Wasserstein distance and the distributionally robust TSP

John Gunnar Carlsson*, Mehdi Behroozi†, and Kresimir Mihic‡

April 6, 2017

Abstract

Recent research on the robust and stochastic travelling salesman problem and the vehicle routing problem has seen many different approaches for describing the region of ambiguity, such as taking convex combinations of observed demand vectors or imposing constraints on the moments of the spatial demand distribution. One approach that has been used outside the transportation sector is the use of statistical metrics that describe a distance function between two probability distributions. Motivated by a districting problem in multi-vehicle routing, we consider a distributionally robust version of the Euclidean travelling salesman problem in which we compute the worst-case spatial distribution of demand against all distributions whose *Wasserstein distance* to an observed demand distribution is bounded from above. This constraint allows us to circumvent common overestimation that arises when other procedures are used, such as fixing the center of mass and the covariance matrix of the distribution. Numerical experiments confirm that our new approach is useful when used in a decision support tool for dividing a territory into service districts for a fleet of vehicles when limited data is available.

1 Introduction

One of the most common tools for addressing uncertainty in vehicle routing problems (VRPs) is the use of *districting* strategies, in which one divides one's service region into smaller pieces, and then assigns one (or more) vehicle(s) to each piece. Such strategies are desirable for many practical reasons: for example, they provide a simple rule for assigning destination points to vehicles, they decompose the original problem into independent sub-problems, and they enable one to make high-level strategic decisions in the absence of complete information. In fact, for

*Department of Industrial and Systems Engineering, University of Southern California. J. G. C. gratefully acknowledges DARPA Young Faculty Award N66001-12-1-4218, NSF grant CMMI-1234585, and ONR grant N000141210719.

†Department of Mechanical and Industrial Engineering, Northeastern University.

‡Modeling, Simulation and Optimization Group, Oracle Labs

many problems, it can be shown that there exist simple districting strategies that are asymptotically optimal as the number of demand points becomes large [38, 48].

This paper describes a districting strategy developed for an industrial affiliate, a major North American mapping company, whose objective is to route a fleet of vehicles throughout a region to collect sensor data that is ultimately used to assemble full 3-dimensional point clouds of buildings and other city structures. The vehicles must visit specific locations within the region that correspond to anomalies that arise when "scrubbing" an existing city map across many CPUs. We will go into more detail on this in Section 4.2, but for now, the salient features that comprise this problem are as follows:

1. The service region is a compact Euclidean domain $\mathcal{R}$, and distances between points are Euclidean, or close to some "natural" metric such as the ℓ^1 or ℓ^∞ norm.

2. During each service period, a fleet of vehicles must use travelling salesman (TSP) tours to visit a collection of destination points in $\mathcal{R}$ that are sampled from a probability distribution f.

3. The goal is to partition $\mathcal{R}$ into districts, with one district per vehicle, in a way that is "optimal" as the number of destination points becomes large.

These problem attributes form the premises of a sizeable body of literature, with a particularly high concentration in sensor data collection, surveillance, and information retrieval. In this context, it is often the case that points are actually sampled over a long time horizon from a heterogeneous Poisson process whose underlying spatial distribution is f; it turns out that, in the heavy-traffic limit, this "online" routing problem has essentially the same characteristics as the static TSP, as can be seen in (for example) [13]. This heavy-traffic assumption is the basis for the second attribute above. It is worth noting that [13] also establishes that the *low-traffic* case of the "online" problem is more closely related to the k-median problem; this is an equally useful perspective, although it is not relevant in our particular application described in Section 4.2. A few examples of papers that use the three assumptions above include [37] and [15] (which use *power diagrams* and *ham sandwich cuts* respectively to design districts) as well as [41] (which uses *Voronoi partitions* to design districts).

One of the drawbacks to these preceding schemes is that one must have full knowledge of the demand distribution f in order to design districts (in particular, as will be explained in Section 2.1, the function $\sqrt{f(x)}$ turns out to have particular importance). In reality, one rarely has this luxury; the purpose of this paper is to describe a districting strategy when a fourth problem attribute is present, namely that:

4. The distribution f is unknown at the time that the districting decision is to be made, and there is only a set of independent samples from f available to make the districting decision.

There are many ways to handle this last attribute; one could, for example, fix a parametric form for f and perform a maximum likelihood estimation, or build an artificial density non-parametrically using kernel density estimation. In our particular application, described in Section 4.2, the number of samples available may be too small to reliably approximate f in such a way. In addition, distances throughout our region are usually not homogeneous due to traffic or poor road network topology, and thus there may be certain areas in $\mathcal{R}$ in which travel is expensive that are not reflected in our (small) set of samples (in other words, we may suspect that the vehicles will need to enter these expensive areas, even if none of our samples lies in them). Therefore, in this paper, we tackle this lack of knowledge using *distributionally robust optimization*, as has been done earlier for the VRP in [16].

With the above attributes in mind, the focus of this paper is the following problem: our inputs consist of an empirical distribution consisting of some sampled demand points in the service region $\mathcal{R}$, and our objective is to divide $\mathcal{R}$ into sub-regions in a way that balances the anticipated workloads in each sub-region. In order to estimate these workloads, we use distributionally robust optimization; that is, we construct probability distributions on the sub-regions that are "close" to this empirical distribution and are as "badly behaved" as possible, in the sense that the asymptotic length of a TSP tour of points drawn from that distribution should be as large as possible. In order to characterize our ambiguity set of distributions (i.e. what we mean by "close"), we use a statistical metric called the *Wasserstein distance*, which is also known as the *earth mover's* or *Kantorovich* metric. On the practical side, we give an algorithm for partitioning $\mathcal{R}$ that significantly out-performs other approaches when applied to the dataset provided by our industrial affiliate. On the theoretical side, we use Lagrange duality to derive a closed-form expression for these "badly behaved" distributions within each sub-region, and show how to compute them quickly using cutting planes.

Conceptually speaking, the Wasserstein distance is very simple and intuitive: if we visualize two probability distributions μ_1 and μ_2 as being two piles of equal amounts of sand, then the Wasserstein distance between them is simply the minimum amount of work needed to move one pile to take the shape of the other, as suggested in Figure 1a. The Wasserstein distance can also be thought of as an infinite-dimensional generalization of a bipartite matching, as suggested in Figures 1b-1d. A particularly attractive feature of the Wasserstein distance that is not present in many other statistical metrics is the ability to directly compare a discrete distribution and a continuous distribution, as illustrated in Figures 1e-1g. In addition, because the Wasserstein distance is a true metric, the set of all distributions within a certain distance of a reference distribution is a convex set that turns out to admit a simple representation.

This paper is structured as follows: Section 2 describes the structure of the worst-case spatial distribution for the TSP under a Wasserstein distance constraint and gives a primal-dual algorithm that finds this worst-case

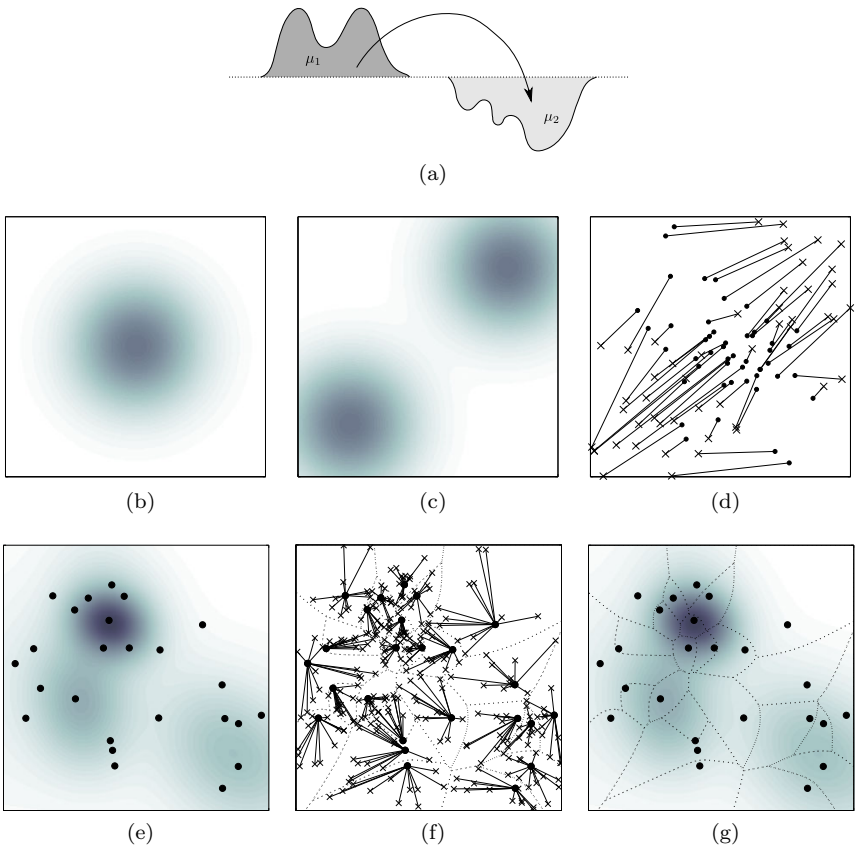


Figure 1: Figure 1a shows a Wasserstein distance problem between two univariate distributions μ_1 and μ_2. Figures 1b-1d show how a Wasserstein mapping can be thought of as an infinite-dimensional generalization of a bipartite matching. Figures 1e-1g show a Wasserstein distance problem when μ_1 is a smooth density and μ_2 is atomic. The two distributions are shown in 1e, and 1f shows the solution to an assignment problem between a large number of samples from μ_1 and the atomic distribution μ_2. Figure 1g shows the optimal mapping between the two distributions: each cell (indicated by the dashed lines) contains $1/n$ of the mass of the density, which is to be transported to the point contained within it.

distribution efficiently. After going over some practical details in Section 3, this algorithm is then implemented in two computational experiments involving both the single-vehicle and multi-vehicle TSP in Section 4.

1.1 Overview of the proposed scheme

This section is a "roadmap" of the scheme that we propose in this paper for partitioning a service region in a distributionally robust fashion. For a distribution f defined on a region $\mathcal{R}$, we let $\Phi(f)$ be a functional that characterizes the expected length of a TSP tour of a collection of independent samples of f. Obviously, the functional Φ depends on the number N of samples drawn, but we will suppress that in the interest of brevity and to emphasize the dependency on the distribution itself. The most obvious form for Φ for a given value of N would be to use $\Phi(f) = \mathbf{E}_f \, \mathsf{TSP}(X_1, \ldots, X_N)$, where $\mathbf{E}_f$ denotes the usual expectation operator with respect to f and

$\mathsf{TSP}(X_1, \ldots, X_N)$ denotes the length of a TSP tour of the points X_i. In fact, as we will see in Section 2.2, it turns out that there exists a sensible Φ that does not depend on N (since the scaling of the TSP tour can be described explicitly as N becomes large) and does not require an expectation operator (since we can actually establish *almost sure* convergence of the length of the TSP tour, rather than mere convergence in expectation).

The focus of this paper is on partitioning $\mathcal{R}$ into m districts, $D_1, \ldots, D_m$, in such a way as to minimize the maximum tour length in each of the districts. In other words, we seek to minimize

$$\max_{i \in \{1,\ldots,m\}} \Phi(f|_{D_i}),$$

where $f|_{D_i}$ denotes the restriction of f to district D_i. Thus, when f is given, the districting problem is written as

$$\begin{aligned} \underset{D_1,\ldots,D_m \subset \mathcal{R}}{\text{minimize}} \ \max_{i \in \{1,\ldots,m\}} \Phi(f|_{D_i}) \qquad & s.t. \\ \bigcup_{i=1}^{m} D_i &= \mathcal{R} \\ D_i \cap D_j &= \emptyset \quad \forall i \neq j \end{aligned} \tag{1}$$

where the two constraints require $D_1, \ldots, D_m$ to be a *partition* of $\mathcal{R}$. It is natural to impose a "shape condition" on the districts as well, such as convexity or contiguity. Problem (1) has been addressed previously in numerous studies, such as [37] and [15], both of which add an additional constraint that the D_i's must be convex (provided that $\mathcal{R}$ is convex to start with).

In this paper, we assume that f is not given, but is merely known to lie in an *ambiguity set* $\mathcal{P}$. Thus, our problem now takes the form

$$\begin{aligned} \underset{D_1,\ldots,D_m \subset \mathcal{R}}{\text{minimize}} \ \max_{i \in \{1,\ldots,m\}} \sup_{f \in \mathcal{P}} \Phi(f|_{D_i}) \qquad & s.t. \\ \bigcup_{i=1}^{m} D_i &= \mathcal{R} \\ D_i \cap D_j &= \emptyset \quad \forall i \neq j\,. \end{aligned}$$

This problem was addressed in [16] for the case where $\mathcal{P}$ is characterized by first and second moment inequalities. In this paper, we address the problem above for the case where $\mathcal{P}$ is defined by the *Wasserstein distance* to a set of given samples (which we will define in Section 2.1): if we let $\mathcal{D}(f, \hat{f})$ denote this distance, where $\hat{f}$ denotes an

empirical distribution on samples $x_1, \ldots, x_n$, then our problem takes the form

$$\begin{aligned} \underset{D_1,\ldots,D_m \subset \mathcal{R}}{\text{minimize}} \max_{i \in \{1,\ldots,m\}} \sup_{f:\mathcal{D}(f,\hat{f}) \leq t} \Phi(f|_{D_i}) \qquad & s.t. \\ \bigcup_{i=1}^{m} D_i &= \mathcal{R} \\ D_i \cap D_j &= \emptyset \quad \forall i \neq j \end{aligned}$$

where t is a fixed distance threshold.

We are initially concerned with the "inner" problem of the above, that is, determining the worst-case distributions f^* that result in large values of Φ. After describing the structure of these distributions in Section 2, we then turn to the matter of determining districts that minimize the worst-case workloads effectively.

1.2 Related work

This paper describes a continuous approximation model that uses robust optimization to describe the worst-case demand distribution for the travelling salesman problem; this model is then applied to solve a districting problem that assigns vehicles to pre-specified zones in a region. As such, there are essentially three bodies of literature from which it stems.

1.2.1 Continuous approximation models

This paper is concerned with a *continuous approximation* model for a transportation problem. The basic premise of the continuous approximation paradigm is that one replaces combinatorial quantities that are difficult to compute with simpler mathematical formulas, which (under certain conditions) provide accurate estimations of the desired quantity [19]. Such approximations exist for many combinatorial problems, such as the travelling salesman problem [7], facility location [30, 36], and any *subadditive Euclidean functional* such as a minimum spanning tree, Steiner tree, or matching [42]. In our computational districting experiment, an approximation of this kind is used as the first level of an optimization problem in which we design service zones that are associated with different vehicles. In a nutshell, we apply the famous square-root law from [7] to estimate the length of a TSP tour of some points in a region.

1.2.2 Districting problems in vehicle routing

The primary application of the theory derived in this paper is in the design of *districts* for allocating a fleet of vehicles to visit a collection of customers when demand is uncertain. The problem of designing such districts is a

foundational one in the continuous approximation literature, as can be seen in Chapter 4 of the seminal book [19], for example. Our problem is motivated by an industrial affiliate, a major North American mapping company, that uses a districting strategy to assign vehicles to scan metropolitan regions. This kind of geographic localization – as opposed to explicitly computing turn-by-turn directions for vehicles – is a common strategy that is also used by Google Street View, for example [28]:

> [Google spokesman Larry] Yu initially stated drivers were given specific routes to follow. But a Street View driver, who asked to remain anonymous for employment reasons, said he was simply told to drive around Sonoma County and collect images. Yu retracted his assertion after learning of the driver's statement.

