

# Distributed Multi-Depot Routing without Communications

Dawsen Hwang  
 Massachusetts Institute of  
 Technology  
 77 Massachusetts Avenue  
 Cambridge, Massachusetts  
 dawsen@mit.edu

Patrick Jaillet  
 Massachusetts Institute of  
 Technology  
 77 Massachusetts Avenue  
 Cambridge, Massachusetts  
 jaillet@mit.edu

Zhengyuan Zhou  
 Stanford University  
 450 Serra Mall  
 Stanford, California  
 zyzhou@stanford.edu

December 8, 2014

## Abstract

We consider and formulate a class of distributed multi-depot routing problems, where servers are to visit a set of requests, with the aim of minimizing the total distance travelled by all servers. These problems fall into two categories: distributed offline routing problems where all the requests that need to be visited are known from the start; distributed online routing problems where the requests come to be known incrementally. A critical and novel feature of our formulations is that communications are not allowed among the servers, hence posing an interesting and challenging question: what performance can be achieved in comparison to the best possible solution obtained from an omniscience planner with perfect communication capabilities? The worst-case (over all possible request-set instances) performance metrics are given by the approximation ratio (offline case) and the competitive ratio (online case).

Our first result indicates that the online and offline problems are effectively equivalent: for the same request-set instance, the approximation ratio and the competitive ratio differ by at most an additive factor of 2, irrespective of the release dates in the online case. Therefore, we can restrict our attention to the offline problem. For the offline problem, we show that the approximation ratio given by the Voronoi partition is  $m$  (the number of servers). For two classes of depot configurations, when the depots form a line and when the ratios between the distances of pairs of depots are upper bounded by a sublinear function  $f(m)$  (i.e.,  $f(m) = o(m)$ ), we give partition schemes with sublinear approximation ratios  $O(\log m)$  and  $\Theta(f(m))$  respectively. We also discuss several interesting open problems in our formulations: in particular, how our initial results (on the two deliberately chosen classes of depots) shape our conjecture on the open problems.

## Keywords

Traveling Salesman, Distributed Algorithms, Multi-Depot Routing, Robotics, Computational Geometry, Worst-case Analysis, Online Optimization

# 1 Introduction

With the advance of technology, it is now possible to deploy a fleet of servers (e.g., UAVs, robots) to visit requests located in a surrounding territory. The problem can be modeled as the classical (uncapacitated) multi-depot vehicle routing problem, in which a team of servers are to collectively visit a set of requests located in an ambient metric space so as to minimize the total travelled distance. Clearly, an optimal solution to this problem will be achieved by centralized planners that know all the requests, or equivalently, by allowing full communications among the servers. However, such a centralized approach suffers from two drawbacks: Practically, deploying a full-communication scheme among all servers is often overly costly and unreasonable, if not, infeasible. Theoretically, even with all the information, computing the optimal assignment (i.e., which server visits which subset of requests) is intractable.

As such, typical existing solutions to this problem are heuristic algorithms that involve some local communications among subsets of servers for relaying requests to each other [18, 20, 26]. While empirical demonstrations indicate the effectiveness of the heuristic algorithms under certain request-set configurations, no worst-case performance guarantees have been presented.

We take an ambitious step back and ask the following question: what if we simply disallow any communications among the servers? We can achieve this by a static partition of the underlying metric space, determined once for all based solely on the depot locations: thereafter, independent of the request-set instance, each server is only responsible for requests in its prescribed region. In this paper, we formulate a class of distributed routing problems and study such static partitions.

## 1.1 Our Contributions

Our contributions are threefold.

First, in Section 2, we formulate two types of distributed multi-depot vehicle routing problems, offline and online, with the critical and novel feature that no communications are allowed among the servers, achieved by a static partition as mentioned above and explained in detail in Definition 2.1. The class of problems considered here are of both theoretical and practical value. Theoretically, it is interesting to understand the role played by communications among servers (or the absence thereof) by quantifying how well a static partition can perform compared to the optimal solution that is dynamic and request set dependent (and hence requires full communication capabilities among the servers). Practically, communications among the servers can be costly (overhead, time lag, inefficiency etc.), and hence a close-to-optimal solution without communications can be a highly attractive alternative in practical deployment. (Our results in this paper suggest, as an initial step, that searching for such alternatives can be worthwhile, discussed more later.) As shown in Theorem 2.1, we can restrict our attention to distributed offline problems and to deriving bounds for approximation ratios therein, since the approximation ratio for the offline problem and the competitive ratio for the online problem differ by at most an additive factor of 2.

Second, for the distributed offline problem, we present several partition schemes that require only polynomial time and space and characterize their approximation ratios. In Section 3.1, we show that the approximation ratio for the Voronoi partition is  $m$ , where  $m$  is the number of servers throughout this paper. In Section 3.2, we consider a class of depots that form a line in any metric space; in this case, we give a partition scheme with an approximation ratio of  $O(\log m)$ . Finally, in Section 3.3, we consider the case where the ratios between the distances of any two pairs of depots are upper bounded by a sublinear function of  $m$  ( $f(m)$ ); in this case, we give a partition scheme with an approximation ratio of  $\Theta(f(m))$ .

Third, adding to the value of the formulations in this paper is a wide array of open problems with rich opportunities for future explorations (discussed in Section 4), with the central one being Open Problem 4. If the answer is positive, then it will have a fundamental impact in how we view

the role played by communications in routing problems. This is so particularly because the prior predominant paradigm has mostly been to use dynamic request-assignment scheme via restricted (i.e., local) communications intelligently, with the implicit assumption that local communications are always needed: a completely natural and reasonable assumption at the outset. Our initial results on the two deliberately chosen classes of depot configurations (see Remark 3.1) indicate that theoretical endeavors along this direction (of finding such a surprisingly good static partition) are worthwhile. We view the proposed open problems as an open invitation towards this goal.

## 1.2 Related Work

There has been extensive research on problems related to ours. Here we place our work in the broader such context by giving a review (in no way complete) of past relevant work, categorized into three classes as follows.

**Single-server.** Given a particular assignment of requests to servers, the problem then reduces to several single-server problems, each of which is the classical *Traveling Salesman Problem* (TSP): the single server needs to find the optimal order to visit the assigned requests and to return to the depot so that the travelled distance is minimized. If  $P \neq NP$ , no polynomial-time algorithms can solve the TSP [25], and they cannot approximate the solution with a ratio better than  $117/116$  [9, 39]. On the other hand, Christofides [16] give an 1.5-approximated algorithm. For cases where the metric space is the Euclidean plane [1, 35] or induced from a unit-weight graph [36, 37, 41, 42],  $(1 + \epsilon)$ -approximated algorithms (for any  $\epsilon > 0$ ) and 1.4-approximated algorithms exist respectively. See [31] for a comprehensive review of hardness results and the classical analysis of heuristic solutions to the TSP and related problems.

Ausiello *et al.* [2] consider the online TSP, in which the requests are revealed incrementally and a server that travels with a unit speed limit is to visit all the requests so as to minimize the returning time (to the original depot). The performance measure is the *competitive ratio*, the worst case ratio between the proposed algorithm and an optimal offline algorithm that knows the locations and release dates of all requests from the start. Ausiello *et al.* [2] have proposed a 2-competitive algorithm and proved that it is the best possible deterministic online algorithm.

**Multi-server Single-depot.** This class of problems also admits the natural division into offline problems and online problems, with the defining feature being that all the servers share the same depot. The offline problems consist primarily of two scenarios. In the first scenario, each server is required to visit at least one request. This is known as the multiple traveling salesman problem (see [8] for a review). In the second scenario, each server has a limited capacity, and thus multiple servers are required to visit all the requests. This is known as the capacitated vehicle routing problem. See [21, 22, 28] for a review and several heuristics for this problem.

For the online problems, the usually studied objective is to minimize the returning time of the last server (i.e., a Min-Max formulation). Jaillet and Wagner [24] have proposed a centralized deterministic algorithm with a competitive ratio of 2 (the best possible solution). Bonifaci and Stougie [11] consider a variant of this problem in which the cost of an algorithm is measured by the time when the last request is visited and the servers are not required to return to the depot. For this problem, when the online algorithm has  $k$  servers and the offline algorithm used to define the competitive ratio has only  $k^*$  servers ( $k^* \leq k$ ), they propose a centralized deterministic online algorithm with competitive ratio  $1 + \sqrt{1 + 1/2^{\lfloor k/k^* \rfloor - 1}}$ .

Another interesting and related online problem is the online  $k$ -server problem, first proposed by Manasse *et al.* [33]. Here, the ambient metric space is typically a network that consists of a finite number of points (say  $n$ ): the servers can be thought of as moving on a discrete graph. The requests are revealed incrementally. When a request is revealed, one of the  $k$  servers must move to the location of the request instantly, before knowing the subsequent requests. In this model,

each request is associated with a release order (as opposed to a release date). The objective is to minimize the total travelled distance of all servers. The performance measure used in the literature is the competitive ratio where additive factors independent of the problem instance are allowed (in this paper, such additive factors are not allowed).

Manasse *et al.* [33] then conjecture that the lowest competitive ratio for a general ( $n$ -point) metric space is  $k$  (the  $k$ -server conjecture). As an initial step, they prove that the lower bound holds, and the upper bound holds for special cases where  $k = 2$  or  $n = k + 1$ . As substantial progress along this line, Koutsoupias and Papadimitriou [27] have proposed the Work Function Algorithm (WFA) and showed that it has a competitive ratio of at most  $2k - 1$ . It is still an open question whether WFA is  $k$ -competitive. For the case where randomization on the actions of online algorithms is allowed, the randomized  $k$ -server conjecture states that the competitive ratio of the best randomized online algorithm is  $\Theta(\log k)$ . The randomized  $k$ -server conjecture holds for paging, a special case of the  $k$ -server problem in which the distance between any pair of points is one. For paging, Borodin *et al.* [12] show that the competitive ratio of the best randomized algorithm is at least the  $k^{\text{th}}$  harmonic number  $H_k$  ( $H_k = 1 + \frac{1}{2} + \dots + \frac{1}{k}$ ). On the other hand, McGeoch and Sleator [34] construct a randomized algorithm that achieves a competitive ratio of  $H_k$ . In addition, using a primal-dual approach,  $O(\log k)$ -competitive randomized online algorithms can be derived for weighted paging [5, 6]. While the lower bound of paging can be applied to the  $k$ -server problem, the competitive ratio of the best randomized online algorithm for the online  $k$ -server problem is still  $2k - 1$ , the same as the deterministic case. For the case where the competitive ratio is allowed to be dependent on the size of the metric space (say  $n$ ), Bansal *et al.* [4] develop a randomized online algorithm with a polylogarithmic competitive ratio  $O(\log^3 n \log^2 k \log \log n)$ .

