Wasserstein distance and the distributionally robust TSP

John Gunnar Carlsson*and Mehdi Behroozi†

September 24, 2015

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. In this paper, 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 as a decision support tool for dividing a territory into service districts for a fleet of vehicle when limited data is available.

1 Introduction

One of the most complex factors that arises in formulating and solving robust travelling salesman problems (TSP) and vehicle routing problems (VRP) is the difficulty of describing one's ambiguity set in a way that is both useful and mathematically tractable. Recent works have seen many different approaches for describing these sets, such as taking convex combinations of observed demand vectors [83], general polyhedral constructions [49], and using mean and covariance information about the spatial distribution of destination points [30]. The choice of one's ambiguity set often yields qualitative insights into what demand patterns affect the outcome most significantly; for example, the worst-case spatial distribution for the Euclidean TSP is that which is as equitably distributed (uniform) as

*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 Industrial and Systems Engineering, University of Minnesota.

possible [79].

In this paper, we consider a *distributionally robust* version of the Euclidean TSP: as input, we are given a compact, contiguous planar region $\mathcal{R}$ and a realization of sampled demand points in that region, and our objective is to construct a probability distribution on $\mathcal{R}$ that is sufficiently "close" to the empirical distribution consisting of the sampled points and is as "spread out" 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, we use a statistical metric called the *Wasserstein distance*, which is also known as the *earth mover's* or *Kantorovich* metric. 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. 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 basic theoretical preliminaries that are needed for the analysis that we perform in Section 3, which describes the structure of the worst-case spatial distribution for the TSP under a Wasserstein distance constraint. Next, Section 4 describes a primal-dual algorithm that finds this worst-case distribution efficiently, and this algorithm is then implemented in two computational experiments involving both the single-vehicle and multi-vehicle TSP in Section 5.

1.1 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.

1.1.1 Continuous approximation models

This paper is concerned with a *continuous approximation* model for a transportation problem, and is therefore philosophically similar to (for example) [22], which analytically determines trade-offs between transportation and inventory costs, [54], which shows how to route emergency relief vehicles to beneficiaries in a time-sensitive manner, and [57], which describes a simple geometric model for determining the optimal mixture of a fleet of vehicles

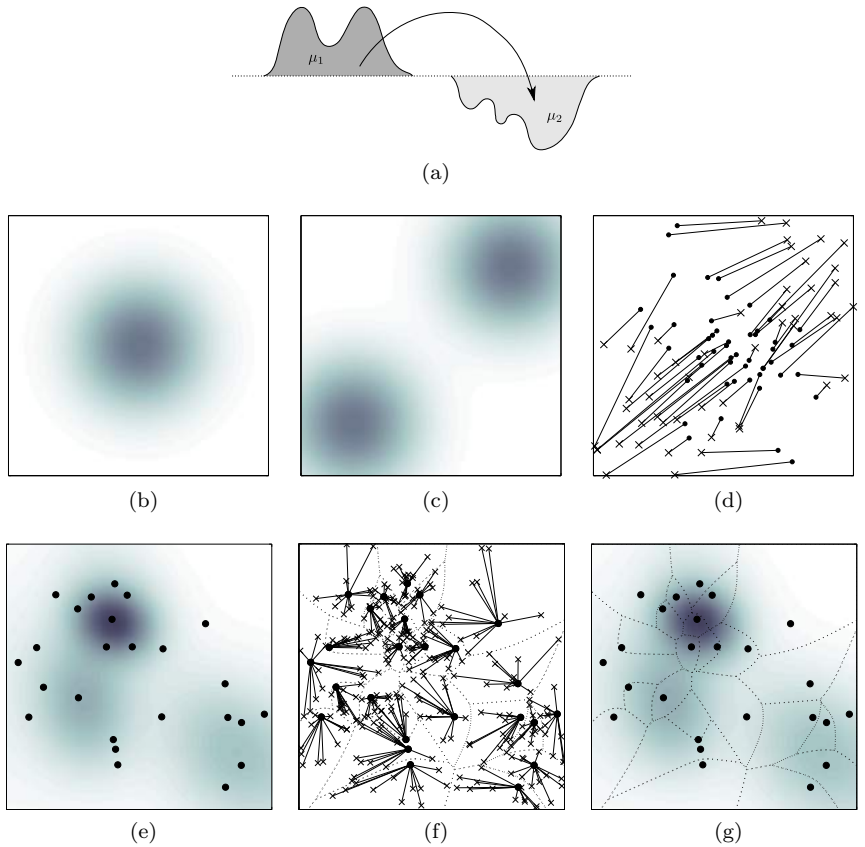


Figure 1: Figure 1a shows a Wasserstein distance problem between two univariate distributions μ_1 and μ_2. Figures 1b-1d show that a Wasserstein mapping can be thought of as an infinite-dimensional generalization of a bipartite matching; here μ_1 and μ_2 are shown in 1b and 1c, and 1d shows a bipartite matching between a large number of samples collected from μ_1 and μ_2. Figures 1e-1g show an interpretation of 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; a side consequence is that the Lagrange multipliers of this assignment induce a partition of the region $\mathcal{R}$, each of whose cells are associated with one of the elements of μ_2. Figure 1g shows this partition together with the atomic distribution μ_2; each cell contains $1/n$ of the mass of the density, which is to be transported to the point contained within it. By using previous results [31], these dashed curves are computationally easy to compute. If we let R_i denote the cell associated with each point x_i in the atomic distribution and f denote the density, then the Wasserstein distance is $\sum_i \iint_{R_i} f(x)\|x - x_i\|\, dA$.

that perform distribution. 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 [27, 46]. Such approximations exist for many combinatorial problems, such as the travelling salesman problem [14, 42], facility location [50, 53, 70], and any *subadditive Euclidean functional* such as a minimum spanning tree, Steiner tree, or matching [74, 80, 81]. 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.

1.1.2 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 [65] or Chapter 4 of the seminal book [35], for example. 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. To give a few examples, [44] uses a multiplicatively-weighted Voronoi partitioning scheme in which district sizes are determined by a set of scalar weights associated with the vehicles, [29] uses so-called "ham sandwich cuts" to recursively partition the region, and [68] uses a "disk model" that allows for explicit control of district sizes and implicit control of district shapes. Further additional examples from the robotics community include [34, 39, 71, 72], which all use various forms of the Voronoi paradigm (such as additive, quadratic, or logarithmic weighting schemes) to partition a geographic region, and place a particular emphasis on "decentralizing" the means by which partitions are constructed.

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 [52], which models the districting problem as a two-stage stochastic optimization program with recourse, by [13], which uses a three-phase procedure that aggregates data points into compact districts using a mixed-integer goal program, and by [45], which uses a steady-state spatial queueing model to simultaneously reason about the optimal locations of emergency response stations and the territories they serve. 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 [61, 62], which use the square-root approximation of [14] 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 5.2 of this paper also shows how to apply a continuous approximation scheme to a

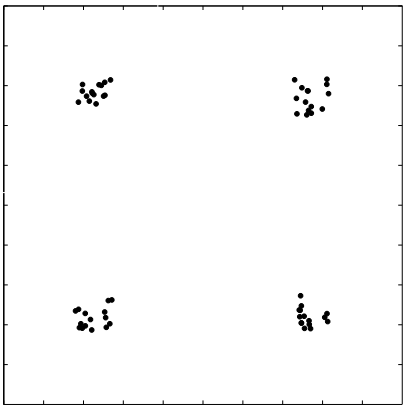


Figure 2: The above point sets in the unit square are extremely clustered and one would expect that their TSP tour should be short. However, because their sample mean and covariance matrix are the same as that of the uniform distribution, any robust methodology that uses only mean and covariance information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

heterogeneous road network, namely, a map of Los Angeles County, to a set of inputs that are non-uniformly distributed.

In practice, one often does not have sufficient information (namely, a probability distribution defined on an entire geographic region) to apply the preceding models. For example, in recent years, numerous start-up companies have emerged that provide "last-minute" delivery of food and groceries such as Good Eggs, DoorDash, BiteSquad, and Caviar [1, 2, 3, 4], and such companies must make high-level strategic allocation decisions without an extensive set of historical data. Another example arises in threat detection and surveillance, in which case a set of vehicles begins with some *a priori* information about the distribution of targets and one seeks a policy for routing these vehicles that takes new information into account as it becomes available [43]. The problem of designing districts in such a data-driven fashion (i.e. when one has an ambiguous distribution setting) is considerably less understood, although the paper [30] 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 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 [30].

1.1.3 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 [5, 49, 83], although the recent paper [6] 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 [26]: 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 [37, 73, 92], and the papers [33, 48] additionally make use of bounds on "directional deviation measures" that isolate stochastically independent components. We have previously made use of first and second moment information for the distributionally robust VRP in the paper [30]; 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. The Wasserstein distance is very commonly used in machine learning and statistics [23, 55, 69, 78], and is also mentioned in the context of robust optimization in [40] for its relationship with the *Prokhorov metric*. To our knowledge, the first direct applications of the Wasserstein distance to optimization problems have occurred very recently in [90, 91]; 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 paper [41] 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 3.