Rather than collecting street-level photographs, our affiliate assembles full 3-dimensional point clouds of buildings and other city structures (among many other data sources, e.g. temperature, radio signals, and sunlight), which necessitates more control over driver behavior than that described by the quote above. As mentioned in the introduction, the most common way that uncertainty is represented is by assuming that demand follows a known probability density function (which is often further assumed to be uniform); this density then informs the districting decision in some way.

An alternative method to the preceding continuous models is to instead assume that demand is present on a known graph, and that vertices on the graph have probability weights. This is the approach taken by [29], which models the districting problem as a two-stage stochastic optimization program with recourse, and by [6], which uses a three-phase procedure that aggregates data points into compact districts using a mixed-integer goal program. It is also possible to apply principles from continuous approximation theory to design districts in graph-based models, provided some basic geometric information is available; this is the case in [32, 33], which use the square-root approximation of [7] in conjunction with a graph-based model, and which show good performance when the inputs are known to be uniformly distributed over a geometric domain. Section 4.2 of this paper also shows how to apply a continuous approximation scheme to a heterogeneous road network to a set of inputs that are non-uniformly distributed.

As stated earlier, the specific contribution of this paper is a method that designs districts in a *data-driven* fashion; rather than having full distribution information, we assume that we are given only a collection of data samples that inform our districting decision. The problem of designing districts in such a way (i.e. when one has an ambiguous distribution setting) is considerably less understood, although the paper [16] describes one approach for doing so when one knows the mean and covariance of the demand distribution. A major deficiency of this approach, which motivates our present work, is its inability to respond to *clustering* or even mere *multi-modality* in

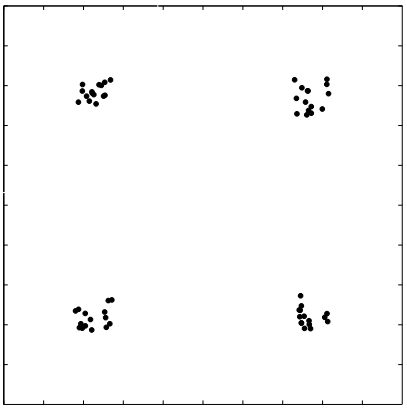


Figure 2: The above point sets in the unit square are extremely clustered and one would expect that their TSP tour should be short. However, because their sample mean and covariance matrix are the same as that of the uniform distribution, any robust methodology that uses only mean and covariance information will fail to recognize the clustering, thereby incurring significant overestimation.

data points. For example, Figure 2 shows a data set that is very clustered, and whose TSP tour should therefore be short relative to (for example) a uniform distribution. However, this clustered data set actually has precisely the same mean and covariance matrix as the uniform distribution; such an approach therefore frequently leads to an over-conservative solution (or more generally, a solution that is not faithful to the true unknown demand distribution), even when a large number of samples is available. This over-conservatism is actually noted in Figure 10 of [16].

1.2.3 Robust optimization and vehicle routing

In most models of the robust VRP, one has a pre-defined ambiguity region and seeks a set of routes that is as good as possible with respect to all of the outcomes; this ambiguity region is usually described as a polyhedral set [1, 26, 43], although the recent paper [2] adopts a "robust mean-variance" approach that minimizes a weighted sum of the average cost and the variance of a route when sampled over many scenarios. In our problem, we are concerned with robustness in the *distributional* sense: we seek the spatial distribution of demand for which the expected cost of a tour is as high as possible, while remaining consistent with some observed data samples or some parameters derived thereof.

By far, the most common parameters used in distributionally robust problems (in general domains, not just those arising in transportation) are the support and the first and second moments of the sample distribution [20, 39]. We have previously made use of first and second moment information for the distributionally robust VRP in the paper [16]; one major drawback of this method is an inability to detect clustering, as we have already noted in the preceding section. In order to remedy this, we propose the use of the Wasserstein distance as a means of defining

the uncertainty region of demand distributions. To our knowledge, the first direct applications of the Wasserstein distance to optimization problems have occurred very recently in [46, 47]; the former uses Carathéodory-type results to reduce the support set of an infinite-dimensional optimization problem to a finite set and the latter uses the Wasserstein distance as one of several statistical metrics to define risk measures for portfolios. Even more recently, the papers [22, 24] shows how to apply complementary slackness principles to solve a very large family of distributionally robust optimization problems subject to Wasserstein distance constraints; their use of convex duality theory is closely related to our own derivation in Section 2.2. For general problems (i.e. not those related specifically to vehicle routing), a variety of other statistical metrics (or *pseudo-metrics*) have been used previously for solving distributionally robust optimization problems; such metrics include the Kullback-Leibler divergence, Hellinger distance, χ^2-distance, total variation distance, or Kolmogorov-Smirnov statistic [8].

There are three reasons why the Wasserstein distance is a particularly appropriate choice for our problem of interest: first, the Wasserstein distance allows one to directly make comparisons between a discrete distribution (such as the empirical distribution consisting of a collection of data points) and a continuous distribution, as we have previously noted in Figure 1; this it not possible in (for example) the Kullback-Leibler divergence, the Hellinger distance, or the total variation distance. Secondly, the Wasserstein distance is in a sense "inherited" from the Euclidean distance, inasmuch as the distance between two distributions is defined as an integral of Euclidean distances. Since we are concerned with obtaining a probability distribution whose induced TSP tour is as long as possible (in an asymptotic limit as many samples are taken), and a TSP tour is also measured using Euclidean distances, the Wasserstein distance is a particularly appropriate choice. The third reason is purely practical: it turns out that the ambiguity set of distributions characterized by a Wasserstein distance threshold gives a very concise, closed-form expression for the worst-case distribution for our problem. As we will later show in Section 3.2, a fourth *a posteriori* justification for the use of the Wasserstein metric is that the worst-case distribution that one obtains for this problem is closely related to that of the classical *geographical gravity model*, which arises in many models of spatial interaction.

1.3 Notational conventions

Our notational conventions throughout this paper are as follows: integrals over regions in $\mathbb{R}^2$ are denoted with the double integral sign $\iint dA$. The diameter of a region $\mathcal{R}$, denoted $\text{diam}(\mathcal{R})$, is the largest possible distance between two points in $\mathcal{R}$, $\sup_{x,y\in\mathcal{R}} \|x-y\|$. The vector consisting of all 1's is written $\mathbf{e}$, whose dimension will always be clear from context, and the indicator function of a particular condition and the Dirac delta function are written as $\mathbb{1}(\cdot)$ and $\delta(\cdot)$ respectively. The Wasserstein distance between two distributions is written $\mathcal{D}(\cdot,\cdot)$ and is defined in

the next section. We will commit a slight abuse of notation and use the expression $\mathsf{TSP}(x_1, \ldots, x_n)$ to represent both the shortest tour that goes through a set of points as well as the length of that shortest tour. Finally, for any univariate function $f(x)$, we say that $f(x) \in o(g(x))$ as $x \to \infty$ if $\lim_{x\to\infty} f(x)/g(x) = 0$.

2 Analysis

Having established our problem of interest and our plan of attack, we now establish some preliminary results in order to solve our problem concretely.

2.1 Preliminaries

In order to retain mathematical rigor, we find the following results useful:

Definition 1 (Wasserstein distance)**.** Let μ_1 and μ_2 denote two probability measures defined on a compact planar region $\mathcal{R}$. The *Wasserstein distance* between μ_1 and μ_2, written $\mathcal{D}(\mu_1, \mu_2)$, is defined as

$$\mathcal{D}(\mu_1, \mu_2) := \inf_{\pi \in \Pi(\mu_1, \mu_2)} \iiiint_{\mathcal{R}\times\mathcal{R}} \|x - y\| \, d\pi(x, y) \,, \tag{2}$$

where $\Pi(\mu_1, \mu_2) \ni \pi(x, y)$ is defined as the set of all probability measures on $\mathcal{R} \times \mathcal{R}$ whose marginals are μ_1 and μ_2, that is, the set of all probability measures that satisfy $\pi(A \times \mathcal{R}) = \mu_1(A)$ and $\pi(\mathcal{R} \times B) = \mu_2(B)$ for all measurable subsets $A, B \subset \mathcal{R}$.

The Wasserstein distance can be thought of as a generalization of an *assignment problem*: for example, when μ_1 and μ_2 are discrete distributions consisting of n points each with equal mass, the Wasserstein distance between the two is simply computed as the cost of a bipartite matching (multiplied by a normalization term of $1/n$). This interpretation is suggested in Figure 1.

Our notion of distributional robustness relies on the following famous theorem, originally stated in [7] and further developed in [42], which relates the length of a TSP tour of some points with the distribution from which they were sampled:

Theorem 2 (Beardwood-Halton-Hammersley (BHH) Theorem)**.** *Suppose that $X = \{X_1, X_2, \ldots\}$ is a sequence of random points i.i.d. according to a probability density function $f(\cdot)$ defined on a compact planar region $\mathcal{R}$. Then with probability one, the length $\mathsf{TSP}(X)$ of the optimal travelling salesman tour through X satisfies*

$$\lim_{N\to\infty} \frac{\mathsf{TSP}(X)}{\sqrt{N}} = \beta \iint_{\mathcal{R}} \sqrt{f_c(x)} \, dA$$

where β is a constant and $f_c(\cdot)$ represents the absolutely continuous part of $f(\cdot)$.

Note in particular that β does not depend on $f(\cdot)$. It is additionally known that $0.6250 \leq \beta \leq 0.9204$ and estimated that $\beta \approx 0.7124$; see [4, 7].

The following classical result from [35] will be useful in confirming the existence of an optimal solution of the problem that we will construct:

Theorem 3 (Lagrange Duality)**.** *Let ϕ be a real-valued convex functional defined on a convex subset Ω of a vector space $\mathcal{X}$, and let $\mathcal{G}$ be a convex mapping of $\mathcal{X}$ into a normed space $\mathcal{Z}$. Suppose there exists an χ_1 such that $\mathcal{G}(\chi_1) < \theta$, where θ is the zero element, and that* $\inf\{\phi(\chi) : \chi \in \Omega,\ \mathcal{G}(\chi) \leq \theta\}$ *is finite. Then*

$$\inf_{\chi \in \Omega,\ \mathcal{G}(\chi) \leq \theta} \phi(\chi) \quad = \quad \max_{z^* \geq \theta} \inf_{\chi \in \Omega} \phi(\chi) + \langle \mathcal{G}(\chi), z^* \rangle$$

and the maximum on the right is achieved by some $z_0^* \in \mathcal{Z}^*$ such that $z_0^* \geq \theta$, where $\mathcal{Z}^*$ denotes the dual space of $\mathcal{Z}$ and $\langle \cdot, \cdot \rangle$ denotes the evaluation of a linear functional, i.e. $z^*(\mathcal{G}(\chi))$. If the infimum on the left is achieved by some $\chi_0 \in \Omega$, then $\langle \mathcal{G}(\chi_0), z_0^* \rangle = 0$, and χ_0 minimizes $\phi(\chi) + \langle \mathcal{G}(\chi), z_0^* \rangle$ over all $\chi \in \Omega$.

Finally, the Wasserstein distance between a discrete distribution consisting of points $\{x_1, \ldots, x_n\}$ with uniform probabilities $1/n$ and a continuous probability density function f defined on a compact planar region $\mathcal{R}$ can be obtained by solving the following infinite-dimensional optimization problem:

$$\begin{aligned}
\underset{I_1(\cdot),\ldots,I_n(\cdot)}{\text{minimize}} \sum_{i=1}^{n} \iint_{\mathcal{R}} \|x - x_i\| f(x) I_i(x)\, dA \qquad & s.t. \\
\iint_{\mathcal{R}} f(x) I_i(x)\, dA &= 1/n \quad \forall i \\
\sum_{i=1}^{n} I_i(x) &= 1 \quad \forall x \in \mathcal{R} \\
I_i(x) &\geq 0 \quad \forall i, \forall x \in \mathcal{R}\,;
\end{aligned}$$

here the value $I_i(x)$ simply describes the amount of the distribution at point $x \in \mathcal{R}$ that should be moved to point x_i. The lemma below summarizes some basic results on the Wasserstein distance between a probability density and an empirical distribution:

Lemma 4. *Let f denote a probability density function on a compact planar region $\mathcal{R}$ and let $\hat{f}$ denote an atomic distribution consisting of distinct points $x_1, \ldots, x_n \in \mathcal{R}$ each having probability mass $1/n$. Then the following statements are true:*

1. *The Wasserstein distance* $\mathcal{D}(f, \hat{f})$ *is the optimal objective value to the concave maximization problem*

$$\begin{aligned} \underset{\boldsymbol{\lambda} \in \mathbb{R}^n}{\text{maximize}} \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\} \, dA \qquad & s.t. \\ \mathbf{e}^T \boldsymbol{\lambda} \quad &= \quad 0 \,, \end{aligned} \tag{3}$$

where $\mathbf{e} \in \mathbb{R}^n$ *denotes a vector whose entries are all* 1*'s.*

2. *For any* $\boldsymbol{\lambda}$*, a valid* supergradient *for the objective function of (3) is the vector* $\mathbf{g} \in \mathbb{R}^n$ *defined by setting*

$$g_i = - \iint_{R_i} f(x) \, dA \,,$$

where each R_i *is a connected piecewise hyperbolic region characterized by*

$$R_i = \{x \in \mathcal{R} : \|x - x_i\| - \lambda_i \leq \|x - x_j\| - \lambda_j \quad \forall j \neq i\} \;;$$

that is, for any other $\boldsymbol{\lambda}'$*, we have*

$$\iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i'\} \, dA \leq \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\} \, dA + \mathbf{g}^T (\boldsymbol{\lambda}' - \boldsymbol{\lambda}) \,.$$

3. *If* $\boldsymbol{\lambda}^*$ *is a maximizer of (3), then an optimal Wasserstein mapping between* f *and* $\hat{f}$ *is obtained by defining*

$$R_i^* = \left\{x \in \mathcal{R} : \|x - x_i\| - \lambda_i^* \leq \|x - x_j\| - \lambda_j^* \quad \forall j \neq i\right\}$$

for each i *and transporting all of the mass of each* R_i^* *to its associated point* x_i*.*

4. *If* $f(x) > 0$ *for all* $x \in \mathcal{R}$*, then there exists a unique maximizer* $\boldsymbol{\lambda}^*$*.*

Proof. Statement 1 is a well-known special case of the *Kantorovich duality theorem*; see for example Theorem 1.3 of [44] for specific details. Proofs of statements 2-4 are routine and can be found in Section A of the Online Supplement. □

The key insight that the above theorem yields is the reduction of an infinite-dimensional problem to a finite-dimensional problem by way of Lagrange duality. Statement 3, for example, is a direct consequence of complementary slackness; it says that if $\|x - x_i\| - \lambda_i^* < \|x - x_j\| - \lambda_j^*$, then the optimal Wasserstein mapping from f and $\hat{f}$ must not transport any mass near x to point x_j. The economic interpretation of the regions R_i^* relative to the dual

variables λ_i^* can be found in [18]; in a nutshell, the sub-regions R_i^* that characterize the mapping are equivalent to market regions induced by a mill pricing scheme at each of the points x_i.

