An Asymptotically Optimal Algorithm for the Dynamic Traveling Repair Problem

UCI-ITS-LI-WP-02-4

Xiangwen Lu ¹
Amelia C. Regan ²
Sandra Irani ³

¹ Institute of Transportation Studies University of California, Irvine

² Department of Civil and Environmental Engineering
and Institute of Transportation Studies, University of California, Irvine, aregan@uci.edu

³ Department of Information and Computer Science
and Institute of Transportation Studies, University of California, Irvine ssirani@uci.edu

February 2002

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu
An Asymptotically Optimal Algorithm for the Dynamic Traveling Repair Problem

Xiangwen Lu
Institute of Transportation Studies
University of California, Irvine

Amelia Regan
Department of Civil and Environmental Engineering
University of California, Irvine

Sandra Irani
Department of Information and Computer Science
University of California, Irvine

Abstract

The dynamic traveling repair problem involves providing service to customers whose locations are uniformly distributed over a bounded area in the Euclidean plane. We assume that customer requests arrive according to a Poisson point process. Earlier research provided a conjecture that the asymptotically optimal algorithm for this problem under very heavy traffic intensity involves partitioning the bounded area into subregions, waiting for sufficient demand to accumulate in the sub-regions and then serving demands in the sub-regions according to the optimal TSP tour and visiting the sub-regions in first-come first-served order as in a GI/G/M queue. Further, the researchers conjectured that the optimal algorithm for the single server case can be extended to the m-server problem by simply partitioning the service region into m sub-regions of the same size. In this paper we prove that both of these conjectures are true under some reasonable conditions and thus we demonstrate the strong connection between optimal solutions for static and dynamic vehicle routing and scheduling problems.
Introduction

The dynamic traveling repair problem falls into a more general class of problems known as stochastic vehicle routing. This class includes the dynamic traveling salesman problem (DTSP), the probabilistic traveling salesman problem (PTSP), probabilistic traveling salesman location problem and other problems such as the probabilistic shortest path problem and the dynamic vehicle allocation problem. Powell, Jaillet and Odoni (1995) provide an excellent review of these problems. In the DTSP, customer locations are known in advance and service requests arrive according to a Poisson point process at each node. The objective is to determine a dispatching strategy which minimizes customers’ expected waiting time. This problem is discussed in Psaraftis (1988). In the PTSP, an a priori tour must be constructed for a network in which each node has a given probability of requiring a visit (Jaillet, 1988). The probabilistic traveling salesman facility location problem involves identifying the optimal location for a depot node in a network in which the probability that customers will require a visit is known (Berman and Simchi-Levi, 1988, Bertsimas, 1989).

In this paper we examine the dynamic traveling repair problem (DTRP) and address a conjecture originally made by Bertsimas and van Ryzin (1993) in a paper titled “Stochastic and Dynamic Vehicle Routing with General Demand and Interarrival Time Distributions”. The problem was also examined by the two researchers in an earlier paper (Bertsimas and van Ryzin, 1991). The problem addressed is the following:
Service requests (demands) arrive over time according to a Poisson process. Upon arrival, the service requests are distributed to a bounded region A in the Euclidean plane independently according to a probability density function $f(x)$. Requests are served by one of $m$ identical mobile servers. The servers spend some time traveling to the location of the requests and some time providing on-site service. Servers move at the same constant speed.

As in Bertsimas and van Ryzin (1991) we assume that demands are uniformly distributed over the area A. The conjecture of interest concerns the asymptotic optimality of the following service policy:

The overall area is partitioned into areas of fixed size. Arrivals are assigned to these areas. When batches of fixed size form in the partitions, they are deposited into a queue in a first come first serve manner as in a GI/G/m queue. In every partition, demands are served in order of the optimal TSP tour.

We define the fraction of time the vehicle spends during on-site service as follows:

$$\rho = \frac{\lambda \bar{s}}{m}$$

where $\bar{s}$ is the average on-site service time.

Letting $A$ represent the service area, $m$ the number of mobile servers, $k$ the number of partitions, $\lambda$ the arrival rate for the area $A$, $v$ the (constant) travel speed of the vehicles, $\beta$ the TSP constant first defined by Beardwood, Halton and Hammersly (1958), Bertsimas and van Ryzin showed that under heavy traffic conditions (as $\rho$ approaches 1), expression (*) below provides an upper bound on the average waiting time.
\[ T_{\text{mod}G/G/m} \leq \frac{\beta^2 \lambda A \left(1 + \frac{m}{\lambda k}\right)}{2m^2 v^2 (1-p)^2} \] (*)

Bertsimas and van Ryzin conjecture that this policy is optimal. In this paper, under some reasonable assumptions, we develop a lower bound on the average service time for the case in which demands are uniformly distributed over the service area and show that as \( \rho \) approaches one, the lower bound has the same limit as the upper bound. Therefore, we prove the conjecture of Bertsimas and van Ryzin. We develop this lower bound without regard for the particular policy used. We simply show that no service policy can have (asymptotically) better performance.

This result shows that the connection frequently observed between static and dynamic problems exists for the DTRP. It demonstrates that if we apply the optimal deterministic static solution properly, we can obtain the asymptotically optimal solution for the dynamic problem.

The paper is organized in the following way. The first part of the paper introduces some definitions. We then present an outline of the proof followed by some concluding remarks. Because the proof is rather long and complex, the full details are provided for interested readers in the appendix.


