Constant Regret Primal-Dual Policy for Multi-way Dynamic Matching

Yehua Wei, Jiaming Xu, Sophie H. Yu
The Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, North Carolina 27708,
yehua.wei@duke.edu, jiaming.xu868@duke.edu, haoyang.yu@duke.edu,

We study a discrete-time dynamic multi-way matching model. There are finitely many agent types that arrive stochastically and wait to be matched. State-of-the-art dynamic matching policies in the literature require the knowledge of all system parameters to determine an optimal basis of the fluid relaxation, and focus on controlling the number of waiting agents using only matches in the optimal basis (Kerimov et al., 2021a,b; Gupta, 2021). In this paper, we propose a primal-dual policy that schedule matches for future arrivals based on an estimator for the dual solution. Our policy does not require the knowledge of optimal bases, and is the first to achieve constant regret at all times under *unknown* arrival rates. In addition, we show that when the arrival rates are known, the primal-dual policy achieves the optimal scaling as the lower-bound described in Kerimov et al. (2021a,b). Furthermore, we find that when the arrival rates are known, the primal-dual policy can significantly outperform alternative dynamic matching policies in numerical simulations.

Contents

1. Introduction

We consider a centralized dynamic matching market. There are finite types of agents. Agents arrive at the market over time and exit it once they are matched. Only certain sets of agent types can be feasibly matched and these matching constraints are described by a network, comprised of agent types (nodes) and possible match types (hyperedges). A match consists of one or several agent types and generates some heterogeneous reward when realized. The key distinguishing feature of the dynamic matching market we consider is that agents act effectively as both demand and supply (i.e., agents arrive to be matched, and the planner needs multiple agent types to match). This differs from the traditional online matching model (e.g., Mehta et al., 2013) in which the supply types are always available upfront while the demand types arrive sequentially.

There is a recent surge of interest in studying the dynamic matching markets under various modeling assumptions (Akbarpour et al., 2020; Aouad and Saritaç, 2020; Kerimov et al., 2021a,b; Gupta, 2021; Blanchet et al., 2022). Centralized dynamic matching markets naturally arise in various applications such as kidney exchange markets, online carpooling platforms, and online labor markets. In a kidney exchange market, patient-donor pairs arrive dynamically, where each pair is

viewed as an agent in our model. After a patient-donor pair arrives, the planner would attempt to match them with another pair or through a multi-way kidney exchange. The matching rewards depend on the biological compatibility among the patient-donor pairs. In a carpooling platform, riders arrive continuously over time. The planner matches riders to share a ride and the reward depends on their locations and destinations. In an online labor market, workers and jobs arrive dynamically, and the planner's task is to match the workers and the jobs, for which the reward depends on the needs of the jobs and the experience and skills of the workers. Note that in an online labor market, the matching network is bipartite, as workers can only be matched with jobs. On the contrary, in a kidney exchange market or an online carpooling platform, the matching network may not be bipartite and may contain hyperedges that allow multi-way matches.

Surprisingly, a recent line of work (Kerimov et al., 2021a,b; Gupta, 2021) reveals the immense power of knowing the basic feasible solutions of the fluid (linear programming) relaxation of the dynamic matching problem. After restricting to matches corresponding to the optimal basic matches and removing the "redundant" matches, various simple designs and greedy-like policies (including periodic resolving, greedy longest-queue, sum-of-squares policies) can achieve bounded regrets *at all times* compared to the best offline policy, under the so-called general position gap (GPG) assumption, i.e., the optimal basic matches are stable under small perturbations of the agent arrival rates. The implications of these breakthrough results are far-reaching, pointing to a new paradigm of policy design for dynamic matching in which the sophisticated dynamic matching problem can be reduced to a simpler queueing control problem after restricting matches to the optimal basic matches.

Despite the theoretical breakthroughs, the policies that restrict matches to the optimal basic matches have some practical drawbacks. First, the planner may not have the exact knowledge of the arrival rates and thus, cannot determine the optimal basic matches. Second, policies relying on optimal basic matches are problematic when multiple optimal solutions exist, as some agents may be unfairly penalized by the tie-breaking rules. These motivated us to propose a new class of primal-dual policy that achieves constant regret at all times for both known and *unknown* arrival rates. Our primal-dual policy avoids restricting matches to only the optimal basic matches thus does unfairly penalize specific agent types by the tie-breaking rules. Also, as the primal-dual policy uses more match types, in numerical simulations, in the case where arrival rates are known, it achieves a lower regret and has fewer waiting agents in the system than the policies of Kerimov et al. (2021b); Gupta (2021).

Our primal-dual policies introduce a Lagrangian multiplier (also can be interpreted as shadow price) for each agent type and pick the best match based on the reduced reward by subtracting

out the total shadow prices associated. The shadow price for a given agent type is dynamically adjusted and serves as a control signal for market thickness: it rises when fewer agents of the given type are waiting and thus decreases the matching rate. The primal-dual policy in the existing literature (Nazari and Stolyar, 2018) sets the shadow price to be proportional to the inventory (difference between the number of arrived agents and agents scheduled to be matched) and only achieves regret of $o(T)$ for a given time T. In contrast, to achieve $O(1)$ regret, our policies compute the shadow price using both the weighted inventories *and* the dual solution of the "sample-averaged approximated" fluid relaxation. In addition, our policies introduce for the inventories a time-varying weight that is useful in establishing constant regrets when the time horizon length is unknown. At a high level, the dual solution enables us to combine the technique of controlling the reward of scheduled matches from Nazari and Stolyar (2018), with the technique of controlling the inventories of agents from (Huang and Neely, 2009). The primal-dual policy of Nazari and Stolyar (2018) suggests that the scheduled matches grow closer to the fluid solution as the weight factor on inventory becomes small, but in their policy, a small weight factor leads to large inventories, which in turn leads to a large gap between the scheduled matches and the matches that can be actually realized. In our primal-dual policies, leveraging on the GPG assumption which turns out to be equivalent to the locally polyhedral condition of Huang and Neely (2009), we find that, the sample-average approximated dual solution provides enough negative drift to ensure that the inventories across agent types stay low, *regardless* of the magnitude of the weight on the inventories. This allows us, in turn, to select an appropriate time-varying weight to attain a constant regret at all times.

1.1. Related Literature

In addition to the aforementioned papers, there is a vast literature on stochastic and dynamic matching models. We next describe several streams of literature and discuss how they differ/relate to our model.

Online Matching and Network Revenue Management. This stream of literature studies a setting with agents arriving online and need to be immediately matched with the available offline resources. In these online models, researchers focus on either the competitive ratio or regret. There is extensive literature analyzing competitive ratio of online matching problems, starting with a seminal paper by Karp et al. (1990). In the interest of space, we refer interested readers to Mehta et al. (2013); Ma and Simchi-Levi (2020) for a more comprehensive review of competitive ratio analysis in online matching and network revenue management models.

Closer to our paper are the regret analysis for network revenue management (NRM) models. In (Talluri and Van Ryzin, 1998), the authors show that the bid-price policy achieves $O(\sqrt{T})$ regret in the (quantity-based) NRM model. The regret is improved by Jasin and Kumar (2012), who show that a re-solving policy achieves $O(1)$ regret, under the assumption that the "fluid relaxation" has a non-degenerate primal optimal solution. The constant regret analysis has since been generalized to online matching problems (Balseiro et al., 2021), and NRM models without distributional knowledge of the arrivals (Jasin, 2015; Chen et al., 2022). We note that the online matching and network revenue management may be viewed as dynamic matching models considered in this paper with less stringent constraints, as offline resources in these models can be allocated at any time during the time horizon, whereas the resources/agents in dynamic matching models can be only matched after they arrive. Indeed, recently, researchers have identified policies achieving constant regret for those dynamic models *without* the non-degeneracy of the fluid relaxation (see, e.g., Arlotto and Gurvich, 2019; Bumpensanti and Wang, 2020; Vera et al., 2021). In contrast, for the dynamic matching model considered in this model, without the non-degeneracy assumption, the lower-bound for the regret is $\Omega(\sqrt{T})$ (Kerimov et al., 2021a). We also note that the recent work of Gupta (2021) suggests that much of the theoretical regret for the sum-of-squares policy established in the dynamic matching models hold in online matching and NRM models with offline resources.

Bipartite (Matching) Queues. Similar to online bipartite matching problems, in the bipartite queueing literature, agents form two partitions, where each match consists of exactly one agent in each partition. The bipartite Queueing models differ from the online bipartite matching model as there are randomness on both sides of partitions. One example of bipartite queueing models is the parallel server system, where one partition of agents acting as customers and the other partition of agents acting as servers. The parallel server system has been extensively studied under different operational settings (see, e.g., Harrison, 1998; Mandelbaum and Stolyar, 2004; Gurvich and Whitt, 2010; Ward and Armony, 2013), and we refer interested readers to the recent work of Afeche et al. (2022) for additional coverages in this line of work. Motivated by ride-sharing platforms, recent papers also consider variations of the parallel server system, where vehicles queue to be matched and exit upon matching that mimics an open network (Özkan and Ward, 2020) or vehicles become busy/available at different locations in a process that mimics a closed network (Banerjee et al., 2018). Closer to our work are the systems with two-sided queueing systems, where all agents not matched stay in a queue. For two-sided queueing systems, researchers have studied matching policies with different objectives, such as minimizing the holding costs (Buỳić and Meyn,

2015), maximizing the match-dependent rewards Kerimov et al. (2021b), and maximizing the total rewards minus the holding costs (Aveklouris et al., 2021).

Dynamic Matching on Random Graphs. While the bipartite queueing literature typically assumes the matching configuration to be fixed, a different stream of literature considers agents arrive over time and form viable matches with existing agents probabilistically. For example, Anderson et al. (2017) and Ashlagi et al. (2019) consider the waiting time for the agents under asymptotic regimes, Akbarpour et al. (2020) consider the total number of matched agents with the additional feature of agent abandonment, while Kanoria (2022) considers the distance between matched pairs in a spatial matching model.

Dynamic Multi-Way Matching with Fixed Graphs. In this paper, we focus on reward maximization in dynamic multi-matching models with fixed matching configurations. Aouad and Saritaç (2020) study a two-way dynamic matching model with agent departures and propose a policy that achieves a constant percent of the upper-bound in steady state. For regret analysis, Nazari and Stolyar (2018) present a primal-dual policy that achieves a regret of $o(T)$. Kerimov et al. (2021a) presents a batching policy to improve the regret to $O(1)$ at all times, under the assumption that the fluid relaxation has an acyclic and non-degenerate primal solution. Gupta (2021) shows that the sum-of-squares policy achieves $O(1)$ regret under the same non-degeneracy condition, but without the acyclic assumption. In this paper, we also propose $O(1)$ regret policies, but our work differs from (Kerimov et al., 2021a; Gupta, 2021) in several significant ways. Both (Kerimov et al., 2021a; Gupta, 2021) requires the knowledge of arrival rates, which they use to find the fluid solution and restrict their policies to use only the match types in the fluid solution. In contrast, we propose primal-dual policies that do not require the knowledge of arrival rates. Moreover, in our policies, any match type m in $\mathcal{M}$ may be utilized depending on the state at the time, and as a result, leverages on the flexibility of additional matching possibilities and achieves superior empirical performances compared to the policies analyzed in (Kerimov et al., 2021a; Gupta, 2021).

2. Model Setup

The setup of our model is similar to (Kerimov et al., 2021a). There is a finite set of agent types $\mathcal{A} = [n] = \{1, 2, \ldots, n\}$, and a finite set of matches $\mathcal{M} = [d]$. We consider discrete time arrivals such that at each time $t \in \mathbb{N}_+$, exactly one agent arrives. Let $\lambda = \{\lambda_i\}_{i=1}^n \in \mathbb{R}_+^n$ and $A_i(t)$ denote the number of arrivals of type i at time t. The arrival of agent type i is with probability $\lambda_i \geq 0$, where $\sum_{i=1}^n \lambda_i = 1$. Once agents arrive, they will wait in the queue until matched by a central planner or decision maker. There is no initial customer in the queue waiting to be matched.

For each match $m \in \mathcal{M}$, let $\mathcal{A}(m)$ denote the set of agent types participating in the match m. Let $M \in \{0,1\}^{n \times d}$ denote the network matrix, where for any $i \in [n]$, let M_i denote its ith row such that for $m \in [d]$, $M_{im} = \mathbf{1}_{\{i \in \mathcal{A}(m)\}}$. Also, we define

$$B = \max_{m \in \mathcal{M}} |\mathcal{A}(m)| + 1\,. \tag{1}$$

At each period t, the central decision maker decides whether to realize one or multiple matches in $\mathcal{M}$, where match m can be realized only if type i agent is waiting for each $i \in \mathcal{A}(m)$. Once a match $m \in \mathcal{M}$ is realized, the agents participating in the match leave the system, and the match itself generates a reward r_m. We define $r_{\max} = \max_{m \in \mathcal{M}} r_m$ and make the following assumption throughout the paper about the rewards.