2.2 Worst-case distributions with Wasserstein distance constraints

The input to our problem is a set of distinct demand points $x_1, \ldots, x_n$ in a compact planar region $\mathcal{R}$, which we assume are sampled from some (unknown) distribution function f. By rearranging the terms of Theorem 2, we can write

$$\mathsf{TSP}(x_1, \ldots, x_n) = \beta\sqrt{n} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA + o(\sqrt{n})$$

with probability one as $n \to \infty$. Since β is a constant and $\sqrt{n}$ is independent of the distribution f by definition, we therefore conclude that the "worst" distribution whose induced TSP workload is as large as possible (subject to whatever other constraints might be present) is precisely that distribution that maximizes $\iint_{\mathcal{R}} \sqrt{f(x)}\, dA$. In the terminology of Section 1.1, this means that we can use $\Phi(f) = \iint_{\mathcal{R}} \sqrt{f(x)}\, dA$ and that $\Phi(f|_{D_i}) = \iint_{D_i} \sqrt{f(x)}\, dA$ for a district $D_i \subset \mathcal{R}$.

We now let $\hat{f}$ denote the empirical distribution on these n points x_i. We will search through all distributions f whose Wasserstein distance to $\hat{f}$ is sufficiently small, i.e. where $\mathcal{D}(f, \hat{f}) \leq t$; here $\mathcal{D}(\cdot, \cdot)$ is the Wasserstein distance from Definition 1 and t is a parameter that will be discussed in Section 3.1. The problem of finding the worst-case TSP distribution, subject to the Wasserstein distance constraint, is then written as the infinite-dimensional convex optimization problem

$$\begin{aligned} \underset{f}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}(f, \hat{f}) &\leq t \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{4}$$

This is our problem of interest throughout this paper. We will embed f in the Banach space $L^1(\mathcal{R})$, which consists of all functions that are absolutely Lebesgue integrable on $\mathcal{R}$.

2.2.1 Structure of the worst-case distribution

Here we describe the solution to (4). To begin, we apply Lemma 4 to express the distance constraint $\mathcal{D}(f, \hat{f}) \leq t$ in (4) differently, obtaining the equivalent formulation

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA \quad &\leq \quad t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0 \\ \iint_{R} f(x)\, dA \quad &= \quad 1 \\ f(x) \quad &\geq \quad 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{5}$$

This is an infinite-dimensional problem with an infinite-dimensional constraint space and is therefore best addressed using Theorem 3; before doing so, we find the following result useful:

Lemma 5. *There exists a unique optimal solution f^* to problem (5), and $f^*(x) > 0$ for all $x \in \mathcal{R}$.*

Proof. The fact that the optimal solution is unique (provided one exists) is an immediate consequence of the fact that the square root function in the integrand of (5) is strictly concave. To prove existence, let $\{f^j\}$ denote a sequence of feasible inputs to (5) whose objective values converge to a supremum. For each f^j, let $\boldsymbol{\lambda}^j$ denote a value of $\boldsymbol{\lambda}$ that induces the optimal Wasserstein mapping between f^j and $\hat{f}$ as described in Lemma 4, i.e. that solves problem (3). It is easy to verify that the iterates $\boldsymbol{\lambda}^j$ lie in the compact set Λ, defined by

$$\Lambda := \left\{ \boldsymbol{\lambda} \in \mathbb{R}^n : \mathbf{e}^T \boldsymbol{\lambda} = 0,\, \lambda_i \leq \operatorname{diam}(\mathcal{R})\, \forall i \right\},$$

because any $\boldsymbol{\lambda}$ lying outside Λ would force some sub-regions to be empty. Therefore, the sequence $\{\boldsymbol{\lambda}^j\}$ must have a convergent subsequence with a limit $\boldsymbol{\lambda}^*$, inducing a partition $R_1^*, \ldots, R_n^*$ as in statement 2 of Lemma 4 that satisfies $\iint_{R_i^*} f(x)\, dA = 1/n$ for all i. Standard arguments then show that the true worst-case distribution f^* is precisely the solution to the problem

$$\begin{aligned}
\underset{f\in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\
\sum_{i=1}^{n} \iint_{R_i^*} \|x - x_i\| f(x)\, dA &\leq t \\
\iint_{R_i^*} f(x)\, dA &= \frac{1}{n} \quad \forall i \\
f(x) &\geq 0 \quad \forall x \in \mathcal{R}
\end{aligned} \tag{6}$$

(this is an immediate consequence of the fact that the optimal objective cost to (6) varies continuously as the vector $\boldsymbol{\lambda}^*$, which defines the partition $R_1^*, \ldots, R_n^*$, is perturbed). This problem has a finite-dimensional constraint space, and it is routine to apply Theorem 3 to (6) to derive the dual problem

$$\begin{aligned}
\underset{\boldsymbol{\nu}\geq \mathbf{0}}{\text{minimize}}\, \frac{1}{4} \sum_{i=1}^{n} \iint_{R_i^*} \frac{1}{\nu_0 \|x - x_i\| + \nu_i}\, dA + \nu_0 t + \frac{1}{n}(\nu_1 + \cdots + \nu_n) \qquad & s.t. \\
\nu_0 \|x - x_i\| + \nu_i &\geq 0 \quad \forall x \in R_i^* \quad \forall i
\end{aligned} \tag{7}$$

whereby we conclude that the optimal solution f^* to (6) must take the form

$$f^*(x) = \frac{1}{4(\nu_0^* \|x - x_i\| + \nu_i^*)^2} \tag{8}$$

on each sub-region R_i^*. This satisfies $f(x) > 0$ for all $x \in \mathcal{R}$ and completes the proof. □

The functional form for the optimal f^* can in fact be simplified further:

Theorem 6. *The worst-case distribution that solves problem (5), and therefore (4), takes the form*

$$f^*(x) = \frac{1}{4(\nu_0^* \min_i\{\|x - x_i\| - \lambda_i^*\} + \nu_1^*)^2} \tag{9}$$

with $\nu_0^*, \nu_1^* \geq 0$ *and* $\mathbf{e}^T \boldsymbol{\lambda}^* = 0$.

Proof. The major difference between the form of f^* as written above and the form described in (8) is the fact that the expression in (8) is not guaranteed to vary continuously as we move from one region R_i^* to another; the

expression (9) is continuous by inspection. We first note that the constraint

$$\iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\} \, dA \leq t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0$$

can be restricted to merely the compact set

$$\Lambda := \left\{ \boldsymbol{\lambda} \in \mathbb{R}^n : \mathbf{e}^T \boldsymbol{\lambda} = 0, \, \lambda_i \leq \operatorname{diam}(\mathcal{R}) \, \forall i \right\}$$

because, if $\lambda_i > \operatorname{diam}(R)$ for some i, then $\|x - x_i\| - \lambda_i < 0$ for all x and the constraint is obviously satisfied. We will apply Theorem 3 where $\mathcal{X} = L^1(\mathcal{R})$, Ω is the subset of the non-negative orthant in $L^1(\mathcal{R})$ that integrates to 1, and $\mathcal{Z}$ consists of all continuous functions on Λ, i.e. $\mathcal{Z} = \mathscr{C}(\Lambda)$ (note that $\mathcal{Z}$ satisfies the interior point requirement of Theorem 3 because inequalities are simply taken elementwise in Λ). We define $\phi(\chi) : \mathcal{X} \to \mathbb{R}$ and $\mathcal{G}(\chi) : \mathcal{X} \to \mathcal{Z}$ as the maps sending

$$f \mapsto \iint_{\mathcal{R}} \sqrt{f(x)} \, dA$$

and

$$f \mapsto \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\} \, dA - t$$

respectively, where the right-hand side of the second expression is regarded as a continuous function of $\boldsymbol{\lambda}$. The dual space $\mathcal{Z}^*$ consists of all *regular signed Borel measures* on Λ (this is the *Riesz representation theorem*; see e.g. [40]). However, Lemma 5 shows that $f^*(x) > 0$ on $\mathcal{R}$, and therefore the optimal $\boldsymbol{\lambda}^*$ that solves problem (3) is unique by statement 4 of Lemma 4. This implies that $\mathcal{G}(\chi) = \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\} \, dA - t < 0$ whenever $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^*$, and therefore, since $\langle \mathcal{G}(\chi), z^* \rangle = 0$ at optimality, it must be the case that z^* is zero everywhere except for (possibly at) $\boldsymbol{\lambda}^*$. Thus, we conclude that z^* is an evaluation functional at $\boldsymbol{\lambda}^*$ (multiplied by a scalar), so that

$$\langle \mathcal{G}(\chi), z^* \rangle = q^* \left(\iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i^*\} \, dA - t \right)$$

for all feasible f, where $q^* \geq 0$ is some scalar. Theorem 3 then says that f^* must also be the solution to the problem

$$\begin{aligned}
\underset{f\in L^1(\mathcal{R})}{\text{maximize}} \iint\limits_{\mathcal{R}} \sqrt{f(x)}\,dA + q^*\left(\iint\limits_{\mathcal{R}} f(x)\min_i\{\|x-x_i\|-\lambda_i^*\}\,dA - t\right) \qquad & s.t. \\
\iint\limits_{\mathcal{R}} f(x)\,dA &= 1 \\
f(x) &\geq 0 \quad \forall x\in\mathcal{R}
\end{aligned}$$

or equivalently, the problem

$$\begin{aligned}
\underset{f\in L^1(\mathcal{R})}{\text{maximize}} \iint\limits_{\mathcal{R}} \sqrt{f(x)}\,dA \qquad & s.t. \\
\iint\limits_{\mathcal{R}} f(x)\min_i\{\|x-x_i\|-\lambda_i^*\}\,dA &\leq t \\
\iint\limits_{\mathcal{R}} f(x)\,dA &= 1 \\
f(x) &\geq 0 \quad \forall x\in\mathcal{R}\,.
\end{aligned}$$

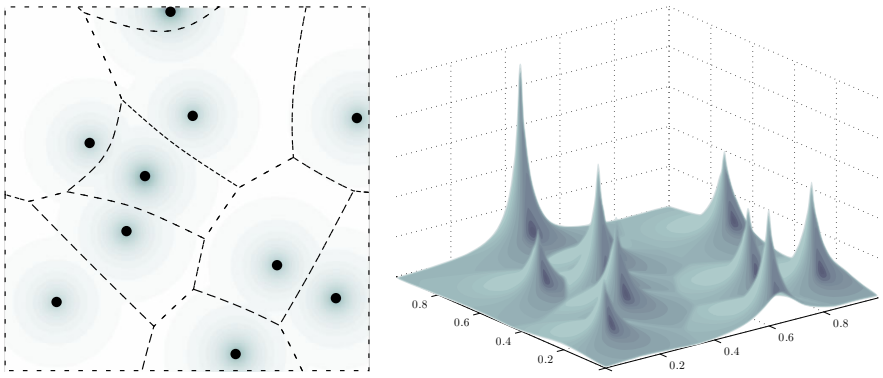
It is routine to verify that the constraint $\iint_{\mathcal{R}} f(x)\,dA = 1$ can be replaced with an inequality (in a nutshell, this is because we are allowed to make $f(x)$ as large as we like when $\|x-x_i\|-\lambda_i^* \leq 0$ for some index i). Thus, we can apply Theorem 3 again to the problem

$$\begin{aligned}
\underset{f\in L^1(\mathcal{R})}{\text{maximize}} \iint\limits_{\mathcal{R}} \sqrt{f(x)}\,dA \qquad & s.t. \\
\iint\limits_{\mathcal{R}} f(x)\min_i\{\|x-x_i\|-\lambda_i^*\}\,dA &\leq t \\
\iint\limits_{\mathcal{R}} f(x)\,dA &\leq 1 \\
f(x) &\geq 0 \quad \forall x\in\mathcal{R}
\end{aligned}$$

to derive the 2-dimensional dual problem

$$\begin{aligned}
\underset{\nu_0,\nu_1}{\text{minimize}} \iint\limits_{\mathcal{R}} \frac{1}{4(\nu_0\min_i\{\|x-x_i\|-\lambda_i^*\}+\nu_1)}\,dA + \nu_0 t + \nu_1 \qquad & s.t. \\
\nu_0\min_i\{\|x-x_i\|-\lambda_i^*\}+\nu_1 &\geq 0 \quad \forall x\in\mathcal{R} \\
\nu_0,\nu_1 &\geq 0\,;
\end{aligned} \tag{10}$$

the optimality conditions of (10) describe precisely the desired form of f^*, which completes the proof. □



(a) (b)

Figure 3: Two views of an example of $f^*(x)$ as described in Theorem 6, where there are $n = 10$ points. In 3a, the shading represents $f^*(x)$ and the dashed lines indicate the optimal Wasserstein map between f^* and $\hat{f}$. By construction, the vector $\boldsymbol{\lambda}^*$ induces those dashed lines, as described in Lemma 4.

Figure 3 shows a surface plot of a worst-case distribution f^*.

Remark 7. Many problems in distributionally robust optimization have objective functions and constraints that are linear in terms of the unknown distribution f (for example, the expectation operator). For such problems, Carathéodory-type theorems imply that the worst-case distribution will consist of a finite number of points, even when one uses ambiguity sets defined by the Wasserstein distance; see for example [46]. As a consequence of this fact, it is sometimes the case that one can determine the worst-case distribution using only one iteration of a finite-dimensional optimization problem; this turns out to hold (for the Wasserstein metric) in [22, 47], for example. Because of the non-linearity in the objective, our worst-case distribution f^* is smooth; we will describe an iterative scheme for finding it in Section 2.3.

Remark 8. One of the salient attributes of the worst-case distribution f^* is that the presence of the square root in the objective of (4) establishes an inverse proportionality between the optimal solution $f^*(x)$ and the *square* of the distance to one of the data points x_i (with some additional additive and multiplicative weights from the dual variables $\boldsymbol{\nu}^*$ and $\boldsymbol{\lambda}^*$). This same inverse proportionality is shared by the classical *geographic gravity model* [3], which is "the most common formulation of the spatial interaction method" and has historically been used to model a wide variety of demographic phenomena such as population migration, spatial utility for retail stores, and trip distributions between cities. This would appear to lend credibility to our solution f^*, inasmuch as it takes a form that closely matches that of distributions for related problems.

Remark 9. The earlier paper [16] considers a problem closely related to (4) in which one has constraints on the

mean and covariance of f instead of the constraint on $\mathcal{D}(f, \hat{f})$; that problem is written as

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} x f(x)\, dA &= \mu \\ \iint_{\mathcal{R}} x x^T f(x)\, dA &\preceq \Sigma + \mu\mu^T \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{11}$$

As we have noted in Figure 2, mean and covariance information may not be sufficient to give any useful information on the ambiguity set of distributions. It is more desirable to describe this ambiguity set in a way that is guaranteed to make better use of sample points as they become available. It is well-known (e.g. Section 2 of [14]) that the Wasserstein distance between the empirical distribution $\hat{f}$ and the true distribution f converges to zero with probability one as $n \to \infty$. This, coupled with some routine analysis, guarantees that the objective cost of our proposed formulation (4) will eventually be the same as that of the ground truth distribution:

Theorem 10. *Let $X = \{X_1, X_2, \dots\}$ be a sequence of random points i.i.d. according to an absolutely continuous probability density function $\bar{f}(\cdot)$ defined on a compact planar region $\mathcal{R}$. For any positive integer n, let $\hat{f}_n$ denote the empirical distribution on points $\{X_1, \dots, X_n\}$. Then with probability one there exists a sequence $\{t_1, t_2, \dots\}$, converging to 0, such that the optimal objective value of the problem*

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}(f, \hat{f}_n) &\leq t_n \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R} \end{aligned} \tag{12}$$

approaches the ground truth (i.e. $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$) as $n \to \infty$.