The aforementioned results (for multi-server single-depot) require centralized algorithms with the knowledge of all the released requests, and hence full communication capabilities among the servers. Bartal and Rosen [7] consider the distributed version of the online  $k$ -server problem, in which communications are allowed but induce some costs. Bartal and Rosen give a translation between a centralized algorithm and a distributed one.

**Multi-server Multi-depot.** This class of problems are most relevant to our current formulations. Due to the computational complexity, few algorithms that find the optimal solution have been given [3, 29, 30] for this case. Researchers have focused more on finding heuristic algorithms that find sub-optimal solutions quickly.

There are primarily three strategies that are commonly used in existing heuristic algorithms. The first commonly-used strategy is to first assign each request to the server at the nearest depot, and then refine the solution in a centralized fashion [15, 17, 38, 40]. The second commonly-used strategy is to first assign each request to a server in a centralized fashion, and then determine the route of each server without modifying the assignment of the requests [23]. The third commonly-used strategy is to start from calculating a sub-optimal TSP tour that visits all requests. The TSP tour is then divided into pieces, and each piece of the tour is assigned to a server. After that, refinement of the solution is applied [19, 32]. There are no performance guarantees for the heuristic algorithms mentioned above. In addition, either a centralized planner or communications between servers are required for these heuristic algorithms.

Probabilistic formulations of the multi-depot vehicle routing problems have also been considered. Bompadre *et al.* [10] give a polynomial-time algorithm in which each request is assigned to the nearest depot, and then apply probabilistic analysis to the algorithm. They assume that the locations of the requests are independently and uniformly distributed in the unit square and all requests have the same demand, and show that the proposed algorithm achieves an approximation ratio strictly less than 2 almost surely as the number of requests goes to infinity. The focus of our work here is rather different: we do not restrict our attention to any particular metric space,

and we do not impose any assumptions on the number of requests nor impose any probabilistic assumptions on the locations of the requests.

The core idea of static partition, a central quantity we study in this paper, lies in dividing the ambient metric space into different regions and letting each server be responsible for requests in the corresponding region. This idea has been applied to balance the load (defined to be the length of the tour in the solution to the TSP) of servers when the ambient metric space is a compact set in the two-dimensional Euclidean space. Given the probability distribution of the locations of the requests, Carlsson [13] find a partition of the metric space such that the load is almost surely the same for all servers as the number of requests approaches infinity. When the precise probability distribution is unknown but some first and second order statistics are given, Carlsson and Delage [14] find a way to partition the metric space such that the load is most balanced under the worst-case distribution. We note that there are several key differences from our formulations. In addition to having different objectives (Min-Max v.s. Min-Sum), the work mentioned above considers only the asymptotic case where the number of requests approaches infinity. Moreover, the partition of the metric space is based on the distribution of the locations of the requests rather than the locations of the depots.

## 2 Problem Formulation

In this section, we formulate two distributed multi-depot routing problems, offline and online, where communications between servers are not allowed. In Section 2.1, we formulate the distributed offline problem. In Section 2.2, we formulate the distributed online problem and state the relation between the competitive ratio of the distributed online problem and the approximation ratio of the distributed offline problem. For simplicity, our formulations belong to the uncapacitated vehicle routing setting. However, all of our results and analysis are applicable for the capacitated vehicle routing setting when replenishment is allowed at every depot for each server. Due to space limitation, we omit the details.

### 2.1 The Distributed Offline Problem

In the distributed offline problem, there are  $m$  servers in the ambient metric space  $(\mathbb{M}, d)$ ; each of them has its initial location in one of the  $m$  corresponding distinct depots  $x_1, x_2, \dots, x_m \in \mathbb{M}$ . A problem instance consists of a finite list  $I$  of requests in the metric space:  $I = \{l_i \in \mathbb{M}\}_{i=1}^n$ , for some positive  $n$  that can vary in different problem instances. Each request needs to be visited by one server. The  $m$  servers, which must start and end at their corresponding depots, aim to visit all requests in the most efficient way: here measured by the sum of all distances travelled by all the servers. Since the entire problem instance is known when finding a solution (as opposed to requests coming to be known incrementally in an online fashion as discussed in Section 2.2), hence the name “offline”.

On the team level, the central question that immediately arises is finding a good partition of the request set, i.e., which server covers which requests; here and onwards, a partition of the request set is understood to be a disjoint collection of  $m$  sets  $S_1, \dots, S_m$  whose union is  $I$ . Note that if server  $i$  is (somehow) assigned to a particular subset  $S_i$  of requests, then  $i$ 's optimal action is to, at least in principle, compute the solution  $TSP_i(S_i)$  to the traveling salesman problem (the shortest tour that visits each of the requests in  $S_i$  subject to the initial and final depot location  $x_i$ ).

The question then, at first sight, becomes how to find an efficient (or perhaps, in some sense, optimal) partition. However, in characterizing an efficient partition, we necessarily need to make assumptions on allowable partition-selection schemes, such as how consensus (on the assignment) is reached among the servers and what communication or collaboration process is allowed. If we allow for full collaboration/communication among the servers, then it effectively becomes a centralized planning problem where a central computational unit decides on the optimal partition  $\{S_i^{OPT}\}_{i=1}^m$

that achieves the minimum total cost  $OPT(I)$ , where

$$OPT(I) \triangleq \min_{S_1, \dots, S_m \mid \bigcup_{i=1}^m S_i = I} TSP_i(S_i), \quad (1)$$

and then distributes the partition to all the servers in some way.

Evidently, the cost achieved in (1) is the best one can hope for. However, there are at least two drawbacks with this formulation. First, it is instantly clear that this problem is computationally intractable, even disregarding the NP-hardness of the TSP. The second drawback lies in the strong assumption on the communications involved on the servers' part; moreover, for the case where the requests are revealed incrementally as defined in Section 2.2, such communications need to happen every time a new request is revealed, since the partition of the requests is intrinsically dependent on the entire problem instance.

Motivated by these two concerns, we naturally wonder if it is possible to find a good static partition of the entire metric space (that depends only on the locations of the depots), which then induces a partition for any request set. Under this setting, each server needs only visit the requests that fall into its assigned region (and hence no communications between servers are required). The following definition formalizes this distributed partition scheme that is static in nature.

**Definition 2.1.** A distributed partition scheme  $\mathbf{par}$  is a function that, given the  $m$  depot locations  $x_1, \dots, x_m$ , assigns each server  $i$  to a region  $M_i^{\mathbf{par}}$ .

$$\mathbf{par}(x_1, x_2, \dots, x_m) = (M_1^{\mathbf{par}}, M_2^{\mathbf{par}}, \dots, M_m^{\mathbf{par}}),$$

where  $M_i^{\mathbf{par}} \subset \mathbb{M}$ ,  $M_i^{\mathbf{par}} \cap M_j^{\mathbf{par}} = \emptyset, \forall i, j$ , and  $\bigcup_{i=1}^m M_i^{\mathbf{par}} = \mathbb{M}$ .

**Remark 2.1.** We note that a static partition is sufficient for achieving the no-communication requirement, but not necessary. In particular, we can consider a time-dependent partition (also independent of the problem instance) as follows.

$$\mathbf{par}(x_1, x_2, \dots, x_m, t) = (M_1^{\mathbf{par}}(t), M_2^{\mathbf{par}}(t), \dots, M_m^{\mathbf{par}}(t)).$$

In this setting, request  $(r_j, l_j)$  is assigned to server  $i$  if and only if  $l_j \in M_i^{\mathbf{par}}(r_j)$ . Since the partition does not depend on the request set, no communications are required.

Each partition scheme  $\mathbf{par}$  then induces a distributed algorithm that has the cost function  $DIS^{\mathbf{par}}$ , given by

$$DIS^{\mathbf{par}}(I) \triangleq \sum_{i=1}^m TSP_i(S_i^{\mathbf{par}})$$

where  $S_i^{\mathbf{par}} = M_i^{\mathbf{par}} \cap I$  is the set of requests assigned to server  $i$  under the partition scheme  $\mathbf{par}$ . For notational convenience, we often drop the dependence on the specific partition scheme used in the regions  $M_i$ , the sets of requests  $S_i$ , and the cost function  $DIS$  when the context shall make it clear which partition scheme we are using. We use  $DIS^{\mathbf{par}}(I)$  as a means to measure the performance of the partition scheme  $\mathbf{par}$  on the problem instance  $I$ . Therefore, we study the value but are not concerned about how the value can be computed. For computing the value, the reader is referred to the related work regarding the TSP, as discussed in Section 1.2.

We note here that the centralized-planning formulation is not entirely useless. The optimal cost  $OPT(I)$  defined in (1) is a good (and ambitious) comparison metric against which we can evaluate the solution quality of a specific partition scheme  $\mathbf{par}$ . We wish the resulting partition from the distributed problem to be "close" to  $OPT(I)$ , and the closer the better, as formalized below.

**Definition 2.2.** A distributed partition scheme  $\mathbf{par}$  is  $\alpha(m)$ -approximated, if for all instances  $I$  of any size  $n$ ,  $DIS(I) \leq \alpha(m)OPT(I)$ . The smaller  $\alpha(m)$ , the better the partition scheme  $\mathbf{par}$ .

We emphasize here that the central problem in the distributed offline problem lies in finding a good partition scheme. Theorem 3.1 gives an  $m$ -approximated partition, hence establishing an initial benchmark. However, it is far from obvious whether any sublinear approximation ratio can be achieved in the general case. We take such a partition scheme to be an ambitious goal.

## 2.2 The Distributed Online Problem

The crucial (and standard) feature in the distributed online problem is that each request is associated with a release date. More precisely, a problem instance  $I$  is

$$I = \{(r_1, l_1), \dots, (r_n, l_n) \mid r_1 \leq r_2 \leq \dots \leq r_n, r_j \in \mathbb{R}_{\geq 0}, l_j \in \mathbb{M}\},$$

where  $r_j$  and  $l_j$  are the released date and the location of request  $j$ , respectively, and the number of requests  $n$  can vary in different problem instances as in the distributed offline problem. All the servers are assumed to have a unit speed limit and are to collectively visit all requests after or at their release dates.

