6.7720/18.619/15.070 Lecture 11
The Stochastic Euclidean Traveling Salesperson Problem

Kuikui Liu

March 10, 2025

Acknowledgements & Disclaimers *In the process of writing these notes, we consulted materials by Eyal Lubetzky and Yuval Peres. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Stochastic Euclidean TSP

In this lecture, we apply martingale-based arguments to study average-case instances of the quintessential combinatorial optimization problem: the *Traveling Salesperson Problem (TSP)* in Euclidean space. In an instance of this problem, we are given n points $\mathcal{P} \subseteq \mathbb{R}^d$, and the goal is to find a *tour*, i.e. a sequence of points $\boldsymbol{p}^{(1)}, \dots, \boldsymbol{p}^{(m)} \in \mathcal{P}$ such that every point of $\mathcal{P}$ is visited at least once (in particular, $m \geq n$), minimizing the total (Euclidean) distance traveled:

$$\mathsf{Cost}\left(\boldsymbol{p}^{(1)}, \dots, \boldsymbol{p}^{(m)}\right) \overset{\mathsf{def}}{=} \sum_{i=1}^{m-1} \left\|\boldsymbol{p}^{(i+1)} - \boldsymbol{p}^{(i)}\right\|_2.$$

We write $\mathsf{OPT} = \mathsf{OPT}(\mathcal{P})$ for the cost of an optimal tour; note that by the Triangle Inequality, we can assume any optimal tour is a permutation of $\mathcal{P}$. Computing OPT and an optimal tour is a classic NP-hard optimization problem, although unlike SAT or the chromatic number, we do have *polynomial-time approximation schemes* [Aro98; Mit99].

Let us now consider average-case instances of this problem, where for convenience, we assume the vectors in $\mathcal{P}$ are drawn independently according to $\mathsf{Unif}[0,1]^d$. Our goal is to study the random variable OPT.

Theorem 1.1 (Beardwood–Halton–Hammersley [BHH59])**.** *For every $d \geq 2$, there is a positive constant $\beta(d)$ such that*

$$\frac{\mathsf{OPT}}{n^{1-\frac{1}{d}}} \overset{\mathsf{a.s.}}{\to} \beta(d), \qquad \textit{as} \qquad n \to \infty.$$

Remark 1. It is known that $\frac{\beta(d)}{\sqrt{d}} \to \frac{1}{\sqrt{2\pi e}}$ as $d \to \infty$ at a rate of $O\left(\frac{\log d}{d}\right)$ [Rhe92].

We note that this result generalizes far beyond the distribution $\mathsf{Unif}[0,1]^d$. The scaling of $n^{1-\frac{1}{d}}$ is fairly intuitive: Imagine an idealized world where the hypercube $[0,1]^d$ is partitioned into a "(hyper)grid" of n subcubes all having side-lengths $\asymp n^{-1/d}$, and the points $\boldsymbol{p}_1, \dots, \boldsymbol{p}_n$ are placed at the vertices of these subcubes in an evenly spaced manner. It is easy to see (e.g. by considering the case $d = 2$ first, and then inducting on d) that the natural tour which traverses the points "linearly" along each dimension has cost $\asymp n^{1-\frac{1}{d}}$, since each step contributes $\asymp n^{-1/d}$ (the side-length of any subcube) to the distance. Based on this intuition, we will prove $\mathbb{E}[\mathsf{OPT}] \asymp n^{1-\frac{1}{d}}$ in Section 2. We further establish concentration for OPT.

Theorem 1.2 (Rhee–Talagrand [RT87]; see also [RT89; Rhe91])**.** *There exists a universal numerical constant $C > 0$ such that for every $d \geq 2$, we have the tail bound*

$$\Pr[|\mathsf{OPT} - \mathbb{E}[\mathsf{OPT}]| \geq t] \leq 2\exp\left(-\frac{t^2}{C(n,d)}\right), \qquad \forall t \geq 0,$$

where

$$C(n,d) = \begin{cases} O(\log n), & \text{if } d = 2 \\ O_d\left(n^{1-\frac{2}{d}}\right), & \text{if } d > 2 \end{cases}.$$

Remarkably, for the plane $d = 2$, Rhee–Talagrand have sharpened the result to true sub-Gaussian tails: For some absolute constant $C > 0$,

$$\Pr\left[|\mathsf{OPT} - \mathbb{E}[\mathsf{OPT}]| \geq t\right] \leq 2\exp\left(-Ct^2\right), \qquad \forall t \geq 0.$$

Theorem 1.2 implies that the typical deviation of OPT is at most of order $n^{\frac{1}{2}-\frac{1}{d}}$ if $d > 2$, and of order $\sqrt{\log n}$ if $d = 2$, which are both much smaller than the expectation $\mathbb{E}[\mathsf{OPT}] \asymp n^{1-\frac{1}{d}}$.

