Mathematical Programming 53 (1992) 127-146
North-Holland

Euclidean semi-matchings of random samples

J. Michael Steele
Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Received 19 August 1988
Revised manuscript received 8 February 1990

A linear programming relaxation of the minimal matching problem is studied for graphs with edge weights determined by the distances between points in a Euclidean space. The relaxed problem has a simple geometric interpretation that suggests the name minimal semi-matching. The main result is the determination of the asymptotic behavior of the length of the minimal semi-matching. It is analogous to the theorem of Beardwood, Halton and Hammersley (1959) on the asymptotic behavior of the traveling salesman problem. Associated results on the length of non-random Euclidean semi-matchings and large deviation inequalities for random semi-matchings are also given.

AMS 1980 Subject Classifications: Primary 05C05; Secondary 60F15.

Key words: Minimal matchings, subadditive processes, complete convergence, large deviations, asymptotic methods, relaxation methods.

1. Introduction

Let $G=(V, E)$ be a graph such that for each edge $e\in E$ there is an associated weight w_e. We are concerned here with the solutions to the following linear program:

$$z=\min_{x}\sum_{e\in E} x_e w_e \tag{1.1a}$$

$$\text{subject to} \quad \sum_{e \text{ meets } v} x_e=1 \quad \text{for all } v\in V, \tag{1.1b}$$

$$x_e\geqslant 0 \quad \text{for all } e\in E. \tag{1.1c}$$

If G is the complete bipartite graph $K_{n,n}$, then (1.1) becomes an assignment problem, and the well-known Integrality Theorem tells us that (1.1) has a solution such that each x_e is an integer, specifically $x_e=0$ or 1 for all e (see, e.g., Chvátal, 1983, p. 327; or Lovász and Plummer, 1986, p. 269).

A less well-known result of Balinski (1965) tells us that if G is the complete graph K_n, then there is a solution of (1.1) in semi-integers, i.e., $x_e=0$, $\frac{1}{2}$, or 1 for each $e\in E$. For this reason (and additional reasons to be reviewed shortly), the linear

Research supported in part by NSF Grant #DMS-8812868, ARO contract DAAL03-89-G-0092.P001, AFOSR-89-08301.A and NSA-MDA-904-89-2034.

program (1.1) will be called the semi-matching problem. A proof of Balinski's theorem is given in Lovász and Plummer (1986, p. 291), and the result is more subtle than one might guess. It is established with help from the classic result of Petersen on the decomposition of regular graphs into 2-factors. One can also find a discussion of the semi-integer nature of the solutions of (1.1), and an indication of proof of Balinski's theorem in Yemelichev et al. (1984, p. 175).

The intention of this paper is to investigate the behavior of the solutions to (1.1) where the vertices of G are points $v_1, v_2, \ldots, v_n$ in $\mathbb{R}^d$ and where the weight w_e associated with edge $e=(v_i, v_j)$ is the Euclidean distance $|v_i - v_j|$. For the main results, we further assume that the points v_i, $1 \leq i \leq n$, are determined by a probability model. In particular, we treat the case where the points are modeled as independent identically distributed random vectors in $\mathbb{R}^d$. In that situation, we are especially concerned with the growth rate of the value of the objective function as n becomes large.

Two branches of optimization theory motivate this investigation. First, the semi-matching problem illustrates the interplay between linear programming and combinatorial optimization. The classic paper of Edmonds (1965a) provided one of the earliest applications of linear programming to combinatorial optimization by showing that (1.1) can be supplemented with additional linear constraints to obtain an LP formulation for the general problem of minimal matching in a weighted graph. It turns out that an exponential number of additional constraints are required, but by appropriate sequential generation of the constraints, an optimal solution can be determined in polynomial time. Intriguingly, Edmonds (1965a) was among the first works to draw attention to the fundamental distinction between polynomial time and exponential time algorithms (see also Edmonds, 1965b, 1970).

The second line of development motivating the investigation of semi-matchings hinges on a geometrical interpretation. Since the value of x_e, here called the *loading factor* of e, can only be $0, \frac{1}{2}$, or 1 in a minimal solution to (1.1), one can easily show that any minimal solution consists of a union of isolated edges with loading 1 and a collection of odd cycles that has all edge loading equal to $\frac{1}{2}$. This interpretation of the LP (1.1) shows that the solutions of (1.1) are close to the usual notion of minimal matching. Still, there are some significant theoretical differences between semi-matching and ordinary matchings. One modest difference that nevertheless makes semi-matchings theoretically more attractive than minimum weight perfect matching is that there is no need to distinguish between odd and even values of n.

The geometric interpretation of semi-matchings also connects them with the theory of subadditive Euclidean functionals. That theory begins with work on the asymptotic behavior of the traveling salesman problem by Beardwood et al. (1959), and its importance for combinatorial optimization is made clear in the work of Karp (1977) that provides a polynomial time probabilistic algorithm for the TSP.

The main result of this paper is the following asymptotic result for S_n that is analogous to the result obtained for the traveling salesman problem by Beardwood et al. (1959).

Theorem 1. *Suppose that V_i, $1 \leq i < \infty$, are independent random variables with the uniform distribution on $[0,1]^d$, and let $S_n = S(V_1, V_2, \ldots, V_n)$ denote the cost of a minimal semi-matching of $\{V_1, V_2, \ldots, V_n\}$. For any $d \geq 2$, there is a constant $c_d > 0$ such that*

$$\lim_{n\to\infty} S_n / n^{(d-1)/d} = c_d \tag{1.2}$$

with probability one.

For $d = 2$ the analogue of Theorem 1 for minimal matchings was first given in Papadimitriou (1977). After developing some basic geometry of semi-matchings, an abstract version of Theorem 1 will be stated in Section 3, and at that point we will consider how the framework of Papadimitriou (1977) differs from the present approach.

The minimal semi-matching functional S is also well-behaved enough to permit the extension of Theorem 1 to more general distributions. In the following result, the constant c_d is the same as that given in the limit (1.2).

Theorem 2. *If V_i, $1 \leq i < \infty$, are independent random variables with distribution μ with compact support and absolutely continuous part $\mathrm{d}\mu_a = f(x)\,\mathrm{d}x$ then with probability one,*

$$\lim_{n\to\infty} S(V_1, V_2, \ldots, V_n)/n^{(d-1)/d} = c_d \int_{\mathbb{R}^d} f(x)^{(d-1)/d}\,\mathrm{d}x. \tag{1.3}$$

The approximation process by which one goes from the asymptotic theory of the uniform distribution to the general distributions of Theorem 2 is technical, and since the required approximation techniques are remote from the potential applications of the result, the proof of Theorem 2 will only be sketched.

Given the current state of development of the theory of subadditive Euclidean functionals, Theorems 1 and 2 rest substantially on non-probabilistic insights, and thus much of the work of this paper is non-probabilistic. In particular, the next section provides three geometric inequalities that serve to articulate the structural parallels between semi-matching and the traveling salesman problem (TSP). In fact, once one makes explicit this underlying structural communality, the techniques of the proof of Theorem 1 are seen to be as pertinent to the theory of the TSP as to the theory of semi-matchings. In this respect it is worth noting that the proof given here contains several twists that contribute to the general theory of subadditive Euclidean functionals. These features of the proof are pointed out as the details are given.