**Problem Definition and Notation**

The m mobile servers are positioned within a bounded region (A) in the Euclidean plane. The vehicles travel at a fixed, constant speed. Service requests (demands) arrive over time according to a Poisson process. When a request arrives, it is distributed to the bounded region A independently according to a uniform distribution. The server has to spend some time to travel to the location of the demand if it is not on the site of the demand and spend some on-site service time. Without loss of generality, we assume the bounded area A is a unit square. Throughout the paper, we number the demands according to the order in which they are served by the vehicles. Let $d_i$ be the distance traveled from demand $(i-1)$ to demand $i$. Let $s_i$ be on-site service time of demand $i$. The total service time includes the travel time $\frac{d_i}{v}$ and the on-site service time $s_i$.

Using the definitions and notation presented by Bertsimas and van Ryzin (1991, 1993), if, during some time $t$, the expected number of waiting requests in the system is uniformly bounded almost surely under a specific policy, we call this a stable policy. For a stable policy, we let $W_i$ denote the expected waiting time for demand $i$. The waiting time is the time between the arrival of demand $i$ and the arrival of the server at the location of demand $i$. The limiting expected values of these random variables are defined as $\bar{W} = \lim_{i \to \infty} E[W_i]$, $\bar{d} = \lim E[d_i]$, $N$, the expected number of requests in the queue is equal to $\lambda \bar{W}$. As do Bertsimas and van Ryzin, we assume these limits exist.
Proof

Bertsimas and van Ryzin have shown that when $\rho$ is less than one, there exists an algorithm that can serve all requests. This implies that $N$, the number of requests in queue is finite. Here we only consider those strategies in which all requests are served.

We make two assumptions about the strategy which serves the demands.

Assumption I:

As $t$ approaches infinity, the average waiting time over any $2N$ consecutive demands satisfies the following condition:

$$\forall \varepsilon > 0, \exists N_t > 0, \text{ s.t. when } n > N_t,$$

$$\text{P} \left\{ \frac{1}{2NW} \sum_{t=2nN}^{i=2nN+2N} W_i < (1+\varepsilon) \right\} > 1-\varepsilon \quad (1)$$

Assumption I simply implies that with high probability, as time approaches infinity, if the system is stable, then the average waiting time of any group of $2N$ consecutively served requests is close to the expected waiting time.

In order to express assumption II clearly we must introduce several new ideas.

For any specific $\rho$, as time approaches infinity, we select an arbitrary set of $2N$ demands that have been consecutively served over time according to an arbitrary service strategy.
Then for any $\varepsilon > 0$, we separate the $2N$ demands into $M$ partitions where $M$, the number of partitions is equal to $\frac{1}{\varepsilon}$. Let each partition hold exactly $2N\varepsilon$ consecutively served demands. Each copy is numbered according to the order in which its demands are served, $A_i$, where $i$ goes from 1 to $M$.

Suppose now that we make $M$ copies of the bounded Euclidean region $A$ and assign each set of $2N\varepsilon$ demands to a copy of $A$. We refer to $A^*$ as the union of the $M$ copies of $A$. (See figure 1).

Figure 1) Diagram of the actual and imagined service region

Let $x_1,...,x_{2N}$ represent the locations of $2N$ consecutively served customers and let the index represent the order in which service is performed. Let $y_{2N\varepsilon+1}...y_{2N(\varepsilon+i)}$ be a random permutation of the points $x_{2N\varepsilon+1}...x_{2N(\varepsilon+i)}$. Since $N\varepsilon$ is arbitrarily large, the distribution over the $Y_i$'s is arbitrarily close to some i.i.d. distribution. Our assumption is when $\rho$
approaches 1, \( Y_{2Nt+1}\ldots Y_{2N(i+1)t} \) is one instance of \( Y_{2Nt+1}\ldots Y_{2N(i+1)t} \), where \( Y_{2Nt+1}\ldots Y_{2N(i+1)t} \) are i.i.d. random variables with the common distribution \( g_i(Y) \). We define \( g(Y) \) based on \( g_i(Y) \) as follows: \( g(Y) = \varepsilon \sum_i g_i(Y) I_{A_i}(Y) \) where \( I_{A_i}(x) = \begin{cases} 1 & \text{if } x \in A_i \\ 0 & \text{otherwise} \end{cases} \).

\( g_i(Y) \) satisfies the following condition: for \( \varepsilon \) as defined in (1) there exists \( K(\varepsilon) \) such that

\[
|g_i(a) - g_i(b)| \leq K(\varepsilon) |a - b|, \quad \forall a, b \in A_i, \forall i. \tag{2}
\]

Referring to requests assigned to the same \( A_i \) as a subtour, and letting \( a \) and \( b \) be customer nodes in the same subtour, the second assumption specifies that the probability density function which describes demands served by the same subtour does not vary widely.

The partition algorithm defined earlier can easily be shown to satisfy conditions (1) and (2) using the smoothing technique described in Bertsimas and van Ryzin (1991).

For the rest of the paper we consider only those strategies which satisfy the two conditions. We wish to provide a lower bound on the average distance per demand served for any such strategy. This in turn provides a lower bound for the expected (heavy traffic) waiting time under such strategy.

We use average case analysis to examine the distance traveled by any algorithm. To do this, we partition the area served into virtual subareas \( A_i \). After that we analyze the average distance between consecutive demands served under the condition that the average waiting time for any \( 2N \) demands cannot exceed the corresponding average waiting time too much (condition (1)). Lemma 1 provides a lower bound for the average
waiting time based on g(Y) while lemma 2 provides a lower bound for the distance traveled per demand served. Lemma 1 says the average waiting time of 2N consecutively demands divided by W can be bounded from below by the integration of g(Y)^2 while lemma 2 says that the total distance traveled while serving the 2N demands divided by (2N)^{1/2} can be bounded below by the integration of g(Y)^{1/2} multiplied by a constant. In lemma 3 we solve an optimization problem needed for the proof of theorem 1. (The proof of these lemmas is provided in the appendix.)