Assumption 1. *For each agent $i \in \mathcal{A}$, there exists a "self-match" for i, i.e., there is $m \in \mathcal{M}$ such that $\mathcal{A}(m) = \{i\}$. Let r_i be the unit reward of self-match agent i, and we have*

1. $r_i \geq 0$.
2. *For each $m \in \mathcal{M}$ where $|\mathcal{A}(m)| > 1$, we have $r_m > \sum_{i \in \mathcal{A}(m)} r_i$.*
3. *There exists at least one $m \in \mathcal{M}$ where $|\mathcal{A}(m)| > 1$.*

We remark that Assumption 1 effectively allows us to focus on all of the non-trivial dynamic matching instances that was studied by Kerimov et al. (2021a); Gupta (2021). Specifically, the first bullet point in Assumption 1 restricts to dynamic matching instances where the decision maker may "discard" the agents without penalty, and this is exactly the set of matching instances studied by Kerimov et al. (2021a); Gupta (2021). The second bullet point removes the match types that are always sub-optimal, and finally, the third bullet point in Assumption 1 eliminates the trivial instance where only self-matches are allowed.

2.1. Performance Measure

Like Kerimov et al. (2021a), we measure the performance of a dynamic matching policy at any time period during the entire time horizon $[0, T]$. Specifically, at time period t, the expected regret of a policy π at time t is measured by

$$\mathbb{E}\left[\mathsf{R}^{*,t} - \mathsf{R}^{\pi,t}\right],$$

where $\mathsf{R}^{*,t}$ and $\mathsf{R}^{\pi,t}$ represent the rewards under the hindsight optimal policy (that optimizes the reward up to time t) and policy π, respectively.

Note that for a fixed t, the hindsight optimal reward can be achieved by a simple policy that waits for all agents to arrive during the first t periods, then solves an optimization problem to determine

the set of matches that maximizes the overall rewards.[1] Thus, a more meaningful performance measure for policy is its regret *at all times*. Formally, for any policy π, its regret at all times is measured by

$$\sup_{0\le t\le T} \mathbb{E}\left[\mathsf{R}^{*,t} - \mathsf{R}^{\pi,t}\right].$$

Throughout the paper, we say that a policy achieves constant regret (at all times) if the above quantity is independent of T. As we shall see, our constant regret (at all times) policy will also keep the expected queue lengths for agents waiting in the system to be small at any time.

3. Main Results

3.1. Fluid Relaxation and General Position Gap

To design an effective policy and analyze its regret at all times, we use a standard fluid relaxation of the dynamic matching problem. Under the fluid relaxation, we have the following deterministic optimization problem:

$$\begin{bmatrix} \max_x & r^\top x \\ \text{s.t.} & Mx = \lambda \\ & x \in \mathcal{X} \end{bmatrix}, \tag{2}$$

where $\mathcal{X} = \{x \,|\, x \in \mathbb{R}^d_{\ge 0}, \sum_{m\in\mathcal{M}} x_m \le 1\}$. Note that the constraint $\sum_{m\in\mathcal{M}} x_m \le 1$ in $\mathcal{X}$ is non-binding and hence redundant, as $\sum_{i\in[n]} \lambda_i = 1$, and Assumption 1 implies that any optimal solution of (2) must contain non-self-matches. Nevertheless, the constraint is useful because it ensures the existence of a bounded optimal solution when we relax $Mx = \lambda$ with Lagrangian multipliers.

Assuming that x^* is an optimal solution of (2), the value $t \cdot r^\top x^*$ forms a natural upper-bound for $\mathbb{E}\left[\mathsf{R}^{*,t}\right]$, the expected rewards for the hindsight optimal policy. To see this, for a fixed t, let $y(\omega)$ denote the matches realized by the offline policy under a random scenario ω, and by Assumption 1, we can assume that $y(\omega)$ match all arrivals up to time t. Thus, we have that $M\mathbb{E}[y(\omega)] = t\lambda$, and by optimality of x^*, we have

$$t \cdot r^\top x^* \ge r^\top \mathbb{E}\left[y(\omega)\right] = \mathbb{E}\left[\mathsf{R}^{*,t}\right]. \tag{3}$$

Finally, this implies that the regret for policy π over the entire time horizon $[0,T]$ is no larger than

$$\sup_{0\le t\le T} \mathbb{E}\left[t \cdot r^\top x^* - \mathsf{R}^{\pi,t}\right]. \tag{4}$$

Next, we define the Lagrangian relaxation of (2) as

$$L_\lambda(U) = \max_{x\in\mathcal{X}} \left\{r^\top x - U^\top (Mx - \lambda)\right\} = \max_{x\in\mathcal{X}} \left(r^\top - U^\top M\right) x + U^\top \lambda. \tag{5}$$

[1] Although the optimization problem is NP-hard, for large t, we can solve a fractional matching problem then round into an integer solution with a bounded loss of reward.

When λ is fixed, we let $L(U) = L_\lambda(U)$ and the corresponding (Lagrangian) dual problem is formulated as

$$\min_{U \in \mathbb{R}^n} \quad L(U)\,. \tag{6}$$

Throughout the paper, we define the notion of *general position gap* (GPG), which measures how much λ may change (in terms of the ℓ_2 norm) without changing the unique optimal solution for the dual formulation.

Definition 1 (General Position Gap)**.** Let $\mathcal{B}_\epsilon(\lambda)$ be the ball centered at λ with radius ϵ under the ℓ_2 norm, i.e., $\mathcal{B}_\epsilon(\lambda) = \{\widehat{\lambda} : \|\widehat{\lambda} - \lambda\|_2 \le \epsilon\}$. We say that λ has a GPG of (at least) ϵ, if there exists U^* that is the unique optimal solution of (6) for any $\widehat{\lambda} \in \mathcal{B}_\epsilon(\lambda)$. If there are multiple optimal solutions of (6), then we say that λ has a GPG of zero.

We remark that a similar notion of GPG has been considered by Gupta (2021) and Kerimov et al. (2021a). In Gupta (2021), GPG is defined using the total variation ball $\mathcal{B}_{\epsilon,\mathrm{TV}}(\lambda) = \{\widehat{\lambda} \in \Delta^n : \|\widehat{\lambda} - \lambda\|_1 \le \epsilon\}$, where Δ^n denote the $(n-1)$-dimensional probability simplex. In Kerimov et al. (2021a), GPG is defined as the smallest value among all variables in a basic optimal solution of the fluid relaxation (2). While the aforementioned definitions of GPG differ slightly, all are describing the stability of the optimal solution for the dual problem subject to change in λ. Indeed, in either Kerimov et al. (2021a), Gupta (2021) or this paper, a positive GPG is equivalent to the uniqueness of the dual solution.

A key motivation for our particular definition of GPG is that having a GPG of ϵ (under Definition 1) is equivalent to $L_\lambda(U) - L_\lambda(U^*) \ge \epsilon \|U - U^*\|_2$ for all $U \in \mathbb{R}^n$, which is known as the locally polyhedral condition (Huang and Neely, 2009). The next proposition formally establishes this equivalence, and furthermore, shows that a similar equivalence holds for GPG defined in ℓ_p norm for any $p \ge 1$.

Proposition 1. *Let $\|\cdot\|$ and $\|\cdot\|_*$ denote any pair of norm and the dual norm. The following two conditions are equivalent for any fixed $\epsilon > 0$:*

1. *U^* is an optimal solution to $\min_U L_{\widehat{\lambda}}(U)$ for all $\left\|\widehat{\lambda} - \lambda\right\| \le \epsilon$;*
2. *$L_\lambda(U) - L_\lambda(U^*) \ge \epsilon \|U - U^*\|_*$ for all $U \in \mathbb{R}^n$.*

The proof of Proposition 1 is in Section C.1 in Appendix C. At a high level, the locally polyhedral condition is key for us to design a primal-dual policy that keeps the dual variables close to U^*, which turns out to be crucial in our regret analysis.

By Hölder's inequality, the difference between the ℓ_1 and ℓ_p for any $p \ge 1$ is no greater than a factor of n. Thus, most of our analyses apply to GPG defined in ℓ_p norm for any $p \ge 1$. This further

illustrates the connection between our GPG definition and the definition in Gupta (2021), which uses the ℓ_1 norm. Finally, the GPG defined in Kerimov et al. (2021a) can be used to establish a lower-bound on our GPG, we refer interested readers to Proposition 1 in Appendix A for further discussions.

3.2. The Primal-Dual Policy

We describe our primal-dual policy, which is summarized as Algorithm 1. At a high level, our primal-dual policy is divided into two components, a scheduling policy that schedule matches based on a Lagrangian relaxation $\max_{x\in\mathcal{X}} (r^\top - U^\top M)\, x$ for some dual estimates U, and a realization policy that realize the scheduled matches when there are sufficient agents. We note that our primal-dual policy bears similarities to the policy studied in Nazari and Stolyar (2018). However, their policy only achieves a regret of $o(T)$, for any $\kappa > 0$, while we identify primal-dual policy that achieves regret independent of T under a wide range of settings.

Next, we provide the intuitions behind our policy by explaining our primal-dual policy during period t. Recall that $A_i(t)$ denotes the number of arrivals of type i at time t. Let $x(t) \in \mathbb{R}^d$ and $y(t) \in \mathbb{R}^d$ and denote the vector of scheduled and realized match at time t, respectively, with $x(0)$ initialized as the zero vector. Let vector $\delta(t) \in \mathbb{R}^n$ denote the difference between the total number of arrivals (of each type) minus the total number of agents (of each type) that are needed for the scheduled matches at time t, with $\delta(0)$ initialized as the zero vector. For any $t \in \mathbb{N}_+$, we update

$$\delta(t) = \delta(t-1) + Mx(t-1) - A(t)\,. \tag{7}$$

It is important to note that $\delta_i(t)$ can be either positive or negative. We will refer to $-\delta(t)$ as the (virtual) *inventory* of agents at period t, as $-\delta_i(t)$ represents the difference between the number of agents of type i that have arrived by time t and the total number of agents of type i required for all matches scheduled before time t.

Let $\widehat{\lambda}(t)$ denote the empirical arrival rate based on the arrivals from periods 1 to t. We solve the linear program

$$\min_{U\in\mathbb{R}^n} U^\top \widehat{\lambda}(t), \text{ s.t. } \sum_{i\in\mathcal{A}(m)} U_i \geq r_m,\ \forall m, \tag{8}$$

and take its optimal solution as $\widehat{U}(t)$. Intuitively, $\widehat{U}(t)$, is an optimal solution of the dual problem (6) when λ is replaced by $\widehat{\lambda}(t)$ (see Appendix B for a more formal discussion). Next, we update $U(t)$, the dual estimates (i.e., Lagrangian multipliers) during period $t \in \mathbb{N}_+$, using $\delta(t)$, the negative inventory of agents, and $\widehat{U}(t)$. Formally, we update $U(t)$ as

$$U(t) = \widehat{U}(t) + \frac{\delta(t)}{V_t}\,, \tag{9}$$

where V_t is some policy (time-varying) parameter that is determined at time zero.

With $U(t)$ updated, the scheduling policy determines vector $x(t)$, with the following simple procedure: if $r - U(t)^\top M \leq 0$, then we set $x(t)$ to be the zero vector; and otherwise, we set $x(t)$ to be the m^*-th standard basis vector (i.e., $x_{m^*}(t) = e_{m^*}$), where m^* is the index of one of the largest entries in $r - U(t)^\top M$ with tie breaks arbitrarily. Note that $x(t)$ corresponds to the optimal solution of the Lagrangian relaxation with multiplier $U(t)$, that is

$$x(t) \in \operatorname*{arg\,max}_{x \in \mathcal{X}} \left(r^\top - U(t)^\top M\right) x + U(t)^\top \lambda = \operatorname*{arg\,max}_{x \in \mathcal{X}} \left(r^\top - U(t)^\top M\right) x. \tag{10}$$

After scheduling matches corresponding to $x(t)$, we next describe the realization policy, which determines $y(t)$, the vector representing the matches realized at time t. We remark that the main focus of our paper is on the scheduling policy, as any myopic realization policy that goes through all scheduled unrealized matches in arbitrary order will suffice in our analysis. For concreteness, we describe one such myopic realization policy.

For each m, let the number of scheduled match m that are not yet realized, to be W_m, where $W_m = \sum_{s=1}^{t} (x_m(s) - y_m(s))$. If $W_m > 0$ and m can be realized with the agents waiting in the system, i.e., $\sum_{s=1}^{t} (A_i(s) - M_i y(s)) > 0$ for any $i \in \mathcal{A}(m)$, then the policy realizes all the matches of type m until the waiting agents for one of the types in $\mathcal{A}(m)$ becomes zero. Once the realization policy goes over each match type m, our policy finishes period t and moves to period $t+1$.

Our primal-dual policy is summarized as Algorithm 1. Some readers may have noted that if $\widehat{U}(t)$ is replaced with an arbitrarily fixed value across t, the primal-dual policy is roughly a stochastic sub-gradient method for approximately solving Lagrangian dual problem (6). More precisely, if we set $\widehat{U}(t) \equiv Z$ instead of the optimal solution of (8), and $V_t \equiv V$; the variables $U(t)$ would be updated according to $U(t) = U(t-1) + \frac{1}{V}(Mx(t-1) - A(t))$ with initialization $U(1) = Z$. This is a stochastic subgradient method with fixed step size $\frac{1}{V}$ for solving the Lagrangian dual problem (6), as $Mx(t-1) - A(t)$ is a stochastic subgradient of $L(U(t-1))$. However, such a policy does not yield constant regret (unless $Z = U^*$), as both $\delta(t)$ (negative inventory) and the actual expected number of agents waiting in queues grows with T. Indeed, one of the key intuitions is that if $\widehat{U}(t)$ is selected to (almost) equal to U^*, then we also have $U(t) \approx U^*$, which facilitates our analysis of δ (negative inventory) and the actual expected number of agents waiting in queues can be bounded effectively.