2. Geometry of semi-matchings

The two geometric properties that contribute the most to the asymptotic analysis of the minimal Euclidean semi-matching are the following: (1) the subadditivity expressed in Lemma 1, and (2) the smoothness property expressed in Lemma 2 that says as vertices are added or deleted the cost of a minimal semi-matching changes only modestly.

We let $Q_0=[0,1]^d$ denote the unit cube in $\mathbb{R}^d$, and we divide this cube into m^d congruent subcubes Q_i, $1\leq i\leq m^d$. If $v_1, v_2, \ldots, v_n$ is a fixed set of points in Q, we write $S(Q_i)$ for the cost of the minimal semi-matching of the points in $\{v_1, v_2, \ldots, v_n\}\cap Q_i$. The first lemma follows from the observation that the union over $1\leq i\leq m^d$ of any feasible semi-matchings of $Q_i\cap\{v_1, v_2, \ldots, v_n\}$ provides a feasible semi-matching of $Q_0\cap\{v_1, v_2, \ldots, v_n\}$.

Lemma 1. *For any n points* $\{v_1, v_2, \ldots, v_n\}$ *and any integer m, we have*

$$S(Q_0)\leq\sum_{i=1}^{m^d} S(Q_i). \qquad \square \tag{2.1}$$

The simplicity of this inequality should not hide its power. One can extract considerable asymptotic information from (2.1) when it is combined with homogeneity and translation invariance.

To make use of probabilistic information on the distribution of random points it is also important to know how S changes as points are added or deleted. The next lemma provides this information and thus provides the required surrogate for the monotonicity that one has for subadditive functionals like the TSP.

Lemma 2. *For any* $n\geq 2$, *we have*

$$|S(v_1, v_2, \ldots, v_n)-S(v_1, v_2, \ldots, v_{n+1})|\leq d_2(v_{n+1}; v_1, v_2, \ldots, v_n) \tag{2.2}$$

where

$$\begin{aligned} &\mathrm{d}_2(v; y_1, y_2, \ldots, y_n)\\ &\quad=\min\{r: |v-y_i|\leq r \text{ and } |v-y_j|\leq r \text{ for some } 1\leq i<j\leq n\}. \end{aligned}$$

Proof. We first show $S(v_1, v_2, \ldots, v_{n+1})$ cannot be much bigger than $S(v_1, v_2, \ldots, v_n)$. A key role in the proof is played by the nearest neighbors of v_{n+1}, and z will denote any such neighbor.

In a minimum weight semi-matching of $\{v_1, v_2, \ldots, v_n\}$, vertex z can occur on an edge $e=(z, z')$ with loading 1, or z can occur on an odd cycle with neighbors z' and z'' such that each of the edges (z, z') and (z, z'') each have loading $\frac{1}{2}$. If (z, z') has loading 1 we build feasible loading of $\{v_1, v_2, \ldots, v_{n+1}\}$ by the following assignments:

$$\begin{aligned} e_1&=(z, v_{n+1}), & x_{e_1}&=\tfrac{1}{2},\\ e_2&=(z', v_{n+1}), & x_{e_2}&=\tfrac{1}{2},\\ e_3&=(z, z'), & x_{e_3}&=\tfrac{1}{2}. \end{aligned}$$

This new loading spans $\{v_1, v_2, \ldots, v_{n+1}\}$, and it exceeds the cost of the minimal loading of $\{v_1, v_2, \ldots, v_n\}$ by, at most,

$$\begin{aligned}\Delta_1 &= \tfrac{1}{2}|v_{n+1} - z| + \tfrac{1}{2}|v_{n+1} - z'| - \tfrac{1}{2}|z - z'| \\ &\leq \tfrac{1}{2}|v_{n+1} - z| + \tfrac{1}{2}\{|v_{n+1} - z| + |z - z'|\} - \tfrac{1}{2}|z - z'| = |v_{n+1} - z|.\end{aligned} \tag{2.3}$$

In the case that z occurs on a triangle (z, z'), (z, z''), (z', z'') we change the minimal loading of $\{v_1, v_2, \ldots, v_n\}$ by setting

$$\begin{aligned} e_1 &= (z, v_{n+1}), & x_{e_1} &= 1, \\ e_2 &= (z', z''), & x_{e_2} &= 1, \\ e_3 &= (z, z'), & x_{e_3} &= 0, \\ e_4 &= (z, z''), & x_{e_4} &= 0. \end{aligned}$$

This new loading increases the cost of the minimal loading of $\{v_1, \ldots, v_n\}$ by, at most,

$$\Delta_2 = |v_{n+1} - z| + \tfrac{1}{2}|z' - z''| - \tfrac{1}{2}|z - z'| - \tfrac{1}{2}|z - z''| \leq |v_{n+1} - z|. \tag{2.4}$$

The third case to consider is that z occurs on an odd cycle containing at least five vertices. If z' and z'' are the neighbors of z on this cycle we change the loading to the following:

$$\begin{aligned} e_1 &= (z, v_{n+1}), & x_{e_1} &= 1, \\ e_2 &= (z, z'), & x_{e_2} &= 0, \\ e_3 &= (z, z''), & x_{e_3} &= 0, \\ e_4 &= (z', z''), & x_{e_4} &= \tfrac{1}{2}. \end{aligned}$$

This again provides feasible loading and, even though the presence of an even cycle marks it as suboptimal, we can check that it is good enough by the following bound on the incremental cost:

$$\Delta_3 = |v_{n+1} - z| - \tfrac{1}{2}|z - z'| - \tfrac{1}{2}|z - z''| + \tfrac{1}{2}|z' - z''| \leq |v_{n+1} - z|. \tag{2.5}$$

Summarizing (2.3), (2.4) and (2.5) we have

$$S(v_1, v_2, \ldots, v_{n+1}) \leq S(v_1, v_2, \ldots, v_n) + \min_{1 \leq i \leq n} |v_i - v_{n+1}|, \tag{2.6}$$

which is a bit sharper than we need to prove the first half of inequality (2.2).

We now need to obtain an upper bound on $S(v_1, v_2, \ldots, v_n)$ in terms of $S(v_1, v_2, \ldots, v_{n+1})$. The simplest case occurs when v_{n+1} is on an odd cycle in a minimal semi-matching of $\{v_1, v_2, \ldots, v_{n+1}\}$. If the neighbors of v_{n+1} are z and z' and the cycle is a triangle, we just give (z, z') a loading factor of one to get our feasible matching of $\{v_1, v_2, \ldots, v_n\}$. Furthermore, if v_{n+1} is on a cycle of cardinality five or greater, we just give (z, z') a loading of $\frac{1}{2}$. In either of these cases, we see there is a feasible semi-matching for $\{v_1, v_2, \ldots, v_n\}$ that has cost bounded by $S(v_1, v_2, \ldots, v_{n+1})$.

We now confront the trickiest case. Suppose v_{n+1} is on an edge (v_{n+1}, z) with loading factor equal to 1. We let w be a vertex other than z that is within a distance d_2 of v_{n+1}. By the definition of d_2, at least one such w must exist.