Using these three lemmas, we prove theorem 1, which provides a lower bound on the distance traveled per demand served. We then provide a lower bound on the average waiting time for service.

Finally, under conditions (1) and (2), we find that if we want to minimize the average waiting time, the distribution g(Y) which describes the spatial distribution of consecutively served customer under the optimal algorithm will be almost uniform. This implies that when the server enters a service area, it will serve almost all the known demands in that area prior to departing for another area.

**Theorem 1:** Let \( d \) be the average distance traveled per demand served and let \( N \) be the average number of demands awaiting service. All strategies satisfying assumptions I and II will also satisfy the following: \( \lim_{p \to 1, t \to \infty} \sqrt{2N} \cdot \bar{d} \geq \beta \). Where \( \beta \) is the TSP constant defined earlier.
Proof: Letting \( L_{TSP}(Y_1,Y_2,\ldots,Y_{2N}) \) be the length of optimal TSP tour over \( Y_1,Y_2,\ldots,Y_{2N} \), for any given \( \varepsilon'>0 \), from lemma 2, we know there exists \( \rho(\varepsilon') > 0 \) such that when \( \rho > \rho(\varepsilon') \) we have

\[
P\left\{ \frac{L_{TSP}(Y_1,Y_2,\ldots,Y_{2N})}{\sqrt{2N}} \geq \beta \int_A \sqrt{g(Y)} dY - \varepsilon' \right\} > 1 - \varepsilon'
\]

For any \( \varepsilon > 0 \), let \( \varepsilon_i = 2(k(\varepsilon)+1)\varepsilon \), using lemma 1 and the facts that \( P(AB) \geq 1 - P(A^c) - P(B^c) \), \( E[X] \geq cP(X \geq c) \), we know that,

\[
\lim_{n \to 1,t \to \infty} \left( \sqrt{2N} \cdot d \right) \geq (1-\varepsilon_i-\varepsilon) \min \left\{ \int_A g(Y)^2 dX + \varepsilon_i \right\} \left\{ \int_A \sqrt{g(Y)} dY - \varepsilon' \right\}.
\]

For any \( \varepsilon'' > 0 \), for any \( g(Y) \) satisfying condition II and \( \int_A g(Y)^2 dX \leq 1 + \varepsilon_i \) we can find step functions \( g_n \geq g \), such that the following three conditions on \( g_n \) are satisfied:

(i) \( \int_A \sqrt{g_n(Y)} dY \leq \int_A \sqrt{g(Y)} dY + \varepsilon_i \),
(ii) \( \int_A g_n(Y)^2 dX \leq 1 + \varepsilon'' + \varepsilon_i \) and, (iii)
\[
1 \leq \int_A g_n(Y) dY \leq 1 + \varepsilon_i.
\]

Because \( \varepsilon'' \) is selected arbitrarily and from Lemma 3, so we have the following fact,

\[
\min \left\{ \int_A \sqrt{g(Y)} dY \right\} \geq \min \left\{ \int_A \sqrt{g_n(Y)} dY - \varepsilon_i \right\} \geq \frac{1}{\sqrt{1 + \varepsilon'' + \varepsilon_i}} - \varepsilon_i.
\]

Which implies that

\[
\lim_{\rho \to 1,t \to \infty} \sqrt{2N} \cdot d \geq \lim_{\varepsilon \to 0, \varepsilon' \to 0} \left( (1-\varepsilon_i-\varepsilon') \beta \min \left\{ \int_A \sqrt{g(Y)} dY - \varepsilon' \right\} \right)
\]

\[
\geq \lim_{\varepsilon \to 0, \varepsilon' \to 0} \beta (1-\varepsilon_i-\varepsilon') \left( \frac{1}{\sqrt{1 + \varepsilon'' + \varepsilon_i}} - \varepsilon_i - \varepsilon' \right) \geq \beta.
\]

This completes the proof of theorem 1.
Theorem 2: Under the same conditions as theorem 1, we have

\[ \lim_{p \to 1, \epsilon \to 0} \{ (1-p)^2 W \} \geq \frac{\lambda \beta^2 A}{2m^2 v^2} \]

(3)

Proof: Recall that \( \frac{\lambda}{m} \) is the average arrival rate for each server and that \( \frac{-d}{v} \) is the actual average service time for each demand. Therefore \( \frac{\lambda}{m} \left( \frac{-d}{s+v} \right) \) must be less than 1.

From theorem 1 we know that \( \frac{\lambda}{m} \left( \frac{s+\beta}{v \sqrt{2N}} + o\left( \frac{1}{\sqrt{2N}} \right) \right) < 1 \). Recalling that \( N = \lambda W \) and that \( \rho = \frac{\lambda s}{m} \), we have \( \lim_{\rho \to 1, \epsilon \to 0} \{ (1-p)^2 W \} \geq \frac{\lambda \beta^2}{2m^2 v^2} \).

Revisiting equation *, we can see that, letting \( k \), the number of partitions, approach infinity, * is equivalent to (3). Therefore, the algorithm which we have shown is optimal for the single vehicle case is also optimal when \( m > 1 \). Therefore we have the following final theorem.