Due to the additional feature of the requests being online, a static partition of the requests is inadequate in specifying the corresponding cost and how each server moves must also be respected. To formalize it, we first consider a feasible algorithm  $ALG$  that determines the location of each server  $i$  at time  $t \in \mathbb{R}_{\geq 0}$ , denoted  $l^{ALG}(i, t)$ , subject to the initial location constraint  $l^{ALG}(i, 0) = x_i$  and the unit speed limit constraint  $d(l^{ALG}(i, t_2), l^{ALG}(i, t_1)) \leq t_2 - t_1$  for all  $t_2 > t_1 \geq 0$ . Under algorithm  $ALG$ , the completion time of request  $j$ , denoted  $c_j^{ALG}$ , is defined to be the earliest time when one of the servers arrives at the location of the request after or at its release date, i.e.,

$$c_j^{ALG} \triangleq \inf_{t \geq r_j} \{t \mid \exists i, l^{ALG}(i, t) = l_j\}.$$

The cost incurred by server  $i$ , denoted  $ALG_i$ , is defined to be the earliest time at which it returns to its depot after all the requests have been served, i.e.,

$$ALG_i \triangleq \inf_{t \geq \max_{j=1}^n c_j^{ALG}} \{t \mid l^{ALG}(i, t) = x_i\}. \quad (2)$$

The objective is to minimize the sum of all the time costs, given by

$$ALG(I) \triangleq \sum_{i=1}^m ALG_i.$$

To meet the no-communication requirement, we now specialize the  $ALG$  described above to the current distributed online setting. We specify a distributed online algorithm  $DOA$  by prescribing the manner in which the trajectories  $l^{DOA}(i, t)$  are determined. A distributed online algorithm  $DOA$  needs to accomplish two tasks. First, as in the offline case,  $DOA$  must specify a partition scheme **par** as described in Definition 2.1. All the requests that appear in a given region will then only be known by the server to which that region is assigned (and hence no communications between servers are involved). Second,  $DOA$  needs to decide, for each server  $i$ , the trajectory  $l^{DOA}(i, t)$  based on the partially revealed problem instance up to time  $t$  that is in  $M_i$ , i.e.,

$$I_t^i \triangleq \{(r_j, l_j) \mid (r_j, l_j) \in I, r_j \leq t, l_j \in M_i\}.$$

Again, we quantify the performance of a distributed offline algorithm  $DOA$  by comparing the cost of a distributed online algorithm to the optimal offline cost  $OPT(I)$  that is obtained with the knowledge of the entire problem instance  $I$  from the start and full communication capabilities, as formalized below. Note that  $OPT(I)$  in the online case is different from the one in the offline case because the release-date constraints need to be satisfied for the online problem.

**Definition 2.3.** A distributed online algorithm *DOA* is  $c(m)$ -competitive, if for all instances  $I$  of any size  $n$ ,  $DOA(I) \leq c(m)OPT(I)$ . The smaller  $c(m)$ , the better the online algorithm *DOA*.

**Remark 2.2.** In this paper, we disallow the approximation ratios and the competitive ratios to depend on  $n$  by insisting that  $DIS(I) \leq \alpha(m)OPT(I)$  and  $DOA(I) \leq c(m)OPT(I)$  hold for all  $n$ . We do so because we focus on the worst-case problem instance  $I$  that may have  $n$  significantly bigger than  $m$ . More generally, one can allow the ratios to depend on both  $m$  and  $n$ . Depending on the particular setting (i.e., the relation between  $m$  and  $n$ ), a good approximation ratio (and competitive ratio) can be decided accordingly.

The following theorem (proved in Appendix A.1) indicates that solving the distributed offline problem is effectively equivalent to solving the distributed online problem. We will therefore restrict our attention to the offline problem in this paper.

**Theorem 2.1.** *If there is an  $\alpha$ -approximated partition scheme **par** for the distributed offline problem, then there exists an  $(\alpha + 2)$ -competitive distributed online algorithm for the distributed online problem.*

### 3 Static Partition Schemes for the Distributed Offline Problem

#### 3.1 General Depot Configuration

We consider the Voronoi partition *VOR* in which a point  $p \in \mathbb{M}$  is in  $M_i^{VOR}$  if  $x_i$  is the nearest depot for  $p$  (ties broken arbitrarily). The following theorem (proved in Appendix A.2) indicates that  $O(m)$  partition schemes exist for general depot configurations.

**Theorem 3.1.** *The approximation ratio of the Voronoi partition *VOR* is exactly  $m$ .*

The central open problem (Open Problem 4) is whether there exists a partition scheme that does better (sublinear in  $m$ ) than the baseline Voronoi partition in the general case. To make progress towards this direction, we consider two special classes of depot configurations and identify such sublinear partition schemes. We note that the Voronoi partition, when specialized to these two classes of depot configurations, still gives an approximation ratio of  $\Omega(m)$  (see Appendix A.3 for examples).

**Remark 3.1.** These two classes of depot configurations are deliberately chosen as they stand on the two extremes of the general case: the line case (Section 3.2) is the most “stretched-out” depot configuration and the bounded-ratio case (Section 3.3) is the most “clustered” depot configuration. The existence of sublinear partition schemes, although different, in the two extremes leads us to conjecture that there exists a sublinear partition scheme for the general case.

#### 3.2 Depots on A Line

Here we consider the case in which the depots form a line in the metric space, i.e.,  $d(x_i, x_j) + d(x_j, x_k) = d(x_i, x_k)$  for any  $1 \leq i < j < k \leq m$ . For this configuration of depots, we give a partition scheme (the Level partition *LEV*) with an approximation ratio of  $O(\log m)$ . It is not clear whether our analysis is tight: the proposed partition scheme may have a lower asymptotic approximation ratio.

**The Level Partition.** For notational convenience, we assume that there are  $2^k + 1$  servers for some integer  $k \geq 0$ , indexed in order as  $0, 1, 2, \dots, m - 1 (= 2^k)$ . If the number of servers is not  $2^k + 1$ , then we can duplicate the last depot for the filling: creating enough copies (at most  $m - 3$ ) of depot  $m - 1$  so that the total number of depots is brought to  $2^k + 1$ . The partition scheme and analysis still apply.

For each integer  $l = 0, 1, \dots, k - 1$ , let  $\mathbb{N}_l$  denote the set of integers between 1 and  $2^k - 1$  that are multiples of  $2^l$  but not  $2^{l+1}$ , i.e.,  $\mathbb{N}_l \triangleq \{2^l(2t + 1) | t = 0, 1, 2, \dots, 2^{k-l-1} - 1\}$ . As special cases, we denote  $\mathbb{N}_k \triangleq \{2^k\}$  and  $\mathbb{N}_{k+1} \triangleq \{0\}$ . We say that server  $i$  is in level  $l$  if  $i \in \mathbb{N}_l$ , hence the name

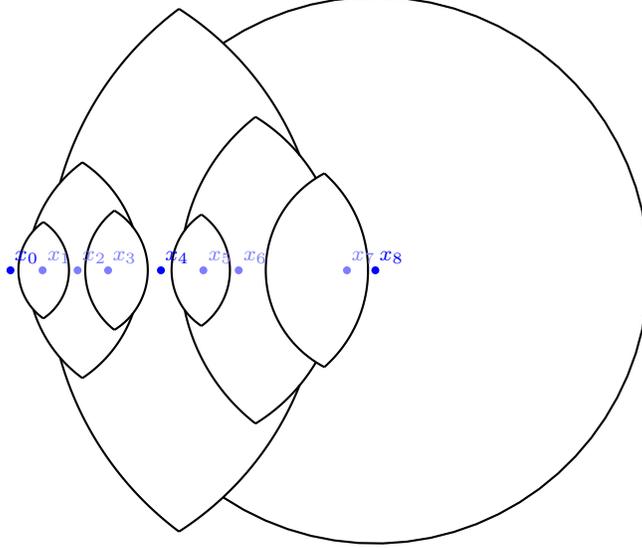


Figure 1: The Level Partition

Level partition.

For  $i = 1, 2, \dots, 2^k - 1$ , we denote by  $\tau_i$  the intersection of two particular closed circular disks:

$$\begin{aligned} \tau_i &\triangleq \{p \in \mathbb{M} \mid d(p, x_{i-2^l}) \leq d(x_{i-2^l}, x_i) + \lambda d(x_{i+2^l}, x_i)\} \\ &\quad \cap \{p \in \mathbb{M} \mid d(p, x_{i+2^l}) \leq d(x_{i+2^l}, x_i) + \lambda d(x_{i-2^l}, x_i)\}, \end{aligned}$$

where  $\lambda \triangleq 3/4$  for simplicity (in fact, any fixed constant in  $(1/2, 1)$  will do) and  $l$  is the integer such that  $i \in \mathbb{N}_l$ . For special cases where  $i = 2^k, 0$ , we define  $\tau_{2^k} \triangleq \{p \in \mathbb{M} \mid d(p, x_{2^k}) \leq \lambda d(x_0, x_{2^k})\}$  and  $\tau_0 \triangleq \mathbb{M}$ .

For each  $i = 0, 1, 2, \dots, 2^k$  with server  $i$  in level  $l$  (i.e.,  $i \in \mathbb{N}_l$ ), we define  $M_i^{LEV}$  to be the points in  $\tau_i$  but not in  $\tau_{i'}$  for any lower-level  $i'$ , i.e.,

$$M_i^{LEV} \triangleq \tau_i \setminus \bigcup_{l'=0}^{l-1} \bigcup_{i' \in \mathbb{N}_{l'}} \tau_{i'}.$$

See Figure 1 for an illustration of the Level partition with  $m = 9$ .

Our main result regarding the Level partition is the following theorem.

**Theorem 3.2.**  $DIS^{LEV}(I) \leq O(\log m)OPT(I)$ .

In order to prove Theorem 3.2, we divide the metric space  $\mathbb{M}$  into  $k+2$  different levels of regions  $L_0, L_1, \dots, L_{k+1}$ , where  $L_l$  is defined to be the union of regions assigned to servers in level  $l$ , i.e.,

$$L_l \triangleq \bigcup_{i \in \mathbb{N}_l} M_i^{LEV}.$$

The definition of  $\{L_l\}_{l=0}^{k+1}$  is used to prove the following lemma.