We can either have w as a vertex in an odd cycle or as a vertex on an isolated edge. If w is on a single edge (w, w') with loading 1, we consider the new loading factors given as follows:

$$e_1 = (z, w), \qquad w_{e_1} = \tfrac{1}{2},$$
$$e_2 = (w, w'), \qquad w_{e_2} = \tfrac{1}{2},$$
$$e_3 = (z, w'), \qquad w_{e_3} = \tfrac{1}{2}.$$

From this loading allocation, we have

$$S(v_1, v_2, \ldots, v_n) \leqslant S(v_1, v_2, \ldots, v_{n+1}) - |v_{n+1} - z| + \tfrac{1}{2}|z - w| + \tfrac{1}{2}|z - w'| - \tfrac{1}{2}|w - w'|, \tag{2.7}$$

and since

$$|z - w| + |z - w'| \leqslant 2|z - w| + |w - w'| \quad \text{and} \quad |z - w| \leqslant |v_{n+1} - z| + |v_{n+1} - w|,$$

we have

$$S(v_1, v_2, \ldots, v_n) \leqslant S(v_1, v_2, \ldots, v_{n+1}) - |v_{n+1} - z| + |z - w| \leqslant S(v_1, v_2, \ldots, v_{n+1}) + |v_{n+1} - w|. \tag{2.8}$$

Next we consider the case that w is on an odd cycle with neighbors (w', w''). If the cycle is a triangle we take new loadings

$$e_1 = (z, w), \qquad w_{e_1} = 1,$$
$$e_2 = (w', w''), \qquad w_{e_2} = 1,$$

and if the cycle is of size five or greater we take the loadings

$$e_1 = (z, w), \qquad w_{e_1} = 1,$$
$$e_2 = (w', w''), \qquad w_{e_2} = \tfrac{1}{2}.$$

In either of these two situations we find that $S(v_1, v_2, \ldots, v_{n+1}) - S(v_1, v_2, \ldots, v_n)$ is bounded by

$$\Delta_4 = -|v_{n+1} - z| + |z - w| - \tfrac{1}{2}|w - w'| - \tfrac{1}{2}|w - w''| + \tfrac{1}{2}|w' - w''|.$$

By the triangle inequality, the sum of the last three terms is bounded by 0 and $|z - w|$ is bounded by $|z - v_{n+1}| + |v_{n+1} - w|$. Thus, we find the bound,

$$\Delta_4 \leqslant |v_{n+1} - w| \leqslant d_2(v_{n+1}; v_1, v_2, \ldots, v_n). \tag{2.9}$$

From inequalities (2.1), (2.4), (2.8) and (2.9), the proof of Lemma 2 is complete. □

Lemma 3. *For any $n \geq 2$, we have*

$$S(v_1, v_2, \dots, v_n) \leq 8d^{1/2} n^{(d-1)/d}. \tag{2.10}$$

Proof. If $m^d < n/2 \leq (m+1)^d$, then one of the cubes Q_i of the decomposition of $[0,1]^d$ into m^d cells of side m^{-1} must contain at least three points. Since $m^{-1}d^{1/2}$ is the diameter of Q_i we see that there is thus a $1 \leq j \leq n$ such that $d_2(v_j; v_1, v_2, \dots, v_{j-1}, v_{j+1}, \dots, v_n) \leq m^{-1}d^{1/2}$. Hence, by Lemma 2, and the bound $m^{-1} \leq 2(m+1)^{-1} \leq 2(n/2)^{-1/d}$, we have

$$\begin{aligned} &S(v_1, v_2, \dots, v_n) \\ &\quad \leq S(v_1, v_2, \dots, v_{j-1}, v_{j+1}, \dots, v_n) + 2^{1+1/d} d^{1/2} n^{-1/d}. \end{aligned} \tag{2.11}$$

Applying the same argument to the $n-1$ set $\{v_1, v_2, \dots, v_{j-1}, v_{j+1}, \dots, v_n\}$ and continuing successively, we find

$$\begin{aligned} &S(v_1, v_2, \dots, v_n) \\ &\quad \leq 2^{1+1/d} d^{1/2} \sum_{n=1}^{n} j^{-1/d} \leq 2^{(d+1)/d} d^{3/2} (d-1)^{-1} n^{(d-1)/d}. \end{aligned} \tag{2.12}$$

Since $2^{(d+1)/d} d/(d-1) \leq 8$ for $d \geq 2$, (2.12) is stronger than the required bound. □

3. Underlying structure

As we have already noted, a close connection exists between the probability theory of semi-matchings and the traveling salesman problem. This connection can be made most explicit by laying out some of the abstract properties of the semi-matching functional that connect semi-matchings with the TSP and other subadditive Euclidean functionals. Very few properties of S are needed to show that the conclusion of Theorem 1 is valid. Explicitly, Theorem 1 will be proved by appeal to only the following:

(A1) $S(\alpha x_1, \alpha x_2, \dots, \alpha x_n) = \alpha S(x_1, x_2, \dots, x_n)$ for all $\alpha > 0$ and $x_i \in \mathbb{R}^d$.
(A2) $S(x_1 + y_1, x_2 + y, \dots, x_n + y) = S(x_1, x_2, \dots, x_n)$ for all y and x_i in $\mathbb{R}^d$.
(A3) $|S(x_1, x_2, \dots, x_n, x_{n+1}) - S(x_1, x_2, \dots, x_n)| \leq d_2(x_{n+1}; x_1, x_2, \dots, x_n)$ for all $n \geq 3$ and x_i in $\mathbb{R}^d$.
(A4) $S(\{x_1, x_2, \dots, x_n\} \cap [0,1]^d) \leq \sum_{i=1}^{m^d} S(\{x_1, x_2, \dots, x_n\} \cap Q_i)$.

If one could replace (A3) by the stronger monotonicity condition $S(x_1, x_2, \dots, x_{n+1}) \geq S(x_1, x_2, \dots, x_n)$, then by results of Steele (1981a) conditions

(A1) through (A4) would be enough to guarantee Theorem 1. Since one can easily provide examples that show S is not monotone, we are compelled to study semi-matchings through surrogate inequalities like (A3). In contrast to its deficient monotonicity, semi-matchings have rather stronger subadditive properties than one typically finds. In the general theory of subadditive Euclidean functionals, one relaxes (A4) to

$$S(\{x_1, x_2, \ldots, x_n\} \cap [0, t]^d) \leq \sum_{i=1}^{m^d} S(\{x_1, x_2, \ldots, x_n\} \cap tQ_i) + Ctm^{d-1}, \tag{3.1}$$

and this relaxation is surely worthwhile since without it the theory would not include the TSP.

We should also note that in Papadimitriou's formulation of the limit theory for the minimal matching $M(x_1, x_2, \ldots, x_n)$ of $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^2$, a key role is played by a lower bound that complements (A4) or (3.1). Specifically, the proof sketched in Papadimitriou (1977, p. 369) called upon the following property of minimal matchings in $\mathbb{R}^2$:

$$M(\{x_1, x_2, \ldots, x_n\} \cap [0, 1]^2) \geq \sum_{i=1}^{m^2} \{M(\{x_1, x_2, \ldots, x_n\} \cap Q_i) - 2|\partial Q_i|\}. \tag{3.2}$$

Here $|\partial Q_i|$ denotes the length of the boundary of the square Q_i. One of the issues that arises when one considers matchings or semi-matchings in $\mathbb{R}^d$, $d > 2$, is the determination of a suitable replacement for the $|\partial Q_i|$ terms. Because of this difficulty, an approach based on (A1)-(A4) lead to a more straightforward theory that is still well suited for applications. The main drawback of the axiomatization (A1)-(A4) is that one can no longer appeal directly to the proof of Beardwood et al. (1959), although, because of the considerable complexity of that proof, perhaps this drawback is really a blessing.