Proof. See Section B of the Online Supplement. □

2.3 Finding the worst-case distribution efficiently

The preceding section established that the worst-case distribution that solves (4) can be expressed in terms of optimal vectors $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ and $\boldsymbol{\nu}^* \in \mathbb{R}^2$. This section describes a simple method for calculating $\boldsymbol{\lambda}^*$ and $\boldsymbol{\nu}^*$ efficiently by way of an *analytic center cutting plane method* [11]. Recall that our problem of interest, as written in (5), is

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA &\leq t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0 \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,; \end{aligned}$$

thus, it is certainly true that if we fix any specific value $\bar{\boldsymbol{\lambda}}$ such that $\mathbf{e}^T \bar{\boldsymbol{\lambda}} = 0$, then the following problem is a relaxation of (5) and hence has an objective value that is at least as large as that of (5):

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \bar{\lambda}_i\}\, dA &\leq t \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned}$$

It is natural to consider the problem of selecting the particular value of $\bar{\boldsymbol{\lambda}}$ that makes the above relaxation as tight as possible. In fact, our proof of Theorem 6 says that there exists a particular value of $\bar{\boldsymbol{\lambda}}$, namely $\boldsymbol{\lambda}^*$, such that the above relaxation is actually tight; in other words, the optimal distribution f^* is the solution to the problem

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i^*\}\, dA &\leq t \\ \iint_{\mathcal{R}} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R} \end{aligned}$$

for an appropriately chosen vector $\boldsymbol{\lambda}^*$. Thus, the problem of finding $\boldsymbol{\lambda}^*$ actually reduces to the optimization problem

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^n}{\text{minimize}} \max_{f\in\Omega(\boldsymbol{\lambda})} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathbf{e}^T\boldsymbol{\lambda} &= 0 \\ \lambda_i &\leq \operatorname{diam}(\mathcal{R}) \quad \forall i \end{aligned} \tag{13}$$

where $\Omega(\boldsymbol{\lambda})$ is the subset of $L^1(\mathcal{R})$ consisting of all functions f such that

$$\begin{aligned} \iint_{\mathcal{R}} f(x) \min_i\{\|x - x_i\| - \lambda_i\}\, dA &\leq t \\ \iint_{\mathcal{R}} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned}$$

Of course, the inner problem of maximizing f given $\boldsymbol{\lambda}$ is easily solved because the gradient vector for the dual problem

$$\begin{aligned} \underset{\nu_0,\nu_1}{\text{minimize}} \iint_{\mathcal{R}} \frac{1}{4(\nu_0 \min_i\{\|x - x_i\| - \lambda_i\} + \nu_1)}\, dA + \nu_0 t + \nu_1 \qquad & s.t. \\ \nu_0 \min_i\{\|x - x_i\| - \lambda_i\} + \nu_1 &\geq 0 \quad \forall x \in \mathcal{R} \\ \nu_0, \nu_1 &\geq 0\,, \end{aligned} \tag{14}$$

as derived in the proof of Theorem 6, can be computed explicitly. Thus, we simply require a better understanding of problem (13):

Lemma 11. *The (outer) objective function of problem (13) is quasiconvex, i.e. its sub-level sets are convex.*

Proof. For notational compactness, let $G(\boldsymbol{\lambda})$ denote the objective function of (13). Recall [12] that $G(\boldsymbol{\lambda})$ is quasiconvex if and only if, for any $\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2$ and any $\theta \in [0,1]$, we have

$$G(\theta\boldsymbol{\lambda}^1 + (1-\theta)\boldsymbol{\lambda}^2) \leq \max\{G(\boldsymbol{\lambda}^1), G(\boldsymbol{\lambda}^2)\}\,.$$

Let $\bar{\boldsymbol{\lambda}} = \theta\boldsymbol{\lambda}^1 + (1-\theta)\boldsymbol{\lambda}^2$ and let $\bar{f}$ denote the distribution that maximizes $\iint_{\mathcal{R}} \sqrt{f(x)}\, dA$ over all $f \in \Omega(\bar{\boldsymbol{\lambda}})$; it will

suffice to prove that either $\bar{f} \in \Omega(\boldsymbol{\lambda}^1)$ or $\bar{f} \in \Omega(\boldsymbol{\lambda}^2)$. By definition, we have

$$\iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \bar{\lambda}_i\} \, dA \leq t \,,$$

and the left-hand side of the above inequality is a concave function in $\bar{\boldsymbol{\lambda}}$ (if we fix the function $\bar{f}$). Thus, if we let $\mathcal{S}$ denote the line segment joining $\boldsymbol{\lambda}^1$ and $\boldsymbol{\lambda}^2$ (which, of course, contains $\bar{\boldsymbol{\lambda}}$), we see that the problem

$$\underset{\boldsymbol{\lambda} \in \mathcal{S}}{\text{minimize}} \iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \lambda_i\} \, dA$$

must realize its minimizer on the boundary of $\mathcal{S}$ (since we are minimizing a *concave* function), i.e. the point $\boldsymbol{\lambda}^1$ or $\boldsymbol{\lambda}^2$. Therefore, it must be the case that $\iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \lambda_i^1\} \, dA \leq t$ or $\iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \lambda_i^2\} \, dA \leq t$, which completes the proof. □

The following theorem describes a cutting plane oracle for the outer problem (13):

Theorem 12. *Let $\bar{\boldsymbol{\lambda}}$ satisfy $\mathbf{e}^T \bar{\boldsymbol{\lambda}} = 0$ and let $\bar{f}$ be the solution to the inner problem of (13) (i.e. $\bar{f}$ maximizes $\iint_{\mathcal{R}} \sqrt{f(x)} \, dA$ over all $f \in \Omega(\bar{\boldsymbol{\lambda}})$). Then the vector $\bar{\mathbf{g}} \in \mathbb{R}^n$ defined by setting*

$$\bar{g}_i = -\iint_{\bar{R}_i} \bar{f}(x) \, dA$$

for all i, where $\bar{R}_i$ is defined as

$$\bar{R}_i = \left\{ x \in \mathcal{R} : \|x - x_i\| - \bar{\lambda}_i \leq \|x - x_j\| - \bar{\lambda}_j \quad \forall j \neq i \right\} ,$$

defines a valid cutting plane for problem (13); that is, if $\bar{\mathbf{g}}^T(\boldsymbol{\lambda}' - \bar{\boldsymbol{\lambda}}) \leq 0$ for some $\boldsymbol{\lambda}'$ satisfying $\mathbf{e}^T \boldsymbol{\lambda}' = 0$, and f' is the solution to the inner problem of (13) associated with $\boldsymbol{\lambda}'$, then $\max_{f \in \Omega(\boldsymbol{\lambda}')} \iint_{\mathcal{R}} \sqrt{f(x)} \, dA \geq \max_{f \in \Omega(\bar{\boldsymbol{\lambda}})} \iint_{\mathcal{R}} \sqrt{f(x)} \, dA$.

Proof. Statement 2 of Lemma 4 says that, for any other $\boldsymbol{\lambda}'$, the assumption that $\bar{\mathbf{g}}^T(\boldsymbol{\lambda}' - \bar{\boldsymbol{\lambda}}) \leq 0$ yields

$$\iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \lambda_i'\} \, dA \leq \iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \bar{\lambda}_i\} \, dA + \bar{\mathbf{g}}^T(\boldsymbol{\lambda}' - \bar{\boldsymbol{\lambda}}) \leq \iint_{\mathcal{R}} \bar{f}(x) \min_i \{\|x - x_i\| - \bar{\lambda}_i\} \, dA \leq t$$

which implies that $\bar{f} \in \Omega(\boldsymbol{\lambda}')$ and therefore that $\max_{f \in \Omega(\boldsymbol{\lambda}')} \iint_{\mathcal{R}} \sqrt{f(x)} \, dA \geq \max_{f \in \Omega(\bar{\boldsymbol{\lambda}})} \iint_{\mathcal{R}} \sqrt{f(x)} \, dA$ as desired. □

We now have a fast method for generating cutting planes associated with problem (13) and thereby recovering

the distribution f^* that solves problem (4); see Algorithm 1 for a formal description.

Input: A compact, planar region $\mathcal{R}$ containing a set of distinct points $x_1, \ldots, x_n$ which are interpreted as an empirical distribution $\hat{f}$, a distance parameter t, and a tolerance ϵ.
Output: An ϵ-approximation of the distribution f^* that maximizes $\iint_{\mathcal{R}} \sqrt{f(x)}\, dA$ subject to the constraint that $\mathcal{D}(f, \hat{f}) \leq t$.
`/* This is a standard analytic center cutting plane method applied to problem (13), which has an` n`-dimensional variable space. */`
Set $\texttt{UB} = \infty$ and $\texttt{LB} = -\infty$;
Set $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^n : \mathbf{e}^T\boldsymbol{\lambda} = 0,\ \lambda_i \leq \operatorname{diam}(\mathcal{R})\ \forall i\}$;
while $\texttt{UB} - \texttt{LB} > \epsilon$ **do**
  Let $\bar{\boldsymbol{\lambda}}$ be the analytic center of Λ;
  `/* Build an upper bounding` $\bar{f}$ `for the original problem (4). */`
  Let $\bar{\nu}_0, \bar{\nu}_1$ be the solution to problem (14) with $\bar{\boldsymbol{\lambda}}$ as an input;
  Let $\bar{f}(x) = \frac{1}{4}(\bar{\nu}_0 \min_i\{\|x - x_i\| - \bar{\lambda}_i\} + \bar{\nu}_1)^{-2}$;
  Let $\texttt{UB} = \iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$;
  `/* Build a lower bounding` $\tilde{f}$ `that is feasible for (4) by construction. */`
  Let $\bar{R}_i = \{x \in \mathcal{R} : \|x - x_i\| - \bar{\lambda}_i \leq \|x - x_j\| - \bar{\lambda}_j \quad \forall j \neq i\}$ for $i = \{1, \ldots, n\}$;
  Let $\tilde{\boldsymbol{\nu}} \in \mathbb{R}^{n+1}$ be the solution to problem (7) with inputs $\bar{R}_1, \ldots, \bar{R}_n$;
  Let $\tilde{f}$ be defined by setting $\tilde{f}(x) = \frac{1}{4}(\tilde{\nu}_0\|x - x_i\| + \tilde{\nu}_i)^{-2}$ on each $\bar{R}_i$;
  Let $\texttt{LB} = \iint_{\mathcal{R}} \sqrt{\tilde{f}(x)}\, dA$;
  Let $g_i = -\iint_{\bar{R}_i} \bar{f}(x)\, dA$ for $i = \{1, \ldots, n\}$;
  Let $\mathcal{H} = \{\boldsymbol{\lambda} \in \mathbb{R}^n : \mathbf{g}^T\boldsymbol{\lambda} \geq \mathbf{g}^T\bar{\boldsymbol{\lambda}}\}$ and set $\Lambda = \Lambda \cap \mathcal{H}$;
end
return $\tilde{f}$;

Algorithm 1: Algorithm WorstTSPDensity takes as input a compact planar region containing a set of n distinct points, a distance threshold t, and a tolerance ϵ.

3 Practical considerations

In solving a distributionally robust districting problem, there are a number of practical considerations that arise; we describe some of them here.

3.1 Selecting the distance parameter t

From the preceding discussion, it is clear that the parameter t in the Wasserstein distance constraint $\mathcal{D}(f, \hat{f}) \leq t$ from our original problem (4) has a significant impact on the problem solution. Of course, in practice, we do not have any way of *a priori* calculating an exact value of t. However, in order to estimate t in a data-driven fashion, the following result is helpful:

Theorem 13. *Let $\hat{f}_1$ and $\hat{f}_2$ denote empirical distributions associated with two sets of independent samples of n points from a distribution f. Then*

$$\frac{1}{2}\mathbf{E}\mathcal{D}(\hat{f}_1, \hat{f}_2) \leq \mathbf{E}\mathcal{D}(f, \hat{f}_1) \leq \mathbf{E}\mathcal{D}(\hat{f}_1, \hat{f}_2)\,.$$

Proof. This is due to [14], and follows from Jensen's inequality and the triangle inequality. □

The above result is useful because the distance between the two empirical distributions $\mathcal{D}(\hat{f}_1, \hat{f}_2)$ is simply the cost of a minimum-weight bipartite matching between the elements of $\hat{f}_1$ and $\hat{f}_2$, multiplied by a factor of $1/n$. Thus, one simple, "back-of-the-envelope" procedure to select the distance parameter t would be to sample two sets of n points each, let c be equal to the cost of the minimum-weight bipartite matching between them, and set $t = \alpha c$ with $\alpha \in [1/2, 1]$.

If we desire rigorous probabilistic bounds on t, more sophisticated machinery is required. Although there exist a number of results of this kind (e.g. Theorem 2 of [23]), nearly all of them depend on additional distributional information such as exponential tail bounds; this was previously noted (in the context of robust optimization) in Section 7.2.B of [22]. One positive example is Theorem 6.15 of [45], which gives a useful bound on the Wasserstein distance between two probability density functions f_1 and f_2 by

$$\mathcal{D}(f_1, f_2) \leq \iint\limits_{\mathcal{R}} \|x_0 - x\| \cdot |f_1(x) - f_2(x)|\, dA$$

for any $x_0 \in \mathcal{R}$. Theorem 1(i) of [10] relates the right-hand side of the above to the *relative entropy* $H(f_1|f_2)$ between f_1 and f_2 by the expression

$$\iint\limits_{\mathcal{R}} \|x_0 - x\| \cdot |f_1(x) - f_2(x)|\, dA \leq \left(\frac{3}{2} + \log \iint\limits_{\mathcal{R}} e^{2\|x - x_0\|} f_2(x)\, dA\right)\left(\sqrt{H(f_1|f_2)} + \frac{1}{2}H(f_1|f_2)\right),$$

where we define

$$H(f_1|f_2) = \iint\limits_{\mathcal{R}} f_1(x) \log \frac{f_1(x)}{f_2(x)}\, dA\,.$$

Let $r = \min_{x_0 \in \mathcal{R}} \max_{x \in \mathcal{R}} \|x - x_0\|$ denote the "radius" of $\mathcal{R}$, whence $\log \iint_{\mathcal{R}} e^{2\|x - x_0\|} f_2(x)\, dA \leq \log e^{2r} = 2r$. Thus,

if $\mathcal{D}(f_1, f_2) \geq t$, we have

$$\begin{aligned} t \leq \mathcal{D}(f_1, f_2) &\leq \iint_{\mathcal{R}} \|x_0 - x\| \cdot |f_1(x) - f_2(x)| \, dA \\ &\leq \left(\frac{3}{2} + 2r\right)\left(\sqrt{H(f_1|f_2)} + \frac{1}{2}H(f_1|f_2)\right) \\ \implies H(f_1|f_2) &\geq \frac{8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6}{3 + 4r}. \end{aligned} \tag{15}$$

Next, the paper [9] shows that, for any distribution f with empirical distribution $\hat{f}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \Pr(\mathcal{D}(f, \hat{f}) \geq t) \leq -\alpha(t),$$

where the function $\alpha(t)$ is defined as

$$\alpha(t) = \inf_{g:\mathcal{D}(f,g)\geq t} H(f|g).$$

The result (15) establishes that $\alpha(t) \geq (8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6)/(3 + 4r)$, and therefore we find that

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log \Pr(\mathcal{D}(f, \hat{f}) \geq t) &\leq -\frac{8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6}{3 + 4r} \\ \implies \Pr(\mathcal{D}(f, \hat{f}) \geq t) &\lesssim \exp\left(-n\frac{8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6}{3 + 4r}\right), \end{aligned} \tag{16}$$

where the notation "$\lesssim$" reflects the approximate inequality that results from dropping the "lim sup" term. Thus, given a desired significance level $1 - \theta$, we can construct a threshold distance t by equating the right-hand side of (16) to $1 - \theta$ and solving for t.

