forall i \in [m]$
  end if
  Update the backlog: $b_{t,i} = (b_{t-1,i} + z_{t,i} - s_{t,i})_+ \quad \forall i \in [m]$
  Update the dual variables for all $i \in [m]$:

$$\theta_{t+1,i} = \min\{\theta_{t,i}\exp(\eta_t(z_{t,i}-\rho_i)), \alpha\} \tag{75}$$

$$\lambda_{t+1,i} = \min\left\{\lambda_{t,i}\exp(\eta_t(\mathbb{1}[i_t^\dagger = 0]z_{t,i} - \rho_i)), \frac{1+2\alpha}{\underline{\rho}}\right\} \tag{76}$$

end for

D.2. Motivating Numerical Example for `CO-DL` (Algorithm 2)

To further motivate our rationale behind the development of Algorithm 2, we illustrate its potential as a viable complement to Algorithm 1 through a numerical example. Consider a scenario where the agency lacks access to accurate information about the backlogs. However, the agency has imperfect knowledge of service rates. In this situation, the agency can simulate the backlog dynamics for each affiliate using its (imperfect) knowledge of the service rates. In the following example, we demonstrate the limitation of such an approach.

Concretely, consider a numerical example with a single affiliate ($m=1$), $T=1000$ arrivals with no tied cases, and rewards following a uniform distribution on the interval $(0,1)$. The capacity ratio is set at $\rho = 0.5$, and the *actual* service rate is $r = 0.51$ (i.e., $\epsilon = 0.01$). However, assume that the agency lacks knowledge of both the actual backlog process and the precise service rate. Instead, it simulates the backlog process based on a mis-specified service rate of $\hat{r} = 0.6$.

In the table 2, we present the performance of (i) Algorithm 1 under full knowledge of the actual backlog, (ii) Algorithm 1 operated under simulated backlog with the mis-specified service rate $\hat{r} = 0.6$, and (iii) Algorithm 2 (which operates without requiring any knowledge of the backlog or service rate). The objective value is computed based on $\gamma = 10$. We observe that, while Algorithm 1 with the correct backlog outperforms Algorithm 2 in terms of the objective value. However, if Algorithm 1 uses the simulated backlog information with the mis-specified service rates, its performance degrades. Specifically, Algorithm 1 with the mis-specified service rate shows a higher average backlog and a lower reward than Algorithm 2.

D.3. Preliminaries: Properties of the Static Dual Problem

The purpose of this section is to study the static dual problem defined in Section 4.2.1. We prove some properties of the static dual problem which will be useful throughout the proof of Theorem 2 (Appendix Section D.5).

Electronic copy available at: https://ssrn.com/abstract=4748762

Algorithm	Reward	Average backlog	Objective ($\gamma = 10$)
Algorithm 1 with knowledge of backlog	350.42	1.57	334.72
Algorithm 1 with mis-specified service rate	363.62	7.38	289.82
Algorithm 2	365.19	6.58	299.39

Table 2 Numerical illustration: performance of Algorithm 1 (with and without correct knowledge of backlogs) and Algorithm 2 (note: we set $\eta = \zeta = 5/\sqrt{T}$ for Algorithm 1 and $\eta_t = 5/\sqrt{t}$ for Algorithm 2. The result is based on averaging over 1000 sample paths of arrivals and services.)

Recall that we use $\boldsymbol{\nu} = (\boldsymbol{\theta}, \boldsymbol{\lambda}) \in \mathbb{R}_+^{2m}$ to collectively denote the two dual variables. For ease of reference, we recall that the static dual problem (Definition 6 in Section 4.2.1) was defined as

$$D(\boldsymbol{\nu}) := \mathbb{E}\left[\max_{\mathbf{z}(\mathbf{A}) \in \mathcal{Z}(i^\dagger)} (\mathbf{w} - \boldsymbol{\theta} - \mathbb{1}[i^\dagger = 0]\boldsymbol{\lambda}) \cdot \mathbf{z}(\mathbf{A}) + \boldsymbol{\rho} \cdot \boldsymbol{\theta} + \boldsymbol{\rho} \cdot \boldsymbol{\lambda}\right]$$
$$D(\boldsymbol{\nu}^*) := \min_{\boldsymbol{\nu} \in \mathcal{V}} D(\boldsymbol{\nu})$$

where we recall that $\mathcal{V} := [0, \alpha]^m \times [0, \frac{1+2\alpha}{\underline{\rho}}]$. We further define the dual-based primal decision $\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})$ and *matching rate* $\boldsymbol{\mu}(\boldsymbol{\nu})$ as

$$\mathbf{z}(\boldsymbol{\nu}, \mathbf{A}) := \arg\max_{\mathbf{z} \in \mathcal{Z}(i^\dagger)} (\mathbf{w} - \boldsymbol{\theta} - \mathbb{1}[i^\dagger = 0]\boldsymbol{\lambda}) \cdot \mathbf{z}$$
$$\boldsymbol{\mu}(\boldsymbol{\nu}) := \mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})].$$

Note that, under Assumption 1, the dual-based primal decision is unique almost surely for each pair of arrival type and dual variable, and hence the dual-based matching rate is well-defined.

In the following, we show some useful properties of the optimal solution $\boldsymbol{\nu}^*$ for the static dual problem.

Claim 12 **(Properties of Optimal Solution for Static Dual Problem)**. *Under Assumption 1, the following properties hold:*

(a) $\|\boldsymbol{\nu}^*\|_\infty \le 1$.[36]

(b) $\mu_i(\boldsymbol{\nu}^*) \le \rho_i$ *with for all* $i \in [m]$ *with equality if* $\theta_i^* > 0$.

(c) $D(\boldsymbol{\nu}^*) = \mathbb{E}[\mathbf{w} \cdot \mathbf{z}(\boldsymbol{\nu}^*, \mathbf{A})]$

Proof of Claim 12. We first prove part (a). Suppose for a contradiction that $\theta_i^* > 1$ or $\lambda_i^* > 1$ for some $i \in [m]$. We focus on the case of $\theta_i^* > 1$ because the other case follows from the same step. To derive contradiction, we will construct an alternative solution $\tilde{\boldsymbol{\nu}}$ that can strictly decrease the dual objective function. Note that we can write $D(\boldsymbol{\nu})$ as

$$D(\boldsymbol{\nu}) = \sum_{j=1}^m \mathbb{E}[(w_j - \theta_j) \cdot \mathbb{1}[i^\dagger = j]] + \mathbb{E}\left[\max_{j \in [m]} (w_j - \theta_j - \lambda_j)_+ \cdot \mathbb{1}[i^\dagger = 0]\right] + \sum_{j=1}^m \rho_j(\theta_j + \lambda_j).$$

We now define $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta}^* - \delta \mathbf{e}_i$ and $\tilde{\boldsymbol{\lambda}} := \boldsymbol{\lambda}^*$ with infinitesimally small $\delta > 0$. We first observe that

$$\max_{j \in [m]} (w_j - \tilde{\theta}_j - \tilde{\lambda}_j)_+ = \max_{j \in [m]} (w_j - \theta_j^* - \lambda_j^*)_+ \tag{77}$$

for every sample path of arrivals and services. To see this, recall that $w_i \le 1$ (by our assuption in Section 2). Because $\theta_i^* > 1$ and $\lambda_i^* \ge 0$, we have $w_i - \theta_i^* - \lambda_i^* < 0$ and therefore $w_i - \tilde{\theta}_i - \tilde{\lambda}_i < 0$ by choosing sufficiently

[36] Note that $[0, 1]^{2m} \in \mathcal{V}$ due to our assumptions $\alpha \ge 1$ and $\underline{\rho} \le 1$.

Electronic copy available at: https://ssrn.com/abstract=4748762

small $\delta > 0$. Furthermore, $(w_j - \theta_j^* - \lambda_j^*)_+ = (w_j - \tilde{\theta}_j - \tilde{\lambda}_j)_+$ for all $j \neq i$ by definition of $\tilde{\boldsymbol{\theta}}$. Combining the two facts leads to line (77). Consequently, we have:

$$\begin{aligned} D(\tilde{\boldsymbol{\nu}}) - D(\boldsymbol{\nu}^*) &= \delta\mathbb{E}[\mathbb{1}[i^\dagger = i]] + \mathbb{E}\left[\left\{\max_{j\in[m]}(w_j - \tilde{\theta}_j - \tilde{\lambda}_j)_+ - \max_{j\in[m]}(w_j - \theta_j^* - \lambda_j^*)_+\right\}\mathbb{1}[i^\dagger = 0]\right] - \delta\rho_i \\ &= \delta(\mathbb{P}[i^\dagger = i] - \rho_i) \\ &< 0. \end{aligned}$$

The second line follows from (77). The last line is due to our assumption $\mathbb{P}[i^\dagger = i] < \rho_i$ for all $i \in [m]$ (see Section 2). One can follow the same steps to show $\|\boldsymbol{\lambda}^*\|_\infty \leq 1$. Hence, we have proved Claim 12-(a).

We now prove parts (b) and (c). By the Karush-Kuhn-Tucker condition for $\boldsymbol{\nu}^*$, there exists $\underline{u}_i \geq 0$ and $\overline{u}_i \geq 0$ such that

$$\frac{\partial D(\boldsymbol{\nu})}{\partial \theta_i}\bigg|_{\nu=\boldsymbol{\nu}^*} - \underline{u}_i + \overline{u}_i = 0 \tag{78}$$

$$\underline{u}_i \theta_i^* = 0 \tag{79}$$

$$\overline{u}_i(\theta_i^* - \alpha) = 0. \tag{80}$$

Because of our assumption $\alpha > 1$ (Section 2) and $\theta_i^* \leq 1$ by part (a), we must have $\overline{u}_i = 0$. Furthermore, Assumption 1 implies that each $\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})$ is unique almost surely. Therefore, by Theorem 7.44 of (Shapiro et al. 2021), the function $D(\cdot)$ is differentiable and the partial derivative with respect to θ_i is

$$\frac{\partial D(\boldsymbol{\nu})}{\partial \theta_i}\bigg|_{\boldsymbol{\nu}=\boldsymbol{\nu}^*} = \rho_i - \mathbb{E}[z_i(\boldsymbol{\nu}^*, \mathbf{A})]. \tag{81}$$

Plugging this into the preceding Karush-Khun-Tucker conditions (along with $\overline{u}_i = 0$), we have:

$$\rho_i - \mathbb{E}[z_i(\boldsymbol{\nu}^*, \mathbf{A})] \geq 0 \tag{82}$$

$$(\rho_i - \mathbb{E}[z_i(\boldsymbol{\nu}^*, \mathbf{A})])\theta_i^* = 0. \tag{83}$$

Part (b) is a direct consequence of line (82). For part (c), one can follow the similar steps as before to show the complementary slackness $(\rho_i - \mathbb{E}[\mathbb{1}[i^\dagger = 0]z_i(\boldsymbol{\nu}^*, \mathbf{A})])\lambda_i^* = 0$ for all $i \in [m]$. Combining this with line (83), the dual objective at optimal dual $\boldsymbol{\nu}^*$ must be $D(\boldsymbol{\nu}^*) = \mathbb{E}[\mathbf{w} \cdot \mathbf{z}(\boldsymbol{\nu}^*, \mathbf{A})]$. This completes the proof of part (c).

□

D.4. Proof of Proposition 3

General Setup. We begin by describing a general setup of online stochastic mirror descent (OSMD). Let $D : \mathcal{V} \to \mathbb{R}$ be a convex function and (with a slight abuse of notation) $\boldsymbol{\nu}_t$ be a sequence of iterates such that

$$\nabla h(\mathbf{y}_b) = \nabla h(\boldsymbol{\nu}_{b-1}) - \eta_{b-1}\hat{\mathbf{g}}_{b-1}$$

$$\boldsymbol{\nu}_b = \arg\min_{\boldsymbol{\nu}\in\mathcal{V}} V_h(\boldsymbol{\nu}, \mathbf{y}_b)$$

where

1. (Bounded domain): $\mathcal{V}$ is a bounded convex set
2. (Strongly convex mirror map): $h(\cdot) : \mathcal{V} \to \mathbb{R}$ is a σ-strongly convex on domain $\mathcal{V}$ with respect to $\|\cdot\|_1$.
3. (Unbiased and bounded Gradient): $\mathbb{E}[\hat{\mathbf{g}}_t | \mathcal{H}_{t-1}] = \nabla D(\boldsymbol{\nu}_t)$ and $\|\hat{\mathbf{g}}_t\|_\infty \leq G$ almost surely for a constant G.
4. (Bounded Noise): The noise $\hat{\mathbf{u}}_t := \mathbb{E}[\hat{\mathbf{g}}_t | \mathcal{H}_{t-1}] - \hat{\mathbf{g}}_t$ satisfies $\|\hat{\mathbf{u}}\|_1 \leq U$ almost surely for some constant U.

Electronic copy available at: https://ssrn.com/abstract=4748762

5. $\eta_t = \frac{k}{\sqrt{t}}$ for some constant $k > 0$.

The following is the main result we will establish.

THEOREM 4 **(Extending Theorem 3.2 of Harvey et al. (2019) to OSMD).** *Define* $\boldsymbol{\nu}^*$ *such that*

$$\boldsymbol{\nu}^* := \arg\min_{\boldsymbol{\nu} \in \mathcal{V}} D(\boldsymbol{\nu}).$$

Under the conditions (1)-(5) *and for any given* s*, we have*

$$\mathbb{P}\left[D(\boldsymbol{\nu}_s) - D(\boldsymbol{\nu}^*) \leq \mathcal{O}\left(\frac{\log(s)\log(1/\delta)}{\sqrt{s}}\right)\right] \geq 1 - \delta. \tag{84}$$

Connection to Proposition 3. To see the connection of the theorem to Proposition 3, recall that we aimed to study the last-iterate convergence of the dual variables $\{\boldsymbol{\nu}_t\}_{t=1}^{T_A}$ of Algorithm 2. For ease of reference, we recall that the domain of the dual variables was given by $\mathcal{V} := [0, \alpha]^m \times [0, \bar{\lambda}]^m$ where $\bar{\lambda} := \frac{1+2\alpha}{\underline{\rho}}$. The static dual function (Definition 6 in Section 4.2.1) was defined as:

$$\begin{aligned} D(\boldsymbol{\nu}) &:= \mathbb{E}\left[\max_{\mathbf{z} \in Z(i^*)} (\mathbf{w} - \boldsymbol{\theta} - \mathbb{1}[i^* = 0]\boldsymbol{\lambda}) \cdot \mathbf{z} + \boldsymbol{\rho} \cdot \boldsymbol{\theta} + \boldsymbol{\rho} \cdot \boldsymbol{\lambda}\right]. \\ D(\boldsymbol{\nu}^*) &= \min_{\boldsymbol{\nu} \in \mathcal{V}} D(\boldsymbol{\nu}). \end{aligned}$$

The dual variables $\{\boldsymbol{\nu}_t\}_{t=1}^{T}$ is the sequence of OMSD (with the negative entropy mirror map), as we discussed in Section 4. Furthermore, the primitives (U, G, σ) for our variant of OSMD is constant. In particular, it is straightforward to see $G = 1$, $U = 2$, and $\sigma = \frac{\underline{\rho}}{2m(1+2\alpha)}$[37]. Hence, applying Theorem 2, we directly obtain Proposition 3.

D.4.1. Proof of Theorem 4

The proof closely follows the seminal work of Harvey et al. (2019), which establishes the same result for stochastic gradient descent, a special case of stochastic mirror descent with a mirror map denoted as $h(\boldsymbol{\nu}) = \|\boldsymbol{\nu}\|_2^2$. Given our primary use of the multiplicative updating rule, an online mirror descent employing the negative entropy mirror map, we need to extend the result to the general mirror descent setting. Although extending the proof of Harvey et al. (2019) to non-Euclidean geometry is straightforward, the exact result we require is, to the best of our knowledge, not available in the literature. Therefore, for the sake of completeness, we will revisit the main steps of Harvey et al. (2019) and highlight necessary modifications to prove the result for our setting. To avoid unnecessary repetitions, we clearly label any result that can be directly derived from Harvey et al. (2019) and kindly refer readers to the original paper.

With loss of generality, we assume that the primitives (U, G, σ) are given in Section D.4. We begin with the standard result on the average convergence guarantee, which follows from Claim 3.

LEMMA 8 **(Average convergence).** *For any* $\boldsymbol{\nu} \in \mathcal{V}$*,* k*, and* s*, we have*

$$\sum_{t=k}^{s} (D(\boldsymbol{\nu}_t) - D(\boldsymbol{\nu})) \leq \sum_{t=k}^{s} \frac{\eta_t G^2}{2\sigma} + \frac{1}{\eta_k} V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_k) + \sum_{t=k}^{s} \hat{\mathbf{u}}_t \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}) + \sum_{t=k+1}^{s} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) V_h(\boldsymbol{\nu}, \boldsymbol{\nu}_t)$$

The following lemma associates the last-iterate convergence with the average convergence as follows.

[37] The mirror map $h(\cdot)$ is $1/\max(\alpha, \bar{\lambda})$-strongly convex with respect to $\mathcal{L}_2$ norm. Hence, it is $1/(2m\max(\alpha, \bar{\lambda}))$-strongly with respect to $\mathcal{L}_1$ norm because $\|\cdot\|_2 \geq \frac{1}{\sqrt{2m}}\|\cdot\|_1$ in $\mathbb{R}^{2m}$. Finally, $\bar{\lambda} \geq \alpha$ because $\alpha > 1$ and $\underline{\rho} \leq 1$.