The bottom line is that most of the difficulties that arise in the limit theory for semi-matchings come from its failure to be monotone. Moreover, the general theory of subadditive Euclidean functionals benefits by seeing how this difficulty can be overcome.

4. Analysis of the expected value

The proof of Theorem 1 is perhaps most easily explained by dealing separately with the asymptotic behavior of ES_n and the behavior of $S_n - ES_n$. The first step in the analysis of ES_n is the introduction of a Poissonization device that permits the expeditious use of (A1), (A2) and (A4). The second step is to extract the asymptotics of ES_n from the asymptotics of the Poissonized sequence. In this step our analysis leans heavily on (A3) and the fact that the Poisson distribution concentrates almost all of its mass in an interval $[\lambda-\lambda^{\gamma}, \lambda+\lambda^{\gamma}]$ if $\gamma>\frac{1}{2}$.

If we let N be a Poisson random variable with mean λ, and we choose N points in $[0,1]^d$ independently according to the uniform distribution, then each of the m^d subcubes Q_i contains N_i points, where the N_i are independent Poisson random variables with mean λ/m^d. If we now set

$$\phi(\lambda)=ES(X_1, X_2, \ldots, X_N) \tag{4.1}$$

where the X_i are independent and uniformly distributed, then from (2.1) we see that

$$\phi(\lambda)\leq m^{d-1}\phi(\lambda/m^d).$$

After substituting $\lambda_m{}^d$ for λ and dividing by $m^{d-1}\lambda^{\alpha}$ with $\alpha=(d-1)/d$, we have

$$\phi(m^d\lambda)/(m^d\lambda)^{\alpha}\leq\phi(\lambda)/\lambda^{\alpha}, \quad m\geq 1. \tag{4.2}$$

Now, given any $\varepsilon>0$ and $k>0$ we can choose an interval $(a,b)=(a(\varepsilon,k), b(\varepsilon,k))$ such that for all $\lambda\in(a,b)$ we have

$$\phi(\lambda)/\lambda^{\alpha}\leq\varepsilon+\inf_{u\geq k}\phi(u)/u^{\alpha}. \tag{4.3}$$

Applying (4.2) to (4.3), we find

$$\phi(m^d\lambda)/(m\lambda)^{\alpha}\leq\varepsilon+\inf_{u\geq k}\phi(u)/u^{\alpha} \tag{4.4}$$

for all m and $\lambda\in(a,b)$, hence

$$\phi(x)/x^{\alpha}\leq\varepsilon+\inf_{\lambda\geq k}\phi(u)/u^{\alpha} \tag{4.5}$$

for all $x\in\bigcup_{m=1}^{\infty}(m^d a, m^d b)=U$. For $m\geq a^{1/d}/(b^{1/d}-a^{1/d})$ the intervals $(m^d a, m^d b)$ and $((m+1)^d a, (m+1)^d b)$ overlap, so U contains $(m_0^d a, \infty)$ where m_0 is the least integer as great as $a^{1/d}/(b^{1/d}-a^{1/d})$. We can thus conclude from (4.5) that

$$\limsup_{\lambda\to\infty}\phi(\lambda)/\lambda^{\alpha}\leq\liminf_{\lambda\to\infty}\phi(\lambda)/\lambda^{\alpha}+\varepsilon. \tag{4.6}$$

Since (4.6) holds for all $\varepsilon>0$, the limit of $\phi(\lambda)/\lambda^{\alpha}$ exists as $\lambda\to\infty$.

Now, if we let

$$s_n=ES_n=ES(X_1, X_2, \ldots, X_n), \tag{4.7}$$

the asymptotics we have established for $\phi(\lambda)$ can be used to obtain the asymptotics of s_n. If we expand $\phi(\lambda)$ by conditioning on N, we find

$$\phi(\lambda)=\sum_{n=0}^{\infty}E(S(X_1, X_2, \ldots, X_n))P(N=n)=e^{-\lambda}\sum_{n=0}^{\infty}s_n\lambda^n/n!, \tag{4.8}$$

so we have

$$\phi(\lambda)=e^{-\lambda}\sum_{n=0}^{\infty}s_n\lambda^n/n!\sim c\lambda^{\alpha}. \tag{4.9}$$

The extraction of the asymptotics of s_n from that of $\phi(\lambda)$ can be achieved by appeal to classical Tauberian theorems, or by completely elementary means. We will pursue the latter route, but first we need to verify that s_n is reasonably smooth. We use $\|Y\|_p = (EY^p)^{1/p}$ to denote the L^p norm of Y.

Lemma 4. *There is a constant α_d depending only on d such that for independent, uniformly distributed random variables X_i, the variables $Y_n = d_2(X_{n+1}; X_1, X_2, \ldots, X_n)$ satisfy*

$$\|Y_n\|_p \leq \alpha_d (p/n)^{1/d} \tag{4.10}$$

for all $n \geq 1$. Consequently, we also have for $-\frac{1}{2}n \leq h \leq \frac{1}{2}n$,

$$|s_n - s_{n+h}| \leq \alpha_d 2^{1/d} n^{-1/d} |h|. \tag{4.11}$$

Proof. We first bound $P(Y_n \geq y)$ by conditioning on X_{n+1} and applying geometric considerations. For any $0 < y < 1$ and for all $x \in [0, 1]^d = Q$, we have the elementary geometric bound

$$P(|X_i - x| \geq y) \leq 1 - 2^{-d} \omega_d y^d, \tag{4.12}$$

where ω_d is the volume of the unit sphere in $\mathbb{R}^d$. To see why the factor 2^{-d} is needed, just note that (4.12) becomes an equality if x is one of the corners of the cube.

Since $\{Y_n \geq y\}$ if and only if at most one of the events $A_i = \{|X_i - x| < y\}$, $1 \leq i \leq n$, occurs, we have for $0 < y < 1$ that

$$P(Y_n \geq y) \leq (1 - \omega_d 2^{-d} y^d)^n + n\omega_d y^d (1 - 2^{-d} \omega_d y^d)^{n-1} \equiv \phi_n(y), \tag{4.13}$$

and, for $1 \leq y \leq d^{1/2}$ we have $P(Y_n \geq y)$ bounded by $\phi_n(1)$. Naturally, $P(Y_n \geq y) = 0$ if $y \geq d^{1/2}$.

If we multiply these bounds by py^{p-1}, apply the bound $(1 - x) \leq e^{-x}$, and integrate using $\int_0^\infty y^{\alpha-1} e^{-\beta y^d} \, dy = d^{-1} \beta^{-\alpha/d} \Gamma(\alpha/d)$ we can complete the proof of (4.10) by applying the bound $\Gamma(x) \leq x^x$.

To prove (4.11), we only have to note by (4.10) and the restrictions on h that

$$|s_n - s_{n+h}| \leq \alpha_d \sum_{n-|h| \leq k < n} 2^{1/d} k^{-1/d} \leq \alpha_d 2^{1/d} n^{-1/d} |h|. \qquad \square$$

The preliminaries are out of the way, and we can now establish the main result of this section.