Next, we present our main results on the regret of the primal-dual policies that fall under the framework of Algorithm 1.

Algorithm 1 Dynamic match scheduling.

Input: $\{V_t\}_{t=1}^{\infty}$, $\delta(0)=\mathbf{0}$ and $x(0)=\mathbf{0}$
for each $t=1,2,\cdots T$ **do**
  Observe arrival $A(t)$
  The scheduling process:
  Update $\delta(t) = \delta(t-1)+Mx(t-1)-A(t)$
  Let $\widehat{U}(t)$ be an optimal solution of (8)
  Update $U(t)=\widehat{U}(t)+\frac{\delta(t)}{V_t}$
  if $r-U(t)^{\top}M\le 0$ **then**
    Set $x(t)$ to the zero vector
  else
    Set $x(t)$ to m^*-th standard basis vector, where m^* is the index of the maximum value in vector $r-U(t)^{\top}M$
  end if
  The realization process:
  Initialize $y(t)=\mathbf{0}$ and let $W_m=\sum_{s=1}^{t}(x_m(s)-y_m(s))$.
  for each $m\in\mathcal{M}$ with $W_m>0$ **do**
    Let $I_m=\min_{i\in\mathcal{A}(m)}\sum_{s=1}^{t}(A_i(s)-M_i y(s))$
    Update $y_m(t)=y_m(t)+\min\{W_m,I_m\}$
  end for
end for

Theorem 1. *Consider the primal-dual policy π described in Algorithm 1, where $\{V_t\}_{t=1}^{T}$ is monotonically increasing, and $\sum_{t=1}^{T}1/V_t<\infty$. Suppose that λ has a GPG of ϵ, then we have*

$$\sup_{0\le t\le T}\mathbb{E}\left[\mathsf{R}^{*,t}-\mathsf{R}^{\pi,t}\right]\le O\left(B+\frac{\sqrt{n}Br_{\max}}{\epsilon^2}\right), \tag{11}$$

where we recall that $B=\max_{m\in\mathcal{M}}|A(m)|+1$, and $r_{\max}=\max_{m\in\mathcal{M}}r_m$.

Theorem 1 suggests that the primal-dual policy achieves regret independent of T, thus providing the first constant regret dynamic matching policy to the setting with unknown arrival rates. If we fix the network matrix M and reward vector r, then the regret of our policy scales on the order of $1/\epsilon^2$ when ϵ becomes small.

We also remark that there are plenty of choices for $\{V_t\}_{t=1}^{\infty}$ to ensure that Algorithm 1 achieves constant regret. One natural choice is to have $V_t=O(T)$, but this requires some knowledge of

the time horizon. Alternatively, we can choose $V_t = t^2$, then Algorithm 1 achieves constant regret without any knowledge of λ *and* T. Next, we show that the regret can be improved (in terms of $1/\epsilon$), with the knowledge of the arrival rates.

Corollary 1. *Suppose that λ is known in advance. Let U^* be the optimal solution of the Lagrangian dual problem (6), and replace $\widehat{U}(t)$ with U^* for all t in Algorithm 1. Consider the corresponding primal-dual policy π, with $\{V_t\}_{t=1}^{\infty}$ is monotonically increasing, and $\sum_{t=1}^{\infty} 1/V_t < \infty$. Suppose that λ has a GPG of ϵ, then we have*

$$\sup_{0\le t\le T} \mathbb{E}\left[\mathsf{R}^{*,t} - \mathsf{R}^{\pi,t}\right] \le O\left(B + \frac{\sqrt{n}Br_{\max}}{\epsilon}\right). \tag{12}$$

Corollary 1 demonstrates that our policy matches the best regret scaling in terms of ϵ (Kerimov et al., 2021a; Gupta, 2021), when λ is known. We remark that the $O(1/\epsilon)$ scaling is in fact the best possible, as illustrated in an example of (Kerimov et al., 2021a, Fig. 5) with fixed M and r. Furthermore, (Kerimov et al., 2021a, Example 3.1) shows that when λ has a GPG of 0, which is equivalent to (5) having multiple optimal solutions, the regret at all times is at least $\sqrt{T}$. Next, we show that our primal-dual policy, under the appropriate V_t, achieves the regret of $\sqrt{T}$ without the guarantee that λ has a GPG of ϵ.

Corollary 2. *Suppose that λ is known in advance but does not necessarily have a positive GPG. Let U^* be an optimal solution of the Lagrangian dual problem (6), and replace $\widehat{U}(t) = U^*$ for all t in Algorithm 1. Consider the corresponding primal-dual policy π, with $V_t = \sqrt{t}$ then we have*

$$\sup_{0\le t\le T} \mathbb{E}\left[\mathsf{R}^{*,t} - \mathsf{R}^{\pi,t}\right] \le O\left(\sqrt{T}\right). \tag{13}$$

The proof of Theorem 1 and Corollary 1 are given in Section 4, while the proof of Corollary 2 is deferred to Section C.2 in Appendix C. As a by-product of our proofs, we also obtain a bound on the expected number of waiting agents in the system, which has the same magnitude as the regret. This will be further discussed in Section 4.5.

3.3. Alternative Interpretations of Our Primal-Dual Policy

Our new primal-dual policy also has an interesting interpretation under the framework of the drift-plus-penalty method, which is widely for stabilizing a queueing network while minimizing the time average of a network penalty function (See Neely (2010) and the references therein for detailed discussions). Specifically, given a Lyapunov function V which is often defined as the sum of squares of the queue sizes, let $\Delta(t)$ denote the (conditional) Lyapunov drift, that is the expected change of V from time slot t to $t+1$. Then at each time slot t, we take a control action to greedily minimize $\Delta(t) + V \times P(t)$, where $P(t)$ is a given network penalty function, and V is a non-negative weight

parameter, allowing for a tradeoff between the reduction of the queue sizes and minimization of the penalty function.

In our setting, since we want to reduce the inventory of agents, it is natural to define Lyapunov function as the sum of squares of $\delta_i(t)$, that is, $\mathsf{V}(t) = \frac{1}{2}\|\delta(t)\|_2^2$. Using (7), we can readily upper-bound the Lyapunov drift $\Delta(t)$ by $B + \langle \delta(t), Mx(t) - \lambda \rangle$. Further, we aim to maximize the cumulative reward, so we take the negative of the reward as the penalty function. However, here instead of using the reward function $\langle r, x \rangle$ *per se*, we let $P(t) = -L(\widehat{U}(t), x(t))$, where $L(U, x) = \langle r, x \rangle - \langle U, Mx - \lambda \rangle$ is the Lagrangian function and $\widehat{U}(t)$ is the plug-in estimator of the optimal dual variable U^*. Finally, we use a time-varying weight parameter V_t to attain a smooth tradeoff between reducing inventories and maximizing the rewards. Combining all these ingredients together, we choose a scheduling action $x(t)$ to minimize the following drift-plus-penalty bound:

$$x(t) \in \arg\min_{x \in \mathcal{X} \cap \mathbb{N}^d} \langle \delta(t), Mx - \lambda \rangle - V_t \times L\left(\widehat{U}(t), x\right). \tag{14}$$

In view of (9), the above objective function is equal to $-V_t \times L\left(U(t), x\right)$. Therefore, the minimization problem (14) is equivalent to the maximization of $L\left(U(t), x\right)$, which exactly coincides with our primal-dual policy as per (10).

Note that as $t \to \infty$, $\widehat{U}(t)$ will coincide with U^* with high probability and V_t is chosen to diverge. Therefore, as $t \to \infty$, our primal-dual policy rephrased in terms of (14) reduces to the following "restricted" max-weight policy. First, it restricts to the set of matches m achieving the highest reduced reward $r_m - \sum_{i \in A(m)} U_i^*$.[2] Then it schedules a match m in the set with the maximum weight – the largest total inventory $\sum_{i \in A(m)}(-\delta_i(t))$. We remark that this "restricted" max-weight policy bears some similarity to the "restricted" longest-queue policy proposed in Kerimov et al. (2021a), which restricts to the set of matches corresponding to the optimal basic variables and picks a match that contains the longest queue – the agent type with the largest inventory $\max_i(-\delta_i(t))$. Note that in Kerimov et al. (2021a), by focusing on the optimal basic variables, the restricted set of matches can be much smaller and hence the scheduling policy is less flexible. Moreover, the longest-queue policy is often less effective in reducing the total inventories than the max-weight policy (Dimakis and Walrand, 2006).

4. Analysis of Our Primal-Dual Policy

Throughout the section, we consider primal-dual policy π where $\widehat{U}(t)$ and V_t are selected according to Theorem 1. We define $\{\mathcal{F}_t\}$ as the natural filtration associated with the agent arrival process

[2] In fact, it can be shown that the highest reduced reward is at most 0, i.e., $r - M^\top U^* \leq 0$.

up to and including period t. We first present a high level roadmap of our regret analysis. Recall that the regret at time t is no larger than $\mathbb{E}\left[t\left\langle r, x^*\right\rangle - \mathsf{R}^{\pi,t}\right]$. We decompose the quantity as

$$\mathbb{E}\left[t\left\langle r, x^*\right\rangle - \mathsf{R}^{\pi,t}\right] = \underbrace{t\left\langle r, x^*\right\rangle - \sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, x(s)\right\rangle\right]}_{\text{(I)}} + \underbrace{\sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, x(s)\right\rangle - \left\langle r, y(s)\right\rangle\right]}_{\text{(II)}}, \tag{15}$$

and bound (I) and (II) separately. The upper-bound for (I) is based on the celebrated drift-plus-penalty technique in stochastic network optimization by Neely (2010), whereas the upper-bound for (II) is based on the structure of the matching model and the property of the realization policy. Both upper-bounds depend on $\delta(t)$, the inventory of agents, which we analyze through Lyapunov analysis that relies crucially on the condition that λ has a GPG of ϵ.

Before proceeding to the roadmap, we first derive a pair of technical lemmas that will be useful for the analysis. The first lemma bounds the maximum change in the inventory of agents in one period.

Lemma 1. *For any $t \geq 1$, it holds that*

$$\left\|Mx(t) - A(t+1)\right\|_2^2 \leq \left\|Mx(t) - A(t+1)\right\|_1 \leq \max_{m\in[d]}\left|\mathcal{A}(m)\right| + 1 \triangleq B.$$

Proof. Note that

$$\left\|Mx(t)\right\|_1 = \sum_{i=1}^{n}\sum_{m=1}^{d} M_{im}x_m(t) = \sum_{m=1}^{d}\left|\mathcal{A}(m)\right|x_m(t) \leq \max_{m\in[d]}\left|\mathcal{A}(m)\right|\sum_{m=1}^{d}x_m(t) \leq \max_{m\in[d]}\left|\mathcal{A}(m)\right|.$$

Therefore $\left\|Mx(t) - A(t+1)\right\|_1 \leq \max_{m\in[d]}\left|\mathcal{A}(m)\right| + 1 = B$. Moreover, since $\left\|Mx(t) - A(t+1)\right\|_\infty \leq 1$, it follows that $\left\|Mx(t) - A(t+1)\right\|_2^2 \leq \left\|Mx(t) - A(t+1)\right\|_1\left\|Mx(t) - A(t+1)\right\|_\infty \leq B$. Q.E.D.

Next, we derive a lemma that upper-bounds the "violation" of complementary slackness between dual variables $U(t)$ and primal variables $x(t)$ in terms of the instantaneous reward difference.

Lemma 2. *For any $t \geq 1$, it holds that*

$$\left\langle U(t), Mx(t) - \lambda\right\rangle \leq \left\langle r, x(t) - x^*\right\rangle.$$

Proof. By the optimality of $x(t)$ given in (10),

$$\left\langle r - M^\top U(t), x(t)\right\rangle \geq \left\langle r - M^\top U(t), x^*\right\rangle.$$

By rearranging the terms, we deduce that

$$\left\langle r, x(t) - x^*\right\rangle \geq \left\langle U(t), Mx(t) - Mx^*\right\rangle = \left\langle U(t), Mx(t) - \lambda\right\rangle,$$

where the last equality holds by the constraint that $Mx^* = \lambda$. Q.E.D.

4.1. Rewards of Scheduled Matches

In this subsection, we introduce a bound on the first term of (15). Specifically, we present Proposition 2 that bounds $t\langle r, x^*\rangle - \sum_{s=1}^{t}\mathbb{E}[\langle r, x(s)\rangle]$, the difference between the expected rewards of scheduled matches and the fluid relaxation up to period t. The analysis is based on the celebrated drift-plus-penalty technique in stochastic network optimization by Neely (2010), with two additional significant innovations. First, we allow for time-varying V_t, provided that it is monotonically increasing. Second, we incorporate the estimation error of dual estimator $\widehat{U}(t)$ into the analysis. These are critical for establishing a constant regret with an unknown time horizon T and arrival rate λ.

Proposition 2. *Suppose V_t is monotonically increasing. Under the primal-dual policy π, we have*

$$t\langle r, x^*\rangle - \sum_{s=1}^{t}\mathbb{E}[\langle r, x(s)\rangle] \le \sum_{s=1}^{t}\frac{B+1}{2V_s} + \sqrt{n}r_{\max}\mathbb{E}[\|\delta(t)\|_2] + 2Br_{\max}\sum_{s=1}^{t}\mathbb{P}\left\{\widehat{U}(s)\neq U^*\right\}, \tag{16}$$

where $r_{\max}$ is the maximum reward that one match can generate.