Electronic copy available at: https://ssrn.com/abstract=4748762

LEMMA 9 **(Lemma 8.1 of Harvey et al. (2019): From Average to Last-iterate Convergence).** *For any $s \geq 1$,*

$$D(\boldsymbol{\nu}_s) - D(\boldsymbol{\nu}^*) \leq \underbrace{\frac{1}{s/2+1} \sum_{t=s/2}^{s} [D(\boldsymbol{\nu}_t) - D(\boldsymbol{\nu}^*)]}_{\spadesuit} + \mathcal{O}\left(\frac{\log(s)}{\sqrt{s}}\right) + \underbrace{\sum_{t=s/2}^{s} \boldsymbol{\omega}_t \cdot \hat{\mathbf{u}}_t}_{\clubsuit}$$

where

$$\boldsymbol{\omega}_t = \sum_{j=s/2}^{t} \frac{1}{(s-j)(s-j+1)} (\boldsymbol{\nu}_t - \boldsymbol{\nu}_j). \tag{85}$$

The proof mirrors that of Lemma 8.1 in (Harvey et al. 2019) and we omit the proofs for brevity. The result says that the last iterate optimality gap is related to that of the suffix average (♠) and the weighted sum of the noise of the gradients (♣). Hence, it suffices to obtain the high-probability bound for each term. We first begin with the term ♠.

Bounding ♠.

LEMMA 10 **(Lemma 8.2 of Harvey et al. (2019)).** *With probability at least $1-\delta$,*

$$\frac{1}{s/2+1} \sum_{t=s/2}^{s} [D(\boldsymbol{\nu}_t) - D(\boldsymbol{\nu}^*)] \leq \mathcal{O}\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{s}}\right)$$

The proof again mirrors Lemma 8.2 of (Harvey et al. 2019) we kindly refer the readers to the original paper. The idea is to observe that $\hat{\mathbf{u}}_t \cdot (\boldsymbol{\nu}_t - \boldsymbol{\nu}^*)$ is a Martingale-difference sequence with respect to the filtration $\{\mathcal{H}_t\}$ because $\mathbb{E}[\hat{\mathbf{u}}_t | \mathcal{H}_{t-1}] = 0$ and $\boldsymbol{\nu}_t$ is $\mathcal{H}_{t-1}$-measurable. Due to the assumption of boundedness of all the primitives (domain, noise, and the norm of gradients), the magnitude of the difference sequence is bounded by a constant. Hence, a direct application of Azuma's inequality leads to the lemma.

Bounding ♣.

The main technical lemma will be:

LEMMA 11.

$$\sum_{t=s/2}^{s} \boldsymbol{\omega}_t \cdot \hat{\mathbf{u}}_t \leq \mathcal{O}\left(\frac{\log(s)\log(1/\delta)}{\sqrt{s}}\right)$$

with probability at least $1-\delta$.

Combining the previous two lemmas directly implies Theorem 4. The proof of Lemma 11 is fairly intricate. We outline the proof in the following. Similar to the previous lemma, we note that $\boldsymbol{\omega}_t \cdot \hat{\mathbf{u}}_t$ is a Marintagle difference sequence with respect to $\{\mathcal{H}_t\}$. Hence, the standard Azuma-type inequality would require a bound of $\sum_{t=s/2}^{s} \|\boldsymbol{\omega}_t\|^2$ to obtain a high-probability bound of ♣. However, we will see that the bound of $\sum_{t=s/2}^{s} \|\boldsymbol{\omega}_t\|^2$ itself is related to a linear combination of the Martingale difference sequence $\boldsymbol{\omega}_t \cdot \hat{\mathbf{u}}_t$. This non-standard nature requires a more sophisticated concentration inequality than the plain Azuma's inequality. For this purpose, we extract the following concentration result from Harvey et al. (2019).

THEOREM 5 **(Corollary C.5 of Harvey et al. (2019)).** *Let $d_t := \mathbf{a}_t \cdot \mathbf{b}_t$ where $\mathbf{a}_t$ is $\mathcal{H}_t$ measurable and $\mathbf{b}_t$ is $\mathcal{H}_{t-1}$ measurable. Suppose $\mathbb{E}[\mathbf{a}_t | \mathcal{H}_{t-1}] = 0$, $\|\mathbf{a}_t\|_2 \leq 1$ almost surely, and there exists positive values $\{\tilde{\alpha}_t\}_{t=1}^{s}$ and $R > 0$ such that*

Electronic copy available at: https://ssrn.com/abstract=4748762

(a) $\max_{t\in[s]} \tilde{\alpha}_t \le \mathcal{O}(\sqrt{R})$

(b) $\sum_{t=1}^{s} \|\mathbf{b}_t\|_2^2 \le \sum_{t=1}^{s-1} \tilde{\alpha}_t d_t + R\sqrt{\log(1/\delta)}$ *with probability at least* $1-\delta$.

Then $\sum_{t=1}^{T} d_t \le \mathcal{O}(\sqrt{R}\log(1/\delta))$ *with probability at least* $1-\delta$.

To see the connection of Theorem 5 to our task of bounding ♣, let $\mathbf{a}_t = \frac{1}{2}\hat{\mathbf{u}}_t$ and $\mathbf{b}_t = \boldsymbol{\omega}_t$, which are $\mathcal{H}_t$ and $\mathcal{H}_{t-1}$ measurable, respectively. Note that $\mathbb{E}[\hat{\mathbf{u}}_t|\mathcal{H}_{t-1}] = 0$ and $\|\mathbf{a}_t\|_2 \le \|\mathbf{a}_t\|_1 \le 1$ because of $U = 2$ in our variant of OSMD. If we can obtain a bound in the form of Theorem 5-(a) and (b), we are done. Naturally, by the definition of $\boldsymbol{\omega}_t$ in line (85), this requires studying $\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_j\|_2^2$ for arbitrary pair (j,t). The following claim makes this point explicit.

Claim 13 **(Claim E.3 of Harvey et al. (2019))**. *For any* $t \le s$,

$$\|\boldsymbol{\omega}_t\|_2^2 \le \frac{1}{s-t+1}\sum_{j=s/2}^{t-1} \alpha_j \|\boldsymbol{\nu}_t - \boldsymbol{\nu}_j\|_2^2$$

where $\alpha_j := \frac{1}{(s-j)(s-j+1)}$.

The proof is again mainly algebraic and can be found in Claim E.3 of Harvey et al. (2019). In light of Theorem 5 and Claim 13, we obtain the bound on $\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_j\|_2^2$ for arbitrary pair (j,t) in the following.

Lemma 12 **(Extension of Lemma 7.3 of Harvey et al. (2019) to Mirror descent)**. *For any* $a < b$, *we have*

$$\frac{\sigma}{2}\|\boldsymbol{\nu}_a - \boldsymbol{\nu}_b\|_2^2 \le \sum_{i=a}^{b-1}\frac{G^2\eta_i^2}{\sigma} + \sum_{i=a}^{b-1}\eta_i(D(\boldsymbol{\nu}_a) - D(\boldsymbol{\nu}_i)) + \sum_{i=a}^{b-1}\eta_i\hat{\mathbf{u}}_i\cdot(\boldsymbol{\nu}_i - \boldsymbol{\nu}_a)$$

This is the key lemma we extend from Harvey et al. (2019) where the same result was established only for the stochastic gradient descent. The original of (Harvey et al. 2019) is based on a recursive relationship between $\|\boldsymbol{\nu}_a - \boldsymbol{\nu}_b\|_2^2$ and $\|\boldsymbol{\nu}_a - \boldsymbol{\nu}_{b-1}\|_2^2$. The recursion that the authors establish cruicially relies on the fact that the updating rule is performed in the Euclidean space.[38] However, the original the proof cannot be directly applied to a general online mirror descent setting because the dual update is performed through the Bregman distance rather than the Euclidean distance. Fortunately, the Bregman distance shares similar structural properties with the Euclidean distance. Hence, we can establish a recursive relationship between $V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_b)$ and $V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_{b-1})$ and use the strong convexity of $h(\cdot)$ to translate the result to that of the Euclidean distance. We prove the lemma in the following.

Proof of Lemma 12. Throughout the proof, we will use $\|\cdot\|$ to denote $\mathcal{L}_1$-norm as a primal norm. We also use the following standard result.

Fact 1 **(Three point equality: Lemma 5.2 of Bubeck (2011))**. *For any* $\mathbf{a}$, $\mathbf{b}$, *and* $\mathbf{c}$,

$$V_h(\mathbf{a},\mathbf{c}) = V_h(\mathbf{a},\mathbf{b}) + V_h(\mathbf{b},\mathbf{c}) + (\nabla h(\mathbf{b}) - \nabla h(\mathbf{c}))\cdot(\mathbf{a}-\mathbf{b})$$

Fact 2 **(Bregman Projection: Lemma 5.3 of Bubeck (2011))**. *For a convex set* C, *let*

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in C} V_h(\mathbf{x},\mathbf{x}_0).$$

For any $\mathbf{y}\in C$, *we have*

$$V_h(\mathbf{y},\mathbf{x}_0) \ge V_h(\mathbf{y},\mathbf{x}^*) + V_h(\mathbf{x}^*,\mathbf{x}_0)$$

[38] That is, the form of $\boldsymbol{\nu}_{t+1} = \Pi_{\mathcal{V}}(\boldsymbol{\nu} - \eta_t\hat{\mathbf{g}}_t)$ where $\Pi_{\mathcal{V}}(\cdot)$ the Euclidean projection onto set $\mathcal{V}$.

Electronic copy available at: https://ssrn.com/abstract=4748762

FACT 3 **(Lemma 5.1 of Bubeck (2011))**. *Let* $h^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathcal{D}}\{\mathbf{x}\cdot\mathbf{y} - h(\mathbf{x})\}$ *be convex conjugate of a convex function* h. *Then* $\nabla(h^*) = (\nabla(h))^{-1}$. *That is,* $\nabla(h^*)$ *is inverse of* $\nabla(h)$.

FACT 4 **(Theorem 5.26 of Beck (2017))**. *The following is equivalent:*

1. h *is strongly-convex in* $\|\cdot\|$ *with modulus* σ. *That is,* $h(\mathbf{x}) \geq h(\mathbf{y}) + \nabla h(\mathbf{y})\cdot(\mathbf{y}-\mathbf{x}) + \frac{\sigma}{2}\|\mathbf{x}-\mathbf{y}\|^2$
2. h^* *(convex conjugate of* h*) is smooth in* $\|\cdot\|_*$ *with modulus* $1/\sigma$. *That is,* $\|\nabla h^*(\mathbf{x}) - \nabla h^*(\mathbf{y})\| \leq \frac{1}{\sigma}\|\mathbf{x}-\mathbf{y}\|_*$ *where* $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$.

Recall that we can write the updating rule of the dual variable as

$$\nabla h(\mathbf{y}_b) = \nabla h(\boldsymbol{\nu}_{b-1}) - \eta_{b-1}\hat{\mathbf{g}}_{b-1} \tag{86}$$
$$\boldsymbol{\nu}_b = \arg\min_{\boldsymbol{\nu}\in\mathcal{V}} V_h(\boldsymbol{\nu}, \mathbf{y}_b)$$

We first establish the following recursion:

$$\begin{aligned} V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_b) &\leq V_h(\boldsymbol{\nu}_a, \mathbf{y}_b) \\ &\leq V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_{b-1}) + V_h(\boldsymbol{\nu}_{b-1}, \mathbf{y}_b) + (\nabla h(\boldsymbol{\nu}_{b-1}) - \nabla h(\mathbf{y}_b))\cdot(\boldsymbol{\nu}_a - \boldsymbol{\nu}_{b-1}) \qquad (87) \\ &\leq V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_{b-1}) + V_h(\boldsymbol{\nu}_{b-1}, \mathbf{y}_b) + \eta_{b-1}\hat{\mathbf{g}}_{b-1}\cdot(\boldsymbol{\nu}_a - \boldsymbol{\nu}_{b-1}) \end{aligned}$$

The first line follows from Fact 2. In the second line, we used Fact 1. The last line follows from Line (86). We now bound the second term in the following.

$$\begin{aligned} V_h(\boldsymbol{\nu}_{b-1}, \mathbf{y}_b) &\leq V_h(\boldsymbol{\nu}_{b-1}, \mathbf{y}_b) + V_h(\mathbf{y}_b, \boldsymbol{\nu}_{b-1}) \\ &= (\nabla h(\boldsymbol{\nu}_{b-1}) - \nabla h(\mathbf{y}_b))\cdot(\boldsymbol{\nu}_{b-1} - \mathbf{y}_b) \qquad (88) \\ &= \eta_{b-1}\hat{\mathbf{g}}_{b-1}\cdot(\boldsymbol{\nu}_{b-1} - \mathbf{y}_b) \qquad (89) \\ &\leq \eta_{b-1}G\|\boldsymbol{\nu}_{b-1} - \mathbf{y}_b\| \qquad (90) \\ &= \eta_{b-1}G\|\nabla h^*(\nabla h(\boldsymbol{\nu}_{b-1})) - \nabla h^*(\nabla h(\mathbf{y}_b))\| \qquad (91) \\ &\leq \frac{\eta_{b-1}G}{\sigma}\|\nabla h(\boldsymbol{\nu}_{b-1}) - \nabla h(\mathbf{y}_b)\|_* \qquad (92) \\ &\leq \frac{\eta_{b-1}^2 G^2}{\sigma} \qquad (93) \end{aligned}$$

The first line is due to the non-negativity of Bregman distance. Line (88) follows from Fact 1.[39] Line (89) comes from the updating rule (86). In line (90), we used Cauchy Schwartz and the bounded gradient assumption ($\|\hat{\mathbf{g}}_{b-1}\|_* \leq G$). For lines (91) and (92), we applied Facts 3 and 4, respectively. The final line is again due to the updating rule (86) and the bounded gradient assumption. Hence, from the last line of (87), We have

$$V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_b) \leq V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_{b-1}) + \frac{\eta_{b-1}^2 G^2}{\sigma} + \eta_{b-1}\hat{\mathbf{g}}_{b-1}\cdot(\boldsymbol{\nu}_a - \boldsymbol{\nu}_{b-1}) \tag{94}$$

By induction, we obtain

$$\begin{aligned} V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_b) &\leq \sum_{i=a}^{b-1}\frac{\eta_i^2 G^2}{\sigma} + \sum_{i=a}^{b-1}\eta_i\hat{\mathbf{g}}_i\cdot(\boldsymbol{\nu}_a - \boldsymbol{\nu}_i) \\ &= \sum_{i=a}^{b-1}\frac{\eta_i^2 G^2}{\sigma} + \sum_{i=a}^{b-1}\eta_i\mathbf{g}_i\cdot(\boldsymbol{\nu}_a - \boldsymbol{\nu}_i) + \sum_{i=a}^{b-1}\eta_i\hat{\mathbf{u}}_i\cdot(\boldsymbol{\nu}_i - \boldsymbol{\nu}_a) \\ &\leq \sum_{i=a}^{b-1}\frac{\eta_i^2 G^2}{\sigma} + \sum_{i=a}^{b-1}\eta_i(D(\boldsymbol{\nu}_a) - D(\boldsymbol{\nu}_i)) + \sum_{i=a}^{b-1}\eta_i\hat{\mathbf{u}}_i\cdot(\boldsymbol{\nu}_i - \boldsymbol{\nu}_a) \end{aligned}$$

[39] That is, $0 = V_h(\boldsymbol{\nu}_{b-1}, \boldsymbol{\nu}_{b-1}) = V_h(\boldsymbol{\nu}_{b-1}, \mathbf{y}_b) + V_h(\mathbf{y}_b, \boldsymbol{\nu}_{b-1}) + (\nabla h(\mathbf{y}_b) - \nabla h(\boldsymbol{\nu}_{b-1}))\cdot(\boldsymbol{\nu}_{b-1} - \mathbf{y}_b)$

Electronic copy available at: https://ssrn.com/abstract=4748762

The last line is because $\mathbf{g}_i$ is subgradient of D at $\boldsymbol{\nu} = \boldsymbol{\nu}_i$. The proof is complete by noting that (i) $V_h(\boldsymbol{\nu}_a, \boldsymbol{\nu}_b) \geq \frac{\sigma}{2}\|\boldsymbol{\nu}_a - \boldsymbol{\nu}_b\|_1^2$ by strong convexity of h with modulus σ (recall that the primal norm is $\mathcal{L}_1$ norm) with respect to $\|\cdot\|_1$ and (ii) $\|\cdot\|_1^2 \geq \|\cdot\|_2^2$. □

Now that we have successfully extended the above lemma to the general mirror descent setting, we can apply Theorem 5 in the following form.

LEMMA 13 **(Claims E.4-E.6 of Harvey et al. (2019))**. *Let* $d_t = \frac{1}{2}\hat{\mathbf{u}}_t \cdot \boldsymbol{\omega}_t$. *There exists* $R = \mathcal{O}(\log^2(s)/s)$ *and positive values* $\tilde{\alpha}_t > 0$ *such that*

1. $\max_{s/2 \leq t \leq s} \tilde{\alpha}_t \leq \mathcal{O}(\sqrt{R})$
2. $\sum_{t=s/2}^{s} \|\boldsymbol{\omega}_t\|_2^2 \leq \sum_{t=s/2}^{s-1} \tilde{\alpha}_t d_t + R\sqrt{\log(1/\delta)}$ *with probability at least* $1-\delta$.