3.2 Variations and extensions

The analysis of Section 2.3 motivates three possible extensions of the existing framework:

Uneven data weights: By definition, the empirical distribution $\hat{f}$ of the points $x_1, \ldots, x_n$ consists of a collection of n atomic masses at each point, each having mass of $1/n$. It is easy to envision scenarios in which one desires uneven weights: for example, one might use an exponential weighting scheme to emphasize more recent measurements, or one might use different weights to distinguish between activities on weekends versus weekdays (or other seasonal effects). If we require point x_i to have a mass q_i associated with it, then we can find the worst-case distribution f^* by solving problem (5), with the one change that we replace the restriction

that $\mathbf{e}^T\boldsymbol{\lambda} = 0$ with a restriction that $\mathbf{q}^T\boldsymbol{\lambda} = 0$ instead; the form of f^* is otherwise unchanged.

Capacitated vehicles: Our approach can also be adapted to solve problems when vehicles have capacities and originate from a central depot located at the origin. To do so, suppose that each vehicle can visit c destinations before returning to the depot. The following theorem from [27] provides useful upper and lower bounds for the cost of a capacitated vehicle routing tour:

Theorem 14. *Let $X = \{x_1, \ldots, x_n\}$ be a set of demand points in the plane serviced by a fleet of vehicles with capacity κ that originate from a single depot located at the origin. The length of the optimal set of capacitated VRP tours of X, written $\mathsf{VRP}(X)$, satisfies*

$$\max\left\{\frac{2}{\kappa}\sum_{i=1}^{n}\|x_i\|, \mathsf{TSP}(X)\right\} \leq \mathsf{VRP}(X) \leq 2\left\lceil\frac{|X|}{\kappa}\right\rceil\cdot\frac{\sum_{i=1}^{n}\|x_i\|}{|X|} + (1-1/\kappa)\,\mathsf{TSP}(X)\,. \tag{17}$$

The probabilistic version of this, as derived in Section C of the online supplement, uses the BHH Theorem (Theorem 2 of this paper) to characterize the length of the TSP term:

$$\sqrt{n}\cdot\max\left\{\frac{2}{s}\iint_{\mathcal{R}}\|x\|f(x)\,dA,\ \beta\iint_{\mathcal{R}}\sqrt{f_c(x)}\,dA\right\} \lesssim \mathsf{VRP}(X) \lesssim \sqrt{n}\cdot\left(\frac{2}{s}\iint_{\mathcal{R}}\|x\|f(x)\,dA + \beta\iint_{\mathcal{R}}\sqrt{f_c(x)}\,dA\right), \tag{18}$$

where we set $s = \kappa/\sqrt{n}$ and we have adopted the notation "$\lesssim$" to denote an "approximate" inequality, both of which are also explained in Section C. It is immediately obvious that the upper and lower bounds are within a factor of 2 of one another. Applying the same analysis as in Section 2.2, the worst-case distribution that maximizes the right-hand side of (18) subject to a Wasserstein distance constraint takes the form

$$f^*(x) = \frac{1}{4(\nu_0^*\min_i\{\|x-x_i\|-\lambda_i^*\}+\nu_1^*-\frac{2}{s}\|x\|)^2}\,;$$

its level sets, i.e. those curves for which $\nu_0^*(\|x-x_i\|-\lambda_i^*)+\nu_1^*-\frac{2}{s}\|x\|$ is constant, consist of piecewise components of so-called *Cartesian ovals* [34].

Higher dimensions: The BHH Theorem (Theorem 2) is also applicable in higher dimensions; the general form says that, when the service region $\mathcal{R}$ belongs to $\mathbb{R}^d$, we have

$$\lim_{N\to\infty}\frac{\mathsf{TSP}(X)}{N^{(d-1)/d}} = \beta_d\iiint_{\mathcal{R}} f_c(x)^{(d-1)/d}\,dV\,,$$

for dimension-dependent constants β_d. Applying the same analysis as in Section 2.2, the worst-case distribu-

tion that maximizes the right-hand side of the above, subject to a Wasserstein distance constraint, takes the form

$$f^*(x) = \frac{(d-1)^{d-1}}{d^d} \cdot \frac{1}{(\nu_0^* \min_i\{\|x - x_i\| - \lambda_i^*\} + \nu_1^*)^d} \,.$$

4 Computational experiments

In this section, we apply our theoretical results to two computational experiments: the first experiment shows the impact of increasing the number of samples n, and the second is a districting strategy in which we divide a map into pieces so as to to minimize the worst-case workload of any vehicle.

4.1 Varying values of n

In our first experiment, we let $\mathcal{R}$ be the unit square, and as a ground truth distribution $\bar{f}$ we use an even mixture of two truncated Gaussian distributions with means $\mu_1, \mu_2 = (0.400, 0.187), (0.795, 0.490)$ and covariance matrices $\Sigma_1 = \Sigma_2 = \left(\begin{smallmatrix} 0.070 & 0 \\ 0 & 0.070 \end{smallmatrix}\right)$. This mixture was chosen because it satisfies $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA = 0.55$ and therefore represents a compromise between extreme clustering (which would have $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$ close to zero) and a perfect uniform distribution (which would have $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$ equal to one). For $n \in \{2, \ldots, 100\}$, we performed 10 independent experiments where we drew n samples from $\bar{f}$ and then obtained the worst-case TSP distribution f^* by solving problem (4) via Algorithm 1 (hence, 99×10 experiments in total). For each experiment, we defined our distance constraint using Theorem 13 by setting t to be the cost of a minimum-weight bipartite matching between two independent collections of samples of size n from $\bar{f}$ (multiplied by a factor of $1/n$); this corresponds to setting $\alpha = 1$. Figure 4a shows a plot of the worst-case TSP costs $\iint_{\mathcal{R}} \sqrt{f^*(x)}\, dA$ as n varies, and Figure 4b shows the same data, only using the Wasserstein distance threshold t as the independent variable. Not surprisingly, it is clear that the worst-case cost decreases as n increases and as t decreases. Figure 4b suggests that the worst-case cost, measured as a function of t, decreases in a *concave* fashion as $t \to 0$. For purposes of comparison, Figure 4d shows the estimates of $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$ obtained when one uses a uniform kernel density estimator; that is, if we draw n samples $x_1, \ldots, x_n$ from $\bar{f}$, then we define an estimator $\tilde{f}$ by setting $\tilde{f}(x) = \frac{1}{C} \sum_i \mathbb{1}(\|x - x_i\| \leq r)$, where r is a "bandwidth" parameter and C is a normalization constant. We used 5 different values of r between 0.03 and 0.3; note that the estimate $\iint_{\mathcal{R}} \sqrt{\tilde{f}(x)}\, dA$ is highly sensitive to the choice of the bandwidth parameter r.

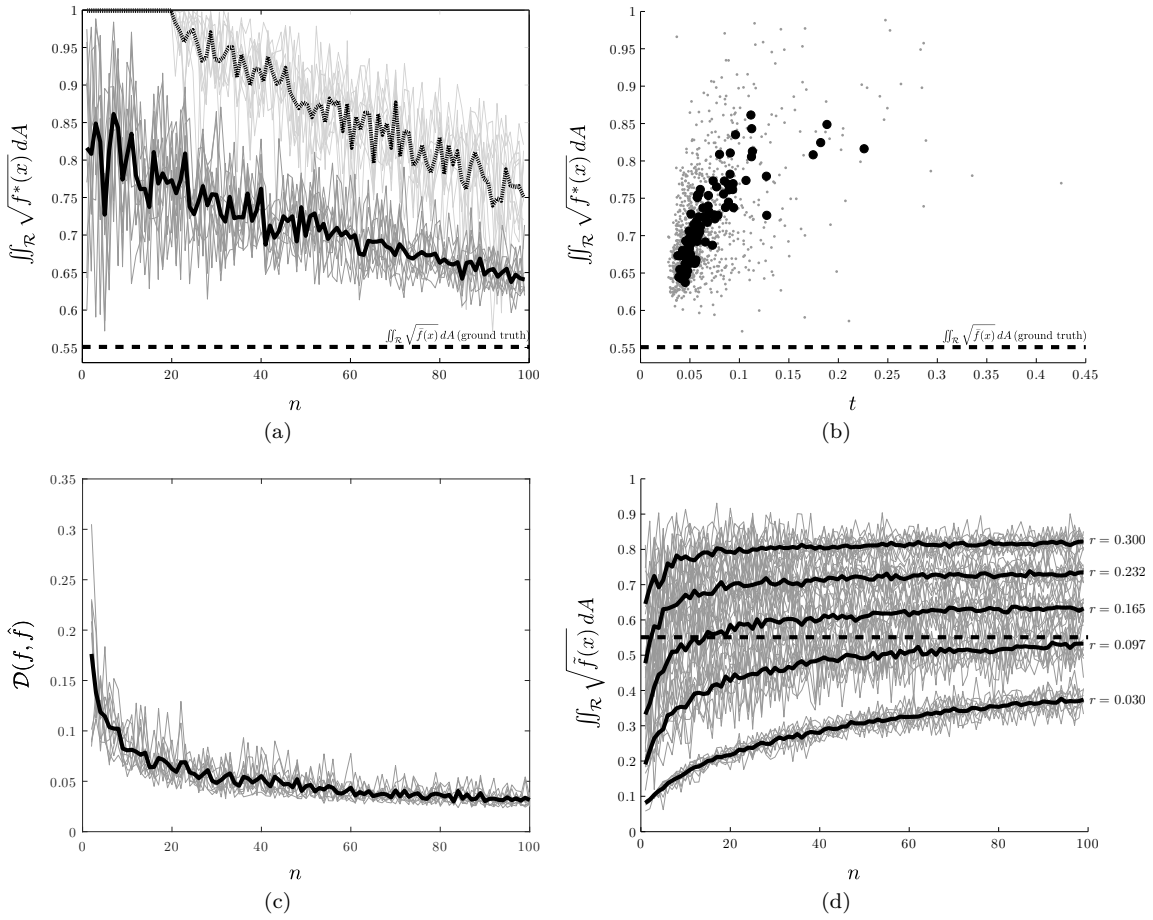
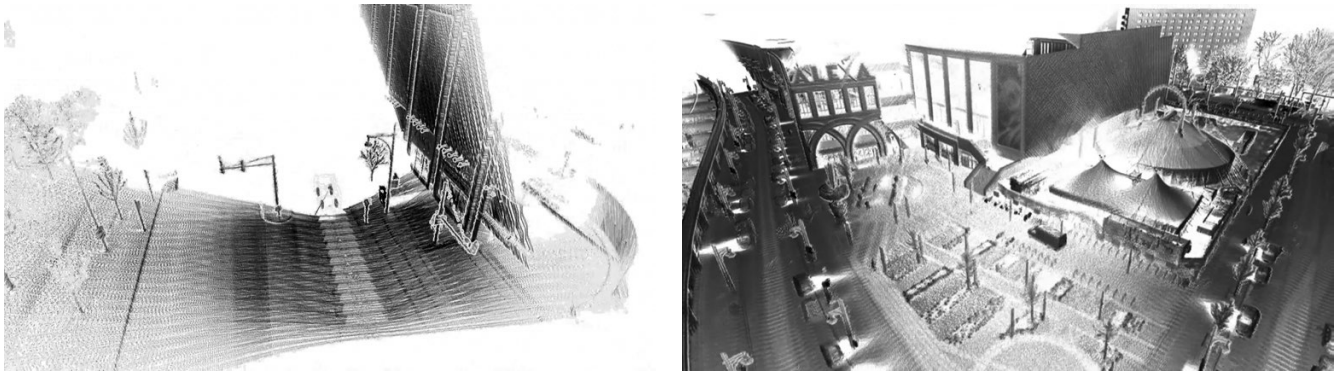


Figure 4: Figure 4a shows the worst-case costs that are computed during the 99×10 executions of our algorithm; the gray plots indicate the results obtained from individual samples and the thick line indicates the sample averages of the 10 trials for fixed n. There are two sets of experiments being represented; the upper plot (indicated with the dotted black line and the light gray lines) corresponds to the worst-case costs when using the upper bound of t from (16), and the lower plot (indicated with the solid black line and the dark gray lines) corresponds to the worst-case costs when using the upper bound of t from Theorem 13. Figure 4b shows the same data set (using only the upper bound of t from Theorem 13), only we plot the worst-case costs as a function of the Wasserstein distance threshold t rather than a function of n; the gray points indicate individual experiments and the dark points again indicate the sample averages of the 10 trials for fixed n. For reference, the actual Wasserstein distance between the samples and the distribution, $\mathcal{D}(f, \hat{f})$, is shown in Figure 4c; we computed this using the algorithm from [17]. Finally, Figure 4d shows the estimates of $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\,dA$ obtained when one uses a uniform kernel density estimator.



(a) (b)

Figure 5: A 3-dimensional LIDAR scan taken from a vehicle travelling along a city street (5a); these measurements are aggregated together to form point cloud representations of entire cities, as shown in 5b.

4.2 A districting experiment with road network data

In this section, we describe an experiment in which we divide a service region into districts so as to allocate the workloads of a fleet of vehicles. This experiment is much more elaborate than that of the preceding section because we compute our TSP tours using data from an actual road network, rather than simply make an assumption that distances are Euclidean. Our dataset comes from an industrial affiliate, a major North American mapping company, that is currently using the algorithm described in this section as part of a routing suite that manages a fleet of 100 vehicles on 3 continents. In a nutshell, the problem faced by the affiliate is to collect map data using vehicles that are equipped with rotating LIDAR, positioning sensors, and high-resolution panoramic cameras. Scans of this kind usually consist of two stages: the first stage, called a *rough scan*, is essentially a large-scale, multi-vehicle *Chinese Postman Problem* in which the goal is to traverse every street (or a large collection of streets) in a region. The second stage, which turns out to be equally time-consuming and is the subject of this experiment, is called a *targeted scan*, and consists of visiting those points in the region for which the measurements in the first stage were not adequate; this might be because the driver missed a street, the sensor's view was obstructed, or because an alternate vantage point of a particular location is needed. It is often the case that, when performing a 3-dimensional scan of a structure (see Figure 5 for an example), multiple trips to that structure are needed to accurately recover the original shape from the measurements.

We describe here an algorithm for designing vehicle districts to minimize the time to completion (i.e. the makespan) of the targeted scan, or equivalently, the driving duration of the longest route. This problem satisfies the four attributes identified in the introduction for the following reasons:

1. **"The service region is a compact Euclidean domain $\mathcal{R}$, and distances between points are Eu-**

clidean, or close to some "natural" metric such as the ℓ^1 or ℓ^∞ norm." Each service region is the metropolitan region surrounding a major city; although distances are measured with respect to a road network and are thus not Euclidean, we will demonstrate that this discrepancy can easily be taken into account.

2. **"During each service period, a fleet of vehicles must use travelling salesman (TSP) tours to visit a collection of destination points in $\mathcal{R}$ that are sampled from a probability distribution f."** Here each "service period" consists of a day, during which time the vehicles drive for 6-8 hours. The destination points are determined as follows: once the rough scan (the first stage) has been completed, the raw data from all vehicles are dumped to an in-house cloud computing service, where they are "scrubbed" to identify those locations that must be re-scanned. Since this scrubbing process is performed in parallel across many CPUs, and because certain errors take longer to find than others (in fact, many errors are explicitly found only by human analysts), these locations can essentially be thought of as occurring randomly; the total time horizon for scrubbing is variable but typically ranges between 2 and 7 days.