Proof. Define the potential function

$$\mathsf{V}(s) = \frac{1}{2}\|\delta(s)\|_2^2. \tag{17}$$

Our proof proceeds by bounding the expected drift of $\mathsf{V}(s)$ in terms of the reward difference $\mathbb{E}[\langle r, x(s)\rangle] - \langle r, x^*\rangle$. Then by taking a certain form of a telescoping sum of the expected drifts across all times and using the non-negativity of the potential function $\mathsf{V}(s)$, we arrive at a bound to the cumulative reward difference.

Recall that $\delta(s+1) = \delta(s) + Mx(s) - A(s+1)$. Therefore, for $s \ge 1$,

$$\begin{aligned}\|\delta(s+1)\|_2^2 - \|\delta(s)\|_2^2 &= \|Mx(s) - A(s+1)\|_2^2 + 2\langle \delta(s), Mx(s) - A(s+1)\rangle \\ &\le B + 2\langle \delta(s), Mx(s) - A(s+1)\rangle \\ &= B + 2V_s\left\langle U(s) - \widehat{U}(s), Mx(s) - A(s+1)\right\rangle,\end{aligned} \tag{18}$$

where the first inequality follows by Lemma 1 and the last equality holds due to $\delta(s) = V_s(U(s) - \widehat{U}(s))$.

Then, we obtain from (18) that

$$\mathbb{E}[\mathsf{V}(s+1) - \mathsf{V}(s) \mid \mathcal{F}_s] \le \frac{1}{2}B + V_s\,\mathbb{E}\left[\left\langle U(s) - \widehat{U}(s), Mx(s) - A(s+1)\right\rangle \mid \mathcal{F}_s\right],$$

where $\mathcal{F}_s$ is the natural filtration associated with the agent arrival process up to and including period s. Since $\mathbb{E}[A(s+1) \mid \mathcal{F}_s] = \lambda$, it follows from Lemma 2 that

$$\mathbb{E}[\langle U(s), Mx(s) - A(s+1)\rangle \mid \mathcal{F}_s] = \langle U(s), Mx(s) - \lambda\rangle \le \langle r, x(s) - x^*\rangle.$$

Combining the last two displayed equations yields that

$$\mathbb{E}\left[\mathsf{V}(s+1)-\mathsf{V}(s)\mid\mathcal{F}_s\right]\le\frac{1}{2}B+V_s\left\langle r,x(s)-x^*\right\rangle-V_s\,\mathbb{E}\left[\left\langle\widehat{U}(s),Mx(s)-A(s+1)\right\rangle\mid\mathcal{F}_s\right].$$

Taking the expectation over $\mathcal{F}_s$ and dividing V_s over both hand sides, we get that

$$\frac{1}{V_s}\mathbb{E}\left[\mathsf{V}(s+1)-\mathsf{V}(s)\right]\le\frac{B}{2V_s}+\mathbb{E}\left[\langle r,x(s)\rangle\right]-\langle r,x^*\rangle-\mathbb{E}\left[\left\langle\widehat{U}(s),Mx(s)-A(s+1)\right\rangle\right]. \tag{19}$$

Now, by summing over $1\le s\le t$ on (19), we get

$$\begin{aligned}\sum_{s=1}^{t}\mathbb{E}\left[\frac{1}{V_s}(\mathsf{V}(s+1)-\mathsf{V}(s))\right]\le&\sum_{s=1}^{t}\frac{B}{2V_s}+\sum_{s=1}^{t}\mathbb{E}\left[\langle r,x(s)\rangle\right]-t\left\langle r,x^*\right\rangle\\&-\sum_{s=1}^{t}\mathbb{E}\left[\left\langle\widehat{U}(s),Mx(s)-A(s+1)\right\rangle\right].\end{aligned}$$

Moreover, we can bound the left hand side from below as

$$\sum_{s=1}^{t}\frac{1}{V_s}(\mathsf{V}(s+1)-\mathsf{V}(s))=\frac{1}{V_t}\mathsf{V}(s+1)+\sum_{s=2}^{t}\left(\frac{1}{V_{s-1}}-\frac{1}{V_s}\right)\mathsf{V}(s)-\frac{1}{V_1}\mathsf{V}(1)\ge-\frac{1}{2V_1},$$

where the last inequality holds because V_s is monotonically increasing and $\mathsf{V}(1)=\frac{1}{2}\left\|\delta(1)\right\|_2^2=\frac{1}{2}$. Combining the last two displayed equations and re-arranging the terms yields that

$$t\left\langle r,x^*\right\rangle-\sum_{s=1}^{t}\mathbb{E}\left[\langle r,x(s)\rangle\right]\le\sum_{s=1}^{t}\frac{B}{2V_t}+\frac{1}{2V_1}-\sum_{s=1}^{t}\mathbb{E}\left[\left\langle\widehat{U}(s),Mx(s)-A(s+1)\right\rangle\right]. \tag{20}$$

It remains to bound the last term in the RHS of (20). Note that

$$\sum_{s=1}^{t}\left\langle\widehat{U}(s),Mx(s)-A(s+1)\right\rangle=\sum_{s=1}^{t}\left\langle U^*,Mx(s)-A(s+1)\right\rangle+\sum_{s=1}^{t}\left\langle\widehat{U}(s)-U^*,Mx(s)-A(s+1)\right\rangle.$$

By the definition of $\delta(t)$,

$$\sum_{s=1}^{t}\left\langle U^*,Mx(s)-A(s+1)\right\rangle=\left\langle U^*,\delta(t)\right\rangle\ge-\left\|U^*\right\|_2\left\|\delta(t)\right\|_2\ge-\sqrt{n}r_{\max}\left\|\delta(t)\right\|_2,$$

where the inequalities hold by the Cauchy-Schwartz inequality and $\left\|U^*\right\|_2\le\sqrt{n}r_{\max}$ in view of Lemma 9. Furthermore,

$$\begin{aligned}\sum_{s=1}^{t}\left\langle\widehat{U}(s)-U^*,Mx(s)-A(s+1)\right\rangle&=\sum_{s=1}^{t}\left\langle\widehat{U}(s)-U^*,Mx(s)-A(s+1)\right\rangle\mathbf{1}_{\left\{\widehat{U}(s)\neq U^*\right\}}\\&\ge-B\sum_{s=1}^{t}\left\|\widehat{U}(s)-U^*\right\|_\infty\mathbf{1}_{\left\{\widehat{U}(s)\neq U^*\right\}}\\&\ge-2Br_{\max}\sum_{s=1}^{t}\mathbf{1}_{\left\{\widehat{U}(s)\neq U^*\right\}},\end{aligned}$$

where the inequalities hold by the Cauchy-Schwartz inequality, $\|Mx(s)-A(s+1)\|_1 \le B$ in view of Lemma 1, and $\left\|\widehat{U}(s)-U^*\right\|_\infty \le \left\|\widehat{U}(s)\right\|_\infty + \|U^*\|_\infty \le 2r_{\max}$ in view of Lemma 9. Combining the last three displayed equations, we get that

$$\sum_{s=1}^{t}\left\langle \widehat{U}(s), Mx(s)-A(s+1)\right\rangle \ge -\sqrt{n}r_{\max}\|\delta(t)\|_2 - 2Br_{\max}\sum_{s=1}^{t}\mathbf{1}_{\{\widehat{U}(s)\neq U^*\}}.$$

Finally, taking expectations over both hand sides of the last displayed equation and substituting it into (20) yields the desired conclusion (16). Q.E.D.

4.2. Realized Matches and Queue Lengths

Next, we establish a bound on the second term of (15). Recall that $y(t)$ denotes the vector of the realized match at time t. By the updating rule of $y(t)$, we have

$$\sum_{s=1}^{t} y(s) \le \sum_{s=1}^{t} x(s), \tag{21}$$

$$\sum_{s=1}^{t} My(s) \le \sum_{s=1}^{t} A(s), \tag{22}$$

where the first inequality holds because our policy cannot realize any match that is not scheduled, and the second inequality holds because the policy cannot match agents that have not yet arrived.

Define $\mathcal{M}_i = \{m : i \in \mathcal{A}(m)\}$. First, we present a lemma showing that for any m in which inequality (21) is strict, then we must have some agent type i such that $m \in \mathcal{M}_i$, and all agents of type i are realized by matches.

Lemma 3. *Fix any $m \in \mathcal{M}$ such that $\sum_{s=1}^{t} y_m(s) < \sum_{s=1}^{t} x_m(s)$. Then there must exist some $i \in \mathcal{A}(m)$ such that $\sum_{s=1}^{t} A_i(s) = \sum_{s=1}^{t} M_i y(s) = \sum_{s=1}^{t}\sum_{m\in\mathcal{M}_i} y_m(s)$, and moreover*

$$\sum_{s=1}^{t-1}\sum_{m\in\mathcal{M}_i} x_m(s) - \sum_{s=1}^{t}\sum_{m\in\mathcal{M}_i} y_m(s) = \delta_i(t). \tag{23}$$

Proof. According to our realization policy, we must have either $\sum_{s=1}^{t} y_m(s) = \sum_{s=1}^{t} x_m(s)$ or $\min_{i\in\mathcal{A}(m)} \sum_{s=1}^{t} A_i(s) - \sum_{s=1}^{t}\sum_{m\in\mathcal{M}_i} y_m(s) = 0$. Thus, by assumption, there must exist some $i \in \mathcal{A}(m)$ such that $\sum_{s=1}^{t} A_i(s) = \sum_{s=1}^{t}\sum_{m\in\mathcal{M}_i} y_m(s)$. Further, it follows that

$$\sum_{s=1}^{t-1}\sum_{m\in\mathcal{M}_i} x_m(s) - \sum_{s=1}^{t}\sum_{m\in\mathcal{M}_i} y_m(s) = \sum_{s=1}^{t-1}\sum_{m\in\mathcal{M}_i} x_m(s) - \sum_{s=1}^{t} A_i(s) = \delta_i(t),$$

where the last equality holds by the definition of $\delta(t)$ in (7). Q.E.D.

Next, we apply Lemma 3 to bound the difference between the total number of scheduled and realized matches, and subsequently, the difference between the virtual reward and the actual reward in terms of $\delta(t)$.

Proposition 3. *For any $t \geq 1$,*

$$\sum_{s=1}^{t} \langle r, x(s) \rangle - \sum_{s=1}^{t} \langle r, y(s) \rangle \leq r_{\max} \left(\|\delta(t)\|_1 + B \right).$$

Proof. Let $\mathcal{M}^0 \subset \mathcal{M}$ denote the set of match m such that $\sum_{s=1}^{t} y_m(s) < \sum_{s=1}^{t} x_m(s)$. Let $\mathcal{A}^0 \subset [n]$ be the set of agent type i such that (23) in Lemma 3 holds. By Lemma 3, for any $m \in \mathcal{M}^0$, there exists $i \in \mathcal{A}(m)$ such that $i \in \mathcal{A}^0$. It follows that $\mathcal{M}^0 \subset \cup_{i \in \mathcal{A}^0} \mathcal{M}_i$. Therefore,

$$\begin{aligned}
\sum_{m \in \mathcal{M}} \sum_{s=1}^{t} (x_m(s) - y_m(s)) &= \sum_{m \in \mathcal{M}^0} \sum_{s=1}^{t} (x_m(s) - y_m(s)) \\
&\leq \sum_{i \in \mathcal{A}^0} \sum_{m \in \mathcal{M}_i} \left(\sum_{s=1}^{t} x_m(s) - \sum_{s=1}^{t} y_m(s) \right) \\
&= \sum_{i \in \mathcal{A}^0} \delta_i(t) + \sum_{i \in \mathcal{A}^0} \sum_{m \in \mathcal{M}_i} x_m(t) \\
&\leq \|\delta(t)\|_1 + B, \qquad (24)
\end{aligned}$$

where the second equality follows because (23) holds for every $i \in \mathcal{A}^0$; and the last inequality holds because $\sum_{i \in \mathcal{A}^0} \sum_{m \in \mathcal{M}_i} x_m(t) \leq \sum_{m \in \mathcal{M}} |\mathcal{A}(m)| x_m(t) \leq B \sum_{m \in \mathcal{M}} x_m(t) \leq B$. It follows that

$$\begin{aligned}
\sum_{s=1}^{t} \langle r, x(s) \rangle - \sum_{s=1}^{t} \langle r, y(s) \rangle &= \sum_{m \in \mathcal{M}} r_m \sum_{s=1}^{t} (x_m(s) - y_m(s)) \\
&\leq r_{\max} \sum_{m \in \mathcal{M}} \sum_{s=1}^{t} (x_m(s) - y_m(s)) \\
&\leq r_{\max} \left(\|\delta(t)\|_1 + B \right),
\end{aligned}$$

where the first inequality holds due to (21) and the last inequality holds due to (24). Q.E.D.

4.3. Inventory of Agents

With the upper-bounds on both terms in (15) via Proposition 2 Proposition 3, we now analyze $\delta(t)$, i.e., the inventory (of agents), which lies critically in both upper-bounds. To bound the expected norm of $\delta(t)$, we first derive a negative drift for $\|\delta(t)\|_2$, similar (in spirit) to the result as (Huang and Neely, 2009, Theorem 1), irrespective of the choice of weight V_t.