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. A few examples follow:

- The paper [15] solves a variety of robust linear programs using *ϕ-divergences*, a wide class of pseudo-metrics to which several of the aforementioned quantities belong.
- The paper [16] gives a highly flexible framework that builds ambiguity sets using classical statistical hypothesis

tests, including the χ^2 test and the Kolmogorov-Smirnov test,

- The paper [25] computes robust financial portfolios using the Kullback-Leibler divergence.
- The paper [56] shows how to solve robust dynamic programs whose distributional ambiguity sets are defined using the Kullback-Leibler divergence.
- The paper [58] proposes a robust lot-sizing model whose distributional ambiguity set is defined via the χ^2 goodness-of-fit test.

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 3.3, 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.

1.2 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 Preliminaries

In order to retain mathematical rigor, we find the following results useful; the first two items are simplified from [86]:

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

$$\mathcal{D}(\mu_1, \mu_2) := \inf_{\pi \in \Pi(\mu_1, \mu_2)} \iiiint_{\mathcal{R} \times \mathcal{R}} \|x - y\| \, d\pi(x, y) \,, \tag{1}$$

where $\Pi(\mu_1, \mu_2)$ 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 [14] and further developed in [80, 81], which relates the length of a TSP tour of some points with the distribution from which they were sampled:

Theorem 2 (BHH Theorem)**.** *Suppose that $X = \{X_1, X_2, \dots\}$ 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 $f(\cdot)$.

It is additionally known that $0.6250 \leq \beta \leq 0.9204$ and estimated that $\beta \approx 0.7124$; see [9, 14].

The following classical result from [64] 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 $x \in \mathcal{S}$ 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*

$$\begin{aligned}
\underset{\boldsymbol{\lambda}\in\mathbb{R}^n}{\text{maximize}} \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA \qquad & s.t. \\
\mathbf{e}^T \boldsymbol{\lambda} &= 0\,,
\end{aligned} \tag{2}$$

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

2. *For any* $\boldsymbol{\lambda}$*, a valid* supergradient *[20] for the objective function of (2) is the vector* $\mathbf{g} \in \mathbb{R}^n$ *defined by setting*

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

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*

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

3. *If* $\boldsymbol{\lambda}^*$ *is a maximizer of (2), 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 [86] or [12, 31, 85] for specific details. In addition, the economic interpretation of the regions R_i^* relative to the dual variables λ_i^* can be found in [32]; 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. Proofs of statements 2-4 are routine and can be found in Section A of the Online Supplement. □

3 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

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

with probability one as $n \to \infty$. Since β is a constant and $\sqrt{n}$ is (presumably) not related to the distribution f, 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 $\iint_{\mathcal{R}} \sqrt{f(x)}\, dA$.

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 4.2. 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

$$\begin{aligned} \underset{f}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}(f, \hat{f}) &\leq t \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{3}$$

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

3.1 Comparison with other approaches

The earlier paper [30] considers a problem closely related to (3) 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

$$\begin{aligned} \underset{f}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} x f(x)\, dA &= \mu \\ \iint_{\mathcal{R}} x x^T f(x)\, dA &\preceq \Sigma + \mu\mu^T \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{4}$$

Since it is well-known (e.g. Section 2 of [28]) that the Wasserstein distance between the empirical distribution $\hat{f}$ and the true distribution f converges to zero with probability one as $n \to \infty$, it is not surprising that our proposed formulation (3) is guaranteed to make better use of sample points as they become available, unlike the problem (4)

written above:

Theorem 5. *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*

$$\begin{aligned} \underset{f}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}(f, \hat{f}_n) &\leq t_n \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R} \end{aligned} \tag{5}$$

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. □

3.2 Structure of the solution to (3)

To begin, we apply Lemma 4 to express the distance constraint $\mathcal{D}(f, \hat{f}) \leq t$ in (3) differently, obtaining the equivalent formulation

$$\begin{aligned} \underset{f \in L_1}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA &\leq t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0 \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{6}$$

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 6. *There exists a unique optimal solution f^* to problem (6), 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 (6) is strictly concave. To prove existence, let $\{f^j\}$ denote a sequence of feasible inputs to (6) 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 (2). 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

$$\begin{aligned} \underset{f \in L_1}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \sum_{i=1}^{n} \iint_{R_i^*} \|x - x_i\| f(x)\, dA &\leq t \\ \iint_{R_i^*} f(x)\, dA &= \frac{1}{n} \quad \forall i \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R} \end{aligned} \tag{7}$$

(this is an immediate consequence of the fact that the optimal objective cost to (7) 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 (7) to derive the dual problem

$$\begin{aligned} \underset{\boldsymbol{\nu} \geq \mathbf{0}}{\text{minimize}}\, \frac{1}{4} \sum_{i=1}^{n} \iint_{R_i^*} \frac{1}{\nu_0 \|x - x_i\| + \nu_i}\, dA + \nu_0 t + \frac{1}{n}(\nu_1 + \cdots + \nu_n) \qquad & s.t. \\ \nu_0 \|x - x_i\| + \nu_i &\geq 0 \quad \forall x \in R_i^* \quad \forall i \end{aligned} \tag{8}$$

whereby we conclude that the optimal solution f^* to (7) must take the form

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

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 7. *The worst-case distribution that solves problem (6), and therefore (3), takes the form*

$$f^*(x) = \frac{1}{4(\nu_0^* \min_i\{\|x - x_i\| - \lambda_i^*\} + \nu_1^*)^2} \tag{10}$$

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 (9) is the fact that the expression in (9) is not guaranteed to vary continuously as we move from one region R_i^* to another; the expression (10) is continuous by inspection. We first note that the constraint

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

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 $\mathcal{X} = L^1$, Ω is the subset of the non-negative orthant in L^1 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

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

and

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

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. [77]). However, Lemma 6 shows that $f^*(x) > 0$ on $\mathcal{R}$, and therefore the optimal $\boldsymbol{\lambda}^*$ that solves problem (2) 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

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

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

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

or equivalently, the problem

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

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

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

to derive the 2-dimensional dual problem

$$\begin{aligned} \underset{\nu_0, \nu_1}{\text{minimize}} \iint_{\mathcal{R}} \frac{1}{4(\nu_0 \min_i \{\|x - x_i\| - \lambda_i^*\} + \nu_1)}\, dA + \nu_0 t + \nu_1 \qquad & s.t. \\ \nu_0 \min_i \{\|x - x_i\| - \lambda_i^*\} + \nu_1 &\geq 0 \quad \forall x \in \mathcal{R} \\ \nu_0, \nu_1 &\geq 0\,; \end{aligned} \tag{11}$$

the optimality conditions of (11) describe precisely the desired form of f^*, which completes the proof. □

Remark 8. 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 [90]. 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 [41, 91], for example. Because of the non-linearity in the objective, our worst-case distribution f^* is smooth and requires an iterative method to solve, which we will describe in Section 4.

3.3 Variations and extensions

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 activity on weekends versus weekdays (or other seasonal effects). If we require that point x_i have a mass q_i associated with it, then we can find the worst-case distribution f^* by solving problem (6), 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 [51] provides useful upper and lower bounds for the cost of a capacitated vehicle routing tour:

Theorem 9. *Let $X = \{x_1, \ldots, x_n\}$ be a set of demand points in the plane serviced by a fleet of vehicles with capacity c 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*

$$\max\left\{\frac{2}{c}\sum_{i=1}^{n}\|x_i\|, \mathsf{TSP}(X)\right\} \leq \mathsf{VRP}(X) \leq 2\left\lceil\frac{|X|}{c}\right\rceil \cdot \frac{\sum_{i=1}^{n}\|x_i\|}{|X|} + (1-1/c)\,\mathsf{TSP}(X)\,. \tag{12}$$

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:

$$\sqrt{n} \cdot \max\left\{ \frac{2}{s} \iint_{\mathcal{R}} \|x\| f(x)\, dA,\ \beta \iint_{\mathcal{R}} \sqrt{f_c(x)}\, dA \right\} \lesssim \mathsf{VRP}(X) \lesssim \sqrt{n} \cdot \left(\frac{2}{s} \iint_{\mathcal{R}} \|x\| f(x)\, dA + \beta \iint_{\mathcal{R}} \sqrt{f_c(x)}\, dA \right), \tag{13}$$

where we set $s = c/\sqrt{n}$ 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 3, the worst-case distribution that maximizes the right-hand side of (13) 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* [63].

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 3, the worst-case distribution that maximizes the right-hand side of the above, subject to a Wasserstein distance constraint, takes the form

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

The gravity model One of the salient attributes of the worst-case distribution f^* as established in Theorem 7 is that the presence of the square root in the objective of (3) 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* [7, 76, 84], which is "the most common formulation of the spatial interaction method" [76] and has historically been used to model a wide variety of demographic phenomena such as population migration [82], spatial utility for retail stores [75], and trip distributions between cities [88]. 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.