The proof is mainly algebraic and can be found in Lemmas E.4-E.6 of Harvey et al. (2019). Here we only give a sketch of the proof. From Claim 13 and Lemma 12, we have $\sum_{t=s/2}^{s} \|\omega_t\|_2^2 \leq \Lambda_1 + \Lambda_2 + \Lambda_3$ where

$$\Lambda_1 := \frac{2G^2}{\sigma^2} \sum_{t=s/2}^{s} \frac{1}{s-t+1} \sum_{j=s/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \eta_i^2$$

$$\Lambda_2 := \frac{2}{\sigma} \sum_{t=s/2}^{s} \frac{1}{s-t+1} \sum_{j=s/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \eta_i (D(\boldsymbol{\nu}_j) - D(\boldsymbol{\nu}_i))$$

$$\Lambda_3 := \frac{2}{\sigma} \sum_{t=s/2}^{s} \frac{1}{s-t+1} \sum_{j=s/2}^{t-1} \alpha_j \sum_{i=j}^{t-1} \eta_i \hat{\mathbf{u}}_i (\boldsymbol{\nu}_i - \boldsymbol{\nu}_j).$$

With some delicate algebras, we can show that $\Lambda_1 \leq O\left(\frac{\log^2(s)}{s}\right)$ (Claim E.4 of Harvey et al. (2019)) and $\Lambda_2 \leq O\left(\frac{\log^2(s)\sqrt{\log(1/\delta)}}{s}\right)$ with proabability $1-\delta$ (Claim E.5 of Harvey et al. (2019)). Interestingly, to prove the bound for Λ_2, we use the result in Lemma 10 again. Finally, note that Λ_3 is essentially the scaled version of term ♣ in Lemma 9, which we were aiming to bound. In fact, one can follow the same algebras taken in Claim E.6 of Harvey et al. (2019) to show $\Lambda_3 \leq \frac{2}{\sigma}\sum_{i=s/2}^{s-1} \alpha_i (\hat{\mathbf{u}}_i \cdot \boldsymbol{\omega})$ where $\alpha_i = \eta_i \sum_{j=i+1}^{s} \frac{1}{s-i+1}$ is the extra scaling factor. Note that $\alpha_i \leq \mathcal{O}(\frac{\log(s)}{\sqrt{s}})$ for all $\frac{s}{2} \leq i \leq s$. Comining the bounds for Λ_1, Λ_2, and Λ_3 will give the desired result. The proof is complete is by defining $\tilde{\alpha}_i = \frac{4}{\sigma}\alpha_i$. Finally, the direct application of Theorem 5 and Lemma 13 leads to Lemma 11. This completes the proof.

D.5. Proof of Theorem 2

We first begin by formally defining the endogenous matching rate of `CO-DL` given the dual variable:

DEFINITION 7 **(Dual-based Matching Rate)**. *The dual-based matching rate for a given* $\boldsymbol{\nu} \in \mathcal{V}$ *is defined as*

$$\boldsymbol{\mu}(\boldsymbol{\nu}) := \mathbb{E}_{\mathbf{A}}[\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})].$$

where we recall from Definition 6 that $\mathbf{z}(\boldsymbol{\nu}, \mathbf{A}) := \arg\max_{\mathbf{z} \in \mathcal{Z}(i^\dagger)} \{(\mathbf{w} - \boldsymbol{\theta} - \mathbb{1}[i^\dagger = 0]\boldsymbol{\lambda}) \cdot \mathbf{z}\}$.

Note that, under Assumption 1, the primal-based decision $\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})$ is uniquely determined for each $\boldsymbol{\nu}$ and $\mathbf{A}$ almost surely, and hence the dual-based matching rate is well defined.

Given all the mentioned ingredients, we now outline the main steps leading to Theorem 2. Similar to Section 3.3, we prove Theorem 2 based on the the same regret decomposition as in line (11) (see Section 3.3) and obtain bounds on (A), the loss regarding the expected net matching reward, and (B), the expected average backlog. The following lemma first establishes the bound on (A).

Electronic copy available at: https://ssrn.com/abstract=4748762

LEMMA 14 **(Bounding Loss of Net Matching Reward of CO-DL).** *For any arrival distribution $\mathcal{F}$ and service slack $\epsilon \geq 0$, CO-DL satisfies*

$$\mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t^*\}_{t=1}^T;\alpha)] - \mathbb{E}[\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T;\alpha)] \leq \mathcal{O}(\sqrt{T})$$

where we recall that $\mathsf{NMR}(\cdot;\alpha)$ is the total net matching reward and $\{\mathbf{z}_t^*\}_{t=1}^T$ is the optimal offline solution.

The proof mirrors that of Lemma 1 and we present the full proof in Appendix D.5.1.

In Section 4.2.2, we stated Lemma 6 that bounds (B), the expected average backlog. For ease of reference, we restate the lemma here again.

LEMMA 6 **(Bounding Average Backlog of CO-DL).** *For any arrival distribution $\mathcal{F}$ and $\epsilon > 0$, we have*

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\Big] \leq \mathcal{O}\Big(\frac{1}{\epsilon}\Big)$$

The proof consists of three steps. In the first step, we establish that the good event $\mathcal{B}_T$ (see (23) in Section 4.2.2) occurs with sufficiently high probability using Proposition 3 and Assumption 1. In the second step, we use the definition of the good event to show an $\mathcal{O}(1/\epsilon)$ bound on the average backlog in expectation up to the stopping time. In the last step, we prove that the average build-up after the stopping time is $\mathcal{O}(1)$.

Step 1: Defining a Good Event and Establishing its High-probability Guarantee.

The following lemma shows that the good event occurs with a high probability.[40]

LEMMA 15 **(High-probability Bound of Good Event).** $\Pr[\mathcal{B}_T] \geq 1 - \mathcal{O}(1/T^2)$.

The proof of this lemma is presented in Section D.5.2. We translate the high-probability convergence in terms of the dual objective (established in Proposition 3) to that of the matching rate for each $t \in \{\sqrt{T}, T_A\}$ using Assumption 1. We then take the union bound of the preceding high probability bound across $t \in \{\sqrt{T}, T_A\}$ to prove the desired result.

Step 2: Using the Good Event to Bound the Average backlog up to the Stopping time.

To utilize the good event for our task of bounding the average backlog, we start with the following decomposition of the expected cumulative backlog.

$$\mathbb{E}\Big[\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\Big] = \mathbb{E}\Big[\sum_{t=1}^{\sqrt{T}}\|\mathbf{b}_t\|_1\mathbb{1}[\mathcal{B}_T]\Big] + \underbrace{\mathbb{E}\Big[\sum_{t=\sqrt{T}+1}^{T_A}\|\mathbf{b}_t\|_1\mathbb{1}[\mathcal{B}_T]\Big]}_{\text{(B-1)}} + \underbrace{\mathbb{E}\Big[\sum_{t=T_A+1}^{T}\|\mathbf{b}_t\|_1\mathbb{1}[\mathcal{B}_T]\Big]}_{\text{(B-2)}} + \mathbb{E}\Big[\sum_{t=1}^{T}\|\mathbf{b}_t\|_1\mathbb{1}[\mathcal{B}_T^c]\Big] \tag{Decomposition}$$

The first three terms are the cumulative backlog under the good event. The last term is the one under the complement of the good event. We first note that the first and the last terms can bounded by $\mathcal{O}(T)$ and $\mathcal{O}(1)$, respectively, because (i) $\|\mathbf{b}_t\|_1 \leq \mathcal{O}(t)$ for every sample path and (ii) $\mathbb{P}[\mathcal{B}_T^c] \leq \mathcal{O}(1/T^2)$ from Lemma 15. Hence, the remaining task is to obtain an upper bound of the terms (B-1) and (B-2). In the following, we bound (B-1) by $\mathcal{O}(T/\epsilon)$.

LEMMA 16 **(Bounding (B-1)).** *The term (B-1) in line (Decomposition) is $\mathcal{O}(T/\epsilon)$.*

[40] More precisely, there exists a constant $\underline{K}$, which only depends on (i) $\epsilon > 0$, (ii) $L > 0$ in Assumption 1, and (iii) the input of the algorithm, such that $\mathbb{P}[\mathcal{B}_T] \geq 1 - \frac{1}{T^2}$ for all $T \geq \underline{K}$.

Electronic copy available at: https://ssrn.com/abstract=4748762

The proof of Lemma 16 can be found in Appendix D.5.3. The approach combines our drift lemma (Lemma 3) with the definition of the good event.

Step 3: Bounding the Average backlog after the Stopping Time.

The last step is to establish that the term (B-2) in (Decomposition), the cumulative backlog after the stopping time, is $\mathcal{O}(T)$. We prove this result in the following lemma:

LEMMA 17 **(Bounding (B-2))**. *The term (B-2) in line* (Decomposition) *is* $O(T)$.

We prove lemma 17 through the two auxiliary results. we first observe that $\|\mathbf{b}_{T_A+\tau}\|_1 \le \|\mathbf{b}_{T_A}\|_1 + \tau$ for all $\tau \ge 1$ for every sample pat, because the backlog can only increase by one per period. This leads to the following decomposition:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=T_A+1}^{T} \|\mathbf{b}_t\|_1 \mathbb{1}[\mathcal{B}_T]\right] &\le \mathbb{E}\left[\sum_{t=T_A+1}^{T} \|\mathbf{b}_t\|_1\right] \\ &\le \mathbb{E}\left[(T-T_A)\|\mathbf{b}_{T_A}\|_1 + (T-T_A)^2\right] \\ &\le \sqrt{\mathbb{E}[(T-T_A)^2]}\sqrt{\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2]} + \mathbb{E}[(T-T_A)^2].\end{aligned}$$

where the last line is due to Cauchy-Schwartz inequality. Hence, it suffices to prove to establish $O(T)$ bounds for each of $\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2]$ and $\mathbb{E}[(T-T_A)^2]$. The following claim first bounds $\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2]$. We prove Claim 14 in Appendix D.5.4.

CLAIM 14. *For* `CO-DL` *(Algorithm 2),* $\mathbb{E}[\|\mathbf{b}_{T_A}\|_1^2] \le O(T)$

Finally, the next proposition bounds $\mathbb{E}[(T-T_A)^2]$ by obtaining a $\mathcal{O}(\sqrt{T})$ high-probability bound on $(T-T_A)$.

PROPOSITION 4 **(High Probability Bound of Stopping Time for `CO-DL`)**. *For* `CO-DL` *(Algorithm 2), there exists constant* $a > 0$ *such that, for any given* $\delta \in (0,1)$*, we have*

$$T - T_A \le \frac{a}{\underline{\rho}}\sqrt{T\log(1/\delta)} + \frac{1}{\underline{\rho}}$$

with probability at least $1-\delta$.

Note that Proposition 4 straightforwardly implies our original goal of proving $\mathbb{E}[(T-T_A)^2] \le O(T)$. To be precise, we have

$$\mathbb{E}[(T-T_A)^2] \le \int_0^{T^2} \mathbb{P}[(T-T_A)^2 \le x] \le O(T).$$

The last inequality follows by inversing and integrating the exponential tail bound in Proposition 4. This completes the proof of Lemma 6.

We prove Proposition 4 in Appendix D.5.5. Here, the time-varying nature of the step size η_t for `CO-DL` prevents us from employing the same techniques used to bound the stopping time of `CA-DL` (see Remark 1 in Appendix C.4). Hence, we take a different approach to establish the result to establish a high-probability bound of $T - T_A$, which may be of independent interest. Specifically, we study a modified version of `CO-DL` that takes the same input as the original one but runs without the capacity constraint. The fact that (a) the modified algorithm is unconstrained and (b) the step size η_t does not depend on T allows us to obtain a high-probability upper bound of the cumulative number of cases at each location for *any fixed time* t. We then extend the bound to that of $T - T_A$ by leveraging the fact that the modified algorithm is identical to the original one until the stopping time T_A. See Appendix D.5.5 for the full proof.

Electronic copy available at: https://ssrn.com/abstract=4748762

D.5.1. Proof of Lemma 14

The proof mirrors the same step of Lemma 1. Here, we only give an outline of the steps and highlight a main difference compared to the proof of Lemma 1. Define the pseudo-reward $\{K_t\}$ as:

$$K_t := \mathbf{w}_t \cdot \mathbf{z}_t + \boldsymbol{\theta}_t \cdot (\boldsymbol{\rho} - \mathbf{z}_t) + \boldsymbol{\lambda}_t \cdot (\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t)$$

We also define the stopping time T_A as line in (17) for `CA-DL` (see Section 3.3). The following lemma lower bounds the expected pseudo-rewards up to the stopping time.

Claim 15 **(Lower bound on Pseudo-rewards for `CO-DL`).** *For any feasible matching profile $\{\hat{\mathbf{z}}_t\}_{t=1}^T$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T_A} K_t\right] \geq \mathbb{E}[\mathsf{NMR}(\{\hat{\mathbf{z}}_t\}_{t=1}^T; \alpha)] - (T - T_A).$$

The proof of this claim follows the identical steps taken in the proof of Lemma 4. We omit the proof for brevity.

The next lemma upper bounds the cumulative pseudo-rewards via the total reward net over-allocation of the algorithm.

Claim 16 **(Upper bound on Pseudo-rewards for `CO-DL`).** *For every sample path $\{\mathbf{A}_t, \mathbf{s}_t\}_{t=1}^T$, we have*

$$\sum_{t=1}^{T_A} K_t \leq \mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T; \alpha) - (T - T_A) + \mathcal{O}(\sqrt{T})$$

Proof of Claim 16. The proof follows the similar steps as that of Lemma 5 except for a minor detail arising from the time-varying learning rates. To avoid repetitions, we only highlight the difference. By applying the similar line of algebras in lines (37)-(39) to the new pseudo-reward (D.5.1), we can deduce that

$$\begin{aligned}\sum_{t=1}^{T_A} K_t \leq \underbrace{\sum_{t=1}^{T_A} \mathbf{w}_t \cdot \mathbf{z}_t - \alpha \sum_{i=1}^{m}\left(\sum_{t=1}^{T} z_{t,i} - c_i\right)_+}_{\mathsf{NMR}(\{\mathbf{z}_t\}_{t=1}^T;\alpha)} + \underbrace{\sum_{t=1}^{T} \boldsymbol{\theta}_t \cdot (\boldsymbol{\rho} - \mathbf{z}_t) - \sum_{t=1}^{T} \boldsymbol{\theta}^\star \cdot (\boldsymbol{\rho} - \mathbf{z}_t)}_{R_\theta} + \\ \underbrace{\sum_{t=1}^{T_A} \boldsymbol{\lambda}_t \cdot (\boldsymbol{\rho} - \mathbb{1}[i_t^\dagger = 0]\mathbf{z}_t)}_{R_\lambda} + 2\alpha(T - T_A).\end{aligned} \tag{95}$$

for some $\boldsymbol{\theta}^\star \in [0, \alpha]^m$. Similar to Claim 6, one can invoke the adversarial guarantee of the online mirror descent (Claim 3) to show that

$$R_\theta \leq 2\alpha \sum_{t=1}^{T} \eta_t + \frac{1}{\eta_1} V_h(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_1) + \sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) V_h(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_1)$$

where we recall that (i) $\eta_t = \frac{k}{\sqrt{t}}$ with the input constant $k > 0$ (ii) $h(\cdot)$ is the negative entropy function, and (iii) $V_h(\cdot, \cdot)$ is the Bregman distance with respect to h. Because $\boldsymbol{\theta}_t \in [0, \alpha]^m$ for all $t \in [T]$, we can bound $\max_{t \in [T]} V_h(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_t) \leq \bar{V}$ for some constant $\bar{V}$. Utilizing the non-negativity of the Bregman divergence and the fact that η_t is non-increasing in t, we can further bound the preceding line as

$$R_\theta \leq 2\alpha \sum_{t=1}^{T} \eta_t + \frac{\bar{V}}{\eta_T}.$$

The last hand side is $\mathcal{O}(\sqrt{T})$ because $\sum_{k=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ and $1/\eta_T = \Theta(\sqrt{T})$. Similarly, One can follow the same algebras and the steps taken in Claim 7 to show that $R_{\boldsymbol{\lambda}} \leq O(\sqrt{T}) + (1 + 2\alpha)(T_A - T)$. Plugging these bound on R_θ and R_λ into line (95) completes the proof. □

Finally, we note that Lemma 14 is a direct consequence of the above two claims, by setting $\hat{\mathbf{z}}_t = \mathbf{z}_t^*$ where $\mathbf{z}_t^*$ is the optimal offline solution (Definition 2) for arrival t for each sample path. This completes the proof of Lemma 14.

Electronic copy available at: https://ssrn.com/abstract=4748762

D.5.2. Proof of Lemma 15

Our first step is to establish that the dual-based matching rate is L-Lipschitz continuous using Assmumption 1, which we prove in the following claim.

Claim 17 **(Lipschitz Continuity of Dual-based Matching Rate)**. *Under Assumption 1, we have* $\|\boldsymbol{\mu}(\boldsymbol{\nu}) - \boldsymbol{\mu}(\boldsymbol{\nu}')\|_\infty \le L\|\boldsymbol{\nu} - \boldsymbol{\nu}'\|_1$ *for all* $\boldsymbol{\nu}, \boldsymbol{\nu}' \in \mathcal{V}$.