Lemma 5. *There is a constant $c_d > 0$, such that as $n \to \infty$,*

$$s_n \sim c_d n^{(d-1)/d}. \tag{4.14}$$

Proof. We first recall a classical bound on the tail of the Poisson distribution (see, e.g., Hardy, 1949, p. 170); if $\frac{1}{2} < \gamma < \frac{2}{3}$ then, as $\lambda \to \infty$,

$$\sum_{k:|k-\lambda|>\lambda^{\gamma}} e^{-\lambda}\lambda^{k}/k! = O(\exp(-\lambda^{\eta})) \tag{4.15}$$

for any $\eta < 2\gamma - 1$. We also note that (4.15) further implies that for each fixed $\beta > 0$,

$$\sum_{k:|k-\lambda|>\lambda^{\gamma}} k^{\beta} e^{-\lambda}\lambda^{k}/k! = O(\exp(-\lambda^{\eta}/2)). \tag{4.16}$$

For easy checking, we will apply (4.15) and (4.16) with $\gamma = \frac{3}{5}$ and $\eta = \frac{1}{6}$. We let λ be any positive integer and use (4.11) together with the crude bound $s_k \leq d^{1/2}k$ to estimate s_λ in terms of $\phi(\lambda)$ as follows:

$$\begin{aligned}
\phi(\lambda) &= s_\lambda + \sum_{k=0}^{\infty} (s_k - s_\lambda)\, e^{-\lambda}\lambda^{k}/k! \\
&= s_\lambda + O\Big(\sum_{k:|k-\lambda|\geq\lambda^{3/5}} (\lambda + k)\, e^{-\lambda}\lambda^{k}/k! \Big) \\
&\quad + O\Big(\sum_{k:|k-\lambda|<\lambda^{3/5}} \lambda^{3/5}\lambda^{-1/d}\, e^{-\lambda}\lambda^{k}/k! \Big) \\
&= s_\lambda + O(\exp(-\lambda^{1/6}/2)) + O(\lambda^{3/5-1/d}).
\end{aligned} \tag{4.17}$$

Since $\phi(\lambda) = c\lambda^{(d-1)/d} + o(\lambda^{(d-1)/d})$ by (4.9), we thus have $s_\lambda = c\lambda^{(d-1)/d} + o(\lambda^{(d-1)/d})$ as well. The fact that $c_d > 0$ follows from the fact that by elementary geometric probability — such as applied in (4.12) — one has a $c > 0$ such that $E\, \min_{1<j\leq n}|X_1 - X_j| > cn^{-1/d}$. □

One finds a familiar pattern in the derivation of the asymptotics of $\phi(\lambda)$ by means of the subadditivity argument given in equations (4.2) through (4.7). On the other hand, the use of Hardy's bound on the Poisson tails has not been used before in the context of subadditive Euclidean functionals, and the introduction of this bound provides a simpler and more powerful approach than the Tauberian arguments used in Karp and Steele (1985) or Steele et al. (1987). From the perspective of the general theory, where there is some benefit to working only with one-sided bounds, the Tauberian arguments retain some benefits; but, where one has two-sided control of $s_n - s_{n+h}$ as given by Lemma 4, the line of argument given by (4.15)-(4.17) seems to be preferable.

5. Finishing the proof of Theorem 1

Now that we know $ES_n \sim c_d n^{(d-1)/d}$, the asymptotic behavior of the random variable S_n is essentially reduced to obtaining good bounds on the central moments $E(S_n - s_n)^k$, or the tail probabilities $P(|S_n - s_n| \geq t)$.

By modifying the methods used for the TSP in Steele (1981b), one can show by means of the inequality of Efron and Stein (1981) that $E(S_n - s_n)^2 = \text{Var } S_n$ is of order $O(n^{1-2/d})$. With such a bound on $\text{Var}(S_n)$, one can use interpolation and a Borel–Cantelli argument to prove Theorem 1.

Still, a cleaner proof of a more precise result can be obtained by using a martingale argument to bound the fourth central moments $E(S_n - s_n)^4$. The latter approach avoids the interpolation argument, and, more pointedly, it provides for so-called complete convergence, rather than just almost sure convergence (see, e.g, Stout, 1974, or Steele, 1981b, for a discussion of the benefits of complete convergence in the context of probabilistic algorithms and of its difference from almost sure convergence).

We recall that a sequence of random variables d_i, $1 \leq i \leq n$, is called a martingale difference sequence provided there is a sequence of sigma fields $\boldsymbol{F}_i$, $1 \leq i \leq n$, such that $\boldsymbol{F}_i \subset \boldsymbol{F}_{i+1}$, where d_i is measurable with respect to $\boldsymbol{F}_i$, and $E(d_i \mid \boldsymbol{F}_{i-1}) = 0$. The martingale differences we use are defined by

$$d_i = E(S_n \mid \boldsymbol{F}_i) - E(S_n \mid \boldsymbol{F}_{i-1}), \quad 1 \leq i \leq n, \tag{5.1}$$

where for $1 \leq i \leq n$ we take $\boldsymbol{F}_i$ to be the sigma field generated by $\{X_1, X_2, \ldots, X_i\}$, and where $\boldsymbol{F}_0$ is the trivial sigma field. Since $E(S_n \mid \boldsymbol{F}_n) = S_n$ and $E(S_n \mid \boldsymbol{F}_0) = s_n$, the telescoping sum of (5.1) lets us express $S_n - s_n$ in terms of the d_i,

$$S_n - s_n = \sum_{i=1}^{n} d_i. \tag{5.2}$$

Although it is not particularly easy to compute with the d_i, they can be made more tractable by introducing some new variables. We let S_n^i denote the length of the minimal semi-matching of the sample with the ith point left out, i.e., $S_n^i = S(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. One benefit of these variables is that $E(S_n^i \mid \boldsymbol{F}_i) = E(S_n^i \mid \boldsymbol{F}_{i-1})$, and thus we have

$$d_i = E(S_n - S_n^i \mid \boldsymbol{F}_i). \tag{5.3}$$

The critical point is that each expectation in (5.3) is well adapted to analysis by means of Lemmas 2 and 4. In fact, by (5.3) and Jensen's inequality, we have for each $p \geq 1$ that

$$|d_i|^p \leq E(|S_n - S_n^i|^p \mid \boldsymbol{F}_i),$$

so taking expectations and using (4.10) gives

$$\|d_i\|_p \leq 2\alpha_d (p/n)^{1/d}, \tag{5.4}$$

where α_d, the constant of Lemma 4, depends only on d.