4 Solving (3) efficiently

The preceding section established that the worst-case distribution that solves (3) 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* [19]. Recall that our problem of interest, as written in (6), is

$$\begin{aligned} \underset{f \in L_1}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA &\leq t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0 \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,; \end{aligned}$$

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 (6) and hence has an objective value that is at least as large as that of (6):

$$\begin{aligned} \underset{f \in L_1}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \bar{\lambda}_i\}\, dA &\leq t \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned}$$

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 7 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

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

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

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^n}{\text{minimize}} \max_{f\in\Omega(\boldsymbol{\lambda})} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathbf{e}^T\boldsymbol{\lambda} &= 0 \\ \lambda_i &\leq \operatorname{diam}(\mathcal{R}) \quad \forall i \end{aligned} \tag{14}$$

where $\Omega(\boldsymbol{\lambda})$ is the subset of L^1 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

$$\begin{aligned} \underset{\nu_0,\nu_1}{\text{minimize}} \iint_{\mathcal{R}} \frac{1}{4(\nu_0 \min_i\{\|x - x_i\| - \lambda_i\} + \nu_1)}\, dA + \nu_0 t + \nu_1 \qquad & s.t. \\ \nu_0 \min_i\{\|x - x_i\| - \lambda_i\} + \nu_1 &\geq 0 \quad \forall x \in \mathcal{R} \\ \nu_0, \nu_1 &\geq 0\,, \end{aligned} \tag{15}$$

as derived in the proof of Theorem 7, can be computed explicitly. Thus, we simply require a better understanding of problem (14):

Lemma 10. *The (outer) objective function of problem (14) is quasiconvex, i.e. its sub-level sets are convex.*

Proof. For notational compactness, let $G(\boldsymbol{\lambda})$ denote the objective function of (14). Recall [21] 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 (14):

Theorem 11. *Let $\bar{\boldsymbol{\lambda}}$ satisfy $\mathbf{e}^T \bar{\boldsymbol{\lambda}} = 0$ and let $\bar{f}$ be the solution to the inner problem of (14) (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 (14); 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 (14) 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 (14) and thereby recovering

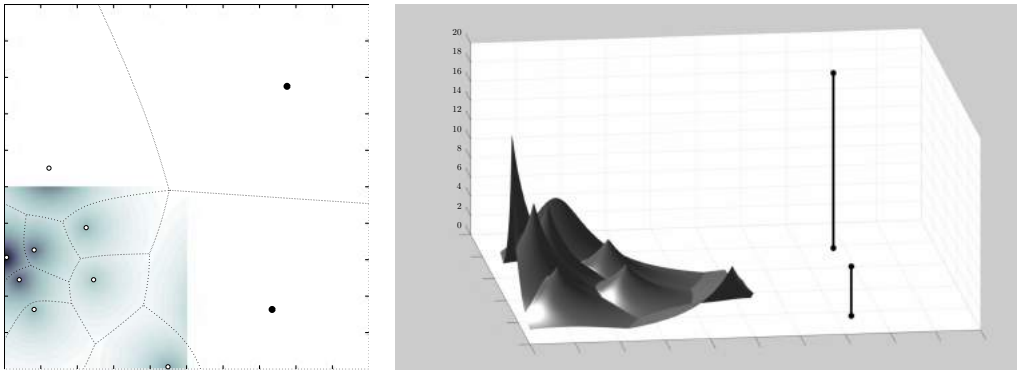
the distribution f^* that solves problem (3); see Algorithm 1 for a formal description.

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 (14), which has an` n`-dimensional variable space. */`
Set UB $= \infty$ and LB $= -\infty$;
Set $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^n : \mathbf{e}^T\boldsymbol{\lambda} = 0,\ \lambda_i \leq \text{diam}(\mathcal{R})\ \forall i\}$;
while UB $-$ LB $> \epsilon$ **do**
  Let $\bar{\boldsymbol{\lambda}}$ be the analytic center of Λ;
  `/* Build an upper bounding` $\bar{f}$ `for the original problem (3). */`
  Let $\bar{\nu}_0, \bar{\nu}_1$ be the solution to problem (15) 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 UB $= \iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$;
  `/* Build a lower bounding` $\tilde{f}$ `that is feasible for (3) 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 (8) 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 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.

4.1 Districting

When one has multiple vehicles available to perform service, a natural strategy for allocating them – especially in the presence of uncertainty – is to use a *districting* strategy in which we divide the region $\mathcal{R}$ into sub-regions, then associate each vehicle with one of these sub-regions [29, 45, 52, 71]. In the context of this paper, the most natural procedure would be to partition $\mathcal{R}$ into districts $D_1, \ldots, D_m$ and calculate the worst-case workloads associated with



(a) (b)

Figure 3: Two views of an example of $f^*(x)$ as described in Theorem 12, where there are $n = 10$ points and the sub-region D_j is the lower quarter of the unit square. At left, the shading represents f^* and the dashed lines indicate the optimal Wasserstein map between f^* and $\hat{f}$; the Dirac delta functions are indicated by the thick black circles in both images.

each district D_j by solving the problem

$$\begin{aligned} \underset{f \in L_1}{\text{maximize}} \iint_{D_j} \sqrt{f(x)}\, dA \qquad & s.t. \\ \iint_{\mathcal{R}} f(x) \min_i \{\|x - x_i\| - \lambda_i\}\, dA \quad &\leq \quad t \quad \forall \boldsymbol{\lambda} : \mathbf{e}^T \boldsymbol{\lambda} = 0 \\ \iint_{R} f(x)\, dA \quad &= \quad 1 \\ f(x) \quad &\geq \quad 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{16}$$

for each sub-region D_j (this is identical to (6) except for the domain of integration in the objective). The worst-case distribution associated with each district D_j is characterized as follows:

Theorem 12. *The worst-case distribution that solves problem (16) 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 7 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. See Figure 3 for an example of such a distribution.

□

In Section 5.2, we will apply the preceding result to a computational experiment in which we seek to divide a service region into districts in such a way as to minimize the maximum workload over any district.

4.2 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 (3) 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 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 [28], 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. Theorem 6.15 of [87] 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 [18] 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 $\log \iint_{\mathcal{R}} e^{2\|x-x_0\|} f_2(x)\, dA \le \log e^{2r} = 2r$. Thus, if $\mathcal{D}(f_1, f_2) \ge t$, we have

$$\begin{aligned} t \le \mathcal{D}(f_1, f_2) &\le \iint_{\mathcal{R}} \|x_0 - x\| \cdot |f_1(x) - f_2(x)|\, dA \\ &\le \left(\frac{3}{2} + 2r\right)\left(\sqrt{H(f_1|f_2)} + \frac{1}{2} H(f_1|f_2)\right) \\ \implies H(f_1|f_2) &\ge \frac{8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6}{3 + 4r}\,. \end{aligned} \tag{17}$$

Next, the paper [17] shows that, for any distribution f with empirical distribution $\hat{f}$, we have

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

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

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

the result (17) establishes that $\alpha(t) \ge (8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6)/(3 + 4r)$, and therefore we find that

$$\begin{aligned} \limsup_{n \to \infty} \frac{1}{n} \log \Pr(\mathcal{D}(f, \hat{f}) \ge t) &\le -\frac{8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6}{3 + 4r} \\ \implies \Pr(\mathcal{D}(f, \hat{f}) \ge t) &\lesssim \exp\left(-n \frac{8r - 2\sqrt{16r^2 + 16rt + 24r + 12t + 9} + 4t + 6}{3 + 4r}\right), \end{aligned} \tag{18}$$

where the notation "$\lesssim$" reflects the approximate inequality that results from dropping the "lim sup" term. Thus, given a desired significance level $1 - \beta$, we can construct a threshold distance t by equating the right-hand side of (18) to $1 - \beta$ and solving for t. Figure 4 shows a plot of these threshold values of t as a function of β, for $n = 100$ samples in the unit square.

5 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 of Los Angeles County into pieces so as to to minimize the worst-case workload of any vehicle.

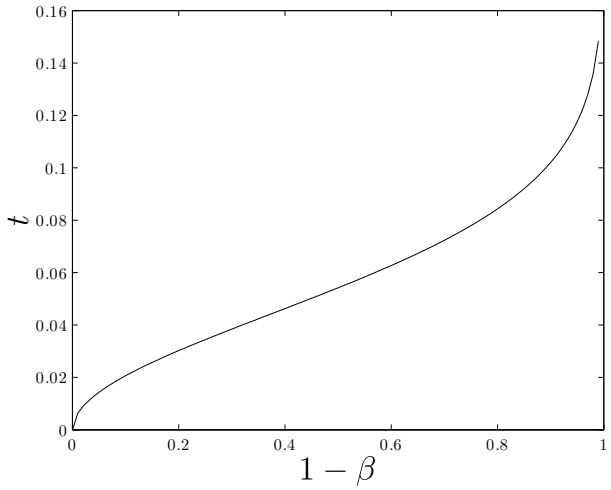


Figure 4: Threshold values of t for significance levels between 0 and 1, where $n = 100$ points are sampled in the unit square. For example, at $1 - \beta = 0.9$, we have $t = 0.102$; this means that, if 100 samples are drawn from any distribution f in the unit square, then there is at least a 90% probability that $\mathcal{D}(f, \hat{f}) \leq 0.102$.

5.1 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 (3) 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 $1/n$). Figure 5a shows a plot of the worst-case TSP costs $\iint_{\mathcal{R}} \sqrt{f^*(x)}\, dA$ as n varies, and Figure 5b 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 5b 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 5c shows the estimates of $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$ obtained when one uses a uniform kernel density estimator.