Proof of Claim 17. Let $\boldsymbol{\xi} := \boldsymbol{\nu}' - \boldsymbol{\nu}$. We will show that, for all $i \in [m]$,

$$|\mu_i(\boldsymbol{\nu}) - \mu_i(\boldsymbol{\nu}')| \le L\|\boldsymbol{\nu} - \boldsymbol{\nu}'\|_1.$$

First, observe that

$$\mu_i(\boldsymbol{\nu}) = \mathbb{P}[i^\dagger = 0] \times \mathbb{P}\Big[(w_i - \nu_i)_+ \ge (w_j - \nu_j)_+,\ \forall j \ne i \Big| i = 0\Big] + \mathbb{P}[i^\dagger = i]. \tag{96}$$

Let $f(\cdot)$ denote the PDF of the reward distribution conditional that $i^\dagger = 0$. we then have:

$$\begin{aligned}
&\mathbb{P}\Big[(w_i - \nu_i)_+ \ge (w_j - \nu_j)_+,\ \forall j \ne i \Big| i = 0\Big] \\
&= \int_{(w_i-\nu_i)_+ \ge (w_j-\nu_j)_+,\ \forall j \ne i} f(\mathbf{w}) d\mathbf{w} \\
&= \int_{(w_i-\nu_i'+\xi_i)_+ \ge (w_j-\nu_j'+\xi_j)_+,\ \forall j \ne i} f(\mathbf{w}) d\mathbf{w} \\
&= \int_{(w_i'-\nu_i')_+ \ge (w_j'-\nu_j')_+,\ \forall j \ne i} f(\mathbf{w}' - \boldsymbol{\xi}) d\mathbf{w}' \\
&\le \int_{(w_i'-\nu_i')_+ \ge (w_j'-\nu_j')_+,\ \forall j \ne i} f(\mathbf{w}') d\mathbf{w}' \quad + \quad L\|\boldsymbol{\xi}\|_1 \int_{(w_i'-\nu_i')_+ \ge (w_j'-\nu_j')_+,\ \forall j \ne i} d\mathbf{w}' \\
&\le \int_{(w_i'-\nu_i')_+ \ge (w_j'-\nu_j')_+,\ \forall j \ne i} f(\mathbf{w}') d\mathbf{w}' \quad + L\|\boldsymbol{\xi}\|_1 \\
&= \mathbb{P}\Big[(w_i - \nu_i')_+ \ge (w_j - \nu_j')_+,\ \forall j \ne i \Big| i = 0\Big] + L\|\boldsymbol{\xi}\|_1
\end{aligned}$$

The fourth line follows from the variable transformation $\mathbf{w}' = \mathbf{w} + \boldsymbol{\xi}$. In the fifth line, we used Assumption 1 that the PDF f is L-Lipschitz with respect to ℓ_1 norm. The sixth line follows because the support of the reward vector is $[0,1]^m$ by our assumption. The rest of the lines are mainly algebraic. Plugging the preceding bound into line (96), we obtain

$$\begin{aligned}
\mu_i(\boldsymbol{\nu}) &\le \mathbb{P}[i^\dagger = 0] \times \Big(\mathbb{P}\Big[(w_i - \nu_i')_+ \ge (w_j - \nu_j')_+,\ \forall j \ne i \Big| i = 0\Big] + L\|\boldsymbol{\xi}\|_1\Big) + \mathbb{P}[i^\dagger = i]. \\
&= \mu_i(\boldsymbol{\nu}') + \mathbb{P}[i^\dagger = 0] L\|\boldsymbol{\xi}\|_1 \\
&\le \mu_i(\boldsymbol{\nu}') + L\|\boldsymbol{\xi}\|_1
\end{aligned}$$

By switching the role of $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$, one can also show that $\mu_i(\boldsymbol{\nu}') - \mu_i(\boldsymbol{\nu}) \le L\|\boldsymbol{\xi}\|_1$. Hence, we have $|\mu_i(\boldsymbol{\nu}) - \mu_i(\boldsymbol{\nu}')| \le L\|\boldsymbol{\nu} - \boldsymbol{\nu}'\|_1$ for all $i \in [m]$. This completes the proof. □

Building on the preceding claim, we now translate the last iterate convergence in Proposition 3 to that of the matching rate. Precisely, we prove the following.

Claim 18. *For any* $\boldsymbol{\nu} \in \mathcal{V}$, *we have*

$$\max_{i \in [m]} \{\mu_i(\boldsymbol{\nu}) - \rho_i\} \le \sqrt{2L(D(\boldsymbol{\nu}) - D(\boldsymbol{\nu}^*))}.$$

Electronic copy available at: https://ssrn.com/abstract=4748762

Proof of Claim 18. We first observe that Claim 17 implies L-smoothness of $D(\cdot)$. That is, for any $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{V}$, we have

$$\|\nabla D(\boldsymbol{\nu}_1) - \nabla D(\boldsymbol{\nu}_2)\|_\infty \le L\|\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2\|_1. \tag{97}$$

To see this, we first recall that Assumption 1 implies that each $\mathbf{z}(\boldsymbol{\nu}, \mathbf{A})$ is unique almost surely. Therefore, by Theorem 7.44 of (Shapiro et al. 2021), the function $D(\cdot)$ is differentiable and the partial derivative of $D(\cdot)$ is

$$\frac{\partial D(\boldsymbol{\nu})}{\partial \theta_i} = \rho_i - \mathbb{E}[z_i(\boldsymbol{\nu}, \mathbf{A})] = \rho_i - \mu_i(\boldsymbol{\nu})$$
$$\frac{\partial D(\boldsymbol{\nu})}{\partial \lambda_i} = \rho_i - \mathbb{E}[\mathbb{1}[i^\dagger = 0] z_i(\boldsymbol{\nu}, \mathbf{A})]$$

Furthermore, note that

$$\frac{\partial D(\boldsymbol{\nu})}{\partial \theta_i} = \rho_i - \mathbb{P}[i^\dagger = i] - \mathbb{E}[\mathbb{1}[i^\dagger = 0] z_i(\boldsymbol{\nu}, \mathbf{A})] = \frac{\partial D(\boldsymbol{\nu})}{\partial \lambda_i} - \mathbb{P}[i^\dagger = i]. \tag{98}$$

where the first inequality is because $\mathbb{E}[z_i(\boldsymbol{\nu}, \mathbf{A})] = \mathbb{E}[\mathbb{1}[i^\dagger = i] z_i(\boldsymbol{\nu}, \mathbf{A})] + \mathbb{E}[\mathbb{1}[i^\dagger = 0] z_i(\boldsymbol{\nu}, \mathbf{A})]$ and $z_i(\mathbf{A}, \boldsymbol{\nu}) = 1$ whenever $i^\dagger = i$. Note that line (98) directly implies $\|\nabla D(\boldsymbol{\nu}_1) - \nabla D(\boldsymbol{\nu}_2)\|_\infty = \|\boldsymbol{\mu}(\boldsymbol{\nu}_1) - \boldsymbol{\mu}(\boldsymbol{\nu}_2)\|_\infty$, leading to the desired result (97).

From Theorem 5.8-(iii) of Beck (2017), the L-smoothness in line (97) is equivalent to

$$D(\boldsymbol{\nu}_1) - D(\boldsymbol{\nu}_2) \ge \nabla D(\boldsymbol{\nu}_2) \cdot (\boldsymbol{\nu}_1 - \boldsymbol{\nu}_2) + \frac{1}{2L}\|\nabla D(\boldsymbol{\nu}_1) - \nabla D(\boldsymbol{\nu}_2)\|_\infty^2$$

for all $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \in \mathcal{V}$. In particular, letting $\boldsymbol{\nu}^* = \arg\min_{\boldsymbol{\nu} \in \mathcal{V}} D(\boldsymbol{\nu})$,

$$D(\boldsymbol{\nu}) - D(\boldsymbol{\nu}^*) \ge \nabla D(\boldsymbol{\nu}^*) \cdot (\boldsymbol{\nu} - \boldsymbol{\nu}^*) + \frac{1}{2L}\|\nabla D(\boldsymbol{\nu}) - \nabla D(\boldsymbol{\nu}^*)\|_\infty^2 \ge \frac{1}{2L}\|\nabla D(\boldsymbol{\nu}) - \nabla D(\boldsymbol{\nu}^*)\|_\infty^2$$

for all $\boldsymbol{\nu} \in \mathcal{V}$. Here the last line follows from the first-order condition of $\boldsymbol{\nu}^*$ that minimizes $D(\cdot)$ over domain $\mathcal{V}$ (see, for example, Lemma 5.4 of Bubeck (2011)). It follows that, for all $i \in [m]$,

$$\mu_i(\boldsymbol{\nu}) \le \mu_i(\boldsymbol{\nu}^*) + \sqrt{2L(D(\boldsymbol{\nu}) - D(\boldsymbol{\nu}^*))}.$$

The proof is complete by noting that $\mu_i(\boldsymbol{\nu}^*) \le \rho_i$ for all $i \in [m]$ from Claim 12-(b). □

We now use the above claim to obtain the high-probability bound of good event $\mathcal{B}_T$. For ease of reference, we recall the (equivalent) definition of good event $\mathcal{B}_T$ from line (23):

$$\mathcal{B}_T := \left\{ \max_{i \in [m]} \{\mu_i(\boldsymbol{\nu}_t) - \rho_i\} \le \frac{\epsilon}{2} \text{ for all } \sqrt{T} \le t \le T_A \right\}.$$

Claim 19. *There exists a constant $\underline{K}$, which only depends on (i) the input of the algorithm, (ii) $L > 0$ in Assumption 1, and (iii) $\epsilon > 0$, such that for all $T \ge \underline{K}$, we have $\mathbb{P}[\mathcal{B}_T] \ge 1 - \frac{1}{T^2}$.*

Proof of Claim 19. By Proposition 3, there exists constant $\kappa > 0$ such that, for any given $t \le T_A$ and $\delta_t \in (0, 1)$,

$$\mathbb{P}\left[D(\boldsymbol{\nu}_t) - D(\boldsymbol{\nu}^*) \le \kappa \frac{\log(t)\log(1/\delta_t)}{\sqrt{t}} \right] \ge 1 - \delta_t.$$

where the constant κ only depends on the input of Algorithm 2. We now set $\delta_t = \frac{1}{T^3}$ for all $t \ge \sqrt{T}$ and use Claim 18 to obtain

$$\mathbb{P}\left[\max_{i \in [m]} \{\mu_i(\boldsymbol{\nu}_t) - \rho_i\} \le \frac{\sqrt{6L\kappa \log(t)\log(T)}}{t^{1/4}} \right] \ge 1 - \frac{1}{T^3}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

for all $t \geq \sqrt{T}$. Note that $f(t) = \frac{\log(t)}{t^{1/2}}$ is decreasing in $t \geq e^2$. Hence, for any $t \geq \sqrt{T} \geq e$, we have

$$\frac{\sqrt{6L\kappa \log(t)\log(T)}}{t^{1/4}} \leq \frac{\sqrt{3L\kappa}\log(T)}{T^{1/8}}$$

The right-hand side converges to zero as $t \to \infty$, meaning that there exists $c(\kappa, L, \epsilon)$ for which the right-hand side is at most $\epsilon/2$ for all $T \geq f(\kappa, L, \epsilon)$ where $f(\kappa, L, \epsilon)$ is a constant that only depends on κ, L, and $\epsilon > 0$. Hence, defining $\underline{K} := \max\{e^4, f(\kappa, L, \epsilon)\}$, we conclude that for all $T \geq \underline{K}$ and $t \geq \sqrt{T}$,

$$\mathbb{P}\left[\max_{i\in[m]}\{\mu_i(\boldsymbol{\nu}_t) - \rho_i\} \leq \frac{\epsilon}{2}\right] \geq 1 - \frac{1}{T^3}.$$

Finally, we use the above bound to lower-bound the good event $\mathcal{B}_T$. For any given T_A, we have:

$$\begin{aligned}
\mathbb{P}[\mathcal{B}_T] &= 1 - \mathbb{P}\left[\max_{i\in[m]}\{\mu_i(\boldsymbol{\nu}_t) - \rho_i\} > \frac{\epsilon}{2} \quad \text{for some } t \in [\sqrt{T}, T_A]\right] \\
&\geq 1 - \sum_{t=\sqrt{T}}^{T_A} \mathbb{P}\left[\max_{i\in[m]}\{\mu_i(\boldsymbol{\nu}_t) - \rho_i\} > \frac{\epsilon}{2}\right] \\
&\geq 1 - \frac{1}{T^2}.
\end{aligned}$$

In the second line, we used the union bound. The third line follows from line (D.5.2) and the fact that $T_A \leq T$ for every sample path.[41] This completes the proof.

□

D.5.3. Proof of Lemma 16

Recall from Lemma 3 that $\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1}) \leq \mathbf{b}_{t-1} \cdot (\mathbf{z}_t - \mathbf{s}_t) + \mathcal{O}(1)$. Hence, for any history $\mathcal{H}_{t-1}$,

$$\mathbb{E}[\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})|\mathcal{H}_{t-1}] \leq \mathbf{b}_{t-1} \cdot (\mathbb{E}[\mathbf{z}_t|\mathcal{H}_{t-1}] - \boldsymbol{\rho} - \boldsymbol{\epsilon}) + \mathcal{O}(1) \tag{99}$$

where we used the fact that (i) $\mathbf{b}_{t-1}$ is $\mathcal{H}_{t-1}$-measurable and (ii) the $\mathbf{s}_t$ is i.i.d random variables with mean $\boldsymbol{\rho} - \boldsymbol{\epsilon}$. Recall that, up to $t \leq T_A + 1$,[42] the primal decision is $\mathbf{z}_t = \mathbf{z}(\mathbf{A}_t, \boldsymbol{\nu}_t)$. Because the dual variables $\boldsymbol{\nu}_t$ is $\mathcal{H}_{t-1}$-measurable, we further have

$$\mathbb{E}[\mathbf{z}_t|\mathcal{H}_{t-1}] = \boldsymbol{\mu}(\boldsymbol{\nu}_t).$$

Hence, one can re-write line (99) as

$$\mathbb{E}[\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})|\mathcal{H}_{t-1}] \leq \mathbf{b}_{t-1} \cdot (\boldsymbol{\mu}(\boldsymbol{\nu}_t) - \boldsymbol{\rho} - \boldsymbol{\epsilon}) + \mathcal{O}(1)$$

for any history $\mathcal{H}_{t-1}$. We now consider two cases. Under the good event $\mathcal{B}_T$, we have

$$\mathbf{b}_{t-1} \cdot (\boldsymbol{\mu}(\boldsymbol{\nu}_t) - \boldsymbol{\rho} - \boldsymbol{\epsilon}) \leq -\frac{\epsilon}{2}\|\mathbf{b}_{t-1}\|_1$$

by the non-negativity of the backlog and by definition of the good event $\mathcal{B}_T$. Otherwise, we have $\mathbf{b}_{t-1} \cdot (\boldsymbol{\mu}(\boldsymbol{\nu}_t) - \boldsymbol{\rho} - \boldsymbol{\epsilon}) \leq \|\mathbf{b}_{t-1}\|_1$ by Cauchy-Schwartz inequality.[43] Combining these two cases, we have

$$\mathbb{E}[\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})|\mathcal{H}_{t-1}] \leq \mathbb{1}[\mathcal{B}_T]\left(-\frac{\epsilon}{2}\|\mathbf{b}_{t-1}\|_1\right) + \mathbb{1}[\mathcal{B}_T^c]\|\mathbf{b}_{t-1}\|_1 + \mathcal{O}(1).$$

[41] Note that our high-probability results hold for any value of T_A.

[42] The definition of the stopping time (see line (17)) allows for at least one more period until the capacity for the hard-constraint is completely depleted.

[43] The service rate $r = \rho_i + \epsilon$ is at most one.

Electronic copy available at: https://ssrn.com/abstract=4748762

Summing the preceding inequality over $\sqrt{T} \le t \le T_A + 1$ and taking the outer expectation, we obtain

$$\mathbb{E}\left[\sum_{t=\sqrt{T}}^{T_A+1} \mathbb{E}[\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})|\mathcal{H}_{t-1}]\right] \le -\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T]\right] + \mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T^c]\right] + \mathcal{O}(\mathbb{E}[T_A - \sqrt{T}]). \tag{100}$$

Now define $X_t := \psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})$ and $Y_t = X_t - \mathbb{E}[X_t|\mathcal{H}_{t-1}]$ with $Y_0 = 0$. It is straightforward to see that Y_t is a martingale difference sequence with respect to $\{\mathcal{H}_t\}$. Since T_A is a bounded stopping time with respect to $\{\mathcal{H}_t\}$, the optional stopping theorem implies

$$\mathbb{E}\left[\sum_{t=\sqrt{T}}^{T_A} \mathbb{E}[\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})|\mathcal{H}_{t-1}]\right] = \mathbb{E}\left[\sum_{t=\sqrt{T}}^{T_A} [\psi(\mathbf{b}_t) - \psi(\mathbf{b}_{t-1})]\right] = \mathbb{E}[\psi(\mathbf{b}_{T_A}) - \psi(\mathbf{b}_{\sqrt{T}-1})].$$

Combining this with (100), we have

$$\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A-1} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T]\right] \le \mathbb{E}[\psi(\mathbf{b}_{\sqrt{T}-1}) - \psi(\mathbf{b}_{T_A})] + \mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A-1} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T^c]\right] + \mathcal{O}(\mathbb{E}[T_A - \sqrt{T}]). \tag{101}$$

Each term on the right-hand side is $\mathcal{O}(T)$. To see this, by taking the worst case growth of the backlog, we have $\mathbf{b}_t \le \mathcal{O}(t)$ and therefore $\psi(\mathbf{b}_{\sqrt{T}-1}) \le \mathcal{O}(T)$. The second term is again $\mathcal{O}(T)$ because $\Pr[\mathcal{B}_T^c] \le \mathcal{O}(1/T)$ from Lemma 15 (along with the worst case growth of the build-up). Finally, the last term in line (101) is $\mathcal{O}(T)$ since $T_A \le T$ for every sample path. The proof is complete by diving both sides of line (101) by $(\epsilon/2)$.

D.5.4. Proof of Claim 14

We will use the inequality (101) in Appendix D.5.3. For ease of reference, we restate it in the following:

$$\frac{\epsilon}{2}\mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A-1} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T]\right] \le \mathbb{E}[\psi(\mathbf{b}_{\sqrt{T}-1}) - \psi(\mathbf{b}_{T_A})] + \mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A-1} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T^c]\right] + \mathcal{O}(\mathbb{E}[T_A - \sqrt{T}]).$$

Note that the left-hand side is non-negative. Hence, moving $\mathbb{E}[\psi(\mathbf{b}_{T_A})]$ to the left-hand side, we have

$$\mathbb{E}[\psi(\mathbf{b}_{T_A})] \le \mathbb{E}[\psi(\mathbf{b}_{\sqrt{T}-1})] + \mathbb{E}\left[\sum_{t=\sqrt{T}-1}^{T_A-1} \|\mathbf{b}_t\|_1 \cdot \mathbb{1}[\mathcal{B}_T^c]\right] + \mathcal{O}(\mathbb{E}[T_A - \sqrt{T}]).$$

In the proof of Lemma 16, we have already shown that each term in the right-hand side is $\mathcal{O}(T)$ (see the paragraph related to line (101)). The proof is complete by recalling the definition of $\psi(\mathbf{b}_{T_A}) = \frac{1}{2}\|\mathbf{b}_{T_A}\|_2^2$ and $\|\mathbf{b}_{T_A}\|_1^2 \le m\|\mathbf{b}_{T_A}\|_2^2$ by Cauchy-Schwartz inequality.