Lemma 4. *Suppose that λ has a GPG of ϵ. Fix any constants η and D satisfying*

$$0 < \eta < \epsilon,\ D \geq \eta,\ \text{and } B - 2(\epsilon - \eta)D \leq \eta^2. \qquad (25)$$

For any $t \geq 1$, whenever $\|\delta(t)\|_2 \geq D$ and $\widehat{U}(t) = U^*$,

$$\mathbb{E}\left[\|\delta(t+1)\|_2 \,|\, \mathcal{F}_t \right] \leq \|\delta(t)\|_2 - \eta.$$

Proof. It follows from (18) that when $\widehat{U}(t) = U^*$,

$$\mathbb{E}\left[\|\delta(t+1)\|_2^2 \mid \mathcal{F}_t\right] \le \|\delta(t)\|_2^2 + B + 2V_t \langle U(t) - U^*, Mx(t) - \lambda \rangle . \tag{26}$$

We next claim that

$$\langle U(t) - U^*, Mx(t) - \lambda \rangle \le L(U^*) - L(U(t)). \tag{27}$$

To see this, on the one hand, by the optimality of $x(t)$ given in (10),

$$L(U(t)) = \langle r, x(t) \rangle + \langle U(t), \lambda - Mx(t) \rangle .$$

On the other hand, by the feasibility of $x(t)$,

$$L(U^*) \ge \langle r, x(t) \rangle + \langle U^*, \lambda - Mx(t) \rangle .$$

Combining the last two displayed equations yields the desired (27).

Combining (26) and (27), we get that when $\widehat{U}(t) = U^*$,

$$\begin{aligned} \mathbb{E}\left[\|\delta(t+1)\|_2^2 \mid \mathcal{F}_t\right] &\le \|\delta(t)\|_2^2 + B - 2V_t \left(L(U(t)) - L(U^*)\right) \\ &\le \|\delta(t)\|_2^2 + B - 2V_t \epsilon \|U(t) - U^*\|_2 \\ &= \|\delta(t)\|_2^2 + B - 2\epsilon \|\delta(t)\|_2 \end{aligned} \tag{28}$$

where the second inequality holds, by the assumption that λ has a GPG of ϵ and Proposition 1.

It follows that, whenever $\|\delta(t)\|_2 \ge D$ and $\widehat{U}(t) = U^*$,

$$\begin{aligned} \mathbb{E}\left[\|\delta(t+1)\|_2^2 \mid \mathcal{F}_t\right] &\le \|\delta(t)\|_2^2 + B - 2\eta \|\delta(t)\|_2 - 2(\epsilon - \eta) \|\delta(t)\|_2 \\ &\le \|\delta(t)\|_2^2 - 2\eta \|\delta(t)\|_2 + \eta^2 \\ &= \left(\|\delta(t)\|_2 - \eta\right)^2 , \end{aligned}$$

where the last inequality holds by (25). By Jensen's inequality and the fact that $D \ge \eta$, the desired result follows. Q.E.D.

With the negative drift for $\|\delta(t)\|_2$ established, we can now bound $\mathbb{E}[\|\delta(t)\|_2]$ using a classical drift analysis. To this end, we need the following general lemma, which is essentially a restatement of the result in Gupta (2021).

Lemma 5. *Let $\Psi(t)$ be an $\{\mathcal{F}_t\}$-adapted stochastic process satisfying:*

- *Bounded variation:* $|\Psi(t+1) - \Psi(t)| \le K$;
- *Expected Decrease:* $\mathbb{E}[\Psi(t+1) - \Psi(t) \mid \mathcal{F}_t] \le -\eta$, *when* $\Psi(t) \ge D$;

- $\Psi(0) \le K + D$.

Then, we have

$$\mathbb{E}\left[\Psi(t)\right] \le K\left(1 + \left\lceil \frac{D}{K} \right\rceil\right) + K\left(\frac{K-\eta}{2\eta}\right). \tag{29}$$

The proof of Lemma 5 is deferred to Section C.3 in Appendix C. Now, combining Lemma 4 with Lemma 5, we are ready to bound $\mathbb{E}\left[\|\delta(t)\|_2\right]$.

Proposition 4. *Suppose that λ has a GPG of ϵ. Then for any $t \ge 1$, we have*

$$\mathbb{E}\left[\|\delta(t)\|_2\right] \le 2\sqrt{B}\sum_{s=1}^{t-1}\mathbb{P}\left\{\widehat{U}_s \ne U^*\right\} + \frac{12B}{\epsilon} + 6\sqrt{B}. \tag{30}$$

Proof. Pick $\eta = \epsilon/2$ and $D = \frac{3B-\eta^2}{2(\epsilon-\eta)} \vee \eta$. For $t \ge 1$, define

$$\gamma(t) = \|\delta(t)\|_2 - 2\sqrt{B}\sum_{s=1}^{t-1}\mathbf{1}_{\left\{\widehat{U}_s \ne U^*\right\}}.$$

Note that

$$\left|\|\delta(t+1)\|_2 - \|\delta(t)\|_2\right| \le \|\delta(t+1) - \delta(t)\|_2 = \|Mx(t) - A(t+1)\|_2 \le \sqrt{B}, \tag{31}$$

where the last inequality holds by Lemma 1. It follows that

$$\left|\gamma(t+1) - \gamma(t)\right| = \left|\|\delta(t+1)\|_2 - \|\delta(t)\|_2 - 2\sqrt{B}\mathbf{1}_{\left\{\widehat{U}(t) \ne U^*\right\}}\right| \le 3\sqrt{B},$$

implying that the variation of $\gamma(t)$ in each period does not exceed $3\sqrt{B}$.

Now, suppose that $\gamma(t) \ge D$. If $\widehat{U}(t) = U^*$, we have

$$\mathbb{E}\left[\gamma(t+1) - \gamma(t) \mid \mathcal{F}_t\right] = \mathbb{E}\left[\|\delta(t+1)\|_2 - \|\delta(t)\|_2 \mid \mathcal{F}_t\right] \le -\eta,$$

where the last inequality follows by the fact that $\|\delta(t)\|_2 \ge \gamma(t) \ge D$ and Lemma 4. If $\widehat{U}(t) \ne U^*$,

$$\gamma(t+1) - \gamma(t) = \|\delta(t+1)\|_2 - \|\delta(t)\|_2 - 2\sqrt{B} \le -\eta,$$

where the last inequality holds due to (31) and $\sqrt{B} \ge 1 \ge \eta$. Therefore, we have that $\mathbb{E}\left[\gamma(t+1) - \gamma(t) \mid \mathcal{F}_t\right] \le -\eta$, when $\gamma(t) \ge D$.

Further, $\gamma(1) = \|\delta(1)\|_2^2 = 1$. By applying (29) in Lemma 5, we have that for all $t \ge 1$

$$\mathbb{E}\left[\gamma(t)\right] \le K\left(1 + \left\lceil \frac{D}{K} \right\rceil\right) + K\left(\frac{K-\eta}{2\eta}\right) \le 2K + D + \frac{K^2}{2\eta},$$

where $K = 3\sqrt{B}$. Hence, for any $t \ge 1$,

$$\mathbb{E}\left[\|\delta(t)\|_2\right] \le 2\sqrt{B}\sum_{s=1}^{t-1}\mathbb{P}\left\{\widehat{U}_s \ne U^*\right\} + \mathbb{E}\left[\gamma(t)\right] \le 2\sqrt{B}\sum_{s=1}^{t-1}\mathbb{P}\left\{\widehat{U}_s \ne U^*\right\} + \frac{12B}{\epsilon} + 6\sqrt{B}.$$

Q.E.D.

4.4. Proof of Main Results

We are almost ready to prove our main results. Next, we present the last technical lemma we need to bound the regret in terms of the GPG and instance primitives. The lemma bounds the estimation error probability of $\widehat{U}(t)$ when λ has a positive GPG.

Lemma 6. *Suppose λ has a GPG of ϵ. Then, we have*

$$\sum_{t=1}^{\infty} \mathbb{P}\left\{\widehat{U}(t) \neq U^*\right\} \leq \frac{16}{\epsilon^2}. \tag{32}$$

The proof of Lemma 6 is deferred to Section C.4 in Appendix C. Now, we complete the proof of Theorem 1, which bounds the regret in the case with unknown arrival rates.

Proof of Theorem 1. Recall that $\mathbb{E}\left[\mathsf{R}^{*,t}\right] \leq t\left\langle r, x^*\right\rangle$ and $\mathsf{R}^{\pi,t} = \sum_{s=1}^{t}\left\langle r, y(s)\right\rangle$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\mathsf{R}^{*,t} - \mathsf{R}^{\pi,t}\right] &\leq t\left\langle r, x^*\right\rangle - \sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, y(s)\right\rangle\right] \\ &= t\left\langle r, x^*\right\rangle - \sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, x(s)\right\rangle\right] + \sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, x(s)\right\rangle - \left\langle r, y(s)\right\rangle\right].\end{aligned}$$

By assumption, $\{V_t\}_{t=1}^{\infty}$ is monotonically increasing. Thus, by Proposition 2,

$$\begin{aligned}t\left\langle r, x^*\right\rangle - \sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, x(s)\right\rangle\right] &\leq \sum_{s=1}^{t}\frac{B+1}{2V_s} + \sqrt{n}r_{\max}\mathbb{E}\left[\left\|\delta(t)\right\|_2\right] + 2Br_{\max}\sum_{s=1}^{t}\mathbb{P}\left\{\widehat{U}(s) \neq U^*\right\} \\ &\leq O\left(B + \sqrt{n}Br_{\max}\left(\frac{1}{\epsilon} + \sum_{s=1}^{t}\mathbb{P}\left\{\widehat{U}(s) \neq U^*\right\}\right)\right),\end{aligned}$$

where the last inequality holds by invoking the assumption $\sum_{t=1}^{\infty} 1/V_t < \infty$ and Proposition 4.

Moreover, by Proposition 3,

$$\begin{aligned}\sum_{s=1}^{t}\mathbb{E}\left[\left\langle r, x(s)\right\rangle - \left\langle r, y(s)\right\rangle\right] &\leq r_{\max}\left(\mathbb{E}\left[\left\|\delta(t)\right\|_1\right] + B\right) \\ &\leq r_{\max}\left(\sqrt{n}\mathbb{E}\left[\left\|\delta(t)\right\|_2\right] + B\right) \\ &\leq O\left(\sqrt{n}Br_{\max}\left(\frac{1}{\epsilon} + \sum_{s=1}^{t}\mathbb{P}\left\{\widehat{U}(s) \neq U^*\right\}\right)\right),\end{aligned}$$

where the second inequality holds due to $\left\|\delta(t)\right\|_1 \leq \sqrt{n}\left\|\delta(t)\right\|_2$ and the last inequality follows by Proposition 4. Combining the last three displayed equations yields that

$$\mathbb{E}\left[\mathsf{R}^{*,t} - \mathsf{R}^{\pi,t}\right] \leq O\left(B + \sqrt{n}Br_{\max}\left(\frac{1}{\epsilon} + \sum_{s=1}^{t}\mathbb{P}\left\{\widehat{U}(s) \neq U^*\right\}\right)\right) \leq O\left(B + \frac{\sqrt{n}Br_{\max}}{\epsilon^2}\right), \tag{33}$$

where the last inequality follows from the first inequality of Lemma 6. Q.E.D.

In the case where the arrival rates are unknown, we set $\widehat{U}(t) = U^*$; hence Corollary 1 readily follows from (33).

4.5. Discussion on the Queue Lengths

In this subsection, we demonstrate that the expected number of agents waiting in the system is bounded, i.e., the queue lengths are stable over time, as a consequence of our analysis of $\delta(t)$. Let $q(t)$ denote the number of agents waiting in the system at the end of time period t, that is,

$$q(t) = \sum_{s=1}^{t} (A(s) - My(s)). \tag{34}$$

The following lemma bounds the total number of waiting agents in terms of $\delta(t)$.

Lemma 7.

$$\|q(t)\|_1 \le (B+1)\|\delta(t)\|_1 + B^2.$$

Proof. Because $q_i(t) \ge 0$, $\|q(t)\|_1 = \sum_{i\in[n]} q_i(t)$ It follows from (34) and (7) that

$$q(t) = -\delta(t) + \sum_{s=1}^{t-1} Mx(s) - \sum_{s=1}^{t} My(s).$$

Therefore,

$$\begin{aligned}
\sum_{i\in[n]} q_i(t) &= -\sum_{i=1}^{n} \delta_i(t) + \sum_{m\in\mathcal{M}} |\mathcal{A}(m)| \left(\sum_{s=1}^{t-1} x_m(s) - \sum_{s=1}^{t} y_m(s) \right) \\
&\le -\sum_{i\in[n]} \delta_i(t) + \max_{m\in\mathcal{M}} |\mathcal{A}(m)| \sum_{m\in\mathcal{M}} \sum_{s=1}^{t} (x_m(s) - y_m(s)) \\
&\le -\sum_{i\in[n]} \delta_i(t) + B(\|\delta(t)\|_1 + B) \le (B+1)\|\delta(t)\|_1 + B^2,
\end{aligned}$$

where the last inequality holds due to (24). Q.E.D.