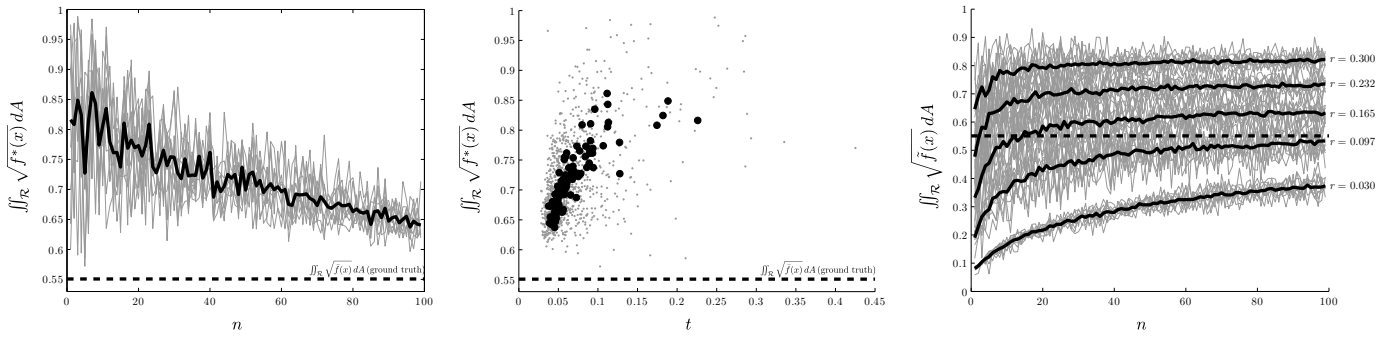


Figure 5: Figure 5a shows the worst-case costs that are computed during the 99×10 executions of our algorithm; the gray plots indicate the results obtained from individual samples and the thick line indicates the sample averages of the 10 trials for fixed n. Figure 5b shows the same data set, only we 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 purposes of comparison, Figure 5c 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. As in the preceding two figures, the gray plots indicate the results from individual samples, the thick lines indicate sample averages of 10 trials for fixed n, and the dashed line indicates the true value of $\iint_{\mathcal{R}} \sqrt{\bar{f}(x)}\, dA$; furthermore, as indicated, 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 r.

5.2 A districting experiment with road network data

In this section, we describe an experiment in which we divide a service region $\mathcal{R}$ into 4 pieces 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. Specifically, our service region $\mathcal{R}$ is a map of Los Angeles County, and distances between points are measured according to the driving distance, as obtained via the Google Distance Matrix API [38]. Our sampled points $x_1, \ldots, x_n$ are the locations of crime reports filed in the first week of July, which were extracted from the "Detailed Report" tool at the CrimeMapping.com website developed by the Omega Group [67], and are shown in Figure 6. The purpose of this experiment is two-fold: first, we must demonstrate that the proposed continuous approximation techniques are actually useful for solving practical problems, and second, we must then show that our proposed methodology is superior to that of existing approaches.

5.2.1 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. As an initial

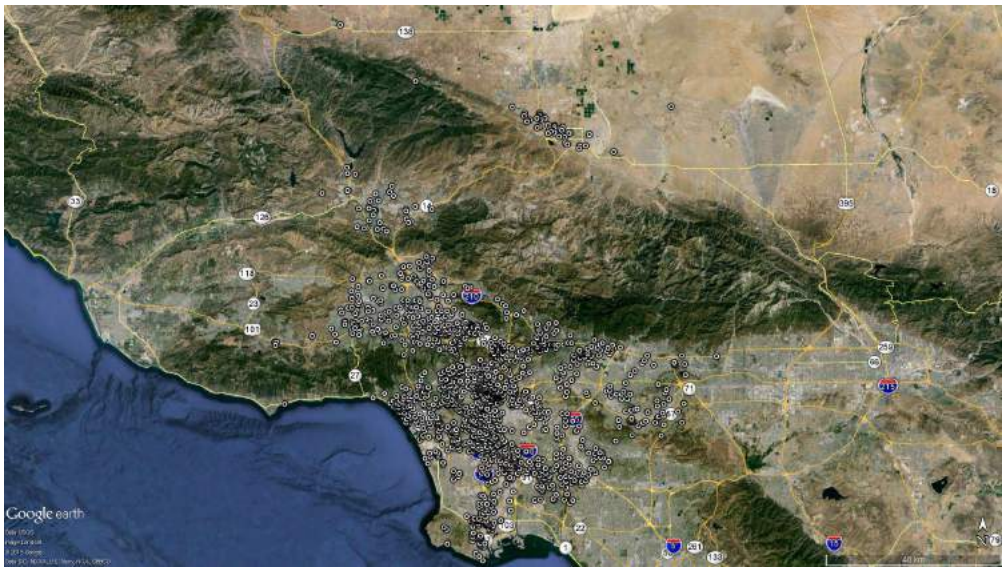
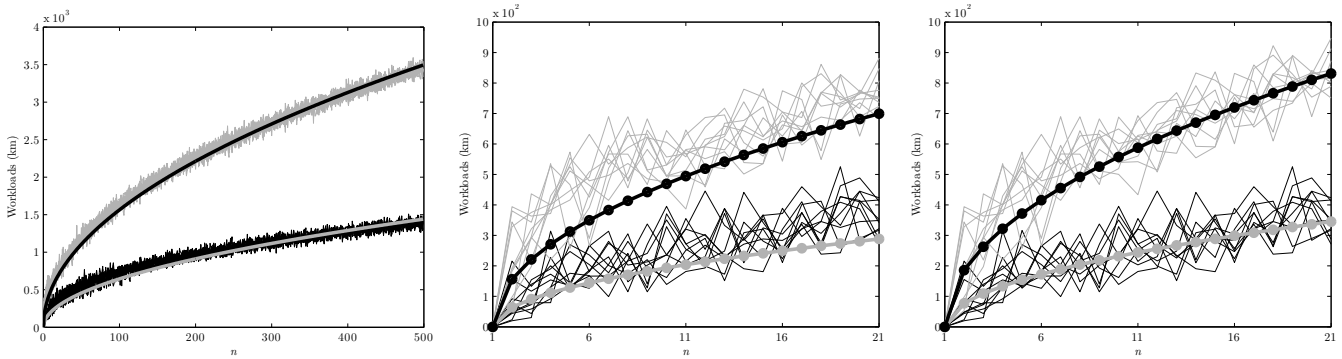


Figure 6: Locations of 1704 crime reports filed in Los Angeles County during the first week of July, obtained from the website [67].

sanity check, we first sample between $n = 2$ and $n = 500$ points uniformly at random within the map $\mathcal{R}$ of Los Angeles County, and solve the TSP for these points where distances are given by the driving distance as obtained from the Google Distance Matrix API [38]. We also sample between $n = 2$ and $n = 500$ points from the non-uniform distribution consisting of the locations of crime reports obtained from [67], as shown in Figure 6. The lengths of these tours, computed via the Concorde TSP solver [8], are shown in Figure 7, and are consistent with the findings of Table 16.7 of [10] for the Euclidean TSP in the unit square: specifically, we see that the square-root approximation tends to slightly *underestimate* the tour length for small values of n. For example, [10] says that a TSP tour of $n \approx 100$ points in the unit square (with Euclidean distances) is approximately $0.78\sqrt{n}$, whereas a TSP tour of $n \approx 1000$ points is approximately $0.73\sqrt{n}$. Figure 7 suggests that, provided one has some vague estimate of the number of destination points n, the square root approximation is indeed a valid one, even when distances are not Euclidean and the spatial distribution is not uniform.

We have thus far established that the length of a TSP tour of n points in $\mathcal{R}$ scales proportionally to $\sqrt{n}$. However, it is necessary to take into account the heterogeneity of the road network as well: for example, point-to-point distances in the middle of downtown Los Angeles are likely to be close to the ℓ_1 metric because the streets form a regular grid, whereas point-to-point distances elsewhere will likely be longer due to sparser road coverage. 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



(a) (b) (c)

Figure 7: The lengths of the TSP tours of n points sampled in Los Angeles County, where point-to-point distances are induced by a road network. In all three diagrams, the upper plot corresponds to uniformly distributed points and the lower plot corresponds to points that are sampled from the CrimeMapping.com website [67]. For each value of $n \in \{1, \ldots, 500\}$ and for each of the two sampling strategies (uniform or non-uniform), we perform 10 independent experiments in which n points are sampled and their TSP tour is calculated using Concorde [8]. Figure 7a shows the tour lengths (the thin lines) together with the best fit of these tour lengths to a curve of the form $C\sqrt{n}$, where we have $C \approx 159$ for uniformly sampled points and $C \approx 64$ for non-uniformly distributed points. Figure 7b shows a close-up of this plot for $n \in \{1, \ldots, 21\}$, where we can see that the fitted curve underestimates the true tour lengths when n is small. A better fit for these small values of n, as shown in Figure 7c, is to set $C \approx 185$ for uniformly sampled points and $C \approx 77$ for non-uniformly distributed points (as an aside, for small values of n, it is clearly common sense to fit a curve of the form $C\sqrt{n-1}$, since the TSP tour of a single point has length 0). This establishes that, provided one has an approximate estimate of the number of points n, the square root approximation is indeed a reasonable one.