Theorem 3: For the DTRP under heavy traffic intensity, if we know the best strategy for the single vehicle case, we can extend this to the \( m \) vehicle case easily by partitioning the service area into \( m \) partitions and assigning each of the \( m \) to serve the demands from one specific partition. Each vehicle simply applies the single vehicle strategy to its service region.
Conclusion

By providing a lower bound on the performance of the asymptotically optimal algorithm for the DTRP we have proven the optimality of the algorithm of Bertsimas and van Ryzin. In addition, we have shown that the algorithm which is asymptotically optimal for the single vehicle case can be extended to the multiple vehicle case. These results demonstrate how powerful the partition algorithm can be. Karp (1985) has shown that a partitioning algorithm can provide asymptotically optimal results for the TSP problem. We show here that it also works for the DTRP by proving the conjecture of Bertsimas and van Ryzin. We provide similar results for the probabilistic traveling salesman problem (PTSP), (Lu, Irani and Regan (2000)). Our results demonstrate the existence of a strong connection between static and dynamic vehicle routing problems that has been observed by many other researchers.

Acknowledgements

This research was partially supported by a grant from the University of California Transportation Center (UCTC) and by the US National Science Foundation (NSF) under Grant No. 9875675. The authors gratefully acknowledge this generous support. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the UCTC or the NSF.
REFERENCES


APPENDIX

Assumption I:
As \( t \) approaches infinity, the average waiting time over any \( 2N \) consecutive demands satisfies the following condition:

\[
\forall \varepsilon > 0, \; \exists C > 0, \; \text{s.t. when } n > C,
\]

\[
P\left( \frac{1}{2NW} \sum_{i=2nN+2N}^{i=2nN+2N} W_i < (1 + \varepsilon) \right) > 1 - \varepsilon \quad (1)
\]

\( Y_1 \ldots Y_{2N} \) are random variables and \( Y_{2N(i)+1} \ldots Y_{2N(i+1)} \) are i.i.d. random variables with the distribution of \( g_i(Y) \). Furthermore, \( Y_{2N(i)+1} \ldots Y_{2N(i+1)} \) is one instance of \( Y_{2N(i)+1} \ldots Y_{2N(i+1)} \).

We define \( g(Y) \) based on \( g_i(Y) \) as follows: \( g(Y) = \varepsilon \sum g_i(Y) I_{A_i}(Y) \) where

\[
I_{A_i}(Y) = \begin{cases} 
1 & \text{if } Y \in A_i \\ 0 & \text{otherwise}
\end{cases} .
\]

Assumption II:
\( g_i(Y) \) satisfies the following condition: for any \( \varepsilon \) defined as in (1) there exists \( K(\varepsilon) \) such that:

\[
|g_i(a) - g_i(b)| \leq K(\varepsilon) ||a-b||, \quad \forall a, b \in A_i, \; \forall i. \quad (2)
\]

Lemma 1. For any strategy satisfying assumptions I and II, for any given \( \varepsilon \) defined as in (1) and (2), for any \( \varepsilon_1 > 2\varepsilon(K(\varepsilon)+1) \),

\[
\lim_{\rho \to 1, \rho \to \infty} P\left( 1 + \varepsilon_1 \geq \int_{\mathcal{A}} g^2(X) dX \right) \geq 1 - \varepsilon_1 .
\]
**Lemma 2.** For any \( \varepsilon_2 > 0 \) and for any \( g(Y) \) satisfying (2) and \( Y_{2tNe+i} \ldots Y_{2(t+1)Ne} \) which are i.i.d. random variables with distribution \( g(Y) \), if we let \( L_{\text{TSP}}(Y_1, Y_2, \ldots, Y_{2N}) \) be the length of optimal TSP tour over \( Y_1, Y_2, \ldots, Y_{2N} \), we can find \( \rho(\varepsilon_2) \) such that when \( \rho > \rho(\varepsilon_2) \), we have

\[
P \left( \frac{L_{\text{TSP}}(Y_1, Y_2, \ldots, Y_{2N})}{\sqrt{2N}} \geq \beta \int_A \sqrt{g(Y)} dY - \varepsilon_2 \right) \geq 1 - \varepsilon_2.
\]

**Lemma 2.1.** Let \( \{B_i\}_{i=1}^{n_0} \) be a grid partition over \( A \) such that each \( B_i \) has the same area.

Let \( f(X) \) be constant within \( B_i \) and \( X_1, X_2, \ldots, X_n \) be i.i.d. random variables distributed according to \( f(X) \). For any \( \varepsilon_3 > 0 \), we can find \( C(\varepsilon_3) \), when \( n > C(\varepsilon_3) \), we have

\[
P \left( \frac{L_{\text{TSP}}(X_1, X_2, \ldots, X_n)}{\sqrt{n}} \geq \beta \int_A \sqrt{f(X)} dX - \varepsilon_3 \right) \geq 1 - \varepsilon_3.
\]

**Lemma 3.** Let \( Z = \min \left\{ \sum_{i=1}^{n} \sqrt{x_i} A_i \right\} \) subject to \( \sum_{i=1}^{n} x_i^2 A_i \leq 1 + \varepsilon \) and \( \sum_{i=1}^{n} x_i A_i = 1 \). We have

\[
Z \geq \frac{1}{\sqrt{1 + \varepsilon}}.
\]

To prove lemma 2 and lemma 2.1, we borrow the method from the classical paper by Beardwood, Halton and Hammersley (1959). To prove lemma 3, we rely on optimization methods and algebraic techniques. First we state the main result from the paper of Beardwood, Halton and Hammersley (1959) that we will use later in our proof.