**Lemma 3.1.** *There exists a constant  $\rho$  such that*

$$LEV(I) \leq \rho OPT(I)$$

*if all requests are in the same level  $L_l$ , i.e.,  $I \subseteq L_l$ , for any integer  $l = 0, 1, \dots, k+1$ .*

Lemma 3.1 is sufficient for proving Theorem 3.2.

*Proof of Theorem 3.2.*

$$LEV(I) = \sum_{l=0}^{k+1} LEV(I \cap L_l) \leq \rho \sum_{l=0}^{k+1} OPT(I \cap L_l) \leq \rho(k+2)OPT(I) = O(\log m)OPT(I)$$

where the last inequality is due to  $I \cap L_l \subseteq I$  for any  $l = 0, 1, \dots, k+1$ .  $\square$

In the rest of this section, we are going to describe the proof sketch for Lemma 3.1 (see Appendix A.4 for the complete proof). To prove Lemma 3.1, we first find the relation between the cost of the optimal centralized algorithm that uses all servers and that of a centralized algorithm that uses only servers with indexes in  $\mathbb{N}_l$ . We call such algorithms *responsible* algorithms because servers in level  $l$  are responsible for all requests located in  $L_l$ . In particular, we have the following lemma (proved in Appendix A.5).

**Lemma 3.2.** *For any integer  $l$ , when  $I \subseteq L_l$ , there exists a responsible algorithm  $RES$  whose cost is at most  $1+1/g$  times of the cost of the optimal solution  $OPT(I)$  for any integer  $l = 0, 1, \dots, k+1$  where  $g$  is defined to be  $1/30$  for simplicity.*

Once we have Lemma 3.2, what is remaining is to find a relation between the cost of the given responsible algorithm ( $RES(I)$ ), and the cost induced by the Level partition ( $DIS^{LEV}(I)$ ). For each server  $i \in \mathbb{N}_l$ , we consider the sequence of the regions  $\tau_j$  with  $j \in \mathbb{N}_l$  visited by server  $i$ . Note that for server  $i$ , the sequence begins and ends at  $\tau_i$ , and if the sequence only consists of one region ( $\tau_i$ ) for all  $i$ , then the responsible algorithm is the same as the optimal distributed algorithm induced by the Level partition. Therefore, we encounter an issue only when the sequence contains multiple regions for some server  $i$ . We say that a responsible algorithm is *non-oscillating* if a zigzag with at least four regions, i.e, pattern  $(\tau_j, \tau_{j'}, \tau_j, \tau_{j'})$ , does not occur in the sequence of the route of any server  $i$ . In a non-oscillating responsible algorithm, it is allowed for each server  $i$  not to travel through the optimal TSP tour given the set of requests that are assigned to it. Given a responsible algorithm  $RES$ , we create a non-oscillating responsible algorithm  $NOS$  as an intermediate step for comparing  $RES(I)$  with  $DIS^{LEV}(I)$ , as shown in the following two lemmas (proved in Appendix A.6 and A.7 respectively).

**Lemma 3.3.** *Given a responsible algorithm  $RES$ , there exists a non-oscillating responsible algorithm  $NOS$  whose cost is at most four times of the cost of  $RES$  when  $I \subseteq L_l$  for any integer  $l = 0, 1, \dots, k+1$ .*

**Lemma 3.4.** *The cost  $DIS^{LEV}(I)$  of the distributed algorithm based on the Level partition is at most  $\frac{25-5\lambda-6\lambda^2}{1-\lambda}$  times of the cost of any given non-oscillating responsible algorithm  $NOS$  when  $I \subseteq L_l$  for any integer  $l = 0, 1, \dots, k+1$ .*

### 3.3 Bounded Ratios between Distances of Depots

Here we consider the case where the ratios between the distances of depots are bounded above by a sublinear function  $f$  of  $m$ , i.e.,

$$\frac{\max_{i,j}\{d(x_i, x_j)\}}{\min_{i,j}\{d(x_i, x_j)\}} \leq f(m).$$

We give a partition scheme (Local partition  $LOC$ ) with an approximation ratio of  $\Theta(f(m))$  (and hence for it to be useful, we assume  $f(m) = o(m)$ ).

**The Local Partition.** Fix an arbitrary ordering of the servers,  $i = 1, 2, \dots, m$ . The region  $M_i^{LOC}$  that each server  $i$  is responsible for centers around, except for server  $m$ , a local region around its depot:

$$M_i^{LOC} \triangleq \left\{ p \in \mathbb{M} \mid d(p, x_i) < \frac{\min_{j',j}\{d(x_{j'}, x_j)\}}{4} \right\}, 1 \leq i \leq m-1,$$

$$M_m^{LOC} \triangleq \mathbb{M} \setminus \bigcup_{i=1}^{m-1} M_i^{LOC}.$$

**Theorem 3.3.** *The Local partition has an approximation ratio of  $\Theta(f(m))$ .*

See Appendix A.8 for the proof.

## 4 Open Problems

Our formulations open up several challenging problems, which we discuss here. We believe that seeking the answers to any of the following open problems, either positive or negative, would be of great theoretical interest and practical use.

**Open Problem 1.** Does there exist a partition scheme with an approximation ratio of  $o(\log(m))$  when the depots form a line?

This open problem is immediate: either by a different partition scheme or tighter analysis of our  $O(\log(m))$ -approximated partition scheme for the line case.

**Open Problem 2.** Does there exist a partition scheme with an approximation ratio of  $o(m)$  when the metric space is the two-dimensional Euclidean space  $\mathbb{R}^2$ ? What about  $\mathbb{R}^3$ ?

Since difficulties can potentially arise when one goes from  $1D$  to  $2D$  and from  $2D$  to  $3D$  (where going from  $3D$  to higher dimensions is typically straightforward), answering this question can be very valuable.

**Open Problem 3.** For a fixed sublinear function  $f$ , under what configurations of depots, does there exist a partition scheme with an approximation ratio of  $\Theta(f(m))$ ?

This problem is a natural generalization of our bounded-ratio result. In particular, the bounded-ratio configuration works for any sublinear  $f$ .

**Open Problem 4.** Does there exist a partition scheme with an approximation ratio of  $o(m)$  for any configuration of depots in any metric space?

Our conjecture here, as mentioned in Remark 3.1, is that such sublinear partition schemes exist.

**Open Problem 5.** Does there exist a partition scheme with an approximation ratio of  $\Theta(1)$  for any configuration of depots in any metric space?

If the answer to Open Problem 5 is yes, then the resulting partition scheme will be a surprisingly good one that completely trivializes the use of communications.

**Variants of The Distributed Online Problem.** Due to the offline-online equivalence result (Theorem 2.1) we have focused exclusively on offline problems. However, there exist variants of the objective function for the online problem where such a reduction is not easily obtained. Below is an example.

Given an online algorithm  $ALG$ , consider the cost incurred by server  $i$  as follows (compare it with Equation (2)).

$$ALG_i \triangleq \inf_{t \geq \max_{(r_j, l_j) \in S_i} \{c_j^{ALG}\}} \{t \mid l^{ALG}(i, t) = x_i\}.$$

For this problem, we are again interested in finding partition schemes for online algorithms that have sublinear competitive ratios. However, it is easy to show that the partition schemes discussed in Definition 2.1 cannot provide such online algorithms (consider the problem instance consisting of  $m$  requests with  $(r_j, l_j) = (TSP(x_1, x_2, \dots, x_m), x_j)$  for all  $j$ ). Therefore, alternative partition schemes that do not require communications between servers need to be taken into consideration, which opens interesting research directions. The time-dependent partition schemes (see Remark 2.1), where the assignment of a request depends on both the location and the release date of the request, are possible candidates for solving this variant of the distributed online problem.

## References

- [1] Arora, S. (1996). Polynomial time approximation schemes for euclidean tsp and other geometric problems. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 2–11. IEEE.
- [2] Ausiello, G., Feuerstein, E., Leonardi, S., Stougie, L., and Talamo, M. (2001). Algorithms for the on-line travelling salesman. *ALGORITHMICA*, 29:2001.
- [3] Baldacci, R., Mingozzi, A., and Calvo, R. W. (2011). An exact method for the capacitated location-routing problem. *Operations research*, 59(5):1284–1296.
- [4] Bansal, N., Buchbinder, N., Madry, A., and Naor, J. (2011). A polylogarithmic-competitive algorithm for the k-server problem. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 267–276. IEEE.
- [5] Bansal, N., Buchbinder, N., and Naor, J. S. (2010). Towards the randomized k-server conjecture: A primal-dual approach. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 40–55. Society for Industrial and Applied Mathematics.
- [6] Bansal, N., Buchbinder, N., and Naor, J. S. (2012). A primal-dual randomized algorithm for weighted paging. *Journal of the ACM (JACM)*, 59(4):19.
- [7] Bartal, Y. and Rosen, A. (1992). The distributed k-server problem—a competitive distributed translator for k-server algorithms. In *Foundations of Computer Science, 1992. Proceedings., 33rd Annual Symposium on*, pages 344–353. IEEE.
- [8] Bektas, T. (2006). The multiple traveling salesman problem: an overview of formulations and solution procedures. *Omega*, 34(3):209–219.
- [9] Bockenhauer, H. J. and Seibert, S. (2000). Improved lower bounds on the approximability of the traveling salesman problem. *RAIRO-Theoretical Informatics and Applications*, 34(03):213–255.
- [10] Bompadre, A., Dror, M., and Orlin, J. B. (2007). Probabilistic analysis of unit-demand vehicle routeing problems. *Journal of Applied Probability*, 44(1):259–278.
- [11] Bonifaci, V. and Stougie, L. (2009). Online k-server routing problems. *Theory Comput.Syst.*, 45(3):470–485.
- [12] Borodin, A., Linial, N., and Saks, M. E. (1992). An optimal on-line algorithm for metrical task system. *Journal of the ACM (JACM)*, 39(4):745–763.
- [13] Carlsson, J. G. (2012). Dividing a territory among several vehicles. *INFORMS Journal on Computing*, 24(4):565–577.
- [14] Carlsson, J. G. and Delage, E. (2013). Robust partitioning for stochastic multivehicle routing. *Operations research*, 61(3):727–744.
- [15] Chao, I.-M., Golden, B. L., and Wasil, E. (1993). A new heuristic for the multi-depot vehicle routing problem that improves upon best-known solutions. *American Journal of Mathematical and Management Sciences*, 13(3-4):371–406.
- [16] Christofides, N. (1976). Worst-case analysis of a new heuristic for the travelling salesman problem. Technical report, DTIC Document.