D.5.5. Proof of Proposition 4

Let us first recall the definition of T_A.

$$T_A := \min\left\{t \le T : \sum_{\tau=1}^{t} \mathbb{1}[i_\tau^\dagger = 0] z_{\tau,i} + 1 \ge c_i \text{ for some } i \in [m]\right\}$$

Applying the proof of Claim 9 for `CA-DL` fails here because of the time-varying step size of `CO-DL` (see Remark 1 in Appendix Section C.4). Hence, we take a different proving the result. In fact, we will prove a stronger result. Define a modified stopping time S_A as

$$S_A := \min\left\{t \le T : \sum_{\tau=1}^{t} z_{\tau,i} + 1 \ge c_i \text{ for some } i \in [m]\right\}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

The difference between T_A and S_A is that S_A accounts for the number of both free and tied cases while T_A only counts the number of free cases. It is straightforward to see that $S_A \le T_A$ for every sample path.[44] Hence, it suffices to obtain the same high probability bound on $T - S_A$.

Toward this goal, let us first define a vector-valued stochastic process $\mathbf{Z}_t$ as

$$\mathbf{Z}_t = \sum_{\tau=1}^{t} \mathbf{z}_\tau$$

This stochastic process tracks the cumulative number of cases assigned to each location. To prove Proposition 4, we would have liked to study the tail probability of the $\mathbf{Z}_t$. Instead, we will study another closely related stochastic process that can be analyzed more simply. Consider a 'modified' version of Algorithm 2, which takes the same input (i.e., initial duals, $\boldsymbol{\rho}$, and $\{\eta_t\}$) but runs *without the hard capacity constraint*. Denoting the decision for arrival t of the modified algorithm by $\mathbf{z}_t^{(\mathsf{M})}$, we define a stochastic process

$$\mathbf{Z}_t^{(\mathsf{M})} = \sum_{\tau=1}^{t} \mathbf{z}_\tau^{(\mathsf{M})}$$

Note that the two algorithms make an identical decision (at least) until S_A. This leads to the crucial observation:

$$\text{If } \mathbf{Z}_t^{(\mathsf{M})} + \mathbf{1} \le \mathbf{c} \text{ for some } t, \text{ then } S_A \ge t \tag{102}$$

where $\mathbf{1}$ is the vector of ones.

The fact that the modified algorithm is unconstrained allows us to establish the concentration bound on $\mathbf{Z}_t^{(\mathsf{M})}$. Formally, the following is the main result we will prove:

LEMMA 18. *there exists a constant $a > 0$ for which, for any fixed $t \in [T]$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left[\mathbf{Z}_t^{(M)} \le t\boldsymbol{\rho} + a\sqrt{t\log(1/\delta)}\mathbf{1}\right] \ge 1 - \delta.$$

where $\mathbf{1}$ is the vector of ones..

The proof of the lemma is fairly intricate and we defer it to the end of this section. But first, we show how we use Lemma 18 to prove Proposition 4. For any given $\delta \in (0,1)$, define

$$T(\delta) = T - \frac{1}{\underline{\rho}}\left(a\sqrt{T\log(1/\delta)} + 1\right) \tag{103}$$

where a is constant specified in (18) and $\underline{\rho} = \min_{i\in[m]} \rho_i$. Then, with probability at least $1-\delta$,

$$\begin{aligned}
\mathbf{Z}_{T(\delta)}^{(\mathsf{M})} + \mathbf{1} &\le T(\delta)\boldsymbol{\rho} + a\sqrt{T(\delta)\log(1/\delta)}\mathbf{1} + \mathbf{1} \\
&= T\boldsymbol{\rho} - \frac{1}{\underline{\rho}}\left(a\sqrt{T\log(1/\delta)} + 1\right)\boldsymbol{\rho} + a\sqrt{T\log(1/\delta)}\mathbf{1} + \mathbf{1} \\
&\le T\boldsymbol{\rho}.
\end{aligned} \tag{104}$$

The first line is from Lemma 18. The second line is due to the definition of $T(\delta)$ in line (103). The last line is because of the definition of $\underline{\rho} = \min_{i\in[m]} \rho_i$. Hence, for any given $\delta \in (0,1)$, we conclude that

$$\mathbb{P}\left[T - S_A \le \frac{a}{\underline{\rho}}\sqrt{T\log(1/\delta)} + \frac{1}{\underline{\rho}}\right] = \Pr\left[S_A \ge T(\delta)\right] \ge \mathbb{P}\left[\mathbf{Z}_{T(\delta)}^{(\mathsf{M})} + 1 \le \mathbf{c}\right] \ge 1 - \delta. \tag{105}$$

[44] By definition of S_A, there exists $i \in [m]$ for which $\sum_\tau^{S_A} z_{\tau,i} + 1 = c_i$ and $\sum_\tau^{S_A} z_{\tau,j} + 1 \le c_j$ for all $j \ne i$. The equality is because the algorithm always makes a binary matching decision. This implies that $\sum_{\tau=1}^{S_A} \mathbb{1}[i_\tau^\dagger = 0] z_{\tau,j} + 1 \le c_j$ for all $j \in [m]$ and hence $T_A \ge S_A$.

Electronic copy available at: https://ssrn.com/abstract=4748762

The first inequality holds from line (102). The next inequality is due to line (104). Hence, we have shown that Lemma 18 implies the statement of Proposition 4.

Preliminaries. We now prove Lemma 18 in the following. First, we provide some preliminaries that will be useful throughout the proof. With a slight abuse of notations, we define the following static problem.

$$D(\boldsymbol{\nu}) := \mathbb{E}\left[\max_{\mathbf{z}(\mathbf{A})\in\mathcal{Z}(i^\dagger)} (\mathbf{w}-\boldsymbol{\nu})\cdot\mathbf{z}(\mathbf{A}) + \boldsymbol{\rho}\cdot\boldsymbol{\nu}\right]$$
$$D(\boldsymbol{\nu}^*) = \min_{\boldsymbol{\nu}\in\mathbb{R}^m_+} D(\boldsymbol{\nu}). \tag{106}$$

Given $\boldsymbol{\nu}\in\mathbb{R}^m_+$, we similarly define the dual-based primal decision with respect to arrival type $\mathbf{A}$ as follows.

$$\mathbf{z}(\boldsymbol{\nu},\mathbf{A}) := \arg\max_{\mathbf{z}\in\mathcal{Z}(i^\dagger)} (\mathbf{w}-\boldsymbol{\nu})\cdot\mathbf{z}$$
$$\boldsymbol{\mu}(\boldsymbol{\nu}) := \mathbb{E}[\mathbf{z}(\boldsymbol{\nu},\mathbf{A})].$$

By following a similar step to the proof of claim 12 (Appendix D.3), we can show the following properties of the above static problem. The proof mirrors that of Claim 12 and we omit it for brevity.

CLAIM 20. *For $D(\cdot)$ and $\boldsymbol{\nu}^*$ defined in line (106), the following property holds:*

(a) $\|\boldsymbol{\nu}^*\|_\infty \le 1$.

(b) $\mu_i(\boldsymbol{\nu}^*) \le \rho_i$ *for all* $i\in[m]$ *with equality if* $\nu^*_i > 0$.

(c) $D(\boldsymbol{\nu}^*) = \mathbb{E}[\mathbf{w}\cdot\mathbf{z}(\boldsymbol{\nu}^*,\mathbf{A})]$

We will also use the following result on the sample average approximation of the stochastic optimization problem.

LEMMA 19 **(Proposition 1 of Guigues et al. (2017))**. *Let $\mathcal{X}$ be a nonempty bounded convex set. Consider the following stochastic optimization problem:*

$$f^* = \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{A}\sim\mathcal{F}}[f(\mathbf{x},\mathbf{A})]$$

where $\mathbf{A}$ is a random vector from probability distribution $\mathcal{F}$ with support $\mathcal{A}$. Given i.i.d samples $\mathbf{A}_1,...,\mathbf{A}_t\sim\mathcal{F}$, consider the following sample-average approximation of f^*:

$$\hat{f}_t := \min_{\mathbf{x}\in\mathcal{X}} \frac{1}{t}\sum_{\tau=1}^{t} f(\mathbf{x},\mathbf{A}_\tau).$$

Assume that there exists positive constant M_1 such that $|f(\mathbf{x},\mathbf{A}) - \mathbb{E}_{A\sim\mathcal{F}}[f(\mathbf{x},\mathbf{A})]| \le M_1$ for all $\mathbf{x}\in\mathcal{X}$ and $\mathbf{A}\in\mathcal{A}$ almost surely. Then, for any given $\delta\in(0,1)$,

$$\mathbb{P}\left[\hat{f}_t \le f^* + 2M_1\sqrt{\frac{\log(1/\delta)}{t}}\right] \ge 1-\delta.$$

Main Proof. Fix an arbitrary time t and a sample path of arrivals $(\mathbf{A}_1,...,\mathbf{A}_t)$.[45] Recall that we use $\mathbf{z}^{(\mathsf{M})}_\tau$ to denote the arrival decision for $\mathbf{A}_\tau$ for $\tau\in[T]$. Because the modified algorithm is unconstrained, for any arbitrary time $\tau\le t$, we can write it as:

$$\mathbf{z}^{(\mathsf{M})}_\tau = \arg\max_{\mathbf{z}\in\mathcal{Z}(i^\dagger_\tau)} (\mathbf{w}_\tau - \boldsymbol{\theta}_\tau - \mathbb{1}[i^\dagger_\tau = 0]\boldsymbol{\lambda}_\tau)\cdot\mathbf{z}. \tag{107}$$

[45] Note that the service sequence never affects the decision of Algorithm 2.

Electronic copy available at: https://ssrn.com/abstract=4748762

We will compare $\mathbf{z}_\tau^{(\mathsf{M})}$ against an alternative decision vector $\mathbf{z}_\tau^*$ defined as follows:

$$\mathbf{z}_\tau^* := \arg\max_{\mathbf{z}\in\mathcal{Z}(i_\tau^\dagger)} (\mathbf{w}_\tau - \boldsymbol{\nu}^*)\cdot\mathbf{z} = \mathbf{z}(\boldsymbol{\nu}^*, \mathbf{A}_\tau) \tag{108}$$

where $\boldsymbol{\nu}^*$ is the minimizer of the static problem (106). Note that $\mathbf{z}_\tau^*$ is i.i.d. random variables because $\boldsymbol{\nu}^*$ only depends on the arrival distribution and the current arrival type. By the optimality criterion of the modified algorithm (107), we have

$$\sum_{\tau=1}^{t}(\mathbf{w}_\tau - \boldsymbol{\theta}_\tau - \mathbb{1}[i_\tau^\dagger = 0]\boldsymbol{\lambda}_\tau)\cdot\mathbf{z}_\tau^{(\mathsf{M})} + \boldsymbol{\rho}\cdot(\boldsymbol{\theta}_\tau + \boldsymbol{\lambda}_\tau) \geq \sum_{\tau=1}^{t}(\mathbf{w}_\tau - \boldsymbol{\theta}_\tau - \mathbb{1}[i_\tau^\dagger = 0]\boldsymbol{\lambda}_\tau)\cdot\mathbf{z}_\tau^* + \boldsymbol{\rho}\cdot(\boldsymbol{\theta}_\tau + \boldsymbol{\lambda}_\tau).$$

From Claim 4 (Appendix C), there exists $\boldsymbol{\theta}^\star \in [0,\alpha]^m$ for which

$$\sum_{\tau=1}^{t}\boldsymbol{\theta}^\star\cdot(\mathbf{z}_\tau^{(\mathsf{M})} - \boldsymbol{\rho}) = \alpha\sum_{i=1}^{m}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)_+.$$

Hence, a straightforward manipulation of line (D.5.5) leads to

$$\begin{aligned}\alpha\sum_{i=1}^{m}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)_+ \leq{}& \underbrace{\sum_{\tau=1}^{t}\mathbf{w}_\tau\cdot\mathbf{z}_\tau^{(\mathsf{M})} - \sum_{\tau=1}^{t}\mathbf{w}_\tau\cdot\mathbf{z}_\tau^*}_{(\mathsf{A})} + \underbrace{\sum_{\tau=1}^{t}\boldsymbol{\theta}_\tau\cdot(\mathbf{z}_\tau^* - \boldsymbol{\rho}) + \sum_{\tau=1}^{t}\boldsymbol{\lambda}_\tau\cdot(\mathbb{1}[i_\tau^\dagger = 0]\mathbf{z}_\tau^* - \boldsymbol{\rho})}_{(\mathsf{B})} + \\ &\underbrace{\sum_{\tau=1}^{t}\boldsymbol{\theta}_\tau\cdot(\boldsymbol{\rho} - \mathbf{z}_\tau^{(\mathsf{M})}) - \sum_{\tau=1}^{t}\boldsymbol{\theta}^\star\cdot(\boldsymbol{\rho} - \mathbf{z}_\tau^{(\mathsf{M})}) + \sum_{\tau=1}^{t}\boldsymbol{\lambda}_\tau\cdot(\boldsymbol{\rho} - \mathbb{1}[i_\tau^\dagger = 0]\mathbf{z}_\tau^{(\mathsf{M})})}_{(\mathsf{C})}\end{aligned} \tag{109}$$

We prove the main result through four claims. In the first three claims (Claim 21-23), we obtain the high-probability bound of each term $(\mathsf{A})-(\mathsf{C})$, respectively. In the final claim (Claim-24), we combine the preceding three claims to prove the main result.

Claim 21 **(Bound (A))**. *There exists constant M_A such that, for any given $\delta\in(0,1)$,*

$$(A) \leq \sum_{i=1}^{m}\nu_{t,i}^*\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(M)} - t\rho_i\right) + M_A\sqrt{t\log(1/\delta)}$$

with probability at least $1-\delta$ and for some $0\leq\nu_{t,i}^*\leq 1$.

Proof of Claim 21. With a slight abuse of notation, define $\overline{\mathsf{OPT}}(t\boldsymbol{\rho})$ defined on the same sequence of arrivals $(\mathbf{A}_1,...,\mathbf{A}_t)$:

$$\overline{\mathsf{OPT}}_t(t\boldsymbol{\rho}) := \max_{\mathbf{z}_\tau\in\mathcal{Z}(i_\tau^\dagger)}\sum_{\tau=1}^{t}\mathbf{w}_\tau\cdot\mathbf{z}_\tau \quad \text{s.t.} \quad \sum_{\tau=1}^{t}\mathbf{z}_\tau \leq t\boldsymbol{\rho} \qquad [\boldsymbol{\nu}_t^*] \tag{110}$$

We the use subscript t in the optimal objective value to emphasize that the optimization problem is defined on the arrival up to time t. In the similar spirit, the optimal value of the above offline depends on the right-hand side of the capacity constraint, which we also emphasized as the argument of $\overline{\mathsf{OPT}}_t(\cdot)$. The dual variable corresponding to the constraint (110) (at the right hand side value $t\boldsymbol{\rho}$) is denoted by $\boldsymbol{\nu}_t^*\geq 0$. The above program has a feasible solution with high probability because the total number of tied cases up to time t for each location i is strictly less than $t\rho_i$ with high-probability, which stems from our assumption (see section 2) that $\mathbb{P}[i_t^\dagger = i] < \rho_i$ for all i. For the sake of simplicity, we operate under this high-probability event. We describe how to make the argument precise in Appendix D.5.6.

Electronic copy available at: https://ssrn.com/abstract=4748762

Note that, by definition, $\{\mathbf{z}_\tau^{(\mathsf{M})}\}_{\tau=1}^t$ is a feasible solution for $\overline{\mathsf{OPT}}\left(\sum_{\tau=1}^t \mathbf{z}_\tau^{(\mathsf{M})}\right)$. Hence, we have

$$\begin{aligned}\sum_{\tau=1}^t \mathbf{w}_\tau \cdot \mathbf{z}_\tau^{(\mathsf{M})} &\le \overline{\mathsf{OPT}}_t\left(\sum_{\tau=1}^t \mathbf{z}_\tau^{(\mathsf{M})}\right)\\ &= \overline{\mathsf{OPT}}_t\left(t\boldsymbol{\rho} + \sum_{\tau=1}^t \mathbf{z}_\tau^{(\mathsf{M})} - t\boldsymbol{\rho}\right)\\ &\le \overline{\mathsf{OPT}}_t(t\boldsymbol{\rho}) + \sum_{i=1}^m \nu_{t,i}^*\left(\sum_{\tau=1}^t z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)\end{aligned}$$

where in the last line we used the fact that (i) $\overline{\mathsf{OPT}}_t(\cdot)$ is concave and increasing in each coordinate of its argument (see, for example, Section 5.6.2 of Boyd et al. (2004)) and (ii) $\boldsymbol{\nu}_t^*$ is a gradient of $\overline{\mathsf{OPT}}_t(\cdot)$ evaluated at the right-hand side value of $t\boldsymbol{\rho}$ in line (110).