**Theorem BHM:** Assume that \( X_1, X_2, \ldots, X_n \) are i.i.d. random variables with distribution \( f(X) \) and let \( L_{\text{TSP}}(X_1, X_2, \ldots, X_n) \) be the length of TSP tour over \( X_1, X_2, \ldots, X_n \).

\[
\lim_{n \to \infty} \left\{ \frac{L_{\text{TSP}}(X_1, X_2, \ldots, X_n)}{\sqrt{n}} \right\} = \beta \int_A \sqrt{f(X)} dX \quad \text{a.s.}
\]
Therefore, for any $\varepsilon$, exists $n(\varepsilon)$, when $n > n(\varepsilon)$,

\[
P\left\{ \left| T_{\text{TSP}}(X_1, X_2, L, X_n) - \frac{\beta}{\sqrt{n}} \int_{\mathbb{A}} \sqrt{\hat{r}(X)}dX \right| \geq \varepsilon \right\} < \varepsilon \quad (**) \]

**Proof of lemma 1:** $M = \frac{1}{\varepsilon}$ is the number of copies of $\mathbb{A}$ (i.e. the number of $A_i$'s). Let $t_{ij}$ be the arrival time for the $j$-th request over $A_i$; and let $t'_{ij}$ be the time this request is served.

We define $\overline{W^i}_{\text{END}} = \frac{1}{2N\varepsilon} \sum_{j=1}^{p=2N} (t'_{L2Nec} - t_{ij})$ and $\overline{W^i} = \frac{1}{2N\varepsilon} \sum_{j=1}^{p=2N} (t'_{ij} - t_{ij})$. \hspace{1cm} (1.1)

Letting $\overline{W} = \sum_{i=1}^{N} \overline{W^i}/M$ and $\overline{W}_{\text{END}} = \sum_{i=1}^{N} \overline{W^i}_{\text{END}}/M$. We have,

\[
\overline{W}_{\text{END}} \leq \overline{W} + \frac{t'_{M2Nec} - t'_{i1}}{M}. \quad (1.2)
\]

For any queuing system, when the system is in a stable state, the average number of customers in the queue is equal to the number of customers served during the average waiting time. We see that $t'_{M2Nec} - t'_{i1}$ represents the duration of service for $2N$ consecutive demands minus the service time for the first demand. Recall that on average we serve $2N$ demands in a period of length $2W$. This means that $\mathbb{E}(t'_{M2Nec} - t'_{i1}) \leq 2W$. By Markov's Inequality\(^1\) we have $P\left\{ t'_{M2Nec} - t'_{i1} \geq 2kW \right\} \leq \frac{1}{2k}$. Noting that $\frac{2kW}{M} = 2k\varepsilon W$, we have

\[
P\left\{ \frac{t'_{M2Nec} - t'_{i1}}{M} \leq \frac{2kW}{M} = 2k\varepsilon W \right\} \geq 1 - \frac{1}{2k} \quad \text{i.e.} \]

\[
P\left\{ \frac{\overline{W}_{\text{END}}}{W} - 2k\varepsilon \leq \frac{\overline{W}}{W} \right\} \geq 1 - \frac{1}{2k} \quad (1.3)
\]

\(^1\) For any non-negative random variable $X$, we have $P\{X \geq a\} \leq \frac{\mathbb{E}\{X\}}{a}$ for any $a > 0$.
The next step is to estimate $\overline{W_{\text{end}}}$ based on $g(Y)$.

Recall that $Y_{2(i-1)N_0 + 1}, \ldots, Y_{2iN_0}$ are i.i.d. random variables with distribution $g_i(Y)$. The situation we need to consider is the following: when the server arrives at an area $A_i$, there are $2iN_0$ requests in the area awaiting service. The distribution of the locations of these requests is $g_i(Y)$. We wish to provide a bound for the average waiting time until the server enters the $A_i$.

To do this we divide each $A_i$ into $1/\varepsilon_4$ grid partitions of equal size such that each $A_i$ has the same partitions. Assume $1/\varepsilon_4$ is an integer. We can view these as temporal partitions of $A_i$. Each individual partition is represented by $B_{ij}$, where $i$ represents the copy of $A$ and $j$ indicates the partition of $A_i$. Note that $B_{ij}$, $i = 1, \ldots, 1/\varepsilon$ corresponds to a single physical region. Referring to this area as $B_j$ for a moment, we number the demands which fall into region $B_j$ according to their arrival time, beginning at time $t = 0$. These requests are $X_{j_1}, X_{j_2}, \ldots, X_{j_n}, \ldots$. The corresponding arrival time is $t(X_{j_1}), t(X_{j_2}), \ldots, t(X_{j_n}), \ldots$. Letting $|B_j|$ represent the area of $B_j$. It is easy to see that the interarrival times $t(X_{j_{i+1}}) - t(X_{j_i}), \ldots, t(X_{j_{i+n-1}}) - t(X_{j_{i+n-2}}), \ldots$ are exponentially distributed with parameter $\lambda|B_j|$. For $B_{ij}$, suppose there are $p$ demands that arrive in $B_{ij}$, these are $X_{j_{i_1}}, X_{j_{i_2}}, \ldots, X_{j_{i_p}}$, ordering according to their arrival time. $X_{j_{i_1}}, X_{j_{i_2}}, \ldots, X_{j_{i_p}}$ is a subsequence of $X_{j_1}, X_{j_2}, \ldots, X_{j_n}, \ldots$. The time between the arrival of two consecutive demands in the subsequence $X_{j_{i_1}}, X_{j_{i_2}}, \ldots, X_{j_{i_p}}$ (for example, $t(X_{j_{i_r}}) - t(X_{j_{i_{r+1}}})$ for $r > 1$ is at least $t(X_{j_{i_r}}) - t(X_{j_{i_{r+1}}})$. We call $t(X_{j_{i_1}}) - t(X_{j_{i_{r+1}}}), X_{j_{i_r}}$
and define \( X_{i,j,i} \) to be equal to zero. We know that \( \{X_{i,j,i}, \forall i,j,r\} \) are independent random variables and \( \{X_{i,j,r}, \forall r\} \) is exponentially distributed with parameter \( \lambda |B_i| \).