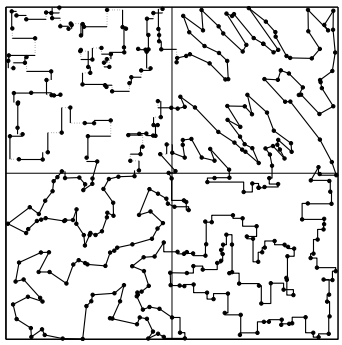


Figure 8: The TSP tour of a collection of points uniformly sampled in the unit square, with varying metrics depending on quadrants. The dashed lines in the paths in the upper left quadrant correspond to the smaller of the two directions (horizontal or vertical) between points, which is relevant because the ℓ_∞ distance is used.

2 holds whenever the underlying structure is a *subadditive Euclidean functional*; see [80, 81] 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:

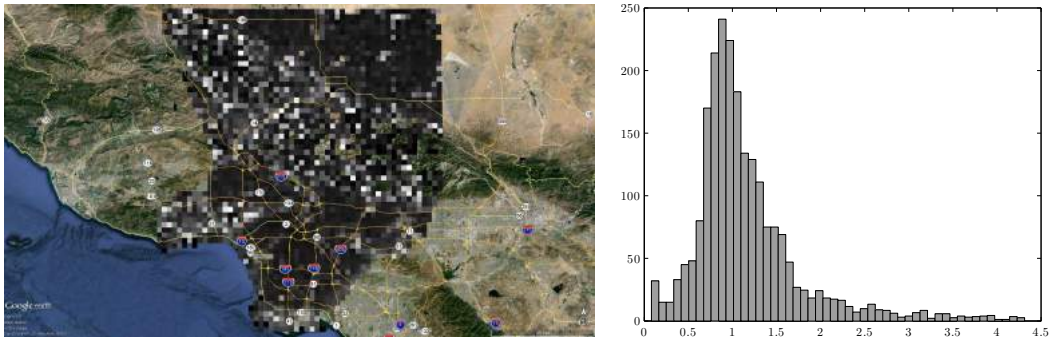
Example 14 (Varying metrics in a region)**.** Consider a set of n points sampled according to a distribution f in the unit square, with distances $d(x_1, x_2)$ between pairs of points $x_1 = (x_1^1, x_1^2)$ and $x_2 = (x_2^1, x_2^2)$ defined as follows:

- 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.
- 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 8. 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 [81].



(a) (b)

Figure 9: The shading in Figure 9a indicates the values of β_i associated with each of the square patches (lighter patches correspond to higher values of β_i). Figure 9b is a histogram of these same values; note that a handful of these values are actually *lower* than the current estimate [9] of the Euclidean TSP coefficient $\beta \approx 0.7124$; this appears to be a combination of statistical noise and an artifact of the Google Maps API [38] that was used to compute driving distances; for example, two locations belonging to the same venue (e.g., two stores on opposite ends of a large shopping mall) report a driving distance of 0.

Example 14 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 length of their TSP tour ℓ_i (using road network distances), 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 $\beta_i = \ell_i / \sqrt{\mathrm{Area}(Q_i) \cdot k}$. We discretized the region $\mathcal{R}$ into a collection of patches Q_i of size $2\,\mathrm{km} \times 2\,\mathrm{km}$, and estimated each coefficient β_i using $k = 10$ samples (a larger number would be preferable, but the Google Maps API [38] imposes a limit of at most $100,000$ queries per day); Figure 9 shows the resulting values. We found a total of $K = 2564$ patches in which road coverage was adequate for distances to be estimated, whence $\mathrm{Area}(\mathcal{R}) = (2\,\mathrm{km} \times 2\,\mathrm{km}) \cdot K = 10256\,\mathrm{km}^2$.

In order to validate these estimates β_i, we revisit the earlier experiment in which we sampled up to 500 points in $\mathcal{R}$, where we found (as suggested in Figure 7a) that the length of the TSP tour of n points sampled uniformly in $\mathcal{R}$ is approximately $159\sqrt{n}$ kilometers. It follows that, for uniformly distributed points X_i in $\mathcal{R}$, the fact that

$f(x) = 1/\operatorname{Area}(\mathcal{R})$ implies that

$$\begin{aligned} 159\sqrt{n} \approx \mathsf{TSP}(X_1, \ldots, X_n) \approx \sqrt{n} \cdot \sum_{i=1}^{K} \beta_i \iint_{Q_i} \sqrt{f_c(x)}\, dA &= \sqrt{n} \cdot \sum_{i=1}^{K} \beta_i \cdot (4\,\mathrm{km}^2) \cdot \sqrt{1/\operatorname{Area}(\mathcal{R})}\, dA \\ &= \sqrt{n} \cdot 0.0395 \sum_{i=1}^{K} \beta_i \,, \end{aligned}$$

and so we expect to find that $0.0395 \sum_{i=1}^{K} \beta_i$ should sum to approximately 159. Indeed, we find that $0.0395 \sum_{i=1}^{K} \beta_i \approx 142$, so that we introduce a relative error of approximately 10%. In order to compensate for this error, we re-scale all terms $\beta_i \mapsto \frac{159}{142}\beta_i$ for all i.

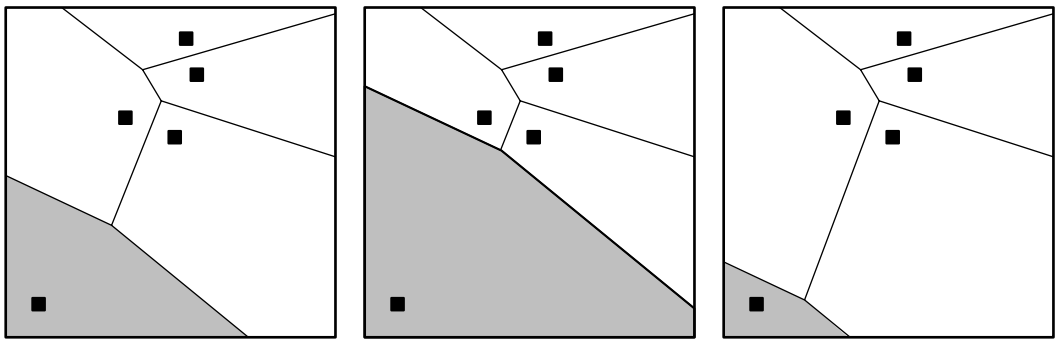
5.2.2 Districting criteria

In this section, we apply the theory developed in this paper to solve a problem in which we seek to partition $\mathcal{R}$ (a map of Los Angeles County) into service districts so as to divide the workloads of a fleet of vehicles in a balanced way. In order to divide $\mathcal{R}$ into districts, we use a computational geometric structure called a *power diagram* [11], which has frequently been applied to districting problems in existing literature on vehicle routing [24, 30, 34, 60, 71]. Given a set of "depot points" $p_1, \ldots, p_m$ in $\mathcal{R}$ and any vector $\mathbf{w} \in \mathbb{R}^m$, the 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

$$D_i = \left\{ x \in \mathcal{R} : \|x - p_i\|^2 - w_i \leq \|x - p_j\|^2 - w_j \; \forall j \neq i \right\} . \tag{19}$$

An example of a power diagram is shown in Figure 10. It is straightforward to verify that the pieces D_i are always convex. In our experiment, we let $p_1, \ldots, p_{m=4}$ be the locations of the 4 major police stations associated with the 4 largest cities in Los Angeles County, namely Los Angeles, Long Beach, Glendale, and Santa Clarita, which are shown in Figure 11. Given these locations p_i, our goal is to select a weight vector $\mathbf{w}$ that determines an "optimal" partition $D_1, \ldots, D_4$ with respect to some cost function that approximates the workloads in these regions; selected cost functions will be described later in this section.

It would of course be desirable to replace the Euclidean distance terms in (19) with the driving distance (which would result in boundaries between sub-regions that are characterized, in some sense, by the underlying road network). Unfortunately, for practical reasons, we are prevented from doing this because of the limit of $100,000$ queries per day of the Google Maps API, which we used to calculate driving distances. When one does not have such a limit (e.g. when driving distances are computed "in-house"), then obviously, such a method is indeed feasible. We next describe three objective functions or attributes associated with the districts D_i that we seek to optimize



(a) (b) (c)

Figure 10: Figure 10a is a Voronoi partition, that is, a power diagram with all weights w_i equal (each district consists of those points that are closer to their associated landmark point than the others). Figures 10b and 10c present the two power diagrams obtained by increasing and decreasing the weight associated with the shaded cell.