We now turn our attention to lower-bounding $\sum_{\tau=1}^t \mathbf{w}_\tau \cdot \mathbf{z}_\tau^*$. Here, we will crucially rely on the properties of the static problem in Claim 20. By definition of $\mathbf{z}_\tau^*$ and i.i.d nature of the arrival sequence, each $\mathbf{w}_\tau \cdot \mathbf{z}_\tau^*$ is i.i.d random variables with $|\mathbf{w}_\tau \cdot \mathbf{z}_\tau^*| \le 1$. Furthermore, from Claim 20-(c), the identical mean is $\mathbb{E}[\mathbf{w}_\tau \cdot \mathbf{z}_\tau^*] = D(\boldsymbol{\nu}^*)$ for all $\tau \in [t]$. Hence, the Azuma-Hoeffding inequality implies that, with probability at least $1-\delta$,

$$\sum_{\tau=1}^t \mathbf{w}_\tau \cdot \mathbf{z}_\tau^* \ge tD(\boldsymbol{\nu}^*) - \sqrt{\frac{t}{2}\log(1/\delta)}.$$

Combining the above bound with line (D.5.5), we have

$$(\mathsf{A}) \le \underbrace{\overline{\mathsf{OPT}}_t(t\boldsymbol{\rho}) - tD(\boldsymbol{\nu}^*)}_{\clubsuit} + \sum_{i=1}^m \nu_{t,i}^*\left(\sum_{\tau=1}^t z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right) + \sqrt{\frac{t}{2}\log(1/\delta)}.$$

To complete the proof of the claim, we obtain the high-probability bound on $\clubsuit$ by invoking Lemma 19. First, let us first define

$$D(\boldsymbol{\nu}; \mathbf{A}) := \max_{\mathbf{z}(\mathbf{A}) \in \mathcal{Z}(i^\dagger)} (\mathbf{w} - \boldsymbol{\nu}) \cdot \mathbf{z}(\mathbf{A}) + \boldsymbol{\rho} \cdot \boldsymbol{\nu}.$$

Given this notation, we first note that $D(\boldsymbol{\nu}^*)$ defined in line (106) can be written as

$$D(\boldsymbol{\nu}^*) = \min_{\boldsymbol{\nu} \in [0,1]^m} \mathbb{E}[D(\boldsymbol{\nu}; \mathbf{A})]$$

where we used Claim 20-(a) to restrict the domain to $[0,1]^m$ without loss of optimality. On the other hand, we have:

$$\frac{\overline{\mathsf{OPT}}_t(t\boldsymbol{\rho})}{t} = \min_{\boldsymbol{\nu}_t \in \mathbb{R}_+^m} \frac{1}{t}\sum_{\tau=1}^t D(\boldsymbol{\nu}; \mathbf{A}_\tau) = \min_{\boldsymbol{\nu}_t \in [0,1]^m} \frac{1}{t}\sum_{\tau=1}^t D(\boldsymbol{\nu}; \mathbf{A}_\tau). \tag{111}$$

The first equality follows from writing the dual of $\overline{\mathsf{OPT}}_t(t\boldsymbol{\rho})$. For the second equality, we again used the fact that the optimal value $\boldsymbol{\nu}_t^*$ is always in $[0,1]^m$ without loss of optimality. [46] Because the arrivals are i.i.d samples, we observe that $\frac{\overline{\mathsf{OPT}}_t(t\boldsymbol{\rho})}{t}$ is the sample average approximation for $D(\boldsymbol{\nu}^*)$ with t i.i.d samples $(\mathbf{A}_1, ..., \mathbf{A}_t)$.

[46] Consider the minimization problem of the first equality in line (111). Let $\boldsymbol{\nu}_t^*$ be the optimal solution. Suppose for a contradiction that there exists i such that $\nu_{t,i}^* > 1$. Such location i is not assigned any free case since the reward is at most one. Hence, one can always consider an alternative solution that reduces $\nu_{t,i}^*$ infinitesimally and strictly decreases the objective value. This argument is valid as long as the total number of tied cases at each affiliate i up to time t is strictly less than $t\rho_i$. As we mentioned earlier, this event occurs with high probability due to our assumption that $\mathbb{P}[i^\dagger = i] < \rho_i$. In Appendix D.5.6, we explain how to make this argument more rigorous by incorporating such high probability.

Electronic copy available at: https://ssrn.com/abstract=4748762

Finally, it is straightforward to show that $|D(\boldsymbol{\nu};\mathbf{A})| \le 3$ for all $\boldsymbol{\nu} \in [0,1]^m$, and $\mathbf{A} \in \mathcal{A}$ almost surely. Hence, $|D(\boldsymbol{\nu};\mathbf{A}) - \mathbb{E}[D(\boldsymbol{\nu};\mathbf{A})]| \le 6$ for all $\boldsymbol{\nu} \in [0,1]^m$ and $\mathbf{A} \in \mathcal{A}$ almost surely. Invoking Lemma 19 with $M_1 = 6$, we obtain

$$\mathbb{P}\left[\frac{\overline{\mathsf{OPT}}_t(t\boldsymbol{\rho})}{t} \le D(\boldsymbol{\nu}^*) + 12\sqrt{\frac{\log(1/\delta)}{t}}\right] \ge 1-\delta.$$

This equivalently implies that, with probability at least $1-\delta$, the term ♣ in line (D.5.5) is at most $12\sqrt{t\log(1/\delta)}$. Plugging this bound in line (D.5.5) and taking the union bound, with probability at least $1-2\delta$, we have

$$(\mathsf{A}) \le \sum_{i=1}^{m} \nu^*_{t,i}\left(\sum_{\tau=1}^{t} z^{(\mathsf{M})}_{\tau,i} - t\rho_i\right) + M_{\mathsf{A}}\sqrt{t\log(1/\delta)}$$

where $M_{\mathsf{A}} = 12 + \sqrt{\frac{1}{2}}$. The proof is complete by replacing δ with $\delta/2$. □

The following lemma shows the high-probability bound of term (B) in line (109).

Claim 22 **(Bound (B))**. *There exists constant M_B such that, for any given $\delta \in (0,1)$, we have*

$$\mathbb{P}[(B) \le M_B\sqrt{t\log(1/\delta)}] \ge 1-\delta.$$

Proof of Claim 22. Let $X_s := \boldsymbol{\theta}_s \cdot (\mathbf{z}^*_s - \boldsymbol{\rho}) + \boldsymbol{\lambda}_s \cdot (\mathbb{1}[i^\dagger_s = 0]\mathbf{z}^*_s - \boldsymbol{\rho})$ and $Y_\tau := \sum_{s=1}^{\tau} X_s$, with $X_0 := 0$ and $Y_0 := 0$. Note that Y_τ is $\mathcal{H}_\tau$-measurable. We first note that Y_τ is supermartingale with respect to $\mathcal{H}_\tau$:

$$\begin{aligned}\mathbb{E}[X_\tau|\mathcal{H}_{\tau-1}] &= \boldsymbol{\theta}_\tau \cdot (\mathbb{E}[\mathbf{z}^*_\tau|\mathcal{H}_{\tau-1}] - \boldsymbol{\rho}) + \boldsymbol{\lambda}_\tau \cdot (\mathbb{E}[\mathbb{1}[i^\dagger_\tau = 0]\mathbf{z}^*_\tau|\mathcal{H}_{\tau-1}] - \boldsymbol{\rho})\\ &= \boldsymbol{\theta}_\tau \cdot (\mathbb{E}[\mathbf{z}^*_\tau] - \boldsymbol{\rho}) + \boldsymbol{\lambda}_\tau \cdot (\mathbb{E}[\mathbb{1}[i^\dagger_\tau = 0]\mathbf{z}^*_\tau] - \boldsymbol{\rho})\\ &\le 0.\end{aligned}$$

The first line is because $\boldsymbol{\theta}_\tau$ and $\boldsymbol{\lambda}_\tau$ is $\mathcal{H}_{\tau-1}$-measurable. The second line is because each $\mathbf{z}^*_\tau$ is i.i.d random variable. The final line is because the identical mean is $\mathbb{E}[\mathbf{z}^*_\tau] \le \boldsymbol{\rho}$ from Claim 20-(b). Furthermore, the Cauchy-Schwartz inequality gives $|X_\tau| \le 2(\alpha + \bar{\lambda})$ almost surely where $\bar{\lambda} := \frac{1+2\alpha}{\underline{\rho}}$. Hence, we use Azuma-Hoeffding inequality for super martingale to obtain

$$\mathbb{P}\left[Y_t \le \sqrt{2t(\alpha+\bar{\lambda})\log(1/\delta)}\right] \ge 1-\delta$$

for any given $\delta \in (0,1)$. This completes the proof. □

The following lemma obtains a deterministic bound on term (C) of line (109).

Claim 23 **(Bound (C))**. *There exists a constant M_C for which $(C) \le M_C\sqrt{t}$ for every sample path.*

Proof of Claim 23. The proof directly follows from the adversarial regret guarantee of the online mirror descent (Claim 3). In particular, let

$$\begin{aligned}R_{\boldsymbol{\theta}}(t) &:= \sum_{\tau=1}^{t} \boldsymbol{\theta}_\tau \cdot (\boldsymbol{\rho} - \mathbf{z}^{(\mathsf{M})}_\tau) - \sum_{\tau=1}^{t} \boldsymbol{\theta}^\star \cdot (\boldsymbol{\rho} - \mathbf{z}^{(\mathsf{M})}_\tau)\\ R_{\lambda}(t) &:= \sum_{\tau=1}^{t} \boldsymbol{\lambda}_\tau \cdot (\boldsymbol{\rho} - \mathbb{1}[i^\dagger_\tau = 0]\mathbf{z}^{(\mathsf{M})}_\tau)\end{aligned}$$

One can follow the same line of proofs as Claim 16 (Appendix D.5.1) to show that

$$\begin{aligned}R_{\boldsymbol{\theta}}(t) &\le 2\alpha\sum_{\tau=1}^{t}\eta_\tau + \frac{\bar{V}_{\boldsymbol{\theta}}}{\eta_t}\\ R_{\boldsymbol{\lambda}}(t) &\le 2\bar{\lambda}\sum_{\tau=1}^{t}\eta_\tau + \frac{\bar{V}_{\lambda}}{\eta_t}\end{aligned}$$

Electronic copy available at: https://ssrn.com/abstract=4748762

where $\bar{\lambda} := \frac{1+2\alpha}{\rho}$, $h(\cdot)$ is the negative entroypy function, $\bar{V}_\theta = \max_{\tau\in[t]} V_h(\boldsymbol{\theta}^\star, \boldsymbol{\theta}_\tau)$, and $\bar{V}_\lambda = \max_{\tau\in[t]} V_h(\mathbf{0}, \boldsymbol{\lambda}_\tau)$. Because the dual variables are all restricted to the bounded domain, both $\bar{V}_\theta$ and $\bar{V}_\lambda$ are bounded by some constant $\bar{V}$. Furthermore, recall that the step size of Algorithm 2 is given by $\eta_\tau = \frac{k}{\sqrt{\tau}}$ for input constant $k > 0$. Hence, both terms above are upper bounded by $M_{\mathsf{C}}\sqrt{t}$ where a positive constant M_{C} only depends on the input of `CO-DL`. This completes the proof. □

We finally combine the preceding lemmas to conclude the proofs.

Claim 24 **(Putting Everything Together)**. *There exists a constant $a > 0$ such that, for any given $\delta \in (0,1)$ and $t \in [T]$, we have*

$$\mathbb{P}\left[\max_{i\in[m]}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(M)} - t\rho_i\right) \le a\sqrt{t\log(1/\delta)}\right] \ge 1-\delta.$$

Proof of Claim 24. We start from line (109) to obtain that, with probability at least $1-\delta$,

$$\begin{aligned}\alpha\sum_{i=1}^{m}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)_+ &\le \sum_{i=1}^{m}\nu_{t,i}^*\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right) + M\sqrt{t\log(1/\delta)} \\ &\le \sum_{i=1}^{m}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)_+ + M\sqrt{t\log(1/\delta)}\end{aligned} \tag{112}$$

In the first line, we applied Claims 21-23 and the union bound to line (109). Here the constant $M > 0$ only depends on $(M_{\mathsf{A}}, M_{\mathsf{B}}, M_{\mathsf{C}})$ specified in Claims 21-23. In the second line, we used the fact that $0 \le \nu_{t,i}^* \le 1$ and $x \le (x)_+$ for any scalar x. Because $\alpha > 1$ by our assumption, we can deduce that, with probability at least $1-\delta$,

$$\sum_{i=1}^{m}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)_+ \le \frac{M}{\alpha-1}\sqrt{t\log(1/\delta)}.$$

The proof is complete by noting that

$$\max_{i\in[m]}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right) \le \sum_{i=1}^{m}\left(\sum_{\tau=1}^{t} z_{\tau,i}^{(\mathsf{M})} - t\rho_i\right)_+$$

and defining $a := \frac{M}{\alpha-1}$. □

D.5.6. Missing Details for Proofs in Appendix D.5.5

In this subsection, we explain how to make the arguments made in Appendix D.5.5 more precise. In particular, in the proof of Claim 21 (Appendix D.5.5), we assumed that the optimization problem (110) is always feasible for the sake of simplicity.[47] Let $\mathcal{G}_t$ be the event that that the (110) admits a feasible solution. Here we show that incorporating the precise high-probability bound of $\mathcal{G}_t$ does not affect our main result (i.e., Proposition 4 and Lemma 17).

We first begin by incorporating the probability of $\mathcal{G}_t$ to the proof of Section D.5.5. The only part we need to modify is Claim 21. Let $N_{t,i}$ be the total number of tied cases at affiliate i up to time t. We first note that the program (110) admits a feasible solution if and only if $N_{t,i} \le \rho_i t$ for all $i \in [m]$. We further recall that we assumed that $\mathbb{P}[i_t^\dagger = i] = \rho_i - d_i$ for some positive constant $d_i > 0$. Hence, by Azuma-Hoeffding inequality and the union bound, we have:

$$\mathbb{P}[\mathcal{G}_t] \ge 1 - \sum_{i=1}^{m}\exp\left(-\frac{d_i^2 t}{2}\right) := 1 - g(t) \tag{113}$$

[47] Note that this is always true in the absence of tied cases.

Electronic copy available at: https://ssrn.com/abstract=4748762

where in the last line we define $g(t) := \sum_{i=1}^{m} \exp\left(-\frac{d_i^2 t}{2}\right)$. Incorporating the above high-probability bound, the precise statement of Claim 21 becomes the following:

CLAIM 25. *There exists constant M_A such that, for any given $\delta \in (0,1)$, the term (A) in line (109) satisfies:*

$$(A) \le \sum_{i=1}^{m} \nu_{t,i}^{*} \left(\sum_{\tau=1}^{t} z_{\tau,i}^{(M)} - t\rho_i \right) + M_A \sqrt{t \log(1/\delta)}$$

with probability at least $1-\delta-g(t)$ and for some $0 \le \nu_{t,i}^{*} \le 1$.

The rest of the proofs in Section D.5.5 does not change. Hence, the statement of Lemma 18 now changes to

LEMMA 20. *For any fixed t, there exists a constant $a > 0$ for which*

$$\mathbb{P}\left[\mathbf{Z}_t^{(M)} \le t\boldsymbol{\rho} + a\sqrt{t\log(1/\delta)}\mathbf{1} \right] \ge 1 - \delta - g(t)$$

where $\mathbf{1}$ is the vector of ones.

By following the same steps from lines (103) to (105) with Lemma 20, the statement of Proposition Proposition 4 accordingly changes to the following.

PROPOSITION 5. *For CO-DL (Algorithm 2), for any given $\delta \in (0,1)$, we have*

$$T - S_A \le \frac{a}{\underline{\rho}}\sqrt{T\log(1/\delta)} + \frac{1}{\underline{\rho}}$$

with probability at least $1-\delta-g(T(\delta))$ where constant $a>0$ is given by Lemma 20 and $T(\delta) := T - \frac{a}{\underline{\rho}}\sqrt{T\log(1/\delta)} - \frac{1}{\underline{\rho}}$.

Finally, we show that Proposition 5 implies $\mathcal{O}(T)$ upper bound on $\mathbb{E}[(T-S_A)^2]$, as we desired for Lemma 17. It suffices to show $\mathbb{E}[(T-S_A-\frac{1}{\underline{\rho}})] \le \mathcal{O}(T)$.[48] We first note that Proposition 5, after some algebras, is equivalent to saying that

$$\mathbb{P}\left[T - S_A - \frac{1}{\underline{\rho}} \ge x\right] \le \exp\left(-\frac{\underline{\rho}^2 x^2}{Ta^2}\right) + \sum_{i=1}^{m} \exp\left(-\frac{d_i^2}{2}(T - x - \frac{1}{\underline{\rho}})\right) \tag{114}$$

for any $x \ge 0$. Assume, for the sake of simplicity, that $1/\underline{\rho}$ is a positive integer (this makes us simplify the algebras). Similar to line (D.5), we have

$$\begin{aligned}
\mathbb{E}\left[(T - S_A - \frac{1}{\underline{\rho}})^2\right] &= \sum_{x=0}^{T^2} \Pr\left[(T - S_A - \frac{1}{\underline{\rho}})^2 \ge x\right] \\
&= \sum_{x=0}^{T^2} \Pr\left[(T - S_A - \frac{1}{\underline{\rho}}) \ge \sqrt{x}\right] \\
&= \sum_{x=0}^{T^2} \left\{ \exp\left(-\frac{\underline{\rho}^2 x}{Ta^2}\right) + \sum_{i=1}^{m} \exp\left(-\frac{d_i^2}{2}(T - \sqrt{x} - \frac{1}{\underline{\rho}})\right) \right\} \\
&\le \underbrace{1 + \int_0^{T^2} \exp\left(-\frac{\underline{\rho}^2 x}{Ta^2}\right) dx}_{(\clubsuit)} + \sum_{i=1}^{m} \underbrace{\int_0^{T^2+1} \exp\left(-\frac{d_i^2}{2}(T - \sqrt{x} - \frac{1}{\underline{\rho}})\right) dx}_{\diamondsuit}
\end{aligned}$$

[48] To see this, we note that $\mathbb{E}[(T-S_A)^2] = \mathbb{E}[(T - S_A - \frac{1}{\underline{\rho}} + \frac{1}{\underline{\rho}})^2] = \mathbb{E}[(T-S_A-\frac{1}{\underline{\rho}})^2] + \frac{2}{\underline{\rho}}\mathbb{E}[T-S_A] + \frac{1}{\underline{\rho}^2} \le \mathbb{E}[(T-S_A-\frac{1}{\underline{\rho}})^2] + \mathcal{O}(T)$ where the last inequality is because $T - S_A \le T$ almost surely and $\underline{\rho}$ is a constant by our assumption.