Let \( Z_{i,j} \) denote the number of requests in \( B_{i,j} \). Using the fact that the distribution of the \( 2N\varepsilon \) requests is \( g_i(Y) \), the distribution of these random variables of \( Z_{i,j} \)'s is given by the following multinomial distribution:

\[
P\left\{Z_{i,j}=z_{i,j}, \ldots, Z_{i,j-1}=z_{i,j-1}\right\} = \frac{2N\varepsilon!}{\prod \left\{g_i(X_{i,j}) |B_{i,j}|\right\}^{z_{i,j}}}
\]

where \( g_i(X_{i,j}) |B_{i,j}| = \int_{B_{i,j}} g_i(X) \, dX \), \( |B_{i,j}| \) is the area of \( B_{i,j} \).

Let \( X_{i,j} = \sum_{r=1}^{Z_{i,j}} (r-1)X_{i,j,r} \), for \( j=1,\ldots,1/\varepsilon_4 \). Let \( X_i^* = \sum_{j=1}^{\frac{j-1}{\varepsilon_i}} X_{i,j} \), for \( i=1,\ldots,1/\varepsilon \). Finally, let

\[
X^* = \sum_{i=1}^{\frac{i-1}{\varepsilon}} X_i^* \quad \frac{1}{2N}.
\]

We define the forming time to be the length of time that until there is \( Z_{i,j} \) demands in the partition of \( B_{i,j} \). \( \overline{W_{\text{END}}} \) is at least equal to the total waiting time during the forming time divided by \( 2N \). So, if there are \( Z_{i,j} \) demands at \( B_{i,j} \) the total waiting time is \( X_{i,j}^* \) plus some non-negative variable. At last we have

\[
\overline{W_{\text{END}}} \geq \frac{X^*}{2N} \quad (1.4)
\]
We can show that

\[
\mathbb{E}\left( \frac{\tilde{X}_i^*}{2N\varepsilon W} \right) = \mathbb{E}\left[ \mathbb{E}(X_i^*)(Z_{i,1}, Z_{i,2}, \ldots Z_{i,\frac{1}{\varepsilon_i}}) \right] \geq \left( \frac{1}{2N\varepsilon} \right) \varepsilon \left[ \int_\mathcal{A} g_i^2(X) dX - 2M_1 \frac{\varepsilon_4}{\varepsilon} \right]
\]

\[
= \varepsilon \left[ \int_\mathcal{A} g_i^2(X) dX - 2M_1 \frac{\varepsilon_4}{\varepsilon} \right] + o(1), \text{ where } M_1 = K(\varepsilon)\sqrt{1 + \frac{1}{\varepsilon^2}} + 1.
\]

We can also show that

\[
\text{var}\left( \frac{X_i^*}{2N\varepsilon W} \right) = \frac{\mathbb{E} \left( X_i^* - \mathbb{E}(X_i^*)(Z_{i,1}, Z_{i,2}, \ldots Z_{i,\frac{1}{\varepsilon_i}}) \right)^2 + \mathbb{E} \left[ \mathbb{E}(X_i^*)(Z_{i,1}, Z_{i,2}, \ldots Z_{i,\frac{1}{\varepsilon_i}}) - \mathbb{E}(X_i^*) \right]^2}{4N^2\varepsilon^2 W^2} = o(1) \text{ as } \rho \to 1.
\]

Since \( \varepsilon_4 \) is arbitrary, we can select \( \varepsilon_4 \) so that \( \frac{\varepsilon_4}{\varepsilon} \to 0 \). By Chebychev's inequality we have

\[
\lim_{\rho \to 1, t \to \infty} \mathbb{P}\left( \frac{X_i^*}{2N\varepsilon W} \geq \varepsilon \left[ \int_\mathcal{A} g_i^2(X) dX \right] + o(1) \right) = 1.
\]

So we have

\[
\lim_{\rho \to 1, t \to \infty} \mathbb{P}\left( \frac{\bar{W}_{\text{END}}}{W} \geq \int_\mathcal{A} g^2(X) dX + o(1) \right) = 1.
\]

Because of (1.4), we have that

\[
\lim_{\rho \to 1, t \to \infty} \mathbb{P}\left( \frac{\bar{W}_{\text{END}}}{W} \geq \int_\mathcal{A} g^2(X) dX + o(1) \right) = 1. \text{ This gives us a lower bound for } \bar{W}_{\text{END}} \text{ based on } g(Y).
\]
Rewriting (1.3) we have \( P( \frac{W_{\text{END}} - 2k\varepsilon}{W} \leq \frac{W}{W}) \geq 1 - \frac{1}{2k} \) we know that for any \( \varepsilon > 2k\varepsilon \),

we have \( \lim_{p \to 1, t \to \infty} P \left\{ \frac{W}{W} \geq \int_{\Lambda} g^2(X) dX - \varepsilon \right\} \geq 1 - \frac{1}{2k} \).

We know that the average waiting time \( W \) equals \( \bar{W} \) minus the average service time.
Because \( k \) can be arbitrarily selected, so from (1), if we let \( \varepsilon = \varepsilon + \varepsilon \), we have

\[
\lim_{p \to 1, t \to \infty} P \left\{ 1 + \varepsilon \geq \int_{\Lambda} g^2(X) dX \right\} \geq 1 - \varepsilon.
\]

This completes the proof of lemma 1.