Now, for any martingale difference sequence, Burkholder's square function inequality (Burkholder, 1973, or Chow and Teicher, 1978, p. 384) tells us for $p > 1$ that

$$\left\| \sum_{i=1}^{n} d_i \right\|_p \leq 18pq^{1/2} \left\| \left(\sum_{i=1}^{n} d_i^2 \right)^{1/2} \right\|_p$$

where $1/p+1/q=1$. For $p=4$, we find

$$E\left(\sum_{i=1}^{n} d_i\right)^4 \leqslant cE\left(\sum_{i=1}^{n} d_i^2\right)^2$$

where $c \leqslant 3^6 \cdot 2^{16}$. By expanding this last sum and applying Hölder's inequality, we find from (5.4) that

$$\begin{aligned} E\left(\sum_{i=1}^{n} d_i\right)^4 &\leqslant cE\sum_{i=1}^{n} d_i^4 + 2cE\sum_{1\leqslant i<j\leqslant n} d_i^2 d_j^2 \\ &\leqslant cE\sum_{i=1}^{n} d_i^4 + 2c\sum_{1\leqslant i<j\leqslant n} (Ed_i^4)^{1/2}(Ed_j^4)^{1/2} \\ &\leqslant K_d n^{2-4/d} \end{aligned} \tag{5.5}$$

for a constant K_d depending only on d. By Markov's inequality, we then have for any $\varepsilon > 0$ that

$$P(|S_n - s_n| \geqslant \varepsilon n^{(d-1)/d}) \leqslant \varepsilon^{-4} K_d n^{-2}. \tag{5.6}$$

Thus, the probabilities in (5.6) are summable for each $\varepsilon > 0$, and since $s_n \sim c_d n^{(d-1)/d}$, this summability is exactly the condition of complete convergence of $s_n^{-1} S_n$ to the constant c_d. Since the Borel-Cantelli lemma tells us complete convergence implies almost sure convergence, the proof of Theorem 1 is complete. □

6. Large deviations

In Section 5 we gave the quickest and cleanest route to the completion of the proof of Theorem 1. More was proved than was needed — complete convergence instead of almost sure convergence — but the stronger result was obtained largely because it came for free. The purpose of this section is to show how martingale theory can be used to obtain much stronger bounds on the tail probabilities of $S_n - s_n$ than we found in (5.6).

We continue to exploit the representation (5.2) and, in particular, it will be used together with a recent martingale inequality to prove the exponential tail bound:

$$P(|S_n - s_n| \geqslant t) \leqslant A\exp(-B_d t^{2d/(2+d)} n^{(2-d)/(2+d)}), \quad 1 \leqslant n < \infty,\ 0 < t < \infty, \tag{6.1}$$

where $A = e$ and B_d is a constant that depends only on the dimension d.

Inequality (6.1) is powerful in any dimension $d \geqslant 2$, but it is especially convenient for $d = 2$. In that case, (6.1) reduces to an ordinary exponential bound that is independent of n. Also, on multiplying (6.1) by pt^{p-1} and integrating, we find by standard bounds on $\Gamma(p)$ that when $d = 2$ all the central moments of S_n are bounded independently of n, i.e.,

$$\|S_n - s_n\|_p \leqslant \beta p, \tag{6.2}$$

where β is a universal constant.

We can also exploit (6.1) in another direction to get a rate result for Theorem 1. For example, when $d=2$ and we take $t=2\log n/B_d$, then (6.1) leads to a summable bound on the tail probabilities. By the Borel-Cantelli lemma, we then have a qualitative refinement of Theorem 1:

$$S_n = s_n + \mathrm{O}(\log n) \tag{6.3}$$

with probability one. One should note, however, that we cannot replace s_n with $c_2 n^{1/2}$ in (6.3), since we have no bound on $s_n - c_n n^{1/2}$ that is sharper than $o(n^{1/2})$.

We will now show that for any $d \geqslant 2$ inequality (6.1) follows easily from the L^p bound on d_i given in (5.4), and a general martingale inequality due to Rhee and Talagrand (1987). The Rhee–Talagrand inequality states that there is a constant K such that for any martingale difference sequence $\{d_i, 1 \leqslant i \leqslant n\}$, and any $t \geqslant 0$ we have

$$P\left(\left|\sum_{1\leqslant i\leqslant n} d_i\right| \geqslant t\right) \leqslant \exp(-\tfrac{1}{2}s), \tag{6.4}$$

provided $s \geqslant 2$ and $Ks(\sum_{i\leqslant n} \|d_i\|_{2s}^2) \leqslant t^2$.

When we apply our L^p bound (5.4) with $p=2s$ in (6.4), we find

$$\sum_{i\leqslant n} \|d_i\|_{2s}^2 \leqslant n 4\alpha_d^2 (2s/n)^{2/d}, \tag{6.5}$$

so we have $Ks(\sum_{i\leqslant n} \|d_i\|_{2s}^2) \leqslant t^2$, provided we take

$$s = K^{-d/(d+2)} \alpha_d^{-2d/(d+2)} 2^{-(2d+2)/(d+2)} n^{(2-d)/(d+2)} t^{2d/(d+2)}. \tag{6.6}$$

For $s \leqslant 2$ the bound on the right in (6.1) is at least one, so if we take $B_d = 2^{-3} K^{-d/(d+2)} \alpha_d^{-2d/(d+2)}$ we have (6.1) for all $d \geqslant 2$, $1 \leqslant n < \infty$ and $0 < t < \infty$.

The intention of this section is to give just a taste of the analytical possibilities that are opened up by martingale methods. For $d=2$, inequality (6.1) is natural and effective, but even then there are many opportunities for additional development. To see how such analyses have gone forward in the theory of the traveling salesman problem one should consult Rhee and Talagrand (1988), where a bound is given that shows for $d=2$ that the tails of the TSP functional are actually sub-Gaussian. It is not known if such bounds hold for $d>2$, but some heuristic considerations suggest that in higher dimensions one would no longer have such rapid decrease of the tail probabilities.

7. Extension to general distributions

The extension of the limit theory of S_n from the uniform case of Theorem 1 to the general case of Theorem 2 rests on the fact that there is a class of probability measures that is rich enough to approximate any probability measure with support in $[0, 1]^d$, yet constrained enough to permit a localized application of Theorem 1.

These are probability measures on $[0, 1]^d$ with the form $g(x)\,dx + d\mu_s$ where (1) $g(x) = \sum \alpha_i 1_{Q_i}$, for m^d congruent disjoint cubes Q_i with edges parallel to the axes, and (2) the measure μ_s is purely singular (i.e., $\mu_s([0, 1]^d) = \mu_s(A)$ for a measurable set A of Lebesgue measure zero). These probability measures — the so-called *blocked distributions* — were used by Beardwood et al. (1959); and since the beginning of the subject, they have remained a regular feature of extension arguments for subadditive Euclidean functionals.

It should be clear that if a functional S is smooth in an appropriate sense then one can carry asymptotic results for the class of blocked distributions to the class of probability measures with bounded support. The following result from Steele (1988) points out a simple continuity condition that suffices.