Lemma 7 immediately implies that

$$\mathbb{E}[\|q(t)\|_1] \le (B+1)\mathbb{E}[\|\delta(t)\|_1] + B^2 \le (B+1)\sqrt{n}\mathbb{E}[\|\delta(t)\|_2] + B^2.$$

Applying the the bounds on $\mathbb{E}[\|\delta(t)\|_2]$ from Proposition 4 and Lemma 6, we have

$$\mathbb{E}[\|q(t)\|_1] \le O\left(\sqrt{n}B \left(\frac{1}{\epsilon} + \sum_{s=1}^{t} \mathbb{P}\left\{ \widehat{U}(s) \ne U^* \right\} \right) + B^2 \right) \le O\left(\frac{\sqrt{n}B}{\epsilon^2} + B^2 \right).$$

5. Numerical Results

In this section, we numerically test our primal-dual policies. In our experiment, we test our primal-dual policy where $\widehat{U}(t)$ is chosen according to Theorem 1 and Corollary 1, respectively. When $\widehat{U}(t)$ is chosen according to Theorem 1, we refer to the primal-dual policy as *primal-dual blind* to reflect that the policy is blind to the actual arrival rate. When $\widehat{U}(t)$ is fixed to U^* as suggested Corollary

1, we refer to it as *primal-dual (with known arrival rates)*. For both primal-dual blind and primal-dual with known arrival rates, we pick $V_t = T$. In our computational experience, the choice of V_t does not play a significant role when V_t is selected to be t^2 or T or even $\sqrt{T}$. It is only when V_t is chosen to be much smaller than T then it starts to change regret and queue length. Because our primal-dual policies are non-batching policies that attempt to realize matches at every period, we simulate two other non-batching policies as benchmarks: (i) the sum-of-squares policy of Gupta (2021) and (ii) the maximum-queue-sum policy, [3] a natural generalization of the greedy policy proposed in Kerimov et al. (2021b) for two-way matching networks. Both of these policies require the knowledge of arrival rates and use match types that only lie in the optimal basis of the fluid relaxation. The maximum-queue-sum policy of Kerimov et al. (2021b) is only guaranteed to have constant regret for two-way matching networks where the residual network is acyclic, while the sum-of-squares policy is shown to achieve constant regret for any matching network with positive GPG.

We test the policies on three instances of dynamic matching problems. For each instance, we simulate 1000 replications to estimate the expected performances of different policies. The first instance in our experiments is a dynamic matching system with 6 types of agents, and 20 types of matches, where each non-self-match is randomly generated where the probability of the agent belonging to the match is 0.5, and the reward for each match m is randomly generated according to a normal distribution with mean $|\mathcal{A}(m)|^2$ and standard deviation 0.1. The second instance in our experiments is taken from Kerimov et al. (2021a), with the matching network and arrival rates shown in Figure 2. The instance was constructed in Kerimov et al. (2021a) as an example that the myopic policy, i.e., a policy that always realizes the available match with the highest reward in each period, does not achieve bounded regret. The third instance in our experiments is a complete bipartite matching system with 5 types of agents in each partite, i.e., all matches consist of at most two agents, and for any two types of agents in different partites, there exists a match containing them. All self-matches have a reward 0 and the reward for any two-way match is generated from a normal distribution with mean 1 and standard deviation 0.001. The arrival rate of each agent is generated from a normal distribution with mean 1 and standard deviation 0.1, then normalized so that the sum of arrival rates is equal to 1. The bipartite instance is created because (i) the maximum-queue-sum policy reduces to the greedy policy proposed in Kerimov et al. (2021b) which was shown to have a constant regret, and (ii) the reward is selected such that although the optimal

[3] When an agent arrives at time t, the policy realizes the match m that maximizes $\sum_{i\in\mathcal{A}(m)} q_i(t)$, among all matches in the optimal basis of the fluid relaxation, with $q_i(t) \geq 0$ for all $i \in \mathcal{A}(m)$.

basis for the fluid solution is unique, there are other solutions close to optimal, thus illustrating the advantage of primal-dual policy which uses all match types instead of just the match types in the optimal basis.

In all three instances, we find that our primal-dual policy with known arrival rates is consistently the best policy in expected regret. Next, we discuss several specific observations we made for each of the instances. In the first problem instance, there are 6 agent types and 20 randomly generated multi-way matches. For this instance, Figure 1 demonstrates that primal-dual blind, despite not knowing the arrival rates, achieves similar regret compared to sum-of-squares, while the primal-dual with known arrival rates achieves about 15-20% lower regret than both. The regret for maximum-queue-sum is much larger. This is expected, as unlike the other policies, the maximum-queue-sum policy is not guaranteed to have constant regret because of multi-way matches. In addition, the two primal-dual policies have the lowest expected number of waiting agents, followed by sum-of-squares and maximum-queue-sum policies.

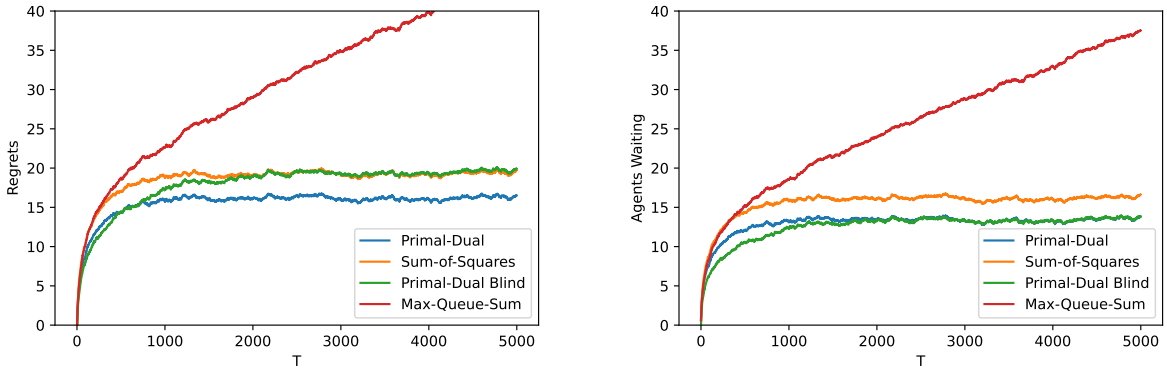

Figure 1 **The dynamic matching instance contains with 6 types of agents and 20 types of matches, where the reward for each match m is randomly generated.**

In the second instance, we use the example of Kerimov et al. (2021a), which has 7 agent types and 4 multi-way matches. As shown in Figure 2, like the first instance, maximum-queue-sum also has the biggest regret, demonstrating that it is not particularly suitable for matching problems with multi-way matches. The primal-dual blind, sum-of-squares, and primal-dual policies all seem to have constant regret, with primal-dual (with known arrival rates) having the lowest, followed by sum-of-squares, then primal-dual blind. More specifically, the regret of the primal-dual policy is about 20-30% lower than sum-of-squares. This difference is mainly attributed to the primal-dual policy having a smaller number of waiting agents, as shown on the second plot in Figure 2. We also note that maximum-queue-sum has the smallest number of agents waiting in the system. This is

because the policy acts as a type of greedy policy that would rarely simultaneously have customers of type 2, type 4 and type 6 waiting in the system, therefore rarely realizes match type 1, which has a far larger reward than the other match types.

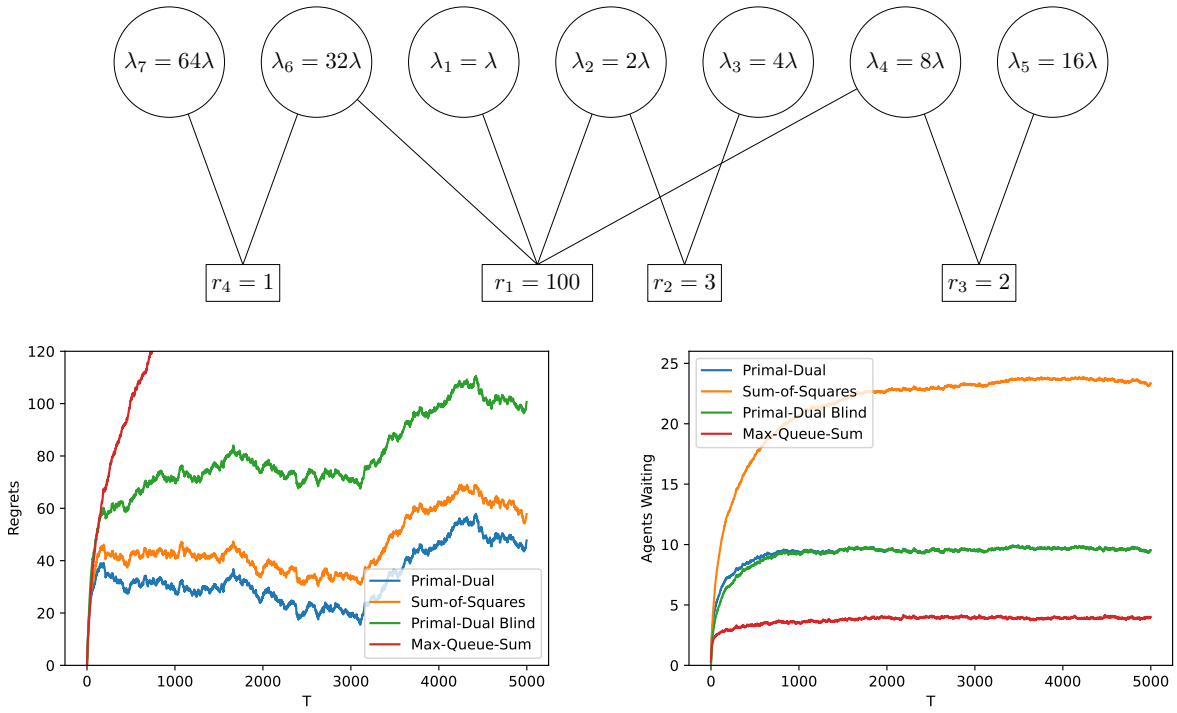

Figure 2 **Multiway matching instance taken from (Kerimov et al., 2021a, Figure 7) and (Gupta, 2021, Figure 2), where λ is a normalizing constant.**

In the third instance, recall that we have a bipartite matching network with only two-way matches. One important feature in the third instance is that due to our setup, the sum of the arrival rates in one partition of the agents is close to that of the other partition. This feature significantly increases the number of periods required for the policies to reach equilibrium. As shown in Figure 3, none of the simulated policies seem to reach equilibrium by period 5000. Because the third instance has a bipartite matching network, the maximum-queue-sum is very effective. It achieves significantly lower regret than both sum-of-squares and primal-dual blind. Nevertheless, the primal-dual policy with known arrival rates still achieves the lowest regret, and in this instance, it is about 40% better than maximum-queue-sum and almost 60% better than sum-of-squares. For the number of agents waiting, the order between primal-dual, maximum-queue-sum, and sum-of-squares remains the same as regret. Interestingly, the primal-dual blind policy has the least number of agents waiting. This is likely because the sample-average-approximated dual solution used in the primal-dual blind policy tends to overestimate the arrival rates for agents waiting in the system at

the beginning of the horizon, leading to the primal-dual blind being more aggressive in matching the agents at the cost of losing future rewards.

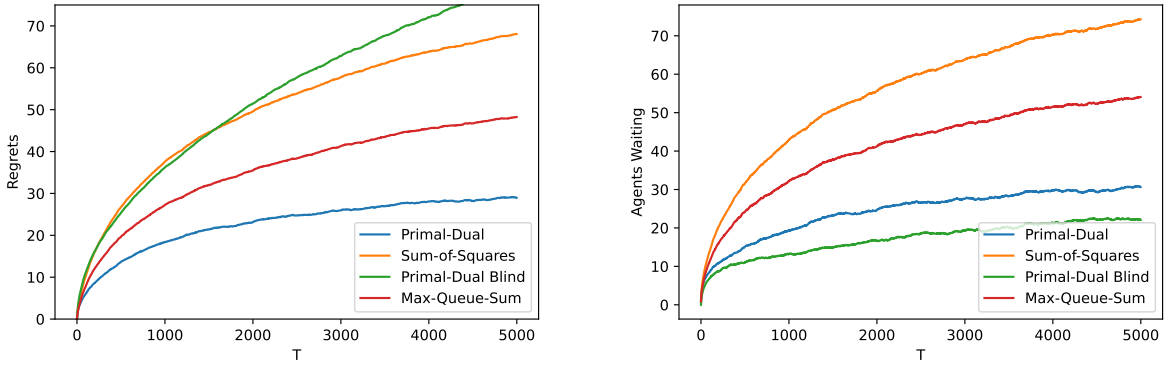

Figure 3 **The dynamic matching instance where the matching network is a complete bipartite graph with 5 agent types in each partite, where the reward of each match type and the arrival rate of each agent are randomly generated.**

6. Conclusion

In this work, we propose primal-dual dynamic matching policies that dynamically adjust a Lagrangian multiplier for each agent type and use it to schedule matches. Compared to the primal-dual policy of Nazari and Stolyar (2018) that achieves $o(T)$ regret, the critical innovation that enabled our policies to achieve constant regret is to design a Lagrangian multiplier combining both the dual solution of the approximated fluid relaxation *and* the inventory. This combination, together with the GPG assumption, allows us to establish a negative drift to effectively control the expected norm of the inventory vector, the pivotal result that leads to constant regret of primal-dual policies.