**Proof of lemma 2.** We know that \( |A^*| \leq \frac{1}{\varepsilon} \) where \( |A^*| \) is the area of \( A^* \).

For any \( \varepsilon > 0 \), let \( \delta = \frac{\varepsilon^2 \varepsilon^2}{2K(\varepsilon)} \). For any \( X,Y \) satisfying \( \|X-Y\| \leq \delta \), we have,

\[
\left| \sqrt{g_i(X)} - \sqrt{g_i(Y)} \right| \leq \varepsilon \varepsilon, \text{ for any } X,Y \in A_i.
\]

We divide \( A^* \) into grid partitions of identical size with diameter \( \delta \). As before, let \( B_{i,j} \) be the \( j \)-th partition in \( A_i \).

Let \( g_i^*(Y) = \min_{\{Y \in B_{i,j}\}} \{ g_i(Y) \} \).

We have \( \int_{A_i} g_i^*(Y) dY \geq 1 - \varepsilon \) and \( \int_{A_i} \sqrt{g_i^*(Y)} dY \geq \int_{A_i} \sqrt{g_i(Y)} dY - \varepsilon \).

We place \( Y_j \) into one of the two set \( (\Psi_{i,1}, \Psi_{i,2}) \) as follows:

For all \( 2N\varepsilon < j < 2N(j+1)\varepsilon + 1 \), if \( g_i^*(Y) = 0 \) let \( Y_j \in \Psi_{i,1} \) with probability one;

otherwise, let \( Y_j \in \Psi_{i,1} \) with probability \( 1 - \frac{g_i^*(Y)}{g_i(Y_j)} \) and \( Y_j \in \Psi_{i,2} \) with probability \( \frac{g_i^*(Y)}{g_i(Y_j)} \).
Let $n_{2i}$ be the number of requests belonging to $\Psi_{i,1}$ and $L_{TSP}(\Psi_{i,2})$ be the length of optimal TSP tour over all the nodes belonging to $\Psi_{i,2}$. We can show that

$$\frac{n_{2i}}{2N\epsilon} \to \int_{A_i} g_i^*(Y) dY \geq 1 - \epsilon_4.$$

We can also show that the random variables in the set of $\Psi_{i,2}$ are i.i.d. random variables with the distribution $g_i^*(Y)$. Using lemma 2.1, we know that for any $\epsilon_3 > 0$, there exists $C(\epsilon_3) > 0$, such that when $n_{2i} > C(\epsilon_3)$,

$$\mathbb{P}\left[ \frac{L_{TSP}(\Psi_{i,2})}{\sqrt{2N\epsilon}} \geq \beta \sqrt{\frac{n_{2i}}{2N\epsilon}} \int_{A_i} \sqrt{g_i^*(Y)} dY \cdot \epsilon_3 \right] > 1 - \epsilon_3.$$

Because exists $M_i > 0$ such that $\sup_{\theta_i(Y)} \int_{A_i} \sqrt{g_i^*(Y)} dY \leq M_i$ and $\frac{n_{2i}}{2N\epsilon} \int_{A_i} g_i^*(Y) dY \to 1$ as $\rho \to 1$, we know that for any $\epsilon' > M_i \epsilon_4 + \epsilon_2$, exists $\rho(\epsilon') > 0$, such that when $\rho > \rho(\epsilon')$,

$$\mathbb{P}\left[ \frac{L_{TSP}(\Psi_{2iN}^{i} \ldots Y_{2(i+1)N})}{\sqrt{2N\epsilon}} \geq \beta \sqrt{\frac{n_{2i}}{2N\epsilon}} \int_{A_i} \sqrt{g_i^*(Y)} dY \cdot \epsilon' \right] > 1 - \epsilon'.$$

Now we consider all the $A_i$'s and the definition of $g(Y)$. Lemma 2 follows from the fact that $\epsilon_4$ and $\epsilon_5$ are arbitrary, $\mathbb{P}\left( \bigcap_{i} A_i \right) \geq 1 - \sum_{i} P\left( A_i^c \right)$ and

$$\frac{L_{TSP}(Y_{1}, \ldots Y_{2N})}{\sqrt{2N}} \geq \sum_{i} \frac{L_{TSP}(Y_{2iN+1}, \ldots Y_{2(i+1)N})}{\sqrt{2N}} - \frac{2}{\sqrt{2N\epsilon}}.$$
Proof of lemma 2.1 We prove lemma 2.1 by induction based on the different values \( f(X) \) can hold. First we assume that \( f(X) \) is equal to zero or any constant over all the \( B_i \), i.e.

\[
f(X) = \sum_{i} c_i I_{B_i}(X), \quad \text{where } I_{B_i}(X) = \begin{cases} 1 & \text{if } x \in B_i \\ 0 & \text{otherwise} \end{cases}
\]

and \( c_i \) equals to zero or \( c \).

Let \( m \) be the number of the partitions on which \( f(X) \) is \( c \). We can see that the probability that any demand falls into any specific partition on which \( f(X) \) is not zero is \( 1/m \). Let \( L_{TSP_i} \) be the length of optimal TSP subtour over all the demands belonging to the \( i \)th piece of the partition. Let \( L_{TSP} \) be the optimal TSP tour over all the demands, remembering that \( n_0 \) is the total number of partitions, we have

\[
L_{TSP} \geq \sum_{i} L_{TSP_i} n_0 x
\]

where \( x \) is the length of the circumference of the partition \( B_i \). For details please see Karp and Steel(1985).

Now we try to estimate \( \frac{\sum_{i} L_{TSP_i} (X_1, X_2, \ldots X_n)}{\sqrt{n}} \), because \( \frac{n_0 x}{\sqrt{n}} = o(1) \) as \( n \) approaches infinity, we focus on \( \frac{\sum_{i} L_{TSP_i}}{\sqrt{n}} \).