- [17] Cordeau, J.-F., Gendreau, M., and Laporte, G. (1997). A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks*, 30(2):105–119.
- [18] Dias, M. B., Zlot, R., Zinck, M., Gonzalez, J. P., and Stentz, A. (2004). A versatile implementation of the traderbots approach for multirobot coordination.
- [19] Escobar, J. W., Linfati, R., Toth, P., and Baldoquin, M. G. (2014). A hybrid granular tabu search algorithm for the multi-depot vehicle routing problem. *Journal of Heuristics*, pages 1–27.
- [20] Franceschelli, M., Rosa, D., Seatzu, C., and Bullo, F. (2013). Gossip algorithms for heterogeneous multi-vehicle routing problems. *Nonlinear Analysis: Hybrid Systems*, 10:156–174.
- [21] Gendreau, M., Hertz, A., and Laporte, G. (1994). A tabu search heuristic for the vehicle routing problem. *Management science*, 40(10):1276–1290.
- [22] Gendreau, M., Laporte, G., and Potvin, J.-Y. (2002). Metaheuristics for the capacitated vrp. *The vehicle routing problem*, 9:129–154.
- [23] Giosa, I., Tansini, I., and Viera, I. (2002). New assignment algorithms for the multi-depot vehicle routing problem. *Journal of the operational research society*, 53(9):977–984.
- [24] Jaillet, P. and Wagner, M. R. (2008). Generalized online routing: new competitive ratios, resource augmentation, and asymptotic analyses. *Operations research*, 56(3):745–757.
- [25] Karp, R. (1972). *Reducibility among combinatorial problems*, pages 85–103. Complexity of Computer Computations. Plenum Press.
- [26] Kivelevitch, E., Cohen, K., and Kumar, M. (2013). A market-based solution to the multiple traveling salesmen problem. *Journal of Intelligent and Robotic Systems*, 72(1):21–40.
- [27] Koutsoupias, E. and Papadimitriou, C. H. (1995). On the k-server conjecture. *Journal of the ACM (JACM)*, 42(5):971–983.
- [28] Laporte, G. (2007). What you should know about the vehicle routing problem. *Naval Research Logistics (NRL)*, 54(8):811–819.
- [29] Laporte, G., Nobert, Y., and Arpin, D. (1984). *Optimal solutions to capacitated multidepot vehicle routing problems*. École des hautes études commerciales.
- [30] Laporte, G., Nobert, Y., and Taillefer, S. (1988). Solving a family of multi-depot vehicle routing and location-routing problems. *Transportation science*, 22(3):161–172.
- [31] Lenstra, J. K. and Kan, A. (1981). Complexity of vehicle routing and scheduling problems. *Networks*, 11(2):221–227.
- [32] Li, C.-L. and Simchi-Levi, D. (1990). Worst-case analysis of heuristics for multidepot capacitated vehicle routing problems. *ORSA Journal on Computing*, 2(1):64–73.
- [33] Manasse, M. S., McGeoch, L. A., and Sleator, D. D. (1990). Competitive algorithms for server problems. *Journal of Algorithms*, 11(2):208–230.
- [34] McGeoch, L. A. and Sleator, D. D. (1991). A strongly competitive randomized paging algorithm. *Algorithmica*, 6(1-6):816–825.

- [35] Mitchell, J. S. B. (1999). Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric tsp, k-mst, and related problems. *SIAM Journal on Computing*, 28(4):1298–1309.
- [36] Momke, T. and Svensson, O. (2011). Approximating graphic tsp by matchings. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 560–569. IEEE.
- [37] Mucha, M. (2012). 13/9-approximation for graphic tsp. *Theory of Computing Systems*, pages 1–18.
- [38] Ombuki-Berman, B. and Hanshar, F. T. (2009). *Using genetic algorithms for multi-depot vehicle routing*, pages 77–99. Bio-inspired algorithms for the vehicle routing problem. Springer.
- [39] Papadimitriou, C. H. and Vempala, S. (2000). On the approximability of the traveling salesman problem. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 126–133. ACM.
- [40] Renaud, J., Laporte, G., and Boctor, F. F. (1996). A tabu search heuristic for the multi-depot vehicle routing problem. *Computers and Operations Research*, 23(3):229–235.
- [41] Sebó, A. and Vygen, J. (2012). Shorter tours by nicer ears: 7/5-approximation for graphic tsp, 3/2 for the path version, and 4/3 for two-edge-connected subgraphs. *arXiv preprint arXiv:1201.1870*.
- [42] Svensson, O. (2013). Overview of new approaches for approximating tsp. In *Graph-Theoretic Concepts in Computer Science*, pages 5–11. Springer.

## A Appendix

### A.1 Proof of Theorem 2.1

*Proof.* To simplify the notation, we denote  $R^i$  the set of the locations of all requests in  $M_i$  and  $R_t^i$  the set of the locations of requests in  $M_i$  with release dates at most  $t$ , i.e.,

$$R^i \triangleq \{l_j : (r_j, l_j) \in I, r_j \in M_i\} \text{ and } R_t^i \triangleq \{l_j : (r_j, l_j) \in I_t^i\}.$$

We consider the following distributed online algorithm *DOA*.

- The online algorithm *DOA* adopts the partition scheme **par**.
- At any time  $t$ , if a request is released in  $M_i$ , then server  $i$  stops traveling through the planned route, returns to its depot  $x_i$ , and then follows the route  $TSP_i(R_t^i)$ .

We first find an upper bound on the cost of the algorithm *DOA*. Let  $j$  denote the index of the request with the maximum release date in  $M_i$ . At the release date  $r_j$ , the distance between the location of server  $i$  and its depot  $x_i$  is at most  $r_j$ , and thus the time when server  $i$  returns at its depot and starts the route  $TSP_i(R_{r_j}^i)$  is at most  $2r_j$ , which is at most  $2r_n$ . Moreover,  $R_{r_j}^i = R^i$ . Therefore,  $DOA_i \leq 2r_n + TSP_i(R^i)$ .

Now let us find a lower bounds on the cost of the optimal centralized offline algorithm  $OPT(I)$ . First we notice that the completion time of request  $n$  is lower bounded by its release date, i.e.,  $c_n^{OPT} \geq r_n$ , and therefore  $OPT_i \geq r_n$  for all  $i = 1, 2, \dots, m$ . As a result,  $OPT(I) \geq mr_n$ . In addition, according to the presumption of this lemma,  $\alpha OPT(I) \geq \sum_{i=1}^m TSP_i(R^i)$ .

Combine the results above, we have

$$DOA(I) = \sum_{i=1}^m DOA_i \leq 2mr_n + \sum_{i=1}^m TSP_i(R^i) \leq 2OPT(I) + \alpha OPT(I).$$

Hence, we conclude that the competitive ratio of the distributed online algorithm *DOA* is at most  $\alpha + 2$ .  $\square$

### A.2 Proof of Theorem 3.1

*Proof.* Bompadre *et al.* [10] give an example demonstrating that the Voronoi partition has an approximation ratio of at least  $m$ . Therefore, it suffices to prove that the approximation ratio of the Voronoi partition is at most  $m$ .

We first consider the route of any server (without loss of generality and for convenience, say Server 1) under the Voronoi partition, and show that the following holds:

$$TSP_1(S_1^{VOR}) \leq OPT(I), \tag{3}$$

where  $S_1^{VOR}$ , as defined in Section 2.1, is the set of requests assigned to Server 1 under the Voronoi partition.

Since Server 1 in Inequality (3) can be replaced by any server, we have

$$DIS^{VOR}(I) = \sum_{i=1}^m TSP_i(S_i^{VOR}) \leq mOPT(I),$$

which proves the theorem.

We now prove Inequality (3) now. Without loss of generality, assume for some integer  $k$ ,  $S_i^{OPT} \cap S_1^{VOR} \neq \emptyset$  for all  $i = 2, 3, \dots, k$ , and  $S_i^{OPT} \cap S_1^{VOR} = \emptyset$  for all  $i = k + 1, k + 2, \dots, m$ ,

where  $S_i^{OPT}$ , as defined in Section 2.1, is the set of requests assigned to Server  $i$  under the optimal partition  $OPT$ .

Therefore, for each  $i \in \{2, 3, \dots, k\}$ , there exists at least one request  $p$  in the set  $S_i^{OPT}$  such that  $d(x_1, p) \leq d(x_i, p)$ . For each  $i$ , among all such “closer-to-deopt  $x_1$ ” requests, we denote the first and last requests visited by  $TSP_i(S_i^{OPT})$  (the route of Server  $i$  under  $OPT$ ) by  $a_i$  and  $b_i$  respectively, where  $a_i$  and  $b_i$  coincide if there is only one such request.

We create as follows a route for Server 1 that visits all requests in  $S_1^{VOR}$  based on the routes of servers  $1, 2, \dots, k$  under the optimal algorithm  $OPT$ . First of all, let Server 1 follow  $TSP_1(S_1^{OPT})$ , the route of Server 1 in the optimal algorithm. For each  $i = 2, 3, \dots, k$ , let Server 1 follow the additional round trip that begins and ends at the depot  $x_1$ .

1. Travel from  $x_1$  to  $a_i$ .
2. Travel from  $a_i$  to  $b_i$  using  $TSP_i(S_i^{OPT})$ .
3. Travel from  $b_i$  to  $x_1$ .

It is clear that by doing so, Server 1 visits all requests in  $S_1^{VOR}$ . Therefore, the length of the route of Server 1 defined above is an upper bound of  $TSP_1(S_1^{VOR})$ . For each  $i \in \{2, 3, \dots, k\}$ , the length of the additional route because of requests in  $S_i^{OPT}$  is at most  $TSP_i(S_i^{OPT})$  because  $d(x_1, a_i) \leq d(x_i, a_i)$  and  $d(x_1, b_i) \leq d(x_i, b_i)$ . Therefore,

$$TSP_1(S_1^{VOR}) \leq \sum_{i=1}^k TSP_i(S_i^{OPT}) \leq OPT(I),$$

which establishes Inequality (3). □

### A.3 $\Omega(m)$ -approximation Examples for The Voronoi Partition

For the line case, we provide the following example. Let the metric space be the two dimensional Euclidean space and  $x_i = (0, i)$  for  $i = 1, 2, \dots, m$ . Let  $I$  to be the problem instance that consists of  $m$  requests where  $l_j = (k, j)$  for some constant  $k$ . Clearly,  $DIS^{VOR}(I) = 2km$  and  $OPT(I) \leq 2k + 2m$ . When  $k \rightarrow \infty$ , the ratio  $DIS^{VOR}(I)/OPT(I)$  approaches  $m$ .