3. **"The goal is to partition $\mathcal{R}$ into districts, with one district per vehicle, in a way that is "optimal" as the number of destination points becomes large."** We seek a districting strategy because it is much simpler to implement than explicitly computing vehicle tours, and also because additional destination points may arrive in an online fashion as they are identified by the cloud computing service. Districting strategies (as opposed to explicit routing strategies) are very common in the mapping industry because of the ease of operation that is afforded. In addition, since the number of vehicles is fixed, and driver wages are relatively cheap, the primary goal is to minimize the amount of time until the scan is completed. Since the actual vehicle tours are traversed over a period of several days (in practice, usually between 5 and 10 days), we have the opportunity to re-optimize the districts as additional data is collected. For simplicity's sake, we do not use any re-optimization because (i) we found that it was only helpful in practice for the most complex cities such as Istanbul or Paris, and (ii) we do not have any information about when each data point was detected in the scrubbing process.

4. **"The distribution f is unknown at the time that the districting decision is to be made, and there is only a set of independent samples from f available to make the districting decision."** The districting decision must be made soon after the rough scan is completed, at which point there are a limited number of sample points available that are obtained from scrubbing the initial dataset. The sample points are generated overnight, after the rough scans are completed. Their distribution is non-uniform throughout the region because of spatial clustering: for example, erroneous data points often arise due to a complex

3-dimensional object (e.g. a sculpture or some elaborate architecture) or if the sun is shining at an unusual angle, so that many of the measurements in a localized area are affected.

The assumption of *independence* is a strong one and requires some care. Simply put, it arises due to the fact that scrubbing takes place across many CPUs simultaneously. As an extreme example, if one were to use a single CPU to process the dataset, it would scan through the measurements and detect errors in the same order that the vehicle visited them. Thus, the samples would appear in an extremely *dependent* way. Conversely, if we distribute the scanning process across many CPUs, then each CPU is responsible for only a small leg of the vehicle's tour, and two CPUs might detect errors in two entirely separate locations at the same time. As more CPUs are used, we see less dependence between (temporally) consecutive samples.

The purpose of this experiment is two-fold: first, we aim to demonstrate that the proposed continuous approximation techniques are actually useful for solving practical problems, and second, we then seek to show that our proposed methodology is superior to that of existing approaches. There are 13 different service regions $\mathcal{R}$ that will be studied, all of which are metropolitan regions in Western Europe.

4.2.1 Validation of continuous approximation methods

In order to apply our results to solve a practical problem, it is necessary to first confirm that the continuous approximation method remains valid and useful even when point-to-point distances are not Euclidean and are instead measured according to a heterogeneous road network. In order to take this factor into account, we should first note that Theorem 2 actually holds under much more general conditions than the Euclidean TSP, and remains valid when one considers the TSP under many "natural" norms or even other combinatorial structures such as the minimum spanning tree or Steiner tree (more precisely, Theorem 2 holds whenever the underlying structure is a *subadditive Euclidean functional*; see [42] for an extensive study thereof). Obviously, the coefficient β depends on the choice of structure. The following example is useful in designing an appropriate framework to handle this disparateness:

Example 15 (Varying metrics in a region)**.** Consider a set of n points sampled according to a distribution f in the unit square, with distances $d(x_1, x_2)$ between pairs of points $x_1 = (x_1^1, x_1^2)$ and $x_2 = (x_2^1, x_2^2)$ defined as follows:

- If x_1 and x_2 are in the lower left quadrant, then $d(x_1, x_2)$ is the Euclidean distance between x_1 and x_2.
- If x_1 and x_2 are in the lower right quadrant, then $d(x_1, x_2)$ is the ℓ^1 distance between x_1 and x_2.
- If x_1 and x_2 are in the upper left quadrant, then $d(x_1, x_2)$ is the ℓ^∞ distance between x_1 and x_2.

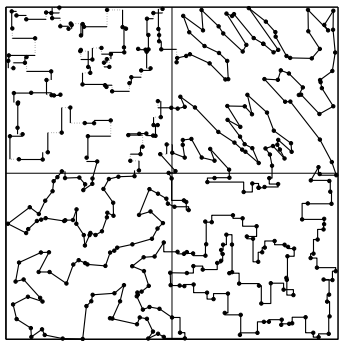


Figure 6: The TSP tour of a collection of points uniformly sampled in the unit square, with varying metrics depending on quadrants. The dashed lines in the paths in the upper left quadrant correspond to the smaller of the two directions (horizontal or vertical) between points, which is relevant because the ℓ^∞ distance is used.

- If x_1 and x_2 are in the upper right quadrant, then $d(x_1, x_2) = \sqrt{(x_1 - x_2)^T A (x_1 - x_2)}$, where A is a symmetric positive definite matrix.
- If x_1 and x_2 are in different quadrants, then $d(x_1, x_2)$ is determined by a tie-breaking rule of some sort (the details of which are not relevant).

The TSP tour of a set of points under these assumptions is shown in Figure 6. If we let $Q_1, \ldots, Q_4$ denote the four quadrants of the square, then it is routine to verify that we in fact have

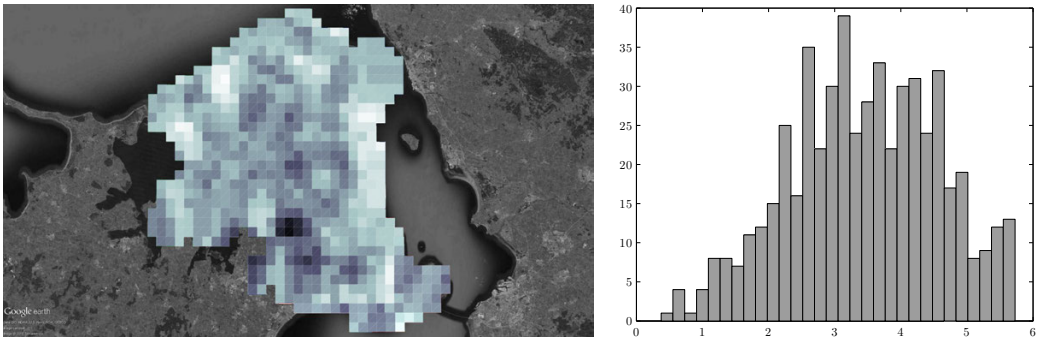
$$\lim_{N \to \infty} \frac{\mathsf{TSP}(X_1, \ldots, X_N)}{\sqrt{N}} = \sum_{i=1}^{4} \beta_i \iint_{Q_i} \sqrt{f_c(x)}\, dA$$

where each β_i is associated with the metric on quadrant Q_i (e.g. β_1 is the Euclidean TSP coefficient); one can verify this by proceeding through the proof of the BHH theorem in (for example) Chapter 2.4 of [42].

Example 15 suggests a general approach that is useful when the service region $\mathcal{R}$ has a heterogeneous road network: if we decompose $\mathcal{R}$ into a collection of "patches" $Q_1, \ldots, Q_K$, then we adopt the approximation

$$\mathsf{TSP}(X_1, \ldots, X_n) \approx \sqrt{n} \cdot \sum_{i=1}^{K} \beta_i \iint_{Q_i} \sqrt{f_c(x)}\, dA\,.$$

One can estimate the values β_i as follows: if we sample a set of k points *uniformly* in Q_i and compute the duration of their TSP tour ℓ_i (using the road network), we would expect to see that $\ell_i \approx \beta_i \sqrt{\mathrm{Area}(Q_i) \cdot k}$; this is simply the uniform case of the BHH theorem applied to points constrained to lie in Q_i. Thus, a sensible estimate of β_i is given by setting $\beta_i = \ell_i / \sqrt{\mathrm{Area}(Q_i) \cdot k}$. In this experiment, we discretize the region $\mathcal{R}$ into a collection of patches Q_i of size $2\,\mathrm{km} \times 2\,\mathrm{km}$, and estimate each coefficient β_i using $k = 10$ samples (a larger number would be preferable,



(a) (b)

Figure 7: The shading in Figure 7a, which shows the Copenhagen metropolitan region, indicates the values of β_i associated with each of the square patches (darker patches correspond to higher values of β_i). Figure 7b is a histogram of these same values. Those values of β_i that are extremely small correspond to regions in which multiple distinct locations are geocoded to the same point (e.g. a large shopping mall, or a square that just barely overlaps with the coastline of the region).

but the mapping API that we use [21] imposes a limit of at most $100,000$ queries per day); Figure 7 shows the resulting values when $\mathcal{R}$ is the Copenhagen metropolitan region. We emphasize that these values β_i are computed with respect to the driving *duration* (not the length) of the TSP tour in each path Q_i, measured in minutes; for example, if we sample $k = 10$ points in a cell Q_i that has $\beta_i = 3$, then we estimate that the duration of the TSP tour of these 10 points will be $\beta_i\sqrt{\text{Area}(Q_i) \cdot k} = 3\sqrt{4 \cdot 10} \approx 19$ minutes.

4.2.2 Worst-case distributions for districts

When dividing $\mathcal{R}$ into districts $D_1, \ldots, D_m$ that minimize the worst-case vehicle workloads, the problem of interest becomes

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint\limits_{D_j} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint\limits_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA \quad &\leq \quad t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0 \\ \iint\limits_{R} f(x)\, dA \quad &= \quad 1 \\ f(x) \quad &\geq \quad 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{19}$$

for each sub-region D_j (this is identical to (5) except for the domain of integration in the objective). The worst-case distribution associated with each district D_j is characterized as follows:

Theorem 16. *The worst-case distribution that solves problem (19) takes the form*

$$f^*(x) = \left[\frac{1}{4(\nu_0^* \min_i\{\|x - x_i\| - \lambda_i^*\} + \nu_1^*)^2}\right] \mathbb{1}(x \in D_j) + \sum_{i=1}^{n} p_i^* \delta(x - x_i)$$

with $\nu_0^*, \nu_1^* \geq 0$, $\mathbf{e}^T \boldsymbol{\lambda}^* = 0$, *and* $0 \leq p_i^* \leq 1$. *Moreover, we have* $p_i^* = 0$ *whenever* $x_i \in D_j$.

Proof. This is almost identical to the proof of Theorem 6 and we omit it here for brevity. The intuition behind the Dirac delta components is not difficult: if any mass of f^* is located outside district D_j, then it does not contribute to the objective and therefore should contribute as little as possible towards the Wasserstein distance constraint. □

4.2.3 Districting criteria

In order to divide each service region $\mathcal{R}$ into districts, we use a computational geometric structure called a *power diagram* [5], which has frequently been applied to districting problems in existing literature on vehicle routing [31, 37]. Given a set of "depot points" $p_1, \ldots, p_m$ in $\mathcal{R}$ and any vector $\mathbf{w} \in \mathbb{R}^m$, the Euclidean power diagram of $\mathcal{R}$ with respect to $p_1, \ldots, p_m$ and $\mathbf{w}$ is a partition of $\mathcal{R}$ into districts $D_1, \ldots, D_m$ defined by

$$D_i = \left\{x \in \mathcal{R} : \|x - p_i\|^2 - w_i \leq \|x - p_j\|^2 - w_j \ \forall j\right\}. \tag{20}$$

In our experiments, the depot points are given to us by the drivers (typically their hotel rooms or apartments), and we substitute road network distances in place of the Euclidean norm. Since the depot points are fixed and given to us, we control the shapes and sizes of the districts D_i by varying the weight vector $\mathbf{w}$; note that if we increase w_i while fixing the other terms w_j, the district D_i expands. In our experiment, we compare four different partitioning criteria:

Wasserstein robust partitioning Problem (19) in Section 4.2.2 of this paper describes the structure of the worst-case distribution that maximizes the asymptotic workload in a particular district, subject to a Wasserstein distance constraint. Thus, it is sensible to seek a weight vector $\mathbf{w}^*$ that results in districts $D_1, \ldots, D_m$ such that the solution to (19) is equal for each district (in other words, the worst-case workloads are the same for all districts). Section 5 of our previous paper [16] describes a branch-and-bound scheme for finding $\mathbf{w}^*$ that is based on a simple set of monotonicity properties that can be directly applied to this problem.

Kernel-based partitioning: Given a set of samples $x_1, \ldots, x_n \in \mathcal{R}$, another natural approach is to build an artificial distribution $\bar{f}$ with a *kernel density estimator*, as we have seen in Section 4.1. For example, if we use

a uniform kernel function with bandwidth parameter r, then we have

$$\bar{f}(x) = \frac{1}{\pi r^2 n} \sum_{i=1}^{n} \mathbb{1}(\|x - x_i\| \leq r)\,.$$

It would then follow that the TSP tour of the additional points should be related to $\sqrt{\bar{f}(x)}$, by the BHH theorem. In order to take the heterogeneous road network into account, the β_i terms are needed. Committing a slight abuse of notation, let $\beta(x)$ be a step function representing the β_i's and their associated Q_i's, that is, $\beta(x) := \sum_{i=1}^{K} \beta_i \mathbb{1}(x \in Q_i)$. The TSP workload in a district D_i would then be estimated as $\iint_{D_i} \beta(x)\sqrt{\bar{f}(x)}\,dA$, and a natural districting criterion would be to construct districts such that this value is equal for all districts. It turns out that this is very simple because the optimal weight vector $\mathbf{w}^*$ is the Lagrange multiplier vector of a semi-infinite assignment problem which is well studied; see for example [17].

Mean-covariance robust partitioning: Section 5 of our earlier paper [16] describes a branch-and-bound method for partitioning a region $\mathcal{R}$ into a power diagram partition in which the worst-case workloads for all districts D_i are equal. Here, the "worst-case workloads" are defined via robust optimization, specifically the solution to (11). In other words, we construct a power diagram partition against all distributions whose mean and covariance matrix are equal to the fixed values obtained from the sampled points $x_1, \ldots, x_n$.

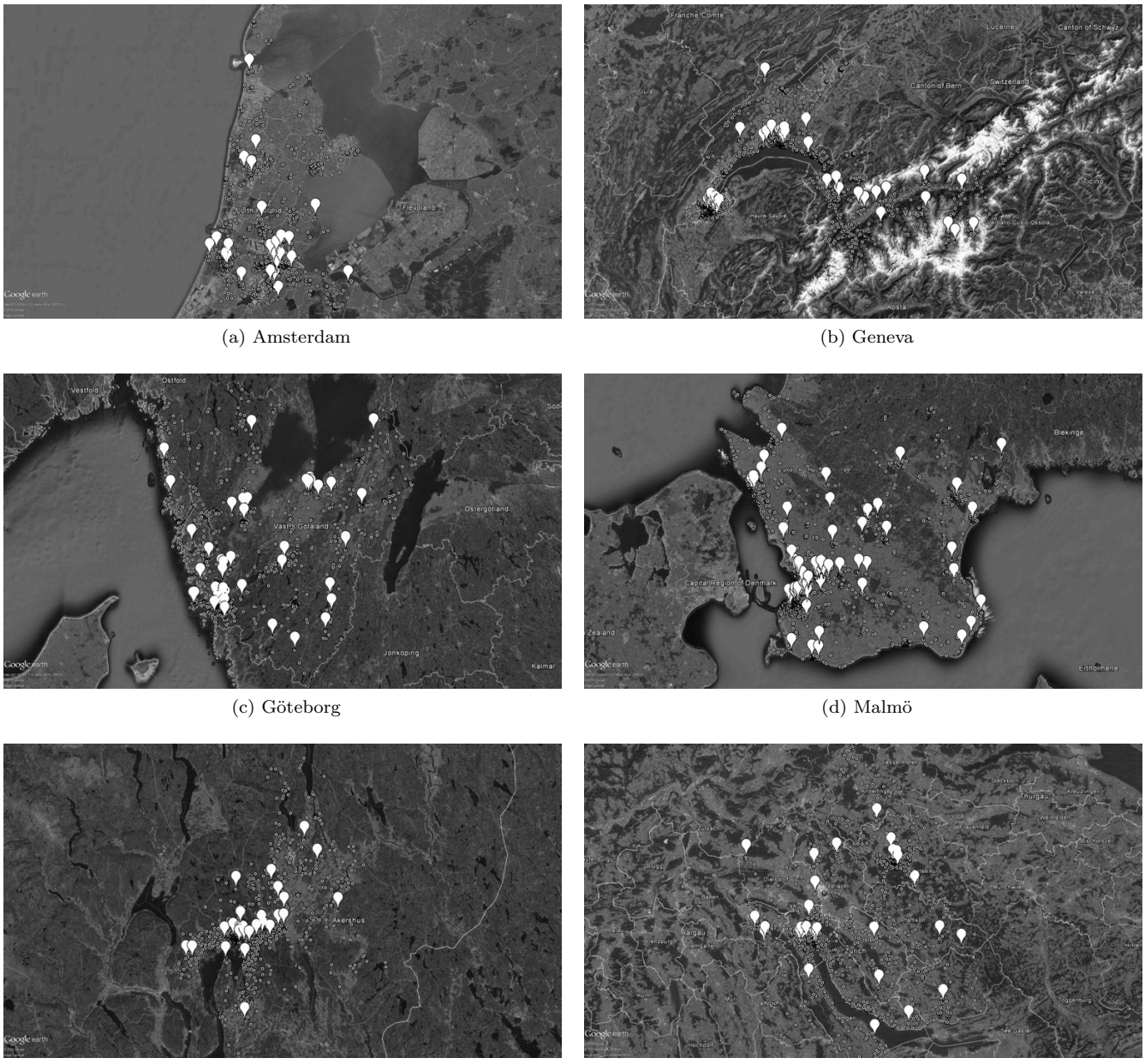
TSP-based partitioning: A very natural, common-sense strategy is to simply construct districts based on the TSP tours of the sample points. In other words, we select weights so as to minimize the maximum TSP tour of each of the districts, that is, to minimize $\max_i \mathsf{TSP}(\{x_1, \ldots, x_n\} \cap D_i)$. We can do this using the same monotonicity property as in our algorithm from [16].