Electronic copy available at: https://ssrn.com/abstract=4748762

The second line is justified because $T - S_A - \frac{1}{\underline{\rho}}$ is non-negative almost surely.[49] The third line follows from the tail bound (114). The last inequality follows since (i) $e(-ax)$ is a decreasing convex function for $x \in \mathbb{R}^+$ and $a > 0$ and (ii) $e(\frac{d_i^2}{2}\sqrt{x})$ is an increasing concave function for all $x \geq 0$ and $d_i \leq 1$. It is straightforward to see that $\clubsuit \leq \mathcal{O}(T)$. For $\diamondsuit$, we have

$$\begin{aligned}
\diamondsuit &= \exp\left(-\frac{d_i^2}{2}(T - \frac{1}{\underline{\rho}})\right) \int_0^{T^2+1} \exp\left(\frac{d_i^2}{2}\sqrt{x}\right) dx \\
&= \exp\left(-\frac{d_i^2}{2}(T - \frac{1}{\underline{\rho}})\right) \int_0^{\frac{d_i^4}{4}(T^2+1)} \exp(\sqrt{u}) \frac{4}{d_i^4} du \\
&= \exp\left(-\frac{d_i^2}{2}(T - \frac{1}{\underline{\rho}})\right) \times \frac{4}{d_i^4} \left\{ 2\exp\left(\sqrt{\frac{d_i^4}{4}(T^2+1)}\right)\left(\sqrt{\frac{d_i^4}{4}(T^2+1)} - 1\right) + 2 \right\} \\
&\leq \exp\left(-\frac{d_i^2}{2}(T - \frac{1}{\underline{\rho}})\right) \times \frac{8}{d_i^4} \left\{ \exp\left(\frac{d_i^2 T}{2} + \frac{d_i^2}{2}\right)\left(\frac{d_i^2 T}{2} + \frac{d_i^2}{2} - 1\right) + 2 \right\} \\
&= \frac{8}{d_i^4}\left[\exp\left(\frac{d_i^2}{2\underline{\rho}} + \frac{d_i^2}{2}\right)\left(\frac{d_i^2 T}{2} + \frac{d_i^2}{2} - 1\right) + 2\exp\left(-\frac{d_i^2}{2}(T - \frac{1}{\underline{\rho}})\right)\right] \\
&= \mathcal{O}(T).
\end{aligned}$$

In the third line, we used $\int \exp(\sqrt{x})dx = 2\exp(\sqrt{x})(\sqrt{x} - 1)$. In the fourth line, we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any nonnegative scalars a and b. The rest of the lines is a straightforward algebra. Hence, the proof is complete.

Appendix E: Proof of Theorem 3

Proof Sketch of Theorem 3: For the proof, we revisit the instance considered in Example 1 with one-affiliate. The reward is drawn from the uniform distribution on interval $(0, 1)$. The capacity ratio is $\rho = 0.5$ and there are no tied cases. For brevity, we omit the subscript for $i = 1$.

The proof consists of two steps. First, we establish that the value of Algorithm 2 (i.e. `CO-DL`) on this instance is at most $\frac{3}{8}T - \Theta(\gamma\sqrt{T})$. To prove this, we show that (i) the expected reward of `CO-DL` is at most $\frac{3}{8}T$ and (ii) the average backlog of `CO-DL` is $\Omega(\sqrt{T})$ in expectation. In the second step, we show that the expected value of the optimal offline (Definition 2) is at least $\frac{3}{8}T - \Theta(\sqrt{\gamma T} + \gamma)$. Combining (and omitting the lower-order terms), we complete the proof of Theorem 3. In the following, we outline each step of the proof.

Step 1: Upper Bound the Value of `CO-DL`.

Let z_t and b_t denote the decision and backlog of `CO-DL`.

LEMMA 21. $\mathbb{E}\left[\sum_{t=1}^{T} w_t z_t\right] \leq \frac{3}{8}T$

Proof of Lemma 21. The proof is a direct consequence of Claim 12-(c) (Appendix D.3). The optimal solution for the static dual problem (Definition 6) is such that $\theta^* + \lambda^* = 0.5$ at which a per-period expected reward is $\mathbb{E}[w_t \mathbb{1}[w_t \geq 0.5]] = \frac{3}{8}$. Hence, the result follows. □

Next, we lower-bound the average backlog of `CO-DL` by $\Omega(\sqrt{T})$ in expectation. We first show the following probability bound on the total number of arrivals to the affiliate:

LEMMA 22. *There exists a constant $\kappa > 0$ such that, for any fixed $t \leq 0.5T$ and $\delta \in (0, 1)$,*

$$\mathbb{P}\left[\sum_{\tau=1}^{t} z_\tau \geq 0.5t - \kappa\sqrt{t\log(1/\delta)}\right] \geq 1 - \delta.$$

[49] Note that $S_A \geq \underline{\rho}T$ for every sample path. Hence, $T - S_A$ must be greater than $\frac{1}{\underline{\rho}}$ for all sufficiently large T.

Electronic copy available at: https://ssrn.com/abstract=4748762

We prove this lemma proof in Appendix E.1. For the proof, we use a similar technique from Lemma 18 (Appendix D.5.5) to study the cumulative number of cases accepted until time $t \le 0.5T$. The difference is that we obtain a "lower-bound" of the number of accepted cases while we were interested in upper-bounding it in Lemma 18.

We now use Lemma 22 to obtain a lower bound on the expected backlog.

LEMMA 23. *There exists constants $a > 0$ and $t(a)$, such that, whenever $T \ge 2t(a)$ and $\epsilon \le \frac{a}{4\sqrt{T}}$, we have:*

$$\mathbb{E}[b_t] \ge \frac{a\sqrt{t}\Phi(-3\sqrt{2}a)}{16}$$

for all $t(a) \le t \le \frac{T}{2}$. Here $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. The exact form of the constant $t(a)$, as a function of the constant $a > 0$, is given in the proof.

The proof of this lemma is presented in Appendix E.2. For the proof, we use the following simple anti-concentration result. Let S_t be the total number of periods that the server is available up to time t. Under the near-critical regime, there is a constant probability that $S_t \le 0.5t - \Theta(\sqrt{t})$. Combining this with the lower bound of the total number of arrivals established in Lemma 22, we can show that at least $\Theta(\sqrt{t})$ cases remain in the queue in expectation for each period $t \le \frac{T}{2}$. Finally, we note that summing the inequality Lemma 23 for all $t(a) \le t \le \frac{T}{2}$ implies the $\Omega(\sqrt{T})$ average backlog of Algorithm 2 in expectation.

Step 2: Lower-bounding the Value of the Optimal Offline. The following lemma lower-bound the value of the offline.

LEMMA 24. $\mathbb{E}[OPT(\gamma)] \ge \frac{3}{8}T - \Theta(\sqrt{\gamma T} + \gamma)$

The proof of this lemma can be found in Appendix E.3. For the proof, we study the expected objective value of the following feasible solution for each sample path: let $\delta := \sqrt{\gamma/T}$[50] and define $z_t^* = \mathbb{1}[w_t \ge 0.5 + \delta]$ up to $t \le T_A$ and $z_t^* = 0$ for all $t > T_A$, where the stopping time T_A is defined as $T_A = \min\{t \le T : \sum_{\tau=1}^{t} z_t^* \ge 0.5T\}$. We then show that (i) the total expected reward of this feasible solution is at least $\frac{3}{8}T - \Theta(\sqrt{\gamma T} + \gamma)$ and (ii) the expected time-average backlog is $O(\sqrt{T/\gamma} + 1)$.

E.1. Proof of Lemma 22

The proof follows similar steps for proving Lemma 18 in Appendix D.5.5. For brevity, we only highlight a main difference from the proof. As we consider a one-dimensional example with capacity ratio $\rho = 0.5$, we omit the subscript for i. Fix a sample path of rewards $w_1, \ldots w_T$. Let z_t be the decision for arrival t. Because the algorithm is unconstrained by the hard capacity constraint until $t \le \rho T$, the optimality criterion of the algorithm implies that

$$\sum_{\tau=1}^{t}(w_\tau - \theta_\tau - \mathbb{1}[i_\tau^\dagger = 0]\lambda_\tau)\cdot z_\tau + \rho\cdot(\theta_\tau + \lambda_\tau) \ge \sum_{\tau=1}^{t}(w_\tau - \theta_\tau - \mathbb{1}[i_\tau^\dagger = 0]\lambda_\tau)z_\tau^* + \rho\cdot(\theta_\tau + \lambda_\tau).$$

for any fixed $t \le \rho T$ and any alternative decision $z_\tau^* \in \mathcal{Z}(i_\tau^\dagger)$. Now take the same alternative decision as defined in line (108). By following the identical steps up to line (112), we can show that, with probability at least $1 - \delta$,

$$\alpha\left(\sum_{\tau=1}^{t} z_\tau - t\rho\right)_+ \le \nu^*\left(\sum_{\tau=1}^{t} z_\tau - t\rho\right) + M\sqrt{t\log(1/\delta)}$$

[50] Note that this is $\delta = o(1)$ because we focus on $\gamma = o(T)$ in light of Proposition 1 (see Section 2.1)

Electronic copy available at: https://ssrn.com/abstract=4748762

where ν^* is the (optimal) dual variable corresponding to the capacity constraint in the optimization problem (110).[51] For this instance, as we discuss in Example 1, the dual variable ν^* for each sample path $w_1, ..., w_t$ is the sample median of the rewards. Because the support of the reward is bounded away from zero, there exists a constant $\underline{\nu} > 0$ such that $\nu^* \geq \underline{\nu}$.[52] Hence, dividing the above inequality by ν^* and defining the constant $\kappa := \frac{M}{\underline{\nu}}$, we conclude that

$$\sum_{\tau=1}^{t} z_\tau \geq t\rho - \kappa\sqrt{t\log(1/\delta)}$$

with probability at least $1-\delta$.

E.2. Proof of Lemma 23

Recall that we are analyzing an instance with $m = 1$. Hence, for brevity, we omit the subscript for $i = 1$. Let $S_t = \sum_{\tau=1}^{t} s_\tau$ and $Z_t = \sum_{\tau=1}^{t} z_\tau$. We first obtain a simple bound on the backlog.

Claim 26. *$b_t \geq (Z_t - S_t)_+$ for all $t \geq 1$ for every sample path.*

Proof of Claim 26. The proof is by induction on t. The base case for $t = 1$ is trivial because $b_0 = 0$. Now suppose that the claim is true for $t \geq 1$. Then we have

$$b_{t+1} = (b_t + z_{t+1} - s_{t+1})_+ \geq (Z_t - S_t + z_{t+1} - s_{t+1})_+ = (Z_{t+1} - S_{t+1})_+$$

where the first inequality follows from the induction hypothesis. This completes the proof. □

Fix a positive constant $a > 0$, which we will specify later. For given constant $a > 0$ and t, define an event

$$\mathcal{Z}_{a,t} = \{Z_t \geq 0.5t - a\sqrt{t}\}$$
$$\mathcal{S}_{a,t} = \{0.5t - a\sqrt{t} - S_t \geq \frac{a}{4}\sqrt{t}\}$$
$$\mathcal{E}_{a,t} := \mathcal{Z}_{a,t} \cap \mathcal{S}_{a,t}$$

First, from Claim 26, we note that

$$\mathbb{E}[b_t] \geq \mathbb{E}[(Z_t - S_t)_+] \geq \mathbb{P}[\mathcal{E}_{a,t}]\mathbb{E}[(Z_t - S_t)_+ | \mathcal{E}_{a,t}] \geq \mathbb{P}[\mathcal{E}_{a,t}] \times \frac{a\sqrt{t}}{4}.$$

where the last inequality follows from the definition of $\mathcal{E}_{a,t}$. Hence, it suffices to lower-bound $\mathbb{P}[\mathcal{E}_{a,t}]$ by $\frac{\Phi(-3\sqrt{2}a)}{4}$ for some constant $a > 0$. First, note that the decision of `CO-DL` *does not* depend on the service (and therefore backlog) process. Hence, the event $\mathcal{Z}_{a,t}$ and $\mathcal{S}_{a,t}$ must be independent for Algorithm 2, implying that $\mathbb{P}[\mathcal{E}_{a,t}] = \mathbb{P}[\mathcal{Z}_{a,t}]\mathbb{P}[\mathcal{S}_{a,t}]$.

Thus, it suffices to separately lower-bound the probability of $\mathcal{Z}_{a,t}$ and $\mathcal{S}_{a,t}$. From Lemma 22, there exists constant $\kappa > 0$ such that, for any given $a > 0$ and $t \leq 0.5T$, we have

$$\mathbb{P}[\mathcal{Z}_{a,t}] \geq 1 - \exp\left(-\frac{a^2}{\kappa^2}\right). \tag{115}$$

[51] The optimization problem admits a feasible solution for every sample path because there is no tied case.

[52] One can make this argument more precise by incorporating the concentration of ν^*. By the property of the uniform random variable on interval $(0,1)$, the sample median of $(w_1, ...w_t)$ follows Beta$(t/2+1, t/2+1)$ distribution (see, for example, Example 5.4.5 of Casella and Berger (2021)). Hence, the mean of ν^* is $1/2$ and the variance is $\Theta(1/t)$. Hence, for large enough t, the variance of ν^* is small enough to ensure that ν^* is bounded above some fixed constant $\underline{\nu}$ with high probability. The result incorporating this detail only changes the result of Theorem 4 in a minor manner. For the sake of brevity, we omit such details in the proof.

Electronic copy available at: https://ssrn.com/abstract=4748762

Let us choose a sufficiently large constant $a>0$ such that the right-hand side is at least $\frac{1}{2}$. For such constant $a>0$, we now turn our attention to lower-bounding $\mathbb{P}[\mathcal{S}_{a,t}]$ as a function of the constant $a>0$. By Berry-Esseen Theorem (see, for example, Theorem 3.4.17 of Durrett (2019)), there exists a constant $d>0$ such that for all $t\geq 1$,

$$\mathbb{P}\Big[S_t \leq \mathbb{E}[S_t] - 3\sqrt{2}a\sqrt{\mathrm{Var}[S_t]}\Big] \geq \Phi(-3\sqrt{2}a) - \frac{d}{\sqrt{t}}.$$

where $\mathbb{E}[S_t]=(0.5+\epsilon)t$, $\mathrm{Var}[S_t]=(0.25-\epsilon^2)t$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. Moving the terms, the preceding inequality is equivalent to

$$\mathbb{P}\left[0.5t - a\sqrt{t} - S_t \geq 3\sqrt{2}a \times \sqrt{\mathrm{Var}[S_t]} - a\sqrt{t} - \epsilon t\right] \geq \Phi(-3\sqrt{2}a) - \frac{d}{\sqrt{t}}.$$

Let us define a constant $t(a)$ as

$$t(a) := \max\{\frac{4d^2}{\Phi^2(-3\sqrt{2}a)}, \frac{a^2}{2}\} \tag{116}$$

We now use the assumption $\epsilon \leq \frac{a}{4\sqrt{T}}$ and $T\geq 2t(a)$ to prove the following:

$$\begin{aligned}
3\sqrt{2}a \times \sqrt{\mathrm{Var}[S_t]} - a\sqrt{t} - \epsilon t &= 3\sqrt{2}a\sqrt{(0.25-\epsilon^2)t} - a\sqrt{t} - \epsilon t \\
&\geq 3\sqrt{2}a\sqrt{\left(0.25 - \frac{a^2}{16T}\right)t} - a\sqrt{t} - \frac{at}{4\sqrt{T}} \\
&\geq \frac{a}{2}\sqrt{t} - \frac{at}{4\sqrt{T}} \\
&\geq \frac{a}{4}\sqrt{t}
\end{aligned}$$

In the second line, we used the assumption $\epsilon \leq \frac{a}{4\sqrt{T}}$. In the third line, we used the assumption $T\geq t(a)\geq \frac{a^2}{2}$. The last line follows from $T\geq t$. Furthermore, for all $t\geq t(a)\geq \frac{4d^2}{(\Phi(-3\sqrt{2}a))^2}$, it is straightforward to check that $\Phi(-3\sqrt{2}a) - \frac{d}{\sqrt{t}} \geq \frac{\Phi(-3\sqrt{2}a)}{2}$. Hence, for any $t(a)\leq t\leq \frac{1}{2}T$, we conclude that

$$\mathbb{P}[\mathcal{S}_{a,t}] = \mathbb{P}\left[0.5t - a\sqrt{t} - S_t \geq \frac{a}{4}\sqrt{t}\right] \geq \mathbb{P}\left[0.5t - a\sqrt{t} - S_t \geq 3\sqrt{2}a \times \sqrt{\mathrm{Var}[S_t]} - a\sqrt{t} - \epsilon t\right] \geq \frac{\Phi(-3\sqrt{2}a)}{2}$$

Combining, we have

$$\mathbb{P}[\mathcal{E}_{a,t}] = \mathbb{P}[\mathcal{Z}_{a,t}]\mathbb{P}[\mathcal{S}_{a,t}] \geq \frac{\Phi(-3\sqrt{2}a)}{4}$$

for some constant $a>0$. This completes the proof.

E.3. Proof of Lemma 24

Let $\delta := \sqrt{\gamma/T}$. Consider the following feasible solution: $z_t^* = \mathbb{1}[w_t \geq 0.5+\delta]$ up to $t\leq T_A$ and $z_t^*=0$ for all $t>T_A$, where the stopping time T_A is defined as $T_A = \min\{t\leq T : \sum_{\tau=1}^t z_t^* \geq 0.5T\}$. We analyze the the expected objective value of this feasible solution, which will be the lower bound on the expected value of the optimal offline solution.