For the bounded-ratio case, we provide the following example with  $f(m) = 1$ . Let the metric space be the  $m$ -dimensional Euclidean space and  $x_i = e_i$  where  $e_i$  the  $m$ -dimensional vector that has the value of 1 in the  $i^{\text{th}}$  dimension and 0 in all other dimensions. Let  $I$  to be the problem instance that consists of  $m$  requests where  $l_j = \epsilon e_j$  for some constant  $\epsilon > 0$  for each  $j = 1, 2, \dots, m$ . Clearly,  $DIS^{VOR}(I) = 2m(1 - \epsilon)$ . On the other hand, the cost of the optimal algorithm is at most that of the algorithm that assigns all requests to the same server, which is upper bounded by  $2 + 2m\epsilon$ . Therefore, the ratio  $DIS^{VOR}(I)/OPT(I)$  approaches  $m$  as  $\epsilon \rightarrow 0^+$ .

### A.4 Proof of Lemma 3.1

*Proof.* With Lemmas 3.2, 3.3, and 3.4, we can prove Lemma 3.1 as follows.

$$\begin{aligned} LEV(I) &\leq \frac{25 - 5\lambda - 6\lambda^2}{1 - \lambda} NOS(I) \leq 4 \times \frac{25 - 5\lambda - 6\lambda^2}{1 - \lambda} RES(I) \\ &\leq \left(1 + \frac{1}{g}\right) \frac{100 - 20\lambda - 24\lambda^2}{1 - \lambda} OPT(I) \leq 9000 OPT(I) \end{aligned}$$

where  $\lambda = 3/4$  and  $g = 1/30$ . Therefore, the lemma holds when  $\rho$  is chosen to be, for example, 9000. □

## A.5 Proof of Lemma 3.2

*Proof.* To simplify the discussion, we prove only the cases in which  $l = 0, 1, 2, \dots, k-1$ . However, it is clear that the proof can be modified and applied to the cases where  $l = k, k+1$ .

We define the following responsible algorithm  $RES$ . For each  $i' = 0, 1, 2, \dots, 2^k$ , if  $S_{i'}^{OPT}$  is not empty, i.e., there are requests served by server  $i'$  in the optimal algorithm  $OPT$ , then we assign all requests in  $S_{i'}^{OPT}$  to the server with the minimum index  $i$  in  $\mathbb{N}_l$  such that  $S_{i'}^{OPT}$  contains a request in  $M_i^{LEV}$ . This is equivalent to the following definition.

$$S_i^{RES} \triangleq \bigcup_{i' \in A_i} S_{i'}^{OPT}$$

where

$$A_i \triangleq \{j : S_j^{OPT} \cap M_i^{LEV} \neq \emptyset \text{ and } S_j^{OPT} \cap M_{i'}^{LEV} = \emptyset \text{ for all } i' < i, i' \in \mathbb{N}_l\}.$$

Now we define the route of each server  $i$  in  $\mathbb{N}_l$  separately. Let us denote  $i_{\max}$  and  $i_{\min}$  the maximum and minimum index in  $A_i$  respectively. The route of server  $i$  travels from  $x_i$  to  $x_{i_{\max}}$ , then to  $x_{i_{\min}}$ , and finally back to  $x_i$  through the shortest path that passes through depots  $x_{i'}$  for all  $i_{\min} \leq i' \leq i_{\max}$ . In addition, server  $i$  travels through  $TSP_{i'}(S_{i'}^{OPT})$  when passing by depot  $x_{i'}$  for the first time before going to the next depot if  $i' \in A_i$ . Clearly, the cost of  $RES$  can be calculated as follows.

$$RES(I) = \sum_{i \in \mathbb{N}_l} \left( \sum_{i' \in A_i} TSP_{i'}(S_{i'}^{OPT}) + 2d(x_{i_{\min}}, x_{i_{\max}}) \right)$$

We will prove the following claim.

$$d(p, x_{i_{\max}}) \geq gd(x_i, x_{i_{\max}}) \tag{4}$$

for any  $p \in M_i^{LEV}$ .

Before proving Claim (4), we first show that Claim (4) implies the lemma.

If Claim (4) is true, then

$$TSP_{i_{\max}}(S_{i_{\max}}^{OPT}) \geq 2d(p, x_{i_{\max}}) \geq 2gd(x_i, x_{i_{\max}})$$

for any point  $p$  in  $S_{i_{\max}}^{OPT} \cap M_i^{LEV}$ . By symmetry, the same result holds for  $i_{\min}$ . Therefore,

$$\frac{1}{g} \sum_{i' \in A_i} TSP_{i'}(S_{i'}^{OPT}) \geq \frac{1}{g} (TSP_{i_{\min}}(S_{i_{\min}}^{OPT}) + TSP_{i_{\max}}(S_{i_{\max}}^{OPT})) \geq 2d(x_{i_{\min}}, x_{i_{\max}})$$

when  $i_{\min} \neq i_{\max}$ . On the other hand, if  $i_{\min} = i_{\max}$ , then  $i_{\min} = i_{\max} = i$ . Hence,

$$\frac{1}{g} \sum_{i' \in A_i} TSP_{i'}(S_{i'}^{OPT}) \geq 0 = 2d(x_{i_{\min}}, x_{i_{\max}}).$$

As a result, given Claim (4), the lemma can be proven as follows.

$$\begin{aligned} RES(I) &= \sum_{i \in \mathbb{N}_l} \left( \sum_{i' \in A_i} TSP_{i'}(S_{i'}^{OPT}) + 2d(x_{i_{\min}}, x_{i_{\max}}) \right) \\ &\leq \left(1 + \frac{1}{g}\right) \sum_{i \in \mathbb{N}_l} \left( \sum_{i' \in A_i} TSP_{i'}(S_{i'}^{OPT}) \right) \\ &= \left(1 + \frac{1}{g}\right) OPT(I). \end{aligned}$$

Let us now prove Claim (4). To simplify the notation, let us denote  $j = i + 2^l$ . First we consider case where  $i_{\max} \geq j$ . In this case,

$$\begin{aligned} d(p, x_{i_{\max}}) &\geq d(x_{i-2^l}, x_{i_{\max}}) - d(x_{i-2^l}, p) \geq d(x_{i-2^l}, x_{i_{\max}}) - (d(x_{i-2^l}, x_i) + \lambda d(x_i, x_j)) \\ &= (1 - \lambda)d(x_i, x_j) + d(x_j, x_{i_{\max}}) \\ &\geq (1 - \lambda)d(x_i, x_{i_{\max}}) \geq gd(x_i, x_{i_{\max}}). \end{aligned} \quad (5)$$

Let us now consider the other case where  $i < i_{\max} < j$  (this case is possible only if  $l \geq 1$ ). Inequality (5) with  $i_{\max} = j$  gives us

$$d(p, x_j) \geq (1 - \lambda)d(x_i, x_j), \quad (6)$$

which motivates us to distinguish cases further based on the ratio between  $d(x_i, x_{i_{\max}})$  and  $d(x_i, x_j)$ . Define the threshold of the ratio to be  $f \triangleq 7/8$  for simplicity. In fact, any fixed constant in  $(\lambda, 1)$  will do (for possibly a different positive real number  $g$ ).

If  $d(x_i, x_{i_{\max}}) > fd(x_i, x_j)$ , according to the triangle inequality and Inequality (6),

$$\begin{aligned} d(p, x_{i_{\max}}) &\geq d(p, x_j) - d(x_j, x_{i_{\max}}) > (1 - \lambda)d(x_i, x_j) - (1 - f)d(x_i, x_j) \\ &= (f - \lambda)d(x_i, x_j) > (f - \lambda)d(x_i, x_{i_{\max}}) \geq gd(x_i, x_{i_{\max}}). \end{aligned}$$

What is remaining is the case where  $1 \leq i_{\max} \leq j - 1$  and  $d(x_i, x_{i_{\max}}) \leq fd(x_i, x_j)$ . This case implies that

$$d(x_{i_{\max}}, x_j) \geq \frac{1 - f}{f}d(x_i, x_{i_{\max}}),$$

which we use frequently in the remaining of proof. We distinguish two cases.

1.  $1 \leq i_{\max} \leq i + 2^{l-1}$ .

In this case, there exists a positive integer  $t' \leq l - 1$  such that  $i + 2^{t'-1} \leq i_{\max} \leq i + 2^{t'}$ . According to the definition of  $M_i^{LEV}$ ,  $p$  is not in  $\tau_{i+2^t}$  for any  $t = 0, 1, \dots, l-1$ . Thus, for each  $t = t' - 1, t', \dots, l-1$ , the point  $p$  in  $M_i^{LEV}$  must violate one of the following two constraints.

$$d(p, x_i) \leq d(x_i, x_{i+2^t}) + \lambda d(x_{i+2^t}, x_{i+2^{t+1}}). \quad (7)$$

$$d(p, x_{i+2^{t+1}}) \leq d(x_{i+2^{t+1}}, x_{i+2^t}) + \lambda d(x_i, x_{i+2^t}). \quad (8)$$

We distinguish three cases.

(a) Constraint (8) is not violated for  $t = l - 1$ .

In this case, Constraint (7) is violated for  $t = l - 1$ . As a result,

$$d(p, x_i) > d(x_i, x_{i+2^{l-1}}) + \lambda d(x_{i+2^{l-1}}, x_{i+2^l}).$$

Therefore,

$$\begin{aligned} d(p, x_{i_{\max}}) &\geq d(p, x_i) - d(x_i, x_{i_{\max}}) \\ &\geq d(x_i, x_{i+2^{l-1}}) + \lambda d(x_{i+2^{l-1}}, x_{i+2^l}) - d(x_i, x_{i_{\max}}) \\ &\geq d(x_{i_{\max}}, x_{i+2^{l-1}}) + \lambda d(x_{i+2^{l-1}}, x_{i+2^l}) \\ &\geq \lambda d(x_{i_{\max}}, x_j) \geq \frac{1 - f}{f} \lambda d(x_i, x_{i_{\max}}) \geq gd(x_i, x_{i_{\max}}). \end{aligned}$$

(b) Constraint (7) is not violated for  $t = t' - 1$ .