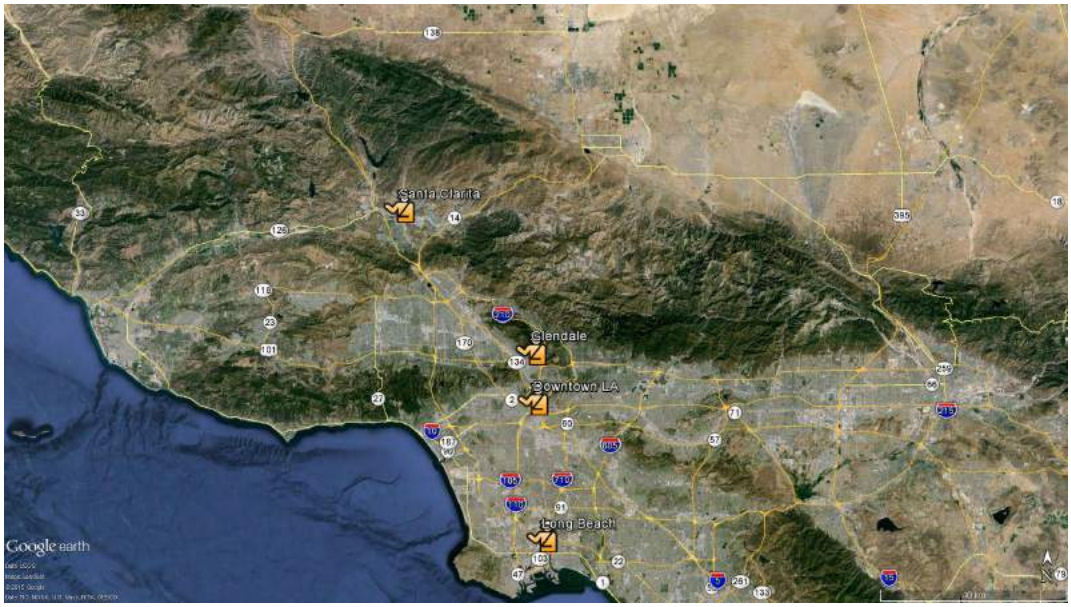


Figure 11: Locations of 4 police stations associated with the 4 largest cities in Los Angeles County.

by selecting the weight vector $\mathbf{w}$.

Equal n_i: A common-sense criterion in designing districts would be to require that the number of customers in each district D_i, which we denote as n_i, be the same. This is eminently sensible when the spatial distribution of customers is uniform, and has been applied previously to large-scale routing problems in [59, 89], for example. Given a collection of points $x_1, \ldots, x_n$ in $\mathcal{R}$, it turns out that it is computationally extremely simple to compute a weight vector $\mathbf{w}^*$ such that each district D_i contains the same number n/m of customers (with the possibility of being off by one if m does not divide n evenly): the desired weight vector is the Lagrange multiplier vector associated with the first set of constraints in the assignment problem

$$\begin{aligned} \underset{z_{ij}}{\text{minimize}} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} z_{ij} \qquad & s.t. \\ \sum_{i=1}^{n} z_{ij} &= \frac{n}{m}\ \forall j \\ \sum_{j=1}^{m} z_{ij} &= 1\ \forall i \\ z_{ij} &\geq 0\ \forall i, j\,, \end{aligned}$$

where we set $c_{ij} = \|x_i - p_j\|^2$. This is solvable as a linear program.

Equal $\sqrt{A_i n_i}$: By far the most popular approximation in designing districts for vehicle routing problems is the estimation

$$\mathsf{TSP}(D_i) \approx \beta\sqrt{A_i n_i}\,,$$

where $A_i = \text{Area}(D_i)$, n_i is the number of customers in D_i as before, and the notation $\mathsf{TSP}(D_i)$ denotes the length of the TSP tour through the n_i points in D_i. This has been used previously in [22, 44, 61, 62, 66], for example, and is predicated on the assumption that the points are uniformly distributed *within* each district D_i (in other words, the distribution may vary over $\mathcal{R}$, but the distribution is assumed to be more or less uniform within each of the districts). This is nothing more than the BHH theorem, applied to a set of points that are uniformly distributed within each district D_i. It is possible to compute a weight vector $\mathbf{w}^*$ such that $\sqrt{A_i n_i}$ is equal for all districts D_i using a simple gradient descent scheme similar to that used in [71] (with the same caveat that an "off by one" error may exist as in the previous example).

Mean-covariance robust partitioning: Section 5 of our earlier paper [30] 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

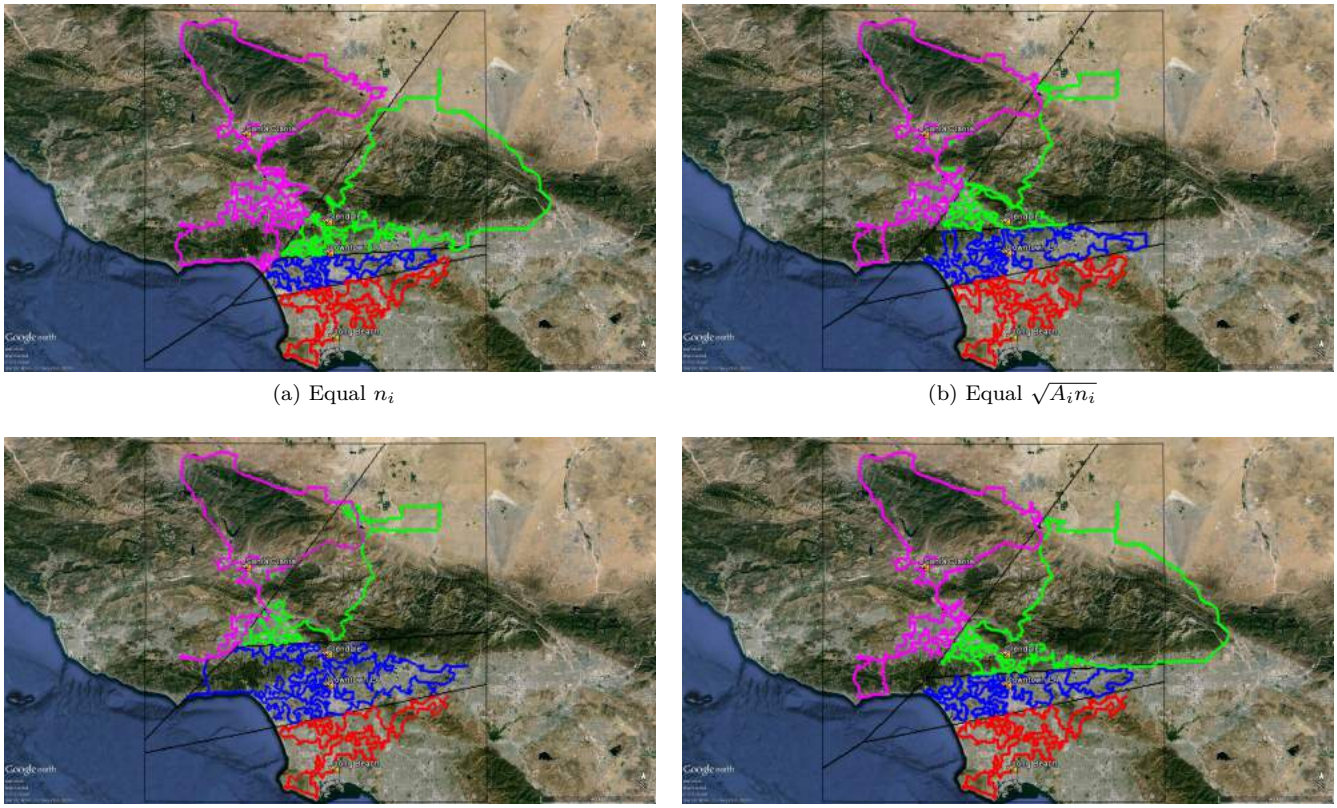
to (4). In other words, we construct a power diagram partition against all distributions whose mean and covariance matrix are equal to the fixed values obtained from the sampled points $x_1, \ldots, x_n$.

Wasserstein robust partitioning Problem (16) in Section 4.1 of this paper describes the structure of the worst-case distribution that maximizes the asymptotic workload in a particular district D_i, 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 (16) is equal for each district (in other words, the worst-case workloads are the same for all districts). The branch-and-bound scheme from Section 5 of [30] is based on a simple set of monotonicity properties and can be applied to find these districts as well.

5.2.3 Results

In order to demonstrate the practicality of our proposed approach, we compare the districts that result from enforcing the four criteria from the preceding section, where the data points are the locations of crime reports filed in the first week of July as previously shown in Figure 6. We give the first three partitioning criteria (i.e. the non-Wasserstein criteria) an advantage by building their districts using knowledge of all 1704 sample points (the mean-covariance partitioning scheme uses exact knowledge of the mean and covariance of the data points). By comparison, the Wasserstein partitions are computed using a sample of only 50 points drawn from the full dataset, and with a threshold distance t calculated according to equation (18) with a 90% confidence level. The four sets of districts, together with the resulting workloads, are shown in Figures 12 and 13. Not surprisingly, we see that the non-uniformity of the samples leads to districts whose workloads are substantially unbalanced. Even more surprising is the fact that the mean-covariance robust partitioning method is by far the *worst* of the three; we attribute this to the fact that the crime locations are distributed in a highly multi-modal, non-uniform fashion, and that the worst-case distribution with given first and second moments as derived in [30] always has a unimodal shape (together with a mixture of Dirac delta components).