We first analyze the reward of this policy. Because (i) T_A is a bounded stopping time with respect to $\{\mathcal{H}_t\}$ and (ii) each $w_t z_t$ is independent with identical mean for $t\leq T_A$, Wald's equation implies that

$$\mathbb{E}\Big[\sum_{t=1}^{T_A} w_t z_t^*\Big] = \mathbb{E}[T_A]\mathbb{E}[w_1 z_1^*] = \mathbb{E}[T_A]\int_{0.5+\delta}^1 w dw \geq \mathbb{E}[T_A]\Big(\frac{3}{8} - \frac{1}{2}\delta - \frac{1}{2}\delta^2\Big). \tag{117}$$

where in the second equality we used the fact that reward follows the uniform distribution on interval $(0,1)$.

Electronic copy available at: https://ssrn.com/abstract=4748762

To further lower-bound $\mathbb{E}[T_A]$, we establish a concentration inequality of $T - T_A$. By Hoeffding's inequality, we have, for any t,

$$\mathbb{P}\Big[\sum_{\tau=1}^{t} z_t^* \le (0.5-\delta)t + \sqrt{\frac{t}{2}\log(1/x)}\Big] \ge 1-x$$

for all $x \in (0,1)$. By plugging $t = T - \frac{1}{0.5}\sqrt{\frac{T}{2}\log(1/x)}$, we can straightforwardly see that, for any $x \in (0,1)$

$$\mathbb{P}\Big[T_A \ge T - \frac{1}{0.5}\sqrt{\frac{T}{2}\log(1/x)}\Big] \ge 1-x \tag{118}$$

which further implies $\mathbb{E}[T_A] \ge T - \Theta(\sqrt{T})$ (by integrating the preceding probability bound). Plugging this bound into line (117) (with $\delta = \sqrt{\gamma/T}$), we obtain

$$\mathbb{E}\Big[\sum_{t=1}^{T_A} w_t z_t^*\Big] \ge (T - \Theta(\sqrt{T}))\Big(\frac{3}{8} - \frac{1}{2}\delta - \frac{1}{2}\delta^2\Big) = \frac{3}{8}T - \Theta(\sqrt{\gamma T} + \gamma).$$

We now turn our attention analyzing the build-up of the proposed policy. Let $\{b_t^*\}_{t=1}^T$ be the induced build-up. From our drift lemma (Lemma 3), for all $t \le T_A$, we have:

$$\mathbb{E}\Big[\frac{(b_t^*)^2 - (b_{t-1}^*)^2}{2}\Big|\mathcal{H}_{t-1}\Big] \le b_{t-1}^*(\mathbb{E}[z_t^*|\mathcal{H}_{t-1}] - 0.5) + O(1) \le -\delta b_{t-1}^* + O(1) \tag{119}$$

where in the last inequality we used $\mathbb{E}[z_t^*|\mathcal{H}_{t-1}] = \mathbb{E}[z_t^*] = \mathbb{P}[w_t \ge 0.5+\delta] = 0.5 - \delta$ because the reward follows the uniform distribution on interval $(0,1)$. Summing up line (119) for $t \le T_A$ and taking the outer expectation, we obtain that

$$\delta\mathbb{E}\Big[\sum_{t=1}^{T_A-1} b_t^*\Big] \le -\frac{1}{2}\mathbb{E}[(b_{T_A}^*)^2] + O(T). \tag{120}$$

Hence, we have

$$\begin{aligned}
\frac{\mathbb{E}[\sum_{t=1}^T b_t^*]}{T} &= \frac{\mathbb{E}[\sum_{t=1}^{T_A-1} b_t^*] + \mathbb{E}[\sum_{t=T_A}^T b_t^*]}{T} \\
&= \frac{\mathbb{E}[\sum_{t=1}^{T_A-1} b_t^*] + \mathbb{E}[(T-T_A+1)b_{T_A-1}^*]}{T} \\
&= \frac{\mathbb{E}[\sum_{t=1}^{T_A-1} b_t^*] + \sqrt{\mathbb{E}[(T-T_A+1)^2]}\sqrt{\mathbb{E}[(b_{T_A-1}^*)^2]}}{T} \\
&\le O(\frac{1}{\delta}+1).
\end{aligned}$$

In the second line, we used the fact that $b_t \le b_{T_A}$ for all $t \ge T_A$ (because this is a one-affiliate example and $\rho < 1$, the build-up cannot increase after $t \ge T_A$). In the third line, we used the Cauchy-Schwartz inequality for expecation. In the last line, we used (i) $\mathbb{E}[\sum_{t=1}^{T_A-1} b_t^*] \le O(T/\delta)$ from line (120), (ii) $\mathbb{E}[(b_{T_A}^*)^2] \le O(T)$ from line (120), and (iii) $\mathbb{E}[(T-T_A)^2] \le O(T)$ from the high-probability bound (118). Combining with $\delta = \sqrt{\gamma/T}$, we have $\frac{\gamma}{T}\mathbb{E}[\sum_{t=1}^T b_t^*] = O(\sqrt{\gamma T} + \gamma)$.

Combining, we conclude that the value of the proposed policy is

$$\mathbb{E}\Big[\sum_{t=1}^{T_A} w_t z_t^*\Big] + \frac{\gamma}{T}\mathbb{E}[\sum_{t=1}^{T} b_t^*] \ge \frac{3}{8}T - \Theta(\sqrt{\gamma T} + \gamma). \tag{121}$$

This completes the proof.

Electronic copy available at: https://ssrn.com/abstract=4748762

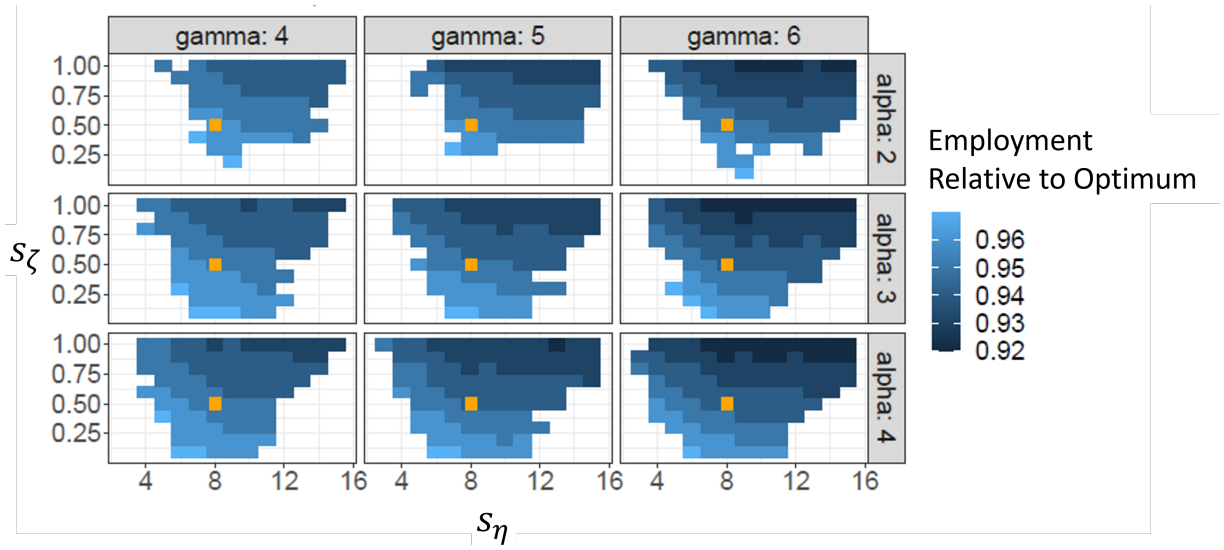


Figure 4 **Acceptable region of (s_η, s_ζ) based on data of year 2014**

Appendix F: Details and Additional Results for Case Study

F.1. Implementation Details of `RO-Learning`

In the following, we present more details on the implementation of `RO-Learning` for our case study. Specifically, Recall that Algorithm 1 (which `RO-Learning` is based on) requires two internal parameters, η and ζ, the step-sizes for $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\boldsymbol{\beta}$, respectively. We set this step sizes as a function of the cost parameters as follows:

$$\eta = \frac{s_\eta \log(\alpha+1)}{\sqrt{T}}$$
$$\zeta = \frac{s_\zeta \gamma}{\sqrt{T}}$$

Intuitively, this makes the algorithm more conservative in terms of the cost metrics as the cost parameters α and γ increase.

We adopt a data-driven approach of tuning the parameters (s_η, s_ζ), to select values that lead to acceptable outcomes for both over-allocation and average backlog. Specifically, we run the algorithm on the dataset of year 2014, obtaining the over-allocation and average backlog for each $1 \le s_\eta \le 15$ and $0.1 \le s_\zeta \le 10$. The resulting *acceptable region* for (s_η, s_ζ) is defined by values that satisfy two criteria: (i) achieving an over-allocation at most $1.2\times$ (the minimum value across all parameter settings) and (ii) maintaining an average backlog at most $1.1\times$ (the value achieved by the actual placement in the year 2014). Figure 4 illustrates the acceptable region for some cost parameters, with the ratio of employment outcome compared to the optimal level of surrogate primal (Definition 5). Upon visual examination, we identify that values within the ranges $5 \le s_\eta \le 10$ and $0.1 \le s_\zeta \le 10$ generally fall within the acceptable region across a broad range of the cost parameters. Consequently, we opt for $s_\eta = 8$ and $s_\zeta = 0.5$, roughly representing a central point within the established acceptable region. This selection aims to manage cost metrics without compromising employment outcomes too severely. The choice is further motivated to ensure the robustness of our case study results with respect to variations in step sizes.

F.2. Implementation Details of `Sampling`

In the following section, we present the details of `Sampling` for the sake of completeness. The original version of `Sampling` (Bansak and Paulson 2022) was designed for a setting without tied cases. Here, we formally describe the extension of the original version to accommodate scenarios with tied cases.

Electronic copy available at: https://ssrn.com/abstract=4748762

As we discussed in Section 5, `Sampling` *requires* a sample trajectory of future arrivals. The future arrivals are drawn from a *sampling pool*, denoted by $\mathcal{P}$, with replacement. Following Bansak and Paulson (2022), we use the previous year's arrival data as the sampling pool. At each time t, let $c_{t,i}$ and $c_{t,i}^{(\text{free})}$ denote the remaining aggregate capacity $c_{t,i}$ and the capacity for the number of free cases, respectively, for each affiliate i. Upon arrival of case t and observing $(\mathbf{w}_t, i_t^\dagger)$, `Sampling` first samples a future arrival $(\mathbf{w}_\tau, i_\tau^\dagger)_{\tau=t+1}^T$ from its sampling pool $\mathcal{P}$. In our simulation, this means that we sample future arrivals of time $t+1$ and onward from the previous year's arrival data. Given this sampled trajectory of future arrivals, `Sampling` solves the following optimization problem to obtain the allocation decision:

$$\begin{aligned} \max_{\{\mathbf{z}_\tau\}_{\tau=t}^T} \quad & \tilde{\mathbf{w}}_t \cdot \mathbf{z}_t + \sum_{\tau=t+1}^{T} \mathbf{w}_t \cdot \mathbf{z}_t - \alpha \sum_{i=1}^{m} \left(\sum_{\tau=t}^{T} z_{t,i} - c_{t,i} \right)_+ \\ \text{s.t} \quad & \sum_{\tau=t}^{T} \mathbb{1}[i_t^\dagger = 0] z_{t,i} \le c_{t,i}^{\text{free}} \end{aligned}$$

where

$$\tilde{w}_{t,i} = w_{t,i} - \frac{\gamma}{T} \left\lceil \frac{b_{t-1,i} - \rho_i}{\rho_i} \right\rceil \mathbb{1}[b_{t-1,i} > 0]. \tag{122}$$

That is, `Sampling` solves an offline problem that maximizes the net matching reward given the current remaining capacity with a tweak: to further control the congestion, as in line (122), it penalizes the (observed) employment probability at affiliate i for the *current* case t by the waiting time the case will experience upon placement computed based on the first-come-first-served manner (see Lemma 1 of Bansak and Paulson (2022) for discussion on such penalty function), with the penalty factor increasing with γ. The `Sampling` algorithm repeats this process for K simulated trajectories of future arrivals and assigns case t to the location where it was allocated most often. In our case study, we set $K = 5$ for these simulations.

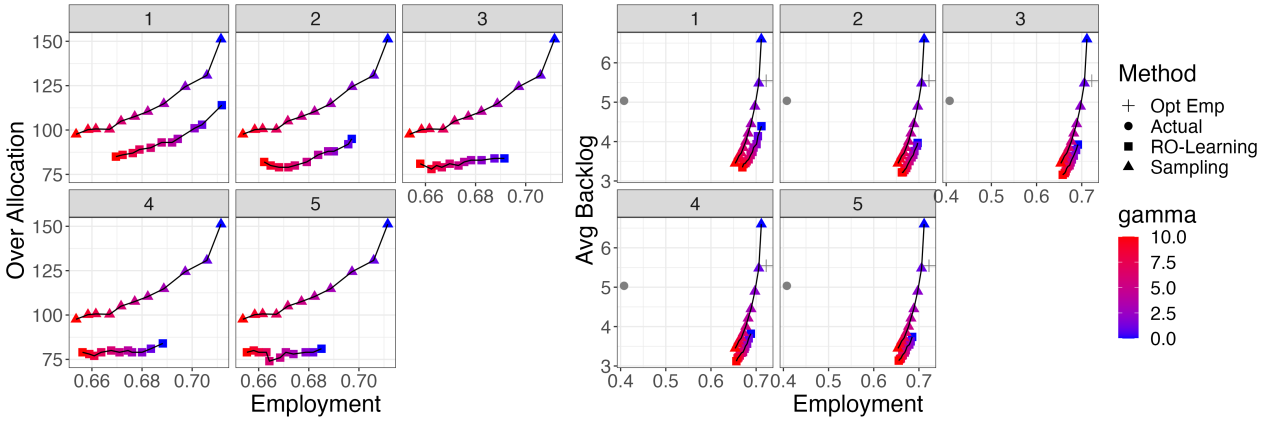
F.3. Additional Numerical Results for `RO-Learning` versus `Sampling`

In this subsection, we present A more extensive case study for a broader range of the cost parameters α and γ. In Figure 5, we display the results of the placement outcome for $\alpha \in \{1, 2, .., 5\}$ and $\gamma \in \{0, 1, ..., 10\}$. In each panel, we display the trajectory of the outcome metrics of fixed α (label of the penal) and varying γ. Here, the point `Opt-Emp` is the outcome achieved by the surrogate primal problem (see Definition 5), which does not incur any over-allocation when $\alpha \ge 1$.

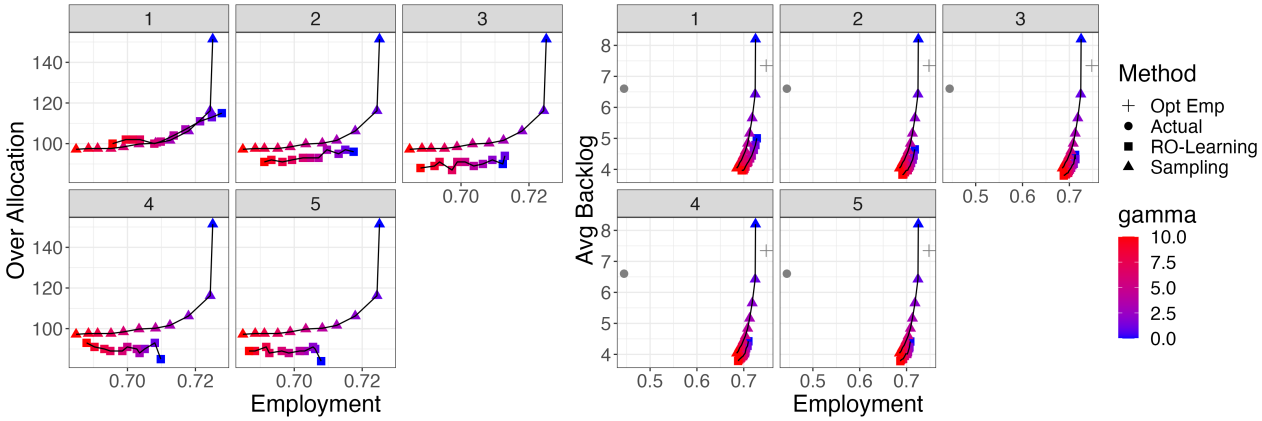
We first note that `RO-Learning` consistently improves upon `Actual` in terms of the employment rate and the average backlog (recall that `Actual` does not incur over-allocation due to the definition of the capacity profile we chose for our case study). When compared with `Sampling`, for $\alpha \ge 2$, the trajectory of outcome metrics for `RO-Learning` generally dominates that of `Sampling`. Specifically, for a given value of the employment outcome, `RO-Learning` achieves lower over-allocation and average backlog compared to `Sampling`. This shows that the our learning-based approach better balances the three (potentially) conflicting outcome metrics. Indeed, `RO-Learning` improves the objective function (defined in (2) in Section 2) compared to `Sampling` by 31% (16%, respectively) in year 2015 (2016, respectively) on average, with improvements ranging from 10% to 50% across the considered range of cost parameters.

Finally, we note that even when $\gamma = 0$, the average backlog of `RO-Learning` is reasonably low compared to both the actual placement and `Sampling`. For example, focusing on the panel $\alpha = 3$, the average backlog of `RO-Learning` at $\gamma = 0$ is roughly equivalent to that of `Sampling` at $\gamma = 5$. As discussed in Section 5, this represents the inherent balancedness of the Learning-based approach.

Electronic copy available at: https://ssrn.com/abstract=4748762



(a) Year 2015



(b) Year 2016

Figure 5 **Additional results for case Study with $\alpha \in \{1, 2, .., 5\}$ and $\gamma \in \{0, 1, ..., 10\}$. In each panel, we display the trajectory of the outcome metrics of fixed α (label of the penal) and varying γ.**

Electronic copy available at: https://ssrn.com/abstract=4748762