In this case, Constraint (8) is violated for  $t = t' - 1$ . Therefore,

$$d(p, x_{i+2^{t'}}) > d(x_{i+2^{t'}}, x_{i+2^{t'-1}}) + \lambda d(x_i, x_{i+2^{t'-1}}).$$

As a result,

$$\begin{aligned} d(p, x_{i_{\max}}) &\geq d(p, x_{i+2^{t'}}) - d(x_{i_{\max}}, x_{i+2^{t'}}) \\ &> d(x_{i_{\max}}, x_{i+2^{t'-1}}) + \lambda d(x_i, x_{i+2^{t'-1}}) \\ &> \lambda d(x_i, x_{i_{\max}}) \geq \frac{f}{1-f} g d(x_i, x_{\max}) \end{aligned}$$

where  $\frac{f}{1-f} > 1$ . We keep the constant  $\frac{f}{1-f}$  to simplify the proof for case 2.

(c) Constraint (8) is violated for  $t + 1$  and Constraint (7) is violated for  $t$  for some  $t = t' - 1, t', \dots, l - 2$ . In this case,

$$\begin{aligned} d(p, x_{i+2^{t+2}}) &> d(x_{i+2^{t+2}}, x_{i+2^{t+1}}) + \lambda d(x_i, x_{i+2^{t+1}}) \text{ and} \\ d(p, x_i) &> d(x_i, x_{i+2^t}) + \lambda d(x_{i+2^t}, x_{i+2^{t+1}}). \end{aligned}$$

Therefore,

$$\begin{aligned} 2d(p, x_{i_{\max}}) &\geq d(p, x_{i+2^{t+2}}) - d(x_{i_{\max}}, x_{i+2^{t+2}}) + d(p, x_i) - d(x_{i_{\max}}, x_i) \\ &> \lambda d(x_i, x_{i+2^t}) + (2\lambda - 1)d(x_{i+2^t}, x_{i+2^{t+1}}) > (2\lambda - 1)d(x_i, x_{i+2^{t+1}}) \\ &\geq (2\lambda - 1)d(x_i, x_{i_{\max}}) \geq 2\frac{f}{1-f} g d(x_i, x_{\max}) \end{aligned}$$

where  $\frac{f}{1-f} > 1$ . We keep the constant  $\frac{f}{1-f}$  to simplify the proof for case 2.

2.  $i + 2^{l-1} < i_{\max} \leq j - 1$ .

In this case, there exists a positive integer  $t' \leq l - 1$  such that  $j - 2^{t'} < i_{\max} \leq j - 2^{t'-1}$ . According to the definition of  $M_i^{LEV}$ ,  $p$  is not in  $\tau_{j-2^t}$  for any  $t = t' - 1, t', \dots, l - 1$ . Note that we did not use  $p \in \tau_i$  in the proof of case 1 where  $i_{\max} \leq i + 2^{l-1}$ . Therefore, the cases 1 and 2 are symmetric to each other (with  $i$  and  $j$  swapped). In the case symmetric to case 1a, we have

$$d(p, x_{i_{\max}}) \geq \lambda d(x_{i_{\max}}, x_i) \geq g d(x_i, x_{i_{\max}}).$$

For the cases symmetric to cases 1b and 1c, we have

$$d(p, x_{i_{\max}}) \geq \frac{f}{1-f} g d(x_j, x_{i_{\max}}) \geq g d(x_i, x_{i_{\max}}).$$

Since we have covered all possible cases, the proof is completed.  $\square$

## A.6 Proof of Lemma 3.3

*Proof.* We define the route of each server  $i$  in *NOS* separately, and prove that for each server  $i$ , the non-isolating route that we defined is at most four times of the route defined in the given algorithm *RES*.

To describe the non-oscillating route for server  $i$ , we first denote the sequence of the regions visited by server  $i$  to be  $(\tau_{i_1}, \tau_{i_2}, \dots, \tau_{i_q})$  where  $q$  is the length of the sequence. According to this definition,  $i_1 = i_q = i$ , and  $i_j \neq i_{j+1}$  for any  $j = 1, 2, \dots, q - 1$ .

For each  $j = 1, 2, \dots, q$ , denote  $a_j$  and  $b_j$  the first and last points (each point could be a request or the depot  $x_i$ ) corresponding to the region  $\tau_{i_j}$  visited by server  $i$  under the algorithm *RES*. Note that according to the definition,  $a_1 = b_q = x_i$ .

Now we consider the case where the route is oscillating between the  $t^{\text{th}}$  region and the  $t'^{\text{th}}$  region and  $t' - t \geq 3$ , i.e.,  $i_j = i_{j+2}$  for  $j = t, \dots, t' - 1$ , but  $i_j \neq i_{j+2}$  for  $j = t - 1, t'$ . For the case where  $t' - t$  is odd, we will define an alternative routes that starts at  $a_t$ , ends at  $b_{t'}$ , and does not oscillate at all. For the case where  $t' - t$  is even, we adopt the alternative route for the sequence  $(\tau_{i_t}, \tau_{i_{t+1}}, \dots, \tau_{i_{t'-1}})$ , and then the new route oscillates only in three regions starting at point  $a_t$  and ends at point  $b_{t'}$ .

Let us now define an alternative route for the case where  $t' - t$  is odd. To simplify the notation, we denote  $r \rightarrow r'$  to be the route described in *RES* that travels between the two requests (or between a request and a depot)  $r$  and  $r'$ , and  $r \dashrightarrow r'$  to be the shortest path to travel from location  $r$  to location  $r'$ . In *NOS*, the alternative path of server  $i$  follows the following three steps.

1. Travel through the requests in  $\tau_{i_t}, \tau_{i_{t+2}}, \dots, \tau_{i_{t'-1}}$  through the order that is the same as *RES*, i.e.,

$$a_t \rightarrow b_t \dashrightarrow a_{t+2} \rightarrow b_{t+2} \dashrightarrow a_{t+4} \rightarrow b_{t+4} \cdots \dashrightarrow a_{t'-1} \rightarrow b_{t'-1}.$$

2. Go to location  $a_{t+1}$ , i.e.,

$$b_{t'-1} \dashrightarrow a_{t+1}.$$

3. Travel through the requests in  $\tau_{i_{t+1}}, \tau_{i_{t+3}}, \dots, \tau_{i_{t'}}$  through the order that is the same as *RES*, i.e.,

$$a_{t+1} \rightarrow b_{t+1} \dashrightarrow a_{t+3} \rightarrow b_{t+3} \dashrightarrow a_{t+5} \rightarrow b_{t+5} \cdots \dashrightarrow a_{t'} \rightarrow b_{t'}.$$

It is clear that the alternative route does not oscillate. Note that the solid arrows are also in the original route of *RES* and they do not duplicate. Therefore, the solid arrows do not increase the length of the route. The length of the dashed arrows in each of the three steps is smaller than the length of the route that travels from  $b_t$  to  $a_{t'}$  in the route of *RES*. In addition, the route between  $b_t$  and  $a_{t'}$  does not intersect with different parts of the route that oscillate, even for other servers. Therefore, the total additional length due to all dashed arrows for all servers is at most  $3RES(I)$ . As a result, we conclude that  $NOS(I) \leq 4RES(I)$ .  $\square$

## A.7 Proof of Lemma 3.4

*Proof.* Because

$$\sum_{i=1}^m DIS^{LEV}(S_i) \geq DIS^{LEV}\left(\bigcup_{i=1}^m S_i\right)$$

for any sets of requests  $S_1, \dots, S_m$ , it is sufficient to prove the following inequality.

$$DIS^{LEV}(S_i^{NOS}) \leq \frac{25 - 5\lambda - 6\lambda^2}{1 - \lambda} (\text{length of the route of server } i \text{ in } NOS)$$

for each  $i = 1, 2, \dots, m$ .

Given the route of server  $i$  in *NOS*, we will create an algorithm *LEV'* such that any request in  $S_i^{NOS} \cap M_j^{LEV}$  is assigned to server  $j$  where  $S_i^{NOS}$ , as defined in Section 2.1, is the set of the requests that are assigned to server  $i$  under the algorithm *NOS*.

After describing the algorithm *LEV'*, we will prove the following claim.

$$LEV'(S_i^{NOS}) \leq \frac{25 - 5\lambda - 6\lambda^2}{1 - \lambda} (\text{length of the route of server } i \text{ in } NOS). \quad (9)$$

If Claim (9) holds, then the lemma holds because each server in  $DIS^{LEV}(S_i^{NOS})$  travels through the optimal TSP tour, and thus the cost  $DIS^{LEV}(S_i^{NOS})$  is not greater than  $LEV'(S_i^{NOS})$ .

Let us now describe how we create the algorithm  $LEV'$ . Given the route of server  $i$  in  $NOS$ , we define  $q$  and  $\{a_j, b_j, i_j\}_{j=1}^q$  in the same way as we defined in the proof of Lemma 3.3. If  $q = 1$ , set  $LEV' = NOS$ , then we are done. Therefore, we assume  $q > 1$ . We first note that when  $q > 1$ ,  $q \geq 3$  because  $i_1 = i_q = i$  but  $i_1 \neq i_2$ . For  $j = 1, q$ , we add segment  $(b_1, x_i)$  and  $(x_i, a_q)$  respectively. For each  $j = 2, 3, \dots, q-1$ , we add two segments  $(x_{i_j}, a_j)$  and  $(b_j, x_{i_j})$ . By doing so, each server  $j \in \mathbb{N}_l$  can follow a route that begins and ends at the depot  $x_j$  and visit all requests in  $I \cap M_j^{LEV}$ . Hence, we have successfully defined a valid algorithm  $LEV'$ .