4.2.4 Results

We applied the four aforementioned districting criteria to problem instances defined on 13 different metropolitan regions using data provided by our industrial affiliate, where the goal is to partition each region into $m = 3$ districts. When using the Wasserstein partitioning criterion, we set the distance threshold t using the result from Theorem 13 with $\alpha = 1$. Specifically, we split the sample points into two equally-sized (random) sets $\hat{f}_1$ and $\hat{f}_2$, and computed a minimum-weight bipartite matching between the two; this process was repeated over 100 independent trials. We found this bound preferable to the rigorous result from (16) because, even for very low significance levels $1 - \theta$, the distance threshold becomes unrealistically high (more specifically, the uniform distribution lies in the distributional ambiguity region, which is plainly not the case as the plots make clear). When using the kernel density estimator, we used bandwidth values of $r \in \{3\,\text{km}, 9\,\text{km}, 18\,\text{km}\}$. Six of these regions are shown in Figure 8, and Figure 9 shows

the output of our method (as well as the other districting criteria) when applied to a map of the metropolitan region surrounding Copenhagen. In each of these experiments, the sample points (i.e. the larger markers in Figures 8 and 9) were provided by our industrial affiliate, and were identified during the overnight scrubbing process that occurred after the vehicles' rough scans were finished. These sample points were identified automatically by software; most of the remaining destinations were detected by human analysts. The complete results of all 13 metropolitan regions are shown in Table 1, in which the "costs" are determined by computing the duration of the longest of the $m = 3$ TSP tours. Table 1 also shows two additional pieces of information: one is the column labelled "Best possible cost", which is obtained by computing the duration of the TSP tour of all of the points, and dividing that by 3 (the number of districts); this represents the best cost that could be realized if the workloads were distributed perfectly evenly. The other is the column labelled "Worst-case cost", which is the worst-case cost obtained when partitioning according to the Wasserstein criterion.

Note that the Wasserstein partitioning method out-performs the other five criteria in a plurality of instances (6 out of 13), as compared to 3 out of 13 regions for which the TSP-based partitioning is the most desirable. Moreover, the total makespan for all 13 regions, as shown in the last row of the table, is more than 50 hours less when using the Wasserstein criterion than the next best method (363.6 hours compared to 418.6). Since a typical workday for drivers does not exceed 8 hours, this would suggest that the total time needed to complete the scans of all 13 regions is reduced by a week. Note also that the Wasserstein partitioning criterion (among others) makes better use of data as it becomes available; this is apparent by comparing the result from (for example) Bergen, which has only 19 samples, with the result from Malmö, which has 60 samples (note that both the worst-case cost and the Wasserstein cost are closer to the ground truth cost and the best possible cost respectively when there are more samples available).



(a) Amsterdam

(b) Geneva

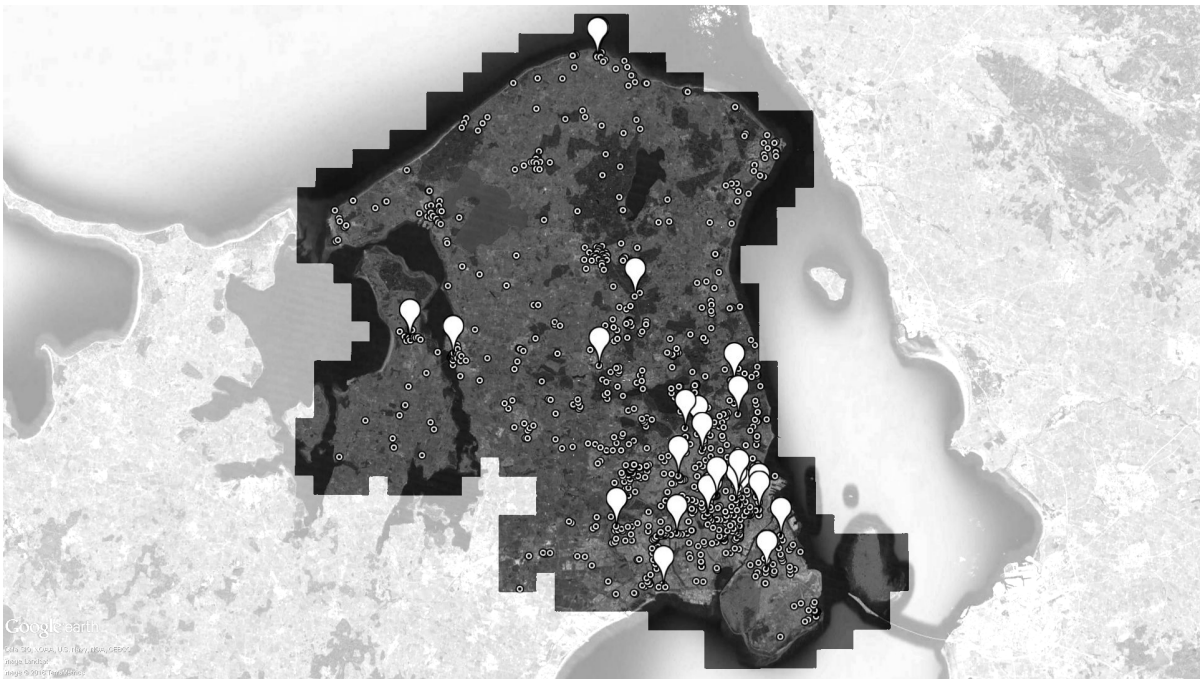
(c) Göteborg

(d) Malmö

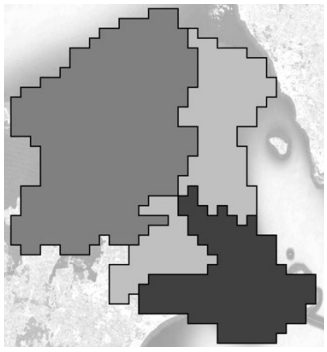
(e) Oslo

(f) Zürich

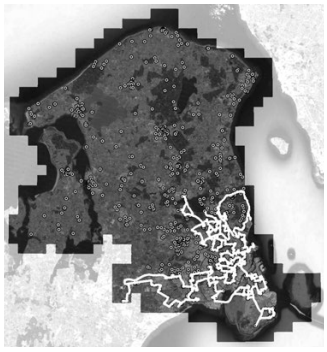
Figure 8: Six of the metropolitan regions surveyed in the computational study; the larger markers indicate data samples x_i and the smaller markers indicate additional locations that were realized later.



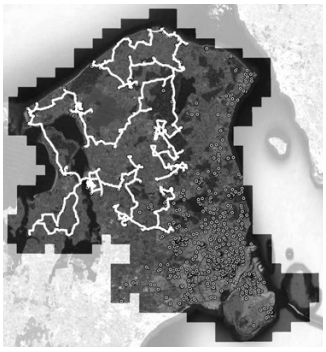
(a) Input



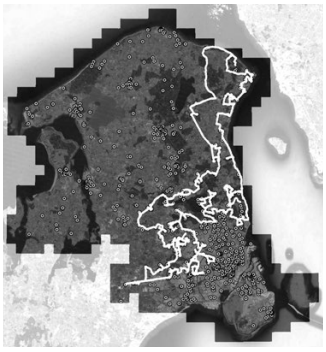
(b) Wasserstein partition



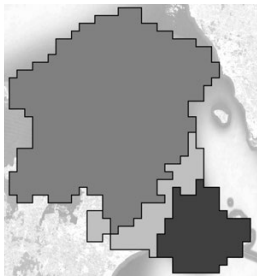
(c) TSP tour



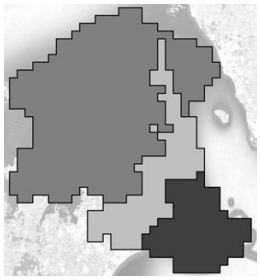
(d) TSP tour



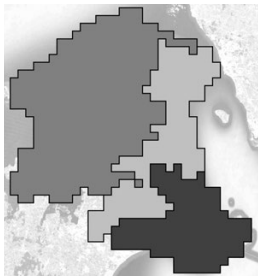
(e) TSP tour



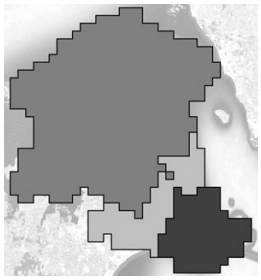
(f) Kernel partition 1, $r = 3\,\text{km}$



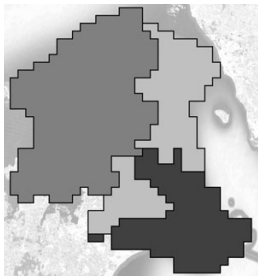
(g) Kernel partition 2, $r = 9\,\text{km}$



(h) Kernel partition 3



(i) Mean-covariance partition



(j) Tour partition

Figure 9: The result of various partitioning schemes applied to a map of Copenhagen and its surrounding metropolitan area. Figure 9a shows the input region, discretized into squares of $2\,\text{km} \times 2\,\text{km}$, with larger markers indicating data samples. Figure 9b shows the optimal robust partition with respect to the Wasserstein metric, and Figures 9c through 9e show the three TSP tours of the points within each district. Figures 9f through 9j show the partitions that result using the remaining criteria. The disconnected sub-region in 9j is somewhat surprising, although not unreasonable: if one substitutes the ℓ_1 norm in (20), then it is easy to construct examples that result in disconnected sub-regions (recall that we are in fact building sub-regions using road network distances).

Region	# points	# samples	Area (km^2)	t (km)	Actual Wass. dist. (km)	Best possible cost (hours)	Worst-case cost (hours)	Partition costs (hours)					
								Wasserstein	Kernel 1	Kernel 2	Kernel 3	Mean-covariance	TSP
Aarhus	1358	36	5904	22.6	21.2	25.4	55.1	29.5	40.4	43.8	43.5	40.1	***28.9***
Amsterdam	808	26	2528	11.1	7.8	17.1	40.0	***18.7***	24.7	22.9	21.2	23.9	25.5
Bergen	972	19	2880	23.1	15.7	29.2	60.2	***33.9***	51.9	51.7	42.7	42.6	34.8
Copenhagen	860	26	1680	30.0	12.0	14.4	24.3	***15.6***	26.8	19.9	16.7	25.2	20.3
Flanders & Brussels	2181	32	12052	9.7	7.2	37.3	112.3	48.7	59.8	62.2	58.6	55.8	***46.1***
Geneva	1194	41	5428	17.7	11.2	33.3	76.3	36.0	55.1	49.3	44.5	53.3	***33.5***
Göteborg	1543	51	9052	22.8	13.1	44.3	74.6	***45.3***	53.1	54.2	51.5	50.0	45.9
Helsinki	867	41	4372	5.9	5.0	15.7	32.9	19.7	***19.1***	19.1	21.4	19.5	39.0
Malmö	1391	60	5164	16.0	10.3	28.0	51.6	***29.3***	34.6	32.7	32.5	30.3	31.7
Oslo	952	29	3460	9.5	5.5	20.4	50.3	***24.8***	26.9	27.4	24.7	34.3	31.2
Rotterdam	881	21	2652	11.5	6.4	15.8	40.3	20.8	27.3	23.9	***19.3***	25.6	29.1
Stockholm	1003	31	4660	12.1	11.3	19.5	69.4	24.9	27.8	28.7	***24.8***	33.1	33.1
Zürich	965	30	1616	7.7	4.2	13.5	30.5	16.4	18.4	***16.2***	17.1	16.9	20.0
Total	14975	443	61448	200.1	131.0	314.2	707.0	***363.6***	465.9	452.0	418.6	450.7	419.1

Table 1: Tour lengths for 13 regions and 6 partitioning criteria; for each region, the optimal criterion is highlighted and underlined. All costs are defined as the maximum duration of the $m = 3$ TSP tours of the destination points.

5 Conclusions

By using the Wasserstein distance to define our region of ambiguity, we have developed a new tool for estimating the worst-case workload that one might face in visiting a sequence of points with a TSP tour. An advantage of this estimate is that it is not affected by problems that would arise if we used only mean and covariance information, as has been previously attempted. Our use of the square root functional $\iint_{\mathcal{R}} \sqrt{f(x)}\, dA$ to approximate lengths of TSP tours is just one possibility; one might also extend our analysis to handle more elaborate routing problems by adopting more sophisticated objective functionals as in [27]. To the best of our knowledge, our use of the Wasserstein distance in such an application is the first of its kind, and may have further applications outside the transportation domain, such as entropy or quantile maximization.

Acknowledgments

The authors thank the area editor, the associate editor, and three anonymous referees.

References

[1] A. Agra, M. Christiansen, R. Figueiredo, L. M. Hvattum, M. Poss, and C. Requejo. The robust vehicle routing problem with time windows. *Computers & Operations Research*, 40(3):856–866, 2013.

[2] M. Allahviranloo, J. Y. J. Chow, and W. W. Recker. Selective vehicle routing problems under uncertainty without recourse. *Transportation Research Part E: Logistics and Transportation Review*, 62:68–88, 2014.

[3] J. E. Anderson. The gravity model. Technical report, National Bureau of Economic Research, 2010.

[4] D. Applegate, W. Cook, D. S. Johnson, and N. J. A. Sloane. Using large-scale computation to estimate the Beardwood-Halton-Hammersley TSP constant. Presentation at 42 SBPO, 2010.

[5] F. Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[6] J. F. Bard and A. I. Jarrah. Large-scale constrained clustering for rationalizing pickup and delivery operations. *Transportation Research Part B: Methodological*, 43(5):542–561, 2009.

[7] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55(4):299–327, 1959.

[8] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[9] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007.

[10] F. Bolley and C. Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pages 331–352, 2005.