We prove Theorem 1.2 in Section 3.

2 Bounding the Expectation

In this section, we bound the order of the expectation.

Theorem 2.1. *For every $d \geq 2$, we have $\mathbb{E}[\mathsf{OPT}] \asymp n^{1-\frac{1}{d}}$. More precisely, there are constants $A_d, B_d > 0$ (depending only on d), such that $A_d \cdot n^{1-\frac{1}{d}} \leq \mathbb{E}[\mathsf{OPT}] \leq B_d \cdot n^{1-\frac{1}{d}}$ for all $n, d \geq 2$.*

For convenience, we define

$$\operatorname{dist}(\boldsymbol{p}, \mathcal{P}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{q} \in \mathcal{P}} \|\boldsymbol{p} - \boldsymbol{q}\|_2$$

for any subset $\mathcal{P} \subseteq [0,1]^d$ and any point $\boldsymbol{p} \in [0,1]^d$. The key technical result we will need to prove Theorem 2.1, as well as Theorem 1.2, is the following.

Proposition 2.2. *Fix an arbitrary point $\boldsymbol{p} \in [0,1]^d$. If $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n \sim \mathsf{Unif}[0,1]^d$ are drawn independently and we set $\mathcal{P} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}$, then*

$$\mathbb{E}\left[\operatorname{dist}(\boldsymbol{p}, \mathcal{P})\right] \asymp \frac{1}{n^{1/d}},$$

where $O_d(1)$ is a constant depending only on d.

We prove Proposition 2.2 at the end of the section.

Proof of Theorem 2.1. For the lower bound, observe that since every point $\boldsymbol{p}_i$ must be visited, we have the lower bound

$$\mathsf{OPT} \geq \sum_{i=1}^{n} \operatorname{dist}\left(\boldsymbol{p}_i, \mathcal{P} \setminus \{\boldsymbol{p}_i\}\right).$$

Taking expectations of both sides yields

$$\begin{aligned} \mathbb{E}[\mathsf{OPT}] &\geq \sum_{i=1}^{n} \mathbb{E}\left[\operatorname{dist}\left(\boldsymbol{p}_i, \mathcal{P} \setminus \{\boldsymbol{p}_i\}\right)\right] \\ &\gtrsim n \cdot (n-1)^{-1/d} && \text{(Proposition 2.2)} \\ &\gtrsim n^{1-\frac{1}{d}}. \end{aligned}$$

For the upper bound, we prove the following stronger claim: For *any* set of n points in $[0,1]^d$, there exists a tour with total cost at most $\lesssim n^{1-\frac{1}{d}}$. To show this, imagine we partition the hypercube $[0,1]^d$ into a "(hyper)grid" of $\asymp n$ subcubes $\mathcal{C}_1, \ldots, \mathcal{C}_n$, each of side-length $\asymp n^{-1/d}$. At the center of each subcube $\mathcal{C}_i$, we place a new point $\boldsymbol{q}_i$; note the n new points $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are evenly spaced throughout $[0,1]^d$. We will construct a tour for the points $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ with cost at most $n^{1-\frac{1}{d}}$; this is enough for our purposes by the Triangle Inequality.

Without loss of generality, assume the points $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are ordered in such a way that

$$\sum_{i=1}^{n-1} \|\boldsymbol{q}_i - \boldsymbol{q}_{i+1}\|_2 \lesssim n^{1-\frac{1}{d}}.$$

It is not difficult to see that such a tour always exists. For instance, in dimension 2, one can choose the "snake" tour, i.e. the one which alternates between left-to-right and right-to-left traversal within each row of the grid. One can inductively construct analogous tours in higher dimensions. Given this, we build a tour for $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ as follows:

- Within each $\mathcal{C}_i$, we construct an arbitrary tour $\mathcal{T}_i$ of the points $\{\mathcal{P} \cap \mathcal{C}_i\} \cup \{\boldsymbol{q}_i\}$ which begins and ends at $\boldsymbol{q}_i$. This determines how we visit points within each subcube.
- In the order $i = 1, \ldots, n$, we alternate between completely traversing the "subcube tour" $\mathcal{T}_i$, and moving from $\boldsymbol{q}_i$ to $\boldsymbol{q}_{i+1}$.

For each $i = 1, \ldots, n$, let $k_i = |\mathcal{P} \cap \mathcal{C}_i|$. The cost of the tour we've constructed is upper bounded by

$$\sum_{i=1}^{n} \mathsf{Cost}\left(\mathcal{T}_i\right) + \sum_{i=1}^{n-1} \|\boldsymbol{q}_i - \boldsymbol{q}_{i+1}\|_2 \lesssim \sqrt{d} \cdot \sum_{i=1}^{n} \frac{k_i + 1}{n^{1/d}} + n^{1-\frac{1}{d}} \lesssim n^{1-\frac{1}{d}}.$$

The first inequality follows from the fact that each