Let \( n_i \) be the number of demands in \( B_i \), let \( D_i = \left\{ n_i > \frac{n}{m} - k_i \sqrt{\frac{n}{m}} \right\} \). Using Chebychev's Inequality again,

\[
P\{D_i\} \geq 1 - \frac{1}{k_i^2}.
\] (2.1.1)

By De Morgan's rule, \( P\{\cap_i D_i\} = 1 - P\{\cup_i D_i^c\} \)

\[
1 - P\{\cup_i D_i^c\} \geq 1 - \sum_{i} [1 - P\{D_i\}]
\]

Finally, \( 1 - \sum_{i} [1 - P\{D_i\}] = 1 - \frac{m}{k_i^2} \rightarrow 1 \) as \( k_i = (\frac{n}{m})^{\frac{1}{2}} \rightarrow \infty \). (2.1.2)
Let $|B|$ be the size of the area of any partition and $|A|$ be the size of the whole area, we use the result of Beardwood et al (1959), please refer the (***) and combining with (2.1.1), (2.1.2), we have,

$$\lim_{n \to \infty} P \left\{ \frac{\sum L_{TSP_i}}{\sqrt{n}} \geq (\beta - \frac{\varepsilon_3}{2|A|}) \sqrt{|B| \sum \sqrt{n_i}} \right\} = 1$$

therefore,

$$\lim_{n \to \infty} P \left\{ \frac{\sum L_{TSP_i}}{\sqrt{n}} \geq (\beta - \frac{\varepsilon_3}{2|A|}) \sqrt{|B| \left( 1 - k_i \sqrt{\frac{1}{n}} \right)} \right\} = 1.$$  

Because $1 - k_i \sqrt{\frac{1}{n}} = 1 - \left( \frac{1}{m} \right)^{\frac{1}{2}} = 1 + o(1)$ as $n \to \infty$, so we can find $n^*$, such that when $n > n^*$,

$$P \left\{ \frac{L_{TSP}}{\sqrt{n}} \geq \beta \sqrt{mA_i - \varepsilon_j} \right\} \geq 1 - \varepsilon_j$$

holds.

The last step in this induction proof is to assume the lemma is true when $f(X)$ has $k$ different values and to show that the lemma holds when $f(X)$ have $k+1$ different values. The proof is based on similar ideas and is rather tedious, so we omit the rest of proof here.

**Proof of lemma 3**  By adding some additional unbounded constraints we can show that the original problem can be translated into the following problem: 

$$\min \left\{ \sum_{i=1}^{m} \sqrt{x_i A_i} \right\}$$

subject to: 

$$\sum_{i=1}^{m} x_i A_i = 1, \quad \sum_{i=1}^{m} x_i A_i \leq 1 + \varepsilon, \quad x_i \geq \varepsilon_i, \quad \forall i$$

for some small number of $\varepsilon_i$. The reason that the constraints $\{x_i \geq \varepsilon_i\}_{i=1}^{m}$ do not affect the solution of the problem when $\varepsilon_i$ is small enough because these constraints are not bounded.

22
For this new problem, we use a standard optimization techniques as follows:

Let \( L(\bar{t},\lambda,\mu,\gamma) = \sum_{i=1}^{\infty} \sqrt{x_i} A_i^* + \left( \lambda \sum_{i=1}^{\infty} x_i A_i^* - 1 \right) + \mu \left( 1 + \varepsilon - \sum_{i=1}^{\infty} x_i^2 A_i^* \right) + \sum_{i=1}^{\infty} \gamma_i (x_i - e_i). \)

By considering the Kuhn-Tucker conditions we know that the optimal solution must satisfy the following: \( \frac{\partial L}{\partial x_i} = 0, \frac{\partial L}{\partial \lambda} = 0, \frac{\partial L}{\partial \mu} = 0, \gamma_i = 0 \quad \forall i. \) Therefore, for the optimal solution \( 1 + 2\lambda x_i^2 - 4\mu x_i^2 = 0, \forall i \) must hold.

Considering the set of equations \( 1 + 2\lambda x_i^2 - 4\mu x_i^2 = 0, \forall i. \) From the theory of algebraic equations, we know the following facts,

Case One: The equations have the same nonnegative solution i.e. \( x_i = x_j \) for any i,j.

Case Two: There are at most two different position solutions.

Assume that the positive solutions are a and b respectively and that \( a \leq b. \) Let

\[
Z = \min \left\{ \sum_{i=1}^{\infty} \sqrt{x_i} A_i^* \right\} \quad \text{subject to:}
\]

\[
\sum_{i=1}^{\infty} x_i A_i^* = 1
\]

\[
\sum_{i=1}^{\infty} x_i^2 A_i^* \leq 1 + \varepsilon
\]

\( x_i \in \{a, b\} \)

\[
Z = \min \left\{ \sum_{\{i|a_i=a\}} \sqrt{a} A_i^* + \sum_{\{i|a_i=b\}} \sqrt{b} A_i^* \right\} \quad \text{subject to}
\]

\[
\sum_{\{i|a_i=a\}} a A_i^* + \sum_{\{i|a_i=b\}} b A_i^* = 1
\]

\[
\sum_{\{i|a_i=a\}} a^2 A_i^* + \sum_{\{i|a_i=b\}} b^2 A_i^* \leq 1 + \varepsilon
\]
So if we let $x = \sum_{\{i | x_i = a\}} A_i^*$ and $y = \sum_{\{i | x_i = b\}} A_i^*$, we have,

$Z \geq \min\left\{\sqrt{ax} + \sqrt{by}\right\}$ subject to:

$ax + by = 1$

$a^2 x + b^2 y \leq 1 + \varepsilon$

$a > 0, \ b > 0, \ b \geq a$.

Therefore, $Z \geq \frac{1}{\sqrt{1+\varepsilon}}$. We complete the proof of lemma 3.