We are now ready to prove Claim (9). The total length that we added is

$$d(b_1, x_i) + d(x_i, a_q) + \sum_{j=2}^{q-1} d(x_{i_j}, a_j) + d(b_j, x_{i_j}).$$

It is sufficient to prove that this quantity is at most  $\frac{24-6\lambda-6\lambda^2}{1-\lambda}$  times of the length of the route of server  $i$  in  $NOS$ . To prove this, we consider each  $d(x_{i_j}, a_j)$  and  $d(x_{i_j}, b_j)$  separately. For each  $j = 3, 4, \dots, q-2$ , we prove the following claim.

$$d(x_{i_j}, a_j) \leq \frac{4-\lambda-\lambda^2}{1-\lambda}(d(a_{j-1}, a_j) + d(a_j, a_{j+1}) + d(a_{j+1}, a_{j+2})). \quad (10)$$

We skip the proof of the cases where  $j = 1, 2, q-1, q$  but similar results hold for those cases. By symmetry, similar results hold when each of the  $a$  appeared in Claim (10) is replaced with  $b$ . If we have Claim (10), we can prove the lemma by summing  $d(x_{i_j}, a_j) + d(x_{i_j}, b_j)$  over all  $j$ .

Before proving Claim (10), we first prove the following two useful inequalities. Given three integers  $t_1, t_2, t_3$  in  $\mathbb{N}_l$  such that  $t_1 < t_2 < t_3$  and three points  $p_1 \in \tau_{t_1}$ ,  $p_2 \in \tau_{t_2}$ , and  $p_3 \in \tau_{t_3}$ , we have

$$d(x_{t_2}, p_2) \leq \frac{2+\lambda}{2}d(p_1, p_3) \quad (11)$$

$$d(x_{t_3}, p_3) \leq \frac{4-\lambda-\lambda^2}{2-2\lambda}(d(p_1, p_3) + d(p_2, p_3)) \quad (12)$$

Let us first prove Inequality (11). To prove this, we first find the following upper bound for  $d(x_{t_2}, p_2)$ .

$$\begin{aligned} 2d(x_{t_2}, p_2) &\leq d(x_{t_2+2^l}, p_2) + d(x_{t_2+2^l}, x_{t_2}) + d(x_{t_2-2^l}, p_2) + d(x_{t_2-2^l}, x_{t_2}) \\ &\leq (2+\lambda)d(x_{t_2+2^l}, x_{t_2-2^l}). \end{aligned}$$

Then, we find the following lower bound for  $d(p_1, p_3)$ .

$$\begin{aligned} d(p_1, p_3) &\geq d(x_{t_1-2^l}, x_{t_3+2^l}) - d(x_{t_1-2^l}, p_1) - d(x_{t_3+2^l}, p_3) \\ &> d(x_{t_1+2^l}, x_{t_3-2^l}) + (1-\lambda)d(x_{t_1+2^l}, x_{t_1}) + (1-\lambda)d(x_{t_3-2^l}, x_{t_3}) \\ &> d(x_{t_2+2^l}, x_{t_2-2^l}). \end{aligned} \quad (13)$$

Combine the two inequalities above, we obtain Inequality (11).

Inequality (12) is a direct result of the following inequality.

$$\begin{aligned} d(x_{t_3}, p_3) &\leq d(x_{t_3}, x_{t_2}) + d(x_{t_2}, p_2) + d(p_2, p_3) \\ &\leq \frac{1}{1-\lambda}d(p_1, p_3) + \frac{2+\lambda}{2}d(p_1, p_3) + d(p_2, p_3) \end{aligned}$$

where the last inequality follows from Inequalities (11) and (13).

We are now ready to prove Claim (10). Without loss of generality, we assume that  $i_{j-1} < i_j$ . We distinguish three cases.

1.  $i_{j+1} < i_j$  and  $i_{j-1} \neq i_{j+1}$ . In this case, according to Inequality (12),

$$d(x_{i_j}, a_j) \leq \frac{4 - \lambda - \lambda^2}{2 - 2\lambda} (d(a_{j-1}, a_j) + d(a_j, a_{j+1})).$$

2.  $i_j < i_{j+1}$ . In this case, according to Inequality (11),

$$d(x_{i_j}, a_j) \leq \frac{2 + \lambda}{2} d(a_{j-1}, a_{j+1}) \leq \frac{2 + \lambda}{2} (d(a_{j-1}, a_j) + d(a_j, a_{j+1})).$$

3.  $i_{j-1} = i_{j+1}$ . In this case, we consider  $i_{j+2}$ . Because the route is non-oscillating,  $i_{j+2} \neq i_j$ . We further distinguish two cases.

- (a)  $i_{j+2} < i_j$ . In this case, according to Inequality (12),

$$\begin{aligned} d(x_{i_j}, a_j) &\leq \frac{4 - \lambda - \lambda^2}{2 - 2\lambda} (d(a_j, a_{j+2}) + d(a_j, a_{j+1})) \\ &\leq \frac{4 - \lambda - \lambda^2}{1 - \lambda} (d(a_{j+1}, a_{j+2}) + d(a_j, a_{j+1})). \end{aligned}$$

- (b)  $i_{j+2} > i_j$ . In this case, according to Inequality (11),

$$d(x_{i_j}, a_j) \leq \frac{2 + \lambda}{2} d(a_{j+1}, a_{j+2}).$$

Hence the proof is completed. □

### A.8 Proof of Theorem 3.3

*Proof.* Without loss of generality, let us assume that the minimum distance between a pair of depots is one, i.e.,  $\min_{i,j} \{d(x_i, x_j)\} = 1$ .

Let us first prove that the approximation ratio given by the partition scheme *LOC* is  $\Omega(f(m))$ . Consider the following example. Let the metric space be  $\mathbb{R}$ ,  $m = 3$  and  $x_1 = 0$ ,  $x_2 = 1$ ,  $x_3 = f(m) + 1$ . Let there be only one request and the request is located at  $1 + 1/4$ . In this case, the request will be assigned to server 3, and the cost of the distributed algorithm is  $DIS^{LOC}(I) = 2(f(m) - 1/4)$ . On the other hand, the optimal algorithm can assign the request to server 2 and the cost would be  $OPT(I) = 1/2$ . As a result, the ratio  $DIS^{LOC}(I)/OPT(I) = 4f(m) - 1$ , which is  $\Theta(f(m))$ . Therefore, the partition scheme *LOC* has an approximation ratio of  $\Omega(f(m))$ .

Let us prove that the partition scheme *LOC* leads to an approximation ratio of  $O(f(m))$  by showing that it is at most  $2 + 4f(m)$ .

We first divide the requests into two sets  $L_1 = I \cap \bigcup_{i=1}^{m-1} M_i$ , and  $L_2 = I \cap M_m$ . We have  $DIS^{LOC}(I) = DIS^{LOC}(L_1) + DIS^{LOC}(L_2)$ ,  $OPT(I) \geq OPT(L_1)$ , and  $OPT(I) \geq OPT(L_2)$ . Therefore, the following two claims are sufficient to proving the theorem.

$$DIS^{LOC}(L_1) = OPT(L_1) \tag{14}$$

and

$$DIS^{LOC}(L_2) \leq (1 + 4f(m))OPT(L_2). \tag{15}$$

We first prove Claim (14). We prove this by showing that, when  $I = L_1$ , in the optimal algorithm  $OPT$ ,  $S_i^{OPT}$  consists only requests in  $M_i^{LOC}$  for any  $i = 1, 2, \dots, m-1$ . Using the same argument for proving this, it will be evident that  $S_m^{OPT}$  is empty. To prove this, we first note that the distance between two points in different regions is greater than  $1/2$ . It is because if  $p_1 \in M_i^{LOC}$ ,  $p_2 \in M_j^{LOC}$ , and  $i \neq j$ , then

$$d(p_1, p_2) \geq d(x_i, x_j) - d(x_i, p_1) - d(x_j, p_2) > 1 - 1/4 - 1/4 = 1/2.$$

Now let us assume on the contrary that server  $i$  serves at least one request not in  $M_i^{LOC}$  in the optimal algorithm. Assume that after leaving  $M_i^{LOC}$ , server  $i$  visits requests in regions in the order of  $M_{i_1}^{LOC}$ ,  $M_{i_2}^{LOC}$ ,  $\dots$ , and  $M_{i_l}^{LOC}$  and then back to  $M_i^{LOC}$ . For each  $l' = 1, 2, \dots, l$ , denote  $a_{i_{l'}}$  and  $b_{i_{l'}}$  the first and last requests visited in the region  $M_{i_{l'}}^{LOC}$  ( $a_{i_{l'}} = b_{i_{l'}}$  if there is only one such request). To simplify the notation, we denote  $b_{i_0}$  the last point (request or depot) server  $i$  visits before leaving  $M_i^{LOC}$  and  $a_{i_{l+1}}$  the first point (request or depot) server  $i$  visits after returning  $M_i^{LOC}$ .

Now we consider alternative routes of servers by removing the  $(l+1)$  edges  $\{(b_{i_{l'}}, a_{i_{l'+1}})\}_{l'=0}^l$  and adding the  $(2l+2)$  edges  $(x_i, b_{i_0})$ ,  $(x_i, a_{i_{l+1}})$ , and  $\{(x_{i_{l'}}, a_{i_{l'}}), (x_{i_{l'}}, b_{i_{l'}})\}_{l'=1}^l$ .

The length of each removed edge is greater than  $1/2$ , and the length of each added edge is at most  $1/4$ . Therefore, the length of the new solution is smaller than that of the optimal solution, which is a contradiction. Hence, Claim 14 is true.

Let us now prove Claim (15). The inequality obviously holds if  $S_i^{OPT}$  is empty for all  $i = 1, 2, \dots, m-1$ . If  $S_i^{OPT}$  is not empty for any  $i = 1, 2, \dots, m-1$ , let  $a_i$  and  $b_i$  be the first and last such request in the route of server  $i$  in the optimal solution ( $a_i = b_i$  if there is only one such request). For each  $i = 1, 2, \dots, m-1$ , we replace  $(a_i, x_i)$  and  $(b_i, x_i)$  with  $(a_i, x_m)$  and  $(b_i, x_m)$  such that all requests in  $S_i^{OPT}$  are served by server  $m$  in the new solution. Clearly, the length of the new solution is an upper bound of  $TSP_m(L_2)$ , which is the same as  $DIS^{LOC}(L_2)$ .

We note that for any  $p \in L_2$ ,  $\frac{d(p, x_m)}{d(p, x_i)} < 1 + 4f(m)$  because

$$\frac{d(p, x_m)}{d(p, x_i)} \leq \frac{d(p, x_i) + d(x_i, x_m)}{d(p, x_i)} < 1 + 4f(m).$$

Therefore, the length of the new solution is at most  $(1 + 4f(m))OPT(L_2)$ . Thus, Claim (15) holds.  $\square$