This experiment establishes that our approach is useful for balancing workloads of vehicles when one has limited sample information and when the underlying distribution is highly non-uniform. We also conducted another set of experiments in which we sampled 1000 points *uniformly* in $\mathcal{R}$, as shown in Figure 14, and applied the four districting criteria. As Figure 15 shows, the four criteria are roughly indistinguishable in this case.



(a) Equal n_i

(b) Equal $\sqrt{A_i n_i}$

(c) Robust mean-covariance

(d) Wasserstein

Figure 12: The power diagram districts obtained according to the four partitioning criteria.

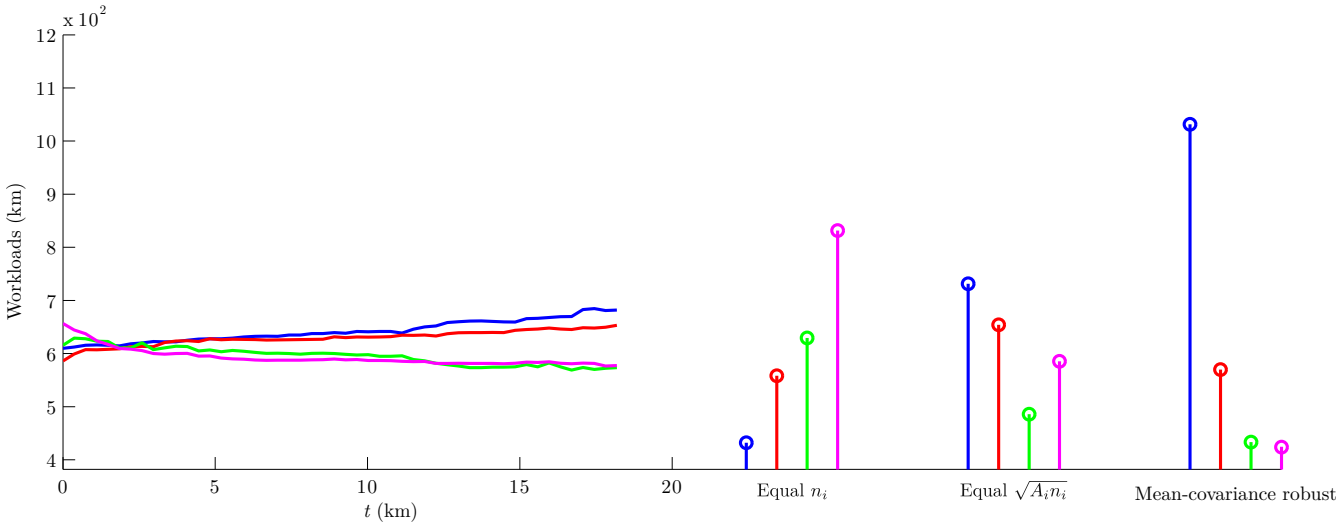


Figure 13: The workloads associated with the four districts and the four partitioning criteria from Figure 12, interpreted as follows: the colors of the various plots correspond to the workloads in the district of the same color (e.g. the magenta plots correspond to the magenta district, which belongs to Santa Clarita). The left set of plots corresponds to the workloads that result when we design districts according to the Wasserstein partitioning criterion: more precisely, for each value of t in the range shown, we construct a weight vector $\mathbf{w}^*$ such that the worst-case workloads in problem (16) are all equal. Thus, different values of t correspond to different values of $\mathbf{w}^*$, and thereby different partitions. The left-hand plot shows the true workloads for each of the districts as t varies; the maximum value of t of 18.3 km is calculated according to equation (18) with a 90% confidence level with $n = 50$ samples. The three sets of stem plots on the right show the workloads in each of the four districts that are obtained when one partitions according to the first three criteria of Section 5.2.2. The Wasserstein partitioning criterion consistently produces districts whose workloads are more balanced than those of the other three criteria, even though the Wasserstein partitions are constructed using a small number of samples, whereas the other three methods are actually permitted to make complete use of all 1704 sample points (the mean-covariance partitioning scheme uses exact knowledge of the mean and covariance of the data points). Surprisingly, we found that the the mean-covariance robust partitioning method is by far the *worst* of the three.

Figure 14: 1000 uniformly sampled points in Los Angeles County.

6 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 that 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 [51]. 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.

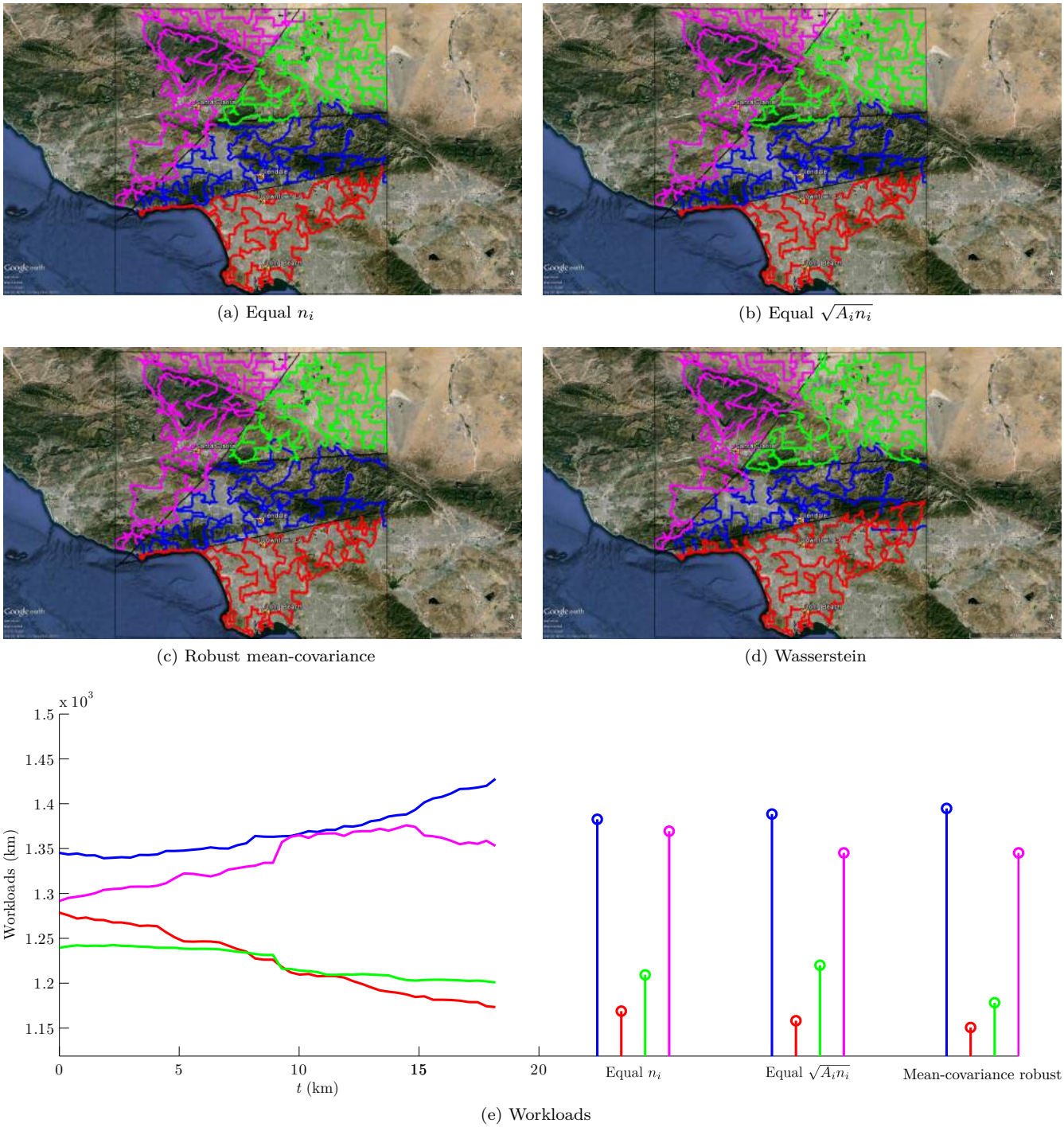
Acknowledgments

The authors thank the editor, three anonymous referees, and Kresimir Mihic and Alan Wood of Oracle Corporation.

References

[1] Caviar. `http://www.trycaviar.com/`. Accessed: 2014-09-27.

[2] DoorDash Food Delivery. `http://www.doordash.com`. Accessed: 2014-10-27.



(a) Equal n_i

(b) Equal $\sqrt{A_i n_i}$

(c) Robust mean-covariance

(d) Wasserstein

(e) Workloads

Figure 15: The power diagram districts obtained according to the four partitioning criteria, and the workloads that result therein, for the case where all samples are uniformly distributed. In this case, all four partitioning criteria perform more or less the same.

[3] Food Delivery & Restaurants Delivery - Order Food Online - BiteSquad.com. `http://www.bitesquad.com`. Accessed: 2014-10-27.

[4] Good Eggs. `http://www.goodeggs.com`. Accessed: 2014-10-27.

[5] 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.

[6] 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.

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

[8] D. Applegate, R. Bixby, V. Chvatal, and W. Cook. Concorde TSP solver, 2006.

[9] 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.