Our primal-dual policies are the first to achieve constant regret at all times under unknown arrival rates and unknown length of time horizon. When the arrival rate is known, our policies match the optimal scaling in terms of GPG from the literature. In contrast to existing constant regret policies that restrict matches to the optimal basis, our primal-dual policies are more flexible in that they would use any match type that maximizes the reduced reward. This explains why our policies enjoy superior numerical performances compared to the other constant regret policies in the literature.

The design and analysis of the primal-dual dynamic matching policies presented in this work lead to multiple interesting future directions. First, the idea of designing a Lagrangian multiplier

combining both the dual solution of the approximated fluid relaxation *and* the inventory may be applicable to other resource allocation models (see, e.g., Balseiro et al., 2021). Second, our analysis of the primal-dual policies in this paper only applies to the model with a finite number of agent types. It would be fascinating to understand how similar primal-dual policies work when the number of agent types is infinite. Finally, it would be interesting to see whether the primal-dual policies can be modified to analyze additional features in dynamic matching, such as unknown reward vector, customer abandonment, fairness considerations, and adversarial arrivals.

Acknowledgement

The authors would like to thank Varun Gupta, Itai Gurvich, and the University of Rochester Operations Management seminar participants for their helpful discussions. J. Xu is also grateful to Xiaojun Lin for suggesting the reference Huang and Neely (2009) which inspired the design of our primal-dual policies.

J. Xu is supported in part by the NSF Grant CCF-1856424 and an NSF CAREER award CCF-2144593. S. H. Yu is supported in part by the NSF Grant CCF-1856424.

References

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

Akbarpour, Mohammad, Shengwu Li, Shayan Oveis Gharan. 2020. Thickness and information in dynamic matching markets. *Journal of Political Economy* **128**(3) 783–815.

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

Aouad, Ali, Ömer Saritaç. 2020. Dynamic stochastic matching under limited time. *Proceedings of the 21st ACM Conference on Economics and Computation*. 789–790.

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

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

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

Balseiro, Santiago, Omar Besbes, Dana Pizarro. 2021. Survey of dynamic resource constrained reward collection problems: Unified model and analysis. *Available at SSRN 3963265* .

Banerjee, Siddhartha, Yash Kanoria, Pengyu Qian. 2018. Dynamic assignment control of a closed queueing network under complete resource pooling. *arXiv preprint arXiv:1803.04959* .

Blanchet, Jose H, Martin I Reiman, Virag Shah, Lawrence M Wein, Linjia Wu. 2022. Asymptotically optimal control of a centralized dynamic matching market with general utilities. *Operations Research* .

Bumpensanti, Pornpawee, He Wang. 2020. A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Science* **66**(7) 2993–3009.

Buỳić, Ana, Sean Meyn. 2015. Approximate optimality with bounded regret in dynamic matching models. *ACM SIGMETRICS Performance Evaluation Review* **43**(2) 75–77.

Chen, Guanting, Xiaocheng Li, Yinyu Ye. 2022. An improved analysis of LP-based control for revenue management. *Operations Research* .

Cox, David Roxbee, Hilton David Miller. 2017. *The theory of stochastic processes*. Routledge.

Dimakis, Antonis, Jean Walrand. 2006. Sufficient conditions for stability of longest-queue-first scheduling: Second-order properties using fluid limits. *Advances in Applied probability* **38**(2) 505–521.

Gupta, Varun. 2021. Greedy algorithm for multiway matching with bounded regret. *arXiv preprint arXiv:2112.04622* .

Gurvich, Itai, Ward Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations research* **58**(2) 316–328.

Harrison, J Michael. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *The Annals of Applied Probability* **8**(3) 822–848.

Huang, Longbo, Michael J Neely. 2009. Delay reduction via lagrange multipliers in stochastic network optimization. *2009 7th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*. IEEE, 1–10.

Jasin, Stefanus. 2015. Performance of an LP-based control for revenue management with unknown demand parameters. *Operations Research* **63**(4) 909–915.

Jasin, Stefanus, Sunil Kumar. 2012. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research* **37**(2) 313–345.

Kanoria, Yash. 2022. Dynamic spatial matching. *Proceedings of the 23rd ACM Conference on Economics and Computation*. 63–64.

Karp, Richard M, Umesh V Vazirani, Vijay V Vazirani. 1990. An optimal algorithm for on-line bipartite matching. *Proceedings of the twenty-second annual ACM symposium on Theory of computing*. 352–358.

Kerimov, Süleyman, Itai Ashlagi, Itai Gurvich. 2021a. Dynamic matching: Characterizing and achieving constant regret. *Available at SSRN 3824407* .

Kerimov, Süleyman, Itai Ashlagi, Itai Gurvich. 2021b. On the optimality of greedy policies in dynamic matching. *Available at SSRN 3918497* .

Ma, Will, David Simchi-Levi. 2020. Algorithms for online matching, assortment, and pricing with tight weight-dependent competitive ratios. *Operations Research* **68**(6) 1787–1803.

Mandelbaum, Avishai, Alexander L Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855.

Mehta, Aranyak, et al. 2013. Online matching and ad allocation. *Foundations and Trends® in Theoretical Computer Science* **8**(4) 265–368.

Nazari, Mohammadreza, Alexander L. Stolyar. 2018. Reward maximization in general dynamic matching systems.

Neely, Michael J. 2010. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks* **3**(1) 1–211.

Özkan, Erhun, Amy R Ward. 2020. Dynamic matching for real-time ride sharing. *Stochastic Systems* **10**(1) 29–70.

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

Vera, Alberto, Siddhartha Banerjee, Itai Gurvich. 2021. Online allocation and pricing: Constant regret via bellman inequalities. *Operations Research* **69**(3) 821–840.

Ward, Amy R, Mor Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* **61**(1) 228–243.

Appendix A: Additional Analysis on General Position Gap

In this section, we provide some additional analysis to complement the discussion of GPG in Section 3.1. In Kerimov et al. (2021a), the authors defined the general position gap (referred to as K-GPG here) using the optimal solution x^* of (2), when the optimal solution is unique and non-degenerate.[4] More specifically, assuming x^* has basis $\mathcal{B}$, then the K-GPG is defined as $\min_{m\in\mathcal{B}} x^*_m$. The next lemma shows that K-GPG can be used to determine a lower-bound on the GPG from our definition.

Lemma 8. *Suppose that (2) has a basic optimal solution x^* with basis $\mathcal{B}$. Let $\epsilon' = \min_{m\in\mathcal{B}} x^*_m$, and $v_m^\top$ be the row vector in $M_{\mathcal{B}}^{-1}$ indexed by m. Then, the GPG for λ is at least*

$$\frac{\epsilon'}{C_1} \wedge \frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2},$$

where $C_1 = \min_{m\in\mathcal{B}} \|v_m\|_2$ and $C_2 = \left\|\sum_{m\in\mathcal{B}} v_m\right\|_2$.

[4] The formulation of Kerimov et al. (2021a) is slightly different with constraint set $Mx \le \lambda$. This is equivalent to formulation (2) with the appropriate self-matches having zero rewards.

Proof. We assume $1-\sum_{m\in\mathcal{B}} x^*_m > 0$, as the lemma is trivially true when $\sum_{m\in\mathcal{B}} x^*_m = 1$. By strong duality, we have

$$\begin{aligned}\min_U L(U) = \min_{U,y}\Big\{U^\top \widehat{\lambda} + y\Big\}\\ y + \sum_{i\in\mathcal{A}(m)} U_i \geq r_m,\ \forall m,\\ y \geq 0.\end{aligned}$$

Let (U^*, y^*) be the optimal solution to the optimization problem above. By complementary slackness, as $1-\sum_{m\in\mathcal{B}} x^*_m > 0$, we must have $y^* = 0$, and $\sum_{i\in\mathcal{A}(m)} U^*_i \geq r_m$ for each $m\in\mathcal{B}$. Also, let $\widehat{x} = M_{\mathcal{B}}^{-1}\widehat{\lambda}$. If $\widehat{x}\geq \mathbf{0}$ and $1-\sum_{m\in\mathcal{B}}\widehat{x}_m \geq 0$, then by complementary slackness, (U^*, y^*) is also the optimal solution if λ is replaced by $\widehat{\lambda}$.

Next, consider an arbitrary $\widehat{\lambda}$ with $\left\|\widehat{\lambda}-\lambda\right\|_2 \leq \epsilon$. Then for any $m\in\mathcal{B}$, by Cauchy-Schwartz,

$$\begin{aligned} v_m^\top(\lambda-\widehat{\lambda}) \leq \left\|v_m^\top\right\|_2 \left\|\widehat{\lambda}-\lambda\right\|_2 \leq C_1\epsilon\\ \implies v_m^\top\widehat{\lambda} \geq v_m^\top\lambda - C_1\epsilon \geq \epsilon' - C_1\epsilon\\ \implies v_m^\top\widehat{\lambda} \geq 0 \text{ if } \epsilon \leq \frac{\epsilon'}{C_1}.\end{aligned}$$

Also by Cauchy-Schwartz, we have

$$\begin{aligned}\sum_{m\in\mathcal{B}} v_m^\top(\widehat{\lambda}-\lambda) \leq \left\|\sum_{m\in\mathcal{B}} v_m^\top\right\|_2 \left\|\lambda-\widehat{\lambda}\right\|_2 \leq C_2\epsilon\\ \implies 1-\sum_{m\in\mathcal{B}} v_m^\top\widehat{\lambda} \geq 1-\sum_{m\in\mathcal{B}} v_m^\top\lambda - C_2\epsilon = 1-\sum_{m\in\mathcal{B}} x^*_m - C_2\epsilon\\ \implies 1-\sum_{m\in\mathcal{B}} v_m^\top\widehat{\lambda} \geq 0 \text{ if } \epsilon \leq \frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2}\end{aligned}$$

Therefore, if

$$\epsilon = \frac{\epsilon'}{C_1} \wedge \frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2},$$

then we have that (U^*, y^*) is still the optimal solution if λ is replaced by $\widehat{\lambda}$. This implies that λ has a GPG of at least

$$\frac{\epsilon'}{C_1} \wedge \frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2}.$$

Q.E.D.

Note that $1-\sum_{m\in\mathcal{B}} x^*_m = 0$ only under the degenerate case where the optimal basic solution only contains self-matches. Also, there is a relatively easy way to improve $\frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2}$ in Lemma 8. Namely, we can change $\mathcal{X}$ by relaxing the $\sum_{m\in\mathcal{M}} x_m \leq 1$ to $\sum_{m\in\mathcal{M}} x_m \leq K$ for any integer $K > 1$, and allow the primal-dual policy defined in Algorithm 1 to schedule K matches instead of 1 match in each period. This change would improve the second term $\frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2}$ to $\frac{K-\sum_{m\in\mathcal{B}} x^*_m}{C_2}$. However, the change is not really practical, as the term $\frac{1-\sum_{m\in\mathcal{B}} x^*_m}{C_2}$ is in general much larger than $\frac{\epsilon'}{C_1}$, and allowing K matches instead of 1 match in each period only increases the variations $\delta(t)$, therefore degrading our policy performance.

In addition, we remark that both C_1 and C_2 are constants that depend on M but are independent of λ. Therefore, one considers matrix M to be fixed as in Kerimov et al. (2021a), then our GPG has at least the same magnitude as the K-GPG.

Appendix B: Properties of the Dual Solution

Here, we provide a lemma showing that any optimal solution for the dual problem without the constraint $\sum_{m\in\mathcal{M}} x_m \le 1$ is also an optimal solution for the dual problem with the constraint.

Lemma 9. *Let λ be any non-negative vector in which $\sum_{i=1}^n \lambda_i = 1$. Let $\widehat{U}$ be an optimal solution of*

$$\min_{U\in\mathbb{R}^n} U^\top \lambda, \quad s.t. \sum_{i\in\mathcal{A}(m)} U_i \ge r_m,\ \forall m.$$

T hen it is also an optimal solution of

$$\min_{U\in\mathbb{R}^n} L_\lambda(U) = \min_{U\in\mathbb{R}^n} \max_{x\in\mathcal{X}} \left(r^\top - U^\top M\right) x + U^\top \lambda,$$

where we recall that $\mathcal{X} = \{x \,|\, x\in\mathbb{R}^d_{\ge 0}, \sum_{m\in\mathcal{M}} x_m \le 1\}$. Furthermore, we have

$$\left\|\widehat{U}\right\|_\infty \le r_{\max} \quad \text{and} \quad \left\|\widehat{U}\right\|_2 \le \sqrt{n} r_{\max}.$$

Proof. First, we show that $\widehat{U}$ is an optimal solution of $\min_{U\in\mathbb{R}^n} L_\lambda(U)$. It suffices to show that

$$\min_{U\in\mathbb{R}^n} L_\lambda(U) = \min_{U\in\mathbb{R}^n : r - M^\top U \le 0} U^\top \lambda.$$

For any $U\in\mathbb{R}^n$ such that $\max_i (r - M^\top U)_i > 0$, define $U'_i = U_i + v$ for all $i\in[n]$, where $v = \max_i (r - M^\top U)_i$. Then $r - M^\top U' \le 0$ and hence $L_\lambda(U') = {U'}^\top \lambda$. Also, note that $L_\lambda(U) = v + U^\top \lambda$. Therefore,

$$L_\lambda(U') - L_\lambda(U) = (U' - U)^\top \lambda - v = v\sum_{i=1}^n \lambda_i - v \le 0.$$

Hence,

$$\min_{U\in\mathbb{R}^n} L_\lambda(U) = \min_{U\in\mathbb{R}^n : r - M^\top U\le 0} L_\lambda(U) = \min_{U\in\mathbb{R}^n : r - M^\top U \le 0} U^\top \lambda.$$

It remains to prove that $\left\|\widehat{U}\right\|_\infty \le r_{\max}$ and $\left\|\widehat{U}\right\|_2 \le \sqrt{n} r_{\max}$. By Assumption 1 and feasibility of $\widehat{U}_i$, we have $\widehat{U}_i \ge r_i \ge 0$. Also, by optimality of $\widehat{U}$, we have that for all $i\in[n]$, $\widehat{U}_i \le r_{\max}$ as otherwise, we can decrease $\widehat{U}_i$ and still obtain a feasible solution. Therefore, we have

$$\left\|\widehat{U}\right\|_\infty \le r_{\max}.$$

For $\left\|\widehat{U}\right\|_2$, we have that $\left\|\widehat{U}\right\|_2^2 \le n \left\|\widehat{U}\right\|_\infty^2 \le n r_{\max}^2$. Q.E.D.