Theorem. *Suppose Z is a measurable real-valued function on the finite subsets of $[0, 1]^d$, and suppose that there is a constant K not depending on n such that Z satisfies the continuity condition*

$$|Z(x_1, x_2, \ldots, x_n) - Z(x'_1, x'_2, \ldots, x'_n)| \leq K|\{i: x_i \neq x'_i\}|^{(d-1)/d}. \tag{7.1}$$

Further, suppose that for every sequence of i.i.d. random variables $\{X_i\}_{1 \leq i < \infty}$ distributed with a a blocked distribution $\mu = \mu_s + g(x)\,dx$ we have with probability one that

$$Z(X_1, X_2, \ldots, X_n) \sim c_d n^{(d-1)/d} \int g(x)^{(d-1)/d}\,dx. \tag{7.2}$$

One then has that with probability one,

$$Z(X'_1, X'_2, \ldots, X'_n) \sim c_d n^{(d-1)/d} \int f(x)^{(d-1)/d}\,dx \tag{7.3}$$

whenever $\{X'_i\}$ are independent and identically distributed with respect to any probability measure on $[0, 1]^d$ with an absolutely continuous part given by $f(x)\,dx$. □

It is easy to justify (7.1) for the semi-matching functional. If $A = \{x_1, x_2, x_2, \ldots, x_n\}$ and $B = \{x'_1, x'_2, \ldots, x'_n\}$ then by (2.11) we have

$$\begin{aligned} |S(A) - S(A \cap B)| &\leq 2^{1+1/d} d^{1/2} \sum_{|A \cap B| < k \leq |A|} k^{-1/d} \\ &\leq 2^{1-1/d} d^{1/2} \sum_{1 \leq k \leq j} k^{-1/d}, \end{aligned} \tag{7.4}$$

where $j = |A| - |A \cap B|$. Applying the same considerations to $S(B) - S(A \cap B)$ and adding that bound to (7.4), we find for $j = |A \cup B| - |A \cap B| = |\{i: X_i \neq X'_i\}|$ that

$$|S(A) - S(B)| \leq 2^{2+1/d} d^{1/2} \sum_{1 \leq k \leq j} k^{-1/d}. \tag{7.5}$$

Finally, (7.1) follows from (7.5) by standard estimates.

To complete the sketch of Theorem 2 we need to indicate why (7.2) holds for blocked distributions. We first consider the purely singular case. We let A denote the support of μ_s, and let $\varepsilon > 0$ be given. We then divide $[0, 1]^d$ into m^d cells Q_i such that m^{-1} is smaller than ε, and a small percentage of the Q_i contain almost all of the mass of μ_s. Specifically, since A has Lebesgue measure zero, we can find an m and a subset $J \subset \{1, 2, \ldots, m^d\}$ such that

$$m^{-1} \leq \varepsilon, \tag{7.6a}$$

$$|J| \leq \varepsilon m^d, \tag{7.6b}$$

and

$$\mu_s\left(\left(\bigcup_{i \in J^c} Q_i\right)^c\right) \leq \varepsilon. \tag{7.6c}$$

Now, by the usual feasibility considerations we have

$$S(Q) \leq \sum_{i \in J} S(Q_i) + S\left(\bigcup_{i \notin J} q_i\right), \tag{7.7}$$

so we only need bounds on the respective sums. By (2.10) any k points in $[0, 1]^d$ have a semi-matching bounded by $8d^{1/2}k^{(d-1)/d}$; so by scaling we see that if k_i is the number of points of $\{X_j : 1 \leq j \leq n\}$ in Q_i then

$$|S(Q_i)| \leq 8d^{1/2}m^{-1}|k_i|^{(d-1)/d}. \tag{7.8}$$

By Hölder's inequality we find

$$\sum_{i \in J} S(Q_i) \leq |J|^{1/d} 8d^{1/2}m^{-1}\left(\sum_{i \in J} k_i\right)^{(d-1)/d} \leq \varepsilon^{1/d} 8d^{1/2} n^{(d-1)/d}, \tag{7.9}$$

and by the direct application of (2.10) we have

$$S\left(\bigcup_{i \notin J} Q_i\right) \leq 8d^{1/2}\left(\sum_{i \notin J} k_i\right)^{(d-1)/d}. \tag{7.10}$$

Since $\mu_s(\bigcup_{i \notin J} Q_i) \leq \varepsilon$, the strong law of large numbers applied to (7.10) combines with (7.7) and (7.9) to give us

$$\limsup_{n \to \infty} n^{-(d-1)/d} S(Q) \leq \varepsilon^{1/d} 8d^{1/2} + 8d^{1/2}\varepsilon^{(d-1)/d}, \tag{7.11}$$

with probability one. Since $\varepsilon > 0$ is arbitrary, we have proved Theorem 1 in the case that μ is purely singular, i.e., $\mu_s([0, 1]^d) = 1$.

Now we consider the purely absolutely continuous case, i.e., $\mu_s([0, 1]^d) = 0$. We let m be fixed and let Q_i, $1 \leq i \leq m^d$, be the usual partition. We consider the set E of edges e of $S(Q)$ such that e has endpoints in two different subcells, i.e., $e \in E$ if e crosses a boundary ∂Q_i for some i. This time, let k_i denote the number of points of Q_i that are endpoints of an edge in E. We claim that there is a K_d such that

$$S\left(\bigcup_{i=1}^{m^d} Q_i\right) \leq \sum_{i=1}^{m^d} S(Q_i) \leq S\left(\bigcup_{i=1}^{m^d} Q_i\right) + K_d \sum_{i=1}^{m^d} m^{-1} k_i^{(d-1)/d}. \tag{7.12}$$

The proof of (7.12) in complete detail would require repetition of some of the considerations of Section 2, but it essentially follows from the idea that one can build a semi-matching on each Q_i by taking the edges of the semi-matching of $S(\bigcup_{i=1}^{m^d} Q_i)$ that are completely interior to Q_i together with a set of at most k_i edges with length at most $m^{-1}d^{1/2}$, i.e., the diameter of Q_i.

To bound the sum of the k_i, we bound the cardinality of E. The key observation is that for each $e \in E$ we either have $|e| > x$, or else an endpoint of e is within x of some ∂Q_i. Thus, if $\nu_n(x)$ is the number of edges in a minimal semi-matching of $\{X_i : 1 \leq i \leq n\}$ that have length at least x, and $n_n(x)$ is the number of points of $\{X_i : 1 \leq i \leq n\}$ that are within x of ∂Q_i for some i, then

$$\sum_{i=1}^{m^d} k_i \leq 2\nu_n(x) + n_n(x). \tag{7.13}$$

Since $x\nu_n(x)$ is bounded by S_n, Lemma 3 and Hölder's inequality tell us that we have

$$\begin{aligned} m^{-1} \sum_{i=1}^{m^d} k_i^{(d-1)/d} &\leq \left(\sum_{i=1}^{m^d} k_i \right)^{(d-1)/d} \\ &\leq (2x^{-1} 8 d^{1/2} n^{(d-1)/d} + n_n(x))^{(d-1)/d}. \end{aligned} \tag{7.14}$$

When we set $x = n^{-1/d^2}$ a standard analysis of the right hand side of (7.14) shows it is $o(n^{(d-1)/d})$ with probability one.

Returning to (7.12), we find

$$S\left(\bigcup_{i=1}^{m^d} Q_i \right) - \sum_{i=1}^{m^d} S(Q_i) = o(n^{(d-1)/d})$$

with probability one. Since the asymptotic behavior of $S(Q_i)$ is determined by Theorem 1, we have completed the proof of Theorem 2 under the assumption of absolute continuity. The proof for mixed distributions follows from the consideration of the two pure cases and standard probabilistic arguments, so those details are safely omitted.

We have thus completed our sketch of the proof of (7.2). By the continuity theorem, the proof of Theorem 2 is complete. □

8. Concluding remarks

The theory developed here brings together two streams of thought that have their origin in the theory of the traveling salesman problem. First, it is widely recognized that one of the most powerful thrusts in contemporary combinatorial optimization is the algorithmic exploitation of the geometry of special polytopes. This point is forcefully made in Crowder and Padberg (1980), Grötschel (1981, 1982), and Lovász and Plummer (1986). Second, for many purposes — especially probabilistic algorithms and the probabilistic analysis of heuristics — there seems to be a sustained contribution to be made by the theory of subadditive Euclidean functionals that has grown out of the work of Beardwood et al. (1959), Karp (1977), Papdimitriou (1977) and Steele (1981a).