[10] D. L. Applegate, R. E. Bixby, V. Chvatal, and W. J. Cook. *The Traveling Salesman Problem: A Computational Study*. Princeton university press, 2011.

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

[12] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[13] 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.

[14] 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.

[15] 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.

[16] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *arXiv preprint arXiv:1401.0212*, 2013.

[17] 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.

[18] 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.

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

[20] S. Boyd. Subgradients. `http://stanford.edu/class/ee364b/lectures/subgradients_slides.pdf`, 2014.

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

[22] L. D. Burns, R. W. Hall, D. E. Blumenfeld, and C. F. Daganzo. Distribution strategies that minimize transportation and inventory costs. *Operations Research*, 33(3):469–490, 1985.

[23] C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. In *Geometric Science of Information*, pages 561–568. Springer, 2013.

[24] A. Caiti, V. Calabro, F. Di Corato, D. Meucci, and A. Munafo. Cooperative distributed algorithm for AUV teams: a minimum entropy approach. In *OCEANS-Bergen, 2013 MTS/IEEE*, pages 1–6. IEEE, 2013.

[25] G. C. Calafiore. Ambiguous risk measures and optimal robust portfolios. *SIAM Journal on Optimization*, 18(3):853–877, 2007.

[26] G. C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[27] J. F. Campbell. Location and allocation for distribution systems with transshipments and transportion economies of scale. *Annals of Operations Research*, 40(1):77–99, 1992.

[28] 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.

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

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

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

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

[33] X. Chen, M. Sim, and P. Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

[34] J. Cortés. Coverage optimization and spatial load balancing by robotic sensor networks. *Automatic Control, IEEE Transactions on*, 55(3):749–754, 2010.

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

[36] C. F. Daganzo. The distance traveled to visit n points with a maximum of c stops per vehicle: An analytic model and an application. *Transportation Science*, 18(4):331–350, 1984.

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

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

[39] J. W. Durham, R. Carli, P. Frasca, and F. Bullo. Discrete partitioning and coverage control for gossiping robots. *Robotics, IEEE Transactions on*, 28(2):364–378, 2012.

[40] E Erdoğan and G Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

[41] 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.

[42] L. Few. The shortest path and the shortest road through n points. *Mathematika*, 2:141–144, 1955.

[43] M. Flint, M. Polycarpou, and E. Fernandez-Gaucherand. Cooperative control for multiple autonomous uav's searching for targets. In *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, volume 3, pages 2823–2828. IEEE, 2002.

[44] L. C. Galvão, A. G. N. Novaes, J. E. Souza De Cursi, and J. C. Souza. A multiplicatively-weighted voronoi diagram approach to logistics districting. *Computers & Operations Research*, 33(1):93–114, 2006.

[45] N. Geroliminis, M. G. Karlaftis, and A. Skabardonis. A spatial queuing model for the emergency vehicle districting and location problem. *Transportation research part B: methodological*, 43(7):798–811, 2009.

[46] J. Geunes, Z.-J. M. Shen, and A. Emir. Planning and approximation models for delivery route based services with price-sensitive demands. *European journal of operational research*, 183(1):460–471, 2007.

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

[48] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917, 2010.

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

[50] M. Haimovich and T. L. Magnanti. Extremum properties of hexagonal partitioning and the uniform distribution in Euclidean location. *SIAM J. Discrete Math.*, 1:50–64, 1988.

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

[52] 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.

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

[54] M. Huang, K. R. Smilowitz, and B. Balcik. A continuous approximation approach for assessment routing in disaster relief. *Transportation Research Part B: Methodological*, 50:20–41, 2013.

[55] A. Irpino and R. Verde. A new wasserstein based distance for the hierarchical clustering of histogram symbolic data. In *Data science and classification*, pages 185–192. Springer, 2006.

[56] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[57] O. Jabali, M. Gendreau, and G. Laporte. A continuous approximation model for the fleet composition problem. *Transportation Research Part B: Methodological*, 46(10):1591–1606, 2012.

[58] D. Klabjan, D. Simchi-Levi, and M. Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.

[59] J. Kytöjoki, T. Nuortio, O. Bräysy, and M. Gendreau. An efficient variable neighborhood search heuristic for very large scale vehicle routing problems. *Computers & Operations Research*, 34(9):2743–2757, 2007.

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

[61] 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.

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

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

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

[65] G. F. Newell and C. F. Daganzo. Design of multiple-vehicle delivery tours - I a ring-radial network. *Transportation Research Part B: Methodological*, 20(5):345–363, 1986.

[66] A. G. N. Novaes and O. D. Graciolli. Designing multi-vehicle delivery tours in a grid-cell format. *European Journal of Operational Research*, 119(3):613–634, 1999.

[67] The Omega Group. Crime Mapping – Building Safer Communities! `http://www.crimemapping.com/map.aspx?aid=3db8cf99-a73b-46d2-b218-bd24cf491577`, 2015.

[68] Y. Ouyang. Design of vehicle routing zones for large-scale distribution systems. *Transportation Research Part B: Methodological*, 41(10):1079–1093, 2007.

[69] E. Pampalk, A. Flexer, G. Widmer, et al. Improvements of audio-based music similarity and genre classificaton. In *ISMIR*, volume 5, pages 634–637. London, UK, 2005.

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

[71] 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.

[72] L. C. A. Pimenta, V. Kumar, R. C. Mesquita, and G. A. S. Pereira. Sensing and coverage for a network of heterogeneous robots. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 3947–3952. IEEE, 2008.

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

[74] C. Redmond and J. E. Yukich. Limit theorems and rates of convergence for euclidean functionals. *The Annals of Applied Probability*, 4(4):pp. 1057–1073, 1994.

[75] W. J. Reilly. *The law of retail gravitation*. WJ Reilly, 1931.

[76] J.P. Rodrigue, C. Comtois, and B. Slack. *The Geography of Transport Systems*. Routledge, 2009.

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

[78] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[79] T. L. Snyder and J. M. Steele. Equidistribution in all dimensions of worst-case point sets for the traveling salesman problem. *SIAM Journal on Discrete Mathematics*, 8(4):678–683, 1995.

[80] J. M. Steele. Subadditive euclidean functionals and nonlinear growth in geometric probability. *The Annals of Probability*, 9(3):pp. 365–376, 1981.

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

[82] J. Q. Stewart. Demographic gravitation: evidence and applications. *Sociometry*, pages 31–58, 1948.

[83] 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.

[84] J. Tinbergen. Shaping the world economy; suggestions for an international economic policy. *Books (Jan Tinbergen)*, 1962.

[85] M. J. Todd. Note-solving the generalized market area problem. *Management Science*, 24(14):1549–1554, 1978.

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

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

[88] A. M. Voorhees. A general theory of traffic movement. In *Proceedings of the Institute of Traffic Engineers*, pages 46–50, 1955.

[89] A. Wade and S. Salhi. An ant system algorithm for the mixed vehicle routing problem with backhauls. In *Metaheuristics: computer decision-making*, pages 699–719. Springer, 2004.

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

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

[92] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.

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 (2), 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 (2) 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 (2), 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 (2) 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 $\lambda_i - \lambda_j > \text{diam}(\mathcal{R})$, 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 5

Purely for ease of exposition, we assume that $\mathcal{R}$ is the unit square. Section 2.1 of [28] 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 (5) 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 (5) – 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

$$\begin{aligned} \underset{f}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\, dA \qquad & s.t. \\ \mathcal{D}(f, \bar{f}) &\leq 2t_n \\ \iint_{R} f(x)\, dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,; \end{aligned} \tag{20}$$

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 (20) 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 [47] 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 (20) given by

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

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

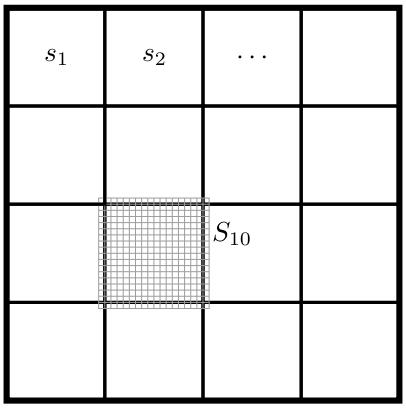


Figure 16: A division of the unit square $\mathcal{R}$ into $N^2 = 16$ grid cells. The larger square S_{10} has side length $1/N + 2/N^3$and contains s_{10}.

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 16). Define $m_i = N^2 \iint_{S_i} \bar{f}(x)\,dA$ for each m_i and consider the relaxation of (21) given by

$$\begin{aligned} \underset{f}{\text{maximize}} \iint_{\mathcal{R}} \sqrt{f(x)}\,dA \qquad & s.t. \\ \iint_{s_i} f(x)\,dA &\leq \frac{m_i}{N^2} + \epsilon \quad \forall i \\ \iint_{R} f(x)\,dA &= 1 \\ f(x) &\geq 0 \quad \forall x \in \mathcal{R}\,. \end{aligned} \tag{22}$$

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 (22) 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 $c = s\sqrt{n}$ is a standard and useful technique that can be seen in Section 4.2 of [35] or the paper [36]. By exchanging the expectation and $\max\{\cdot,\cdot\}$ operators, we can express the bound (12) 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.