Appendix C: Additional Proofs

Here, we provide the proofs for some of our technical results which are fairly standard given the literature.

C.1. Proof of Proposition 1

Proof. By the definition given in (5),

$$L_{\widehat{\lambda}}(U) = \max_{x\in\mathcal{X}} \left\langle r - M^\top U, x\right\rangle + \left\langle U, \widehat{\lambda}\right\rangle = L_\lambda(U) + \left\langle U, \widehat{\lambda} - \lambda\right\rangle.$$

Therefore,

$$L_{\widehat{\lambda}}(U) - L_{\widehat{\lambda}}(U^*) = L_\lambda(U) - L_\lambda(U^*) + \left\langle U - U^*, \widehat{\lambda} - \lambda\right\rangle. \tag{35}$$

Cond.1 $\Rightarrow$ Cond.2: By the duality between $\|\cdot\|$ and $\|\cdot\|_*$, we can choose $\widehat{\lambda}$ such that $\left\|\widehat{\lambda} - \lambda\right\| = \epsilon$ and $\left\langle U - U^*, \widehat{\lambda} - \lambda\right\rangle = -\epsilon \left\|U - U^*\right\|_*$. Thus, in view of $L_{\widehat{\lambda}}(U) \ge L_{\widehat{\lambda}}(U^*)$, we deduce from (35) that

$$L_\lambda(U) - L_\lambda(U^*) \ge \epsilon \left\|U - U^*\right\|_*.$$

Cond.2 $\Rightarrow$ *Cond.1:* Combining (35) with Cond.2 yields that for all $\left\|\widehat{\lambda}-\lambda\right\| \le \epsilon$,

$$L_{\widehat{\lambda}}(U) - L_{\widehat{\lambda}}(U^*) \ge \epsilon \left\|U-U^*\right\| + \left\langle U-U^*, \widehat{\lambda}-\lambda \right\rangle \ge 0,$$

where the last inequality holds by the duality between $\|\cdot\|$ and $\|\cdot\|_*$. Therefore, U^* is an optimal solution to $L_{\widehat{\lambda}}(U)$ for all $\left\|\widehat{\lambda}-\lambda\right\| \le \epsilon$. Q.E.D.

C.2. Proof of Corollary 2

In this section, we prove that the regret is $O(\sqrt{T})$ by setting $V_t = \sqrt{t}$ and $\widehat{U}(t) = U^*$ when the GPG is possibly zero. In particular, suppose $\widehat{U}(t) = U^*$. Recall from (28) that

$$\mathbb{E}\left[\|\delta(t+1)\|_2^2 \mid \mathcal{F}_t\right] \le \|\delta(t)\|_2^2 + B - 2V_t\left(L_\lambda(U(t)) - L_\lambda(U^*)\right). \tag{36}$$

For any given constant $\gamma > 0$, pick $\widehat{\lambda}$ such that $\left\|\lambda-\widehat{\lambda}\right\|_2 = \gamma$ and

$$\left\langle U(t)-U^*, \lambda-\widehat{\lambda}\right\rangle = \gamma \left\|U(t)-U^*\right\|_2.$$

Note that

$$L_\lambda(U(t)) - L_\lambda(U^*) = L_{\widehat{\lambda}}(U(t)) - L_{\widehat{\lambda}}(U^*) + \left\langle U(t)-U^*, \lambda-\widehat{\lambda}\right\rangle. \tag{37}$$

Furthermore, by definition

$$L_{\widehat{\lambda}}(U(t)) \ge \left\langle r - M^\top U(t), \widehat{x}\right\rangle + \left\langle U(t), \widehat{\lambda}\right\rangle = \langle r, \widehat{x}\rangle,$$

where $\widehat{x} \in \arg\max_x \left\{\langle r, x\rangle : Mx = \widehat{\lambda}, x \in \mathcal{X}\right\}$. Also,

$$L_{\widehat{\lambda}}(U^*) = \max_{x\in\mathcal{X}} \left\langle r - M^\top U^*, x\right\rangle + \left\langle U^*, \widehat{\lambda}\right\rangle = \left\langle r - M^\top U^*, x^*\right\rangle + \left\langle U^*, \widehat{\lambda}\right\rangle = \langle r, x^*\rangle + \left\langle U^*, \widehat{\lambda}-\lambda\right\rangle \le \langle r, x^*\rangle + \gamma\left\|U^*\right\|_2.$$

Assembling the above four displayed equations, we get that

$$L_\lambda(U(t)) - L_\lambda(U^*) \ge \gamma\left\|U(t)-U^*\right\|_2 + \langle r, \widehat{x}-x^*\rangle - \gamma\left\|U^*\right\|_2.$$

Since $0 \le r_m \le r_{\max}$ for all match $m \in \mathcal{M}$, it follows that $\langle r, \widehat{x}-x^*\rangle \ge -r_{\max}$. Also, recall that $\|U^*\|_2 \le \sqrt{n} r_{\max}$. We get that

$$L_\lambda(U(t)) - L_\lambda(U^*) \ge \gamma\left\|U(t)-U^*\right\|_2 - \left(1+\sqrt{n}\gamma\right) r_{\max}.$$

In conclusion, we get that

$$\mathbb{E}\left[\|\delta(t+1)\|_2^2 \mid \mathcal{F}_t\right] \le \|\delta(t)\|_2^2 + B + 2\left(1+\sqrt{n}\gamma\right) r_{\max} V_t - 2\gamma\|\delta(t)\|_2$$

Choosing $V_t = \sqrt{t}$, and invoking Lemma 5, we get that

$$\mathbb{E}\left[\|\delta(t)\|_2\right] \le O\left(B + \sqrt{n} r_{\max}\sqrt{T}\right), \quad \forall t \le T.$$

It then follows from Proposition 2 that

$$t\langle r, x^*\rangle - \sum_{s=1}^{t} \mathbb{E}\left[\langle r, x(s)\rangle\right] \le O(\sqrt{T}), \quad \forall t \le T.$$

C.3. Proof of Lemma 5

Proof. Following Gupta (2021), let $\Gamma(t)$ denote a reflected random walk with step size K, reflected boundary $\Gamma_{\min} = \left(1 + \left\lceil \frac{D}{K} \right\rceil\right) K$, initial condition $\Gamma(0) = \Gamma_{\min}$, and

$$\Gamma(t+1) = \max\{\Gamma(t) + \xi(t)K, \Gamma_{\min}\} \quad \text{for } t \geq 0,$$

where $\xi(t)$ are *i.i.d.* random variables taking the value $+1$ with probability $\frac{1}{2} - \frac{\eta}{2K}$, and -1 with probability $\frac{1}{2} + \frac{\eta}{2K}$.

By (Gupta, 2021, Lemma 3), $\Psi(t) \leq_{\text{icx}} \Gamma(t)$ for all $t \geq 0$, where for any $X, Y \in \mathbb{R}^n$, $X \leq_{\text{icx}} Y$ means that X is dominated by Y in the increasing convex order, that is, $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$ for any increasing convex function $f : \mathbb{R}^n \to \mathbb{R}$ with valid expectation. It follows that that for any $t \geq 0$,

$$\mathbb{E}[\Psi(t)] \leq \mathbb{E}[\Gamma(t)]. \tag{38}$$

Define $\rho \triangleq \frac{\frac{1}{2} - \frac{\eta}{2K}}{\frac{1}{2} + \frac{\eta}{2K}} = \frac{K-\eta}{K+\eta}$. By (Cox and Miller, 2017, Section 2.2, eq. (52)), $\{\Gamma(t)\}$ has a unique steady state distribution under which

$$\mathbb{E}[\Gamma_\infty] = \Gamma_{\min} + K\frac{\rho}{1-\rho} = \Gamma_{\min} + K\left(\frac{K-\eta}{2\eta}\right). \tag{39}$$

Next, we construct $\Gamma'(t)$ such that $\Gamma'(0) \overset{\text{d}}{=} \Gamma_\infty$ follows a steady state distribution, and for any $t \geq 0$, $\Gamma'(t+1) = \max\{\Gamma'(t) + \xi(t)K, \Gamma_{\min}\}$. By induction, we have $\Gamma(t) \leq \Gamma'(t)$ for all $t \geq 0$. Therefore,

$$\mathbb{E}[\Gamma(t)] \leq \mathbb{E}\left[\Gamma'(t)\right] = \mathbb{E}[\Gamma_\infty], \tag{40}$$

where the last inequality holds because $\Gamma'(t) \overset{\text{d}}{=} \Gamma_\infty$ by construction. Combining (38), (39) and (40) yields the desired result (29). Q.E.D.

C.4. Proof of Lemma 6

Proof. By Definition 1,

$$\mathbb{P}\left\{\widehat{U}(t) \neq U^*\right\} \leq \mathbb{P}\left\{\left\|\widehat{\lambda}(t) - \lambda\right\|_2 > \epsilon\right\}. \tag{41}$$

It remains to establish the concentration inequality for $\left\|\widehat{\lambda}(t) - \lambda\right\|_2$. First,

$$\mathbb{E}\left[\left\|\widehat{\lambda}(t) - \lambda\right\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^n \left(\frac{1}{t}\sum_{s=1}^t A_i(s) - \lambda_i\right)^2\right] = \frac{1}{t}\sum_{i=1}^n \lambda_i(1-\lambda_i) \leq \frac{1}{t}.$$

Thus, by Jensen's inequality,

$$\mathbb{E}\left[\left\|\widehat{\lambda}(t) - \lambda\right\|_2\right] \leq \sqrt{\mathbb{E}\left[\left\|\widehat{\lambda}(t) - \lambda\right\|_2^2\right]} \leq \sqrt{\frac{1}{t}}.$$

Note that the function $f : (A(1), \ldots, A(t)) \to \left\|\widehat{\lambda} - \lambda\right\|_2$ satisfies the bounded difference property with parameter $2/t$, that is, for any $s \in [t]$ and any $A(1), \ldots, A(t), A'(s)$,

$$\begin{aligned}&\left|f\left(A(1), \ldots, A(s-1), A(s), A(s+1), \ldots, A(t)\right) - f\left(A(1), \ldots, A(s-1), A'(s), A(s+1), \ldots, A(t)\right)\right| \\ &\leq \frac{1}{t}\left\|A(s) - A'(s)\right\|_2 \leq \frac{2}{t}.\end{aligned}$$

Thus, by McDiarmid's inequality, for any $\Delta > 0$,

$$\mathbb{P}\left\{\left\|\widehat{\lambda}(t) - \lambda\right\|_2 \geq \mathbb{E}\left[\left\|\widehat{\lambda}(t) - \lambda\right\|_2\right] + \sqrt{\Delta/t}\right\} \leq \exp\left(-\frac{2\Delta/t}{t(2/t)^2}\right) = \exp(-\Delta/2).$$

Combining the last two displayed equations gives that

$$\mathbb{P}\left\{\left\|\widehat{\lambda}(t)-\lambda\right\|_2 \geq \sqrt{\frac{1}{t}}+\sqrt{\frac{\Delta}{t}}\right\} \leq \exp(-\Delta/2).$$

If $\Delta \geq 1$, then this further implies that

$$\mathbb{P}\left\{\left\|\widehat{\lambda}(t)-\lambda\right\|_2 \geq 2\sqrt{\frac{\Delta}{t}}\right\} \leq 2\exp(-\Delta/2);$$

If $\Delta \leq 1$, then the above is true trivially, as $2\exp(-\Delta/2) \geq 2\exp(-1/2) \geq 1$. Therefore, by letting $\epsilon = 2\sqrt{\Delta/t}$, we deduce that

$$\mathbb{P}\left\{\left\|\widehat{\lambda}(t)-\lambda\right\|_2 \geq \epsilon\right\} \leq 2\exp(-t\epsilon^2/8).$$

Finally, plugging the last displayed equation into (41) and summing over all $t \geq 1$, we get that

$$\sum_{t=1}^{\infty}\mathbb{P}\left\{\widehat{U}(t)\neq U^*\right\} \leq \sum_{t=1}^{\infty} 2\exp(-t\epsilon^2/8) \leq \frac{2}{\exp(\epsilon^2/8)-1} \leq \frac{16}{\epsilon^2}, \tag{42}$$

where the last equality holds because $\exp(x) \geq 1+x$. Q.E.D.