The semi-matching problem arises as the simplest part of Edmonds' early program to understand the matching polytope, and thus it has a special historical distinction. Its claim on our attention is further strengthened by the fruitfulness of Edmonds' conception as developed in the investigations of Padberg and Rao (1982), Grötschel (1977) and Grötschel and Holland (1985, 1987).

The semi-matching problem is also the first problem to be brought into the theory of subadditive Euclidean functionals that has its origin within the framework of linear programming. To be sure, the availability of a second geometrical interpretation made the development of the probability theory of semi-matching more natural than one would have otherwise expected. Still, the geometry of semi-matching is not so simple that its asymptotic theory can be obtained by off-the-shelf tools. The failure of the semi-matching functional to be monotone lead us to a natural extension of the theory of subadditive Euclidean functionals and, in fact, the continuity established in Lemma 4 turned out to be a condition that is in some respects even more effective than monotonicity.

Of the problems that aré left unresolved by the present analysis, the most obvious ones are the determination of c_d and the possibility of a central limit theorem. There is no subadditive Euclidean functional for which these issues have been resolved; and, despite the fact that semi-matching are in some ways very well behaved, there is little hope for quick progress on these problems.

A more promising line of development lies in the direction of general weight functions. The present analysis for $w_e = |x-y|$ can be extended to $w_e = |x-y|^\alpha$, for $0<\alpha<d$, but it does not seem easy to deal with $\alpha = d$. In that case, a natural conjecture is that with probability one we have

$$\lim_{n\to\infty} \min_{S} \sum_{e\in S} |e|^d = c > 0, \tag{8.1}$$

where the minimum is over all semi-matchings of $\{X_i, 1 \leq i \leq n\}$ and the X_i are independent and uniformly distributed on $[0, 1]^d$. The conjecture (8.1) is analogous to a conjecture of R. Bland on minimal spanning trees (MSTs). Bland's conjecture came about from empirical observation of simulation experiments. Its validity has been recently established by David Aldous and the author. The limit (8.1) still offers considerable technical challenge because semi-matchings cannot be obtained via a greedy algorithm, and it is that feature of the MST that proved to be essential in the resolution of Bland's conjecture.

A more conceptual shift is created when one changes to $w_e = d(x, y)$ where d is a non-Euclidean distance. Sharp geometrical facts still abound, and the investigation offers reasonable hope. Nevertheless, the subtlety of tilings that might serve in the same role as Q_i raises many new issues, and the development of results analogous to Theorem 1 seems to call for a new approach.

Acknowledgement

I would like to thank John Mitchell for telling me about the geometric nature of (1.1), and for providing the references to the work of Grötschel and Holland. I also thank Michel Goemans for comments on an earlier draft.

References

M.L. Balinski, "Integer programming: methods, uses, and computation," *Management Science* 12 (1965) 253–313.

J. Beardwood, J.H. Halton and J.M. Hammersley, "The shortest path through many points," *Cambridge Philosophical Society: Proceedings* 55 (1959) 299–327.

L. Burkholder, "Distribution function inequalities for martingales: The 1971 Wald Memorial Lectures," *Annals of Probability* 1 (1973) 19–42.

Y.S. Chow and H. Teicher, *Probability Theory: Independence Interchangeability Martingales* (Springer, New York, 1978).

V. Chvátal, *Linear Programming* (Freeman, New York, 1983).

H. Crowder and M.W. Padberg, "Solving large-scale symmetric travelling salesman problems, "*Management Science* 26 (1980) 495–509.

J. Edmonds, "Matching and a polyhedron with 0–1 vertices," *Journal of Research of the National Bureau of Standards* 69B (1965a) 125–130.

J. Edmonds, "Paths, trees and flowers," *Canadian Journal of Mathematics* 17 (1965b) 449–467.

J. Edmonds, "Submodular functions, matroids and certain polyhedra," in: *Combinatorial Structures and Their Applications* (Gordon and Breach, New York, 1970).

B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics* 9 (1981) 586–596.

M. Grötschel, *Polyedrische Charakterisierungen kombinatorischer Optimierungsprobleme* (Hain, Meisenheim, 1977).

M. Grötschel, "Developments in combinatorial optimization," in: W. Jäger, J. Moser and R. Remmert, eds., *Perspectives in Mathematics, Anniversary of Oberwolfach 1981* (Birkhäuser, Basel, 1981) pp. 249–294.

M. Grötschel, "Approaches to hard combinatorial optimization problems," in: B. Korte, ed., *Modern Applied Mathematics - Optimization and Operations Research* (North-Holland, Amsterdam, 1982) pp. 22–39.

M. Grötschel and O. Holland, "Solving matching problems with linear programming," *Mathematical Programming* 33 (1985) 243–259.

M. Grötschel and O. Holland, "A cutting plane algorithm for minimum perfect 2-matchings," *Computing* 39 (1987) 327–344.

G.H. Hardy, *Divergent Series* (Oxford Press, Oxford, 1949).

R.M. Karp, "Probabilistic analysis of partitioning algorithms for the traveling salesman problem in the plane," *Mathematics of Operations Research* 2 (1977) 209–224.

R.M. Karp and J.M. Steele, "Probabilistic analysis of heuristics," in: E.L. Lawler et al., eds., *The Traveling Salesman Problem: A Guided tour of Combinatorial Optimization* (Wiley, New York, 1985) pp. 181–206.

L. Lovász and M.D. Plummer, *Matching Theory* (Adadémiai Kiadó, Budapest, 1986).

M.W. Padberg and M.R. Rao, "Odd minimum cut-sets and b-matchings," *Mathematics of Operations Research* 7 (1982) 67–80.

C.H. Papadimitriou, "The probabilistic analysis of matching heuristics," *Fifteenth Annual Allerton Conference on Communication, Control and Computing* (1977) pp. 368–378.

W.T. Rhee and M. Talagrand, "Martingale inequalities, interpolations and NP-complete problems," to appear in: *Mathematics of Operations Research.*

W.T. Rhee and M. Talagrand, "A sharp deviation inequality for the stochastic traveling salesman problem," *Annals of Probability* 17 (1989) 1–8.

J.M. Steele, "Subadditive Euclidean functionals and non-linear growth in geometric probability," *Annals of Probability* 9 (1981a) 365–376.

J.M. Steele, "Complete convergence of short paths and Karp's algorithm for the TSP," *Mathematics of Operations Research* 6 (1981b) 374–378.

J.M. Steele, "Growth rates of Euclidean minimal spanning trees with power weighted edges," *Annals of Probability* 16 (1988) 1767–1787.

J.M. Steele, L.A. Shepp and W.F. Eddy, "On the number of leaves of a Euclidean minimal spanning tree," *Journal of Applied Probability* 24 (1987) 809–826.

W. Stout, *Almost Sure Convergence, Journal of Applied Probability* (Wiley, New York, 1976).

V.A. Yemelichev, M.M. Kovalev and M.K. Kravtsov, *Polytopes, Graphs, and Optimization* (Cambridge University Press, New York, 1984).