[11] S. Boyd. Localization and cutting-plane methods. `http://stanford.edu/class/ee364b/lectures/localization_methods_slides.pdf`, 2014.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[13] F. Bullo, E. Frazzoli, M. Pavone, K. Savla, and S. L. Smith. Dynamic vehicle routing for robotic systems. *Proceedings of the IEEE*, 99(9):1482–1504, 2011.

[14] G. Canas and L. Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pages 2492–2500, 2012.

[15] J. G. Carlsson. Dividing a territory among several vehicles. *INFORMS Journal on Computing*, 24(4):565–577, 2012.

[16] J. G. Carlsson and E. Delage. Robust partitioning for stochastic multivehicle routing. *Operations Research*, 61(3):727–744, 2013.

[17] J. G. Carlsson and R. Devulapalli. Dividing a territory among several facilities. *INFORMS Journal on Computing*, 25(4):730–742, 2012.

[18] J.G. Carlsson, E. Carlsson, and R. Devulapalli. Shadow prices in territory division. *Networks and Spatial Economics*, 2015.

[19] C. Daganzo. *Logistics Systems Analysis*. Springer, 2005.

[20] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[21] Google Developers. The Google Distance Matrix API. `https://developers.google.com/maps/documentation/distancematrix/intro`, 2015.

[22] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

[23] N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[24] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[25] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[26] C. E. Gounaris, W. Wiesemann, and C. A. Floudas. The robust capacitated vehicle routing problem under demand uncertainty. *Operations Research*, 2013.

[27] M. Haimovich and A. H. G. Rinnooy Kan. Bounds and heuristics for capacitated routing problems. *Mathematics of Operations Research*, 10(4):527–542, 1985.

[28] N. Halverson. Google claims right to post photos from private land. *The Press Democrat*, August 21, 2008.

[29] D. Haugland, S. C. Ho, and G. Laporte. Designing delivery districts for the vehicle routing problem with stochastic demands. *European Journal of Operational Research*, 180(3):997 – 1010, 2007.

[30] D. S. Hochbaum. When are NP-hard location problems easy? *Annals of Operations Research*, 1:201–214, 1984.

[31] J. Le Ny and G. J. Pappas. Adaptive deployment of mobile robotic networks. *Automatic Control, IEEE Transactions on*, 58(3):654–666, 2013.

[32] H. Lei, G. Laporte, and B. Guo. Districting for routing with stochastic customers. *EURO Journal on Transportation and Logistics*, 1(1-2):67–85, 2012.

[33] H. Lei, G. Laporte, Y. Liu, and T. Zhang. Dynamic design of sales territories. *Computers & Operations Research*, 56:84–92, 2015.

[34] E. H. Lockwood. *A book of curves*. Cambridge University Press, 1967.

[35] D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1968.

[36] C. H. Papadimitriou. Worst-case and probabilistic analysis of a geometric location problem. *SIAM Journal on Computing*, 10:542, 1981.

[37] M. Pavone, A. Arsie, E. Frazzoli, and F. Bullo. Distributed algorithms for environment partitioning in mobile robotic networks. *Automatic Control, IEEE Transactions on*, 56(8):1834–1848, 2011.

[38] M. Pavone, K. Savla, and E. Frazzoli. Sharing the load. *Robotics & Automation Magazine, IEEE*, 16(2):52–61, 2009.

[39] I. Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112, 2007.

[40] H. L. Royden and P. Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

[41] M. Schwager, D. Rus, and J.-J. Slotine. Decentralized, adaptive coverage control for networked robots. *The International Journal of Robotics Research*, 28(3):357–375, 2009.

[42] J.M. Steele. *Probability Theory and Combinatorial Optimization*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1987.

[43] I. Sungur, F. Ordónez, and M. Dessouky. A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty. *IIE Transactions*, 40(5):509–523, 2008.

[44] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[45] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[46] D. Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[47] D. Wozabal. Robustifying convex risk measures for linear portfolios: a nonparametric approach. *Operations Research*, 62(6):1302–1315, 2014.

[48] W. Xie and Y. Ouyang. Optimal layout of transshipment facility locations on an infinite homogeneous plane. *Transportation Research Part B: Methodological*, 75:74–88, 2015.

Online supplement to "Wasserstein distance and the distributionally robust TSP"

A Proof of Lemma 4

Proofs of statements 2 through 4 follow below:

Proof of statement 2 We seek to show that

$$\iint\limits_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i^{'}\}\, dA \leq \iint\limits_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA + \mathbf{g}^T(\boldsymbol{\lambda}^{'} - \boldsymbol{\lambda})\,,$$

which is equivalent to showing that

$$\iint\limits_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i^{'}\}\, dA \leq \sum_{i=1}^{n} \iint\limits_{R_i} f(x)(\|x - x_i\| - \lambda_i)\, dA + g_i(\lambda_i^{'} - \lambda_i)\,.$$

Consider the right-hand side of the above; for each i, we have

$$\begin{aligned} \iint\limits_{R_i} f(x)(\|x - x_i\| - \lambda_i)\, dA + g_i(\lambda_i^{'} - \lambda_i) &= \iint\limits_{R_i} f(x)(\|x - x_i\| - \lambda_i)\, dA - (\lambda_i^{'} - \lambda_i) \iint\limits_{R_i} f(x)\, dA \\ &= \iint\limits_{R_i} f(x)(\|x - x_i\| - \lambda_i^{'})\, dA \end{aligned}$$

and therefore, if we define regions $R_1^{'}, \ldots, R_n^{'}$ in the obvious way by

$$R_i^{'} = \left\{ x \in \mathcal{R} : \|x - x_i\| - \lambda_i^{'} \leq \|x - x_j\| - \lambda_j^{'}\ \forall j \neq i \right\}\,,$$

we see that

$$\iint\limits_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i^{'}\}\, dA = \sum_{i=1}^{n} \iint\limits_{R_i^{'}} f(x)(\|x - x_i\| - \lambda_i^{'})\, dA \leq \sum_{i=1}^{n} \iint\limits_{R_i} f(x)(\|x - x_i\| - \lambda_i^{'})\, dA$$

is obvious because the partition $R_1^{'}, \ldots, R_n^{'}$ is obtained by taking the *minimal* value of $\|x - x_i\| - \lambda_i^{'}$, and is therefore minimal over all partitions of $\mathcal{R}$. This completes the proof.

Proof of statement 3 We observe that the vector $-\frac{1}{n}\mathbf{e} \in \mathbb{R}^n$ must be a supergradient at $\boldsymbol{\lambda}^*$; this simply follows from the KKT conditions of (3), which is a finite-dimensional problem. Therefore, it follows that $\iint_{R_i^*} f(x)\,dA = 1/n$ for all i, and therefore the objective value of problem (3) is

$$\begin{aligned}
\iint\limits_{\mathcal{R}} f(x) \min_i\{\|x - x_i\| - \lambda_i^*\}\,dA &= \sum_{i=1}^n \iint\limits_{R_i^*} f(x)(\|x - x_i\| - \lambda_i^*)\,dA \\
&= \sum_{i=1}^n \iint\limits_{R_i^*} f(x)\|x - x_i\|\,dA - \lambda_i^* \iint\limits_{R_i^*} f(x)\,dA \\
&= \sum_{i=1}^n \iint\limits_{R_i^*} f(x)\|x - x_i\|\,dA - \frac{1}{n}\underbrace{\mathbf{e}^T\boldsymbol{\lambda}^*}_{=0} = \sum_{i=1}^n \iint\limits_{R_i^*} f(x)\|x - x_i\|\,dA
\end{aligned}$$

and therefore the Wasserstein distance between f and $\hat{f}$ as induced by the partition $R_1^*, \ldots, R_n^*$ is the same as that of the optimal objective value of (3), which completes the proof.

Proof of statement 4 We simply note that if $f(x) > 0$ then the supergradient inequality in the proof of statement 2 is actually strict:

$$\sum_{i=1}^n \iint\limits_{R_i'} f(x)(\|x - x_i\| - \lambda_i')\,dA < \sum_{i=1}^n \iint\limits_{R_i} f(x)(\|x - x_i\| - \lambda_i')\,dA\,.$$

The objective function of problem (3) is therefore strictly concave, thus guaranteeing uniqueness of $\boldsymbol{\lambda}^*$. The fact that $\boldsymbol{\lambda}^*$ exists follows from the boundedness of $\mathcal{R}$, because if we were ever to have $\lambda_i - \lambda_j > \mathrm{diam}(\mathcal{R})$, it would imply that $\|x - x_i\| - \lambda_i < \|x - x_j\| - \lambda_j$ for all $x \in \mathcal{R}$, thus rendering R_j to be empty.

B Proof of Theorem 10

Purely for ease of exposition, we assume that $\mathcal{R}$ is the unit square. Section 2.1 of [14] says that $\mathcal{D}(\hat{f}_n, \bar{f}) \to 0$ with probability one because the Wasserstein distance metrizes weak convergence whenever $\mathcal{R}$ is compact. Thus, setting $t_n = \mathcal{D}(\hat{f}_n, \bar{f})$ for all $n \geq 1$ gives us a sequence that converges to 0 with probability one, with the added feature that $\bar{f}$ is feasible for problem (12) by construction. Next, for each n, the triangle inequality says that the set of distributions f on $\mathcal{R}$ such that $\mathcal{D}(f, \bar{f}) \leq 2t_n$ must contain the set of distributions where $\mathcal{D}(f, \hat{f}_n) \leq t_n$. Thus, an upper bound for problem (12) – which is itself always an upper bound for the ground truth cost $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\,dA$ by

our construction of t_n – is given by the problem

$$\begin{aligned} \underset{f\in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}(f,\bar{f}) &\leq 2t_n \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,; \end{aligned} \tag{21}$$

it will therefore suffice to verify that the optimal objective value to this problem approaches the ground truth cost as $t_n \to 0$.

We will relax problem (21) one step further by using an alternate metric to the Wasserstein distance, namely the *Prokhorov metric* $\mathcal{D}_P(\cdot,\cdot)$, defined by

$$\mathcal{D}_P(\mu_1,\mu_2) = \inf\{\epsilon > 0 : \mu_1(B) \leq \mu_2(B^\epsilon) + \epsilon \text{ for all Borel sets } B \text{ on } \mathcal{R}\}$$

where $B^\epsilon = \{x : \inf_{y\in B} d(x,y) \leq \epsilon\}$. Theorem 2 of [25] says that for any two distributions f and g on $\mathcal{R}$, we have $(\mathcal{D}_P(f,g))^2 \leq \mathcal{D}(f,g)$, and therefore we can study the relaxation of (21) given by

$$\begin{aligned} \underset{f\in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}_P(f,\bar{f}) &\leq \sqrt{2t_n} \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{22}$$

as $n \to \infty$, whence $t_n \to 0$ with probability one. For ease of notation, we will define $\epsilon = \sqrt{2t_n}$.

Let N be a positive integer and suppose that $\epsilon = 1/N^3$. We then divide $\mathcal{R}$ into N^2 square grid cells s_i with side length $1/N$. The distance constraint $\mathcal{D}_P(f,\bar{f}) \leq \epsilon$ implies that for each $B = s_i$, we have $\iint_{s_i} f(x)\, dA \leq \iint_{S_i} \bar{f}(x)\, dA + \epsilon$, where S_i is the square of side length $1/N + 2/N^3$ that contains s_i (see Figure 10). Define $m_i = N^2 \iint_{S_i} \bar{f}(x)\, dA$ for each m_i and consider the relaxation of (22) given by

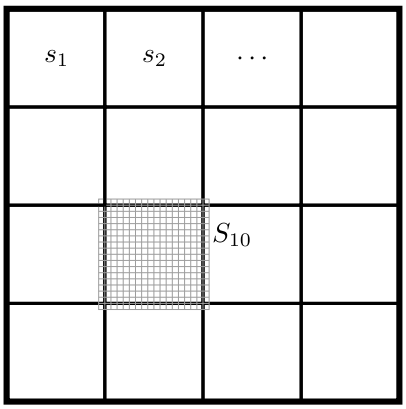


Figure 10: A division of the unit square $\mathcal{R}$ into $N^2 = 16$ grid cells. The larger square S_{10} has side length $1/N + 2/N^3$and contains s_{10}.

$$\begin{aligned} \underset{f \in L^1(\mathcal{R})}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{s_i} f(x)\, dA &\leq \frac{m_i}{N^2} + \epsilon \quad \forall i \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{23}$$

If we ignore the constraint that $\iint_R f(x)\, dA = 1$, then clearly, our optimal solution f^* would simply have $\iint_{s_i} f^*(x)\, dA = m_i/N^2 + \epsilon$ for each i. This problem has a finite-dimensional constraint space and it is straightforward to see that its optimal solution f^* must be piecewise constant on each piece s_i, so that $f^* = q_i^*$ on each s_i, defined by

$$\frac{q_i^*}{N^2} = \frac{m_i}{N^2} + \epsilon$$

or equivalently

$$q_i^* = m_i + 1/N\,.$$

Thus, the optimal objective value of (23) is at most

$$\frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{q_i^*} = \frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{m_i + 1/N} \leq \frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{m_i} + \frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{1/N} = \frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{m_i} + \sqrt{1/N}\,;$$

it is routine to verify that $\frac{1}{N^2}\sum_{i=1}^{N^2}\sqrt{m_i} \to \iint_{\mathcal{R}}\sqrt{\bar{f}(x)}\, dA$ (the only reason that this is not simply the definition of an integral is because the squares S_i that characterize the m_i's have an area of $(1/N + 2/N^3)^2$ rather than $1/N^2$), which thereby completes the proof.

C Probabilistic analysis of the capacitated VRP

We first note that, if n samples are drawn from a distribution f, then $\mathbf{E}(\sum_{i=1}^{n} \|x_i\|) = n \iint_{\mathcal{R}} \|x\| f(x)\, dA$. The representation of capacity constraints via the substitution $\kappa = s\sqrt{n}$ is a standard and useful technique that can be seen in Section 4.2 of [19]. By exchanging the expectation and $\max\{\cdot,\cdot\}$ operators, we can express the bound (17) as

$$\begin{aligned} & \max\left\{ \frac{2\sqrt{n}}{s} \iint_{\mathcal{R}} \|x\| f(x)\, dA,\ \beta\sqrt{n} \iint_{\mathcal{R}} \sqrt{f_c(x)}\, dA \right\} + o(\sqrt{n}) \\ \leq\ & \mathbf{E}\, \mathsf{VRP}(X) \\ \leq\ & 2\left\lceil \frac{\sqrt{n}}{s} \right\rceil \iint_{\mathcal{R}} \|x\| f(x)\, dA + \left(1 - \frac{1}{s\sqrt{n}}\right) \beta\sqrt{n} \iint_{\mathcal{R}} \sqrt{f_c(x)}\, dA + o(\sqrt{n})\,. \end{aligned}$$

Note that $\lceil \sqrt{n}/s \rceil$ is simply the number of vehicles needed to provide service. Since we are interested in the limiting behavior as $n \to \infty$, we have $\lceil \sqrt{n}/s \rceil \sim \sqrt{n}/s$ and $1/(s\sqrt{n}) \to 0$, so that we can write

$$\sqrt{n} \cdot \max\left\{ \frac{2}{s} \iint_{\mathcal{R}} \|x\| f(x)\, dA,\ \beta \iint_{\mathcal{R}} \sqrt{f_c(x)}\, dA \right\} \lesssim \mathsf{VRP}(X) \lesssim \sqrt{n} \cdot \left(\frac{2}{s} \iint_{\mathcal{R}} \|x\| f(x)\, dA + \beta \iint_{\mathcal{R}} \sqrt{f_c(x)}\, dA \right)$$

as desired, where the approximate inequality implied by the "$\lesssim$" terms simply reflects the fact that we have disregarded the $o(\sqrt{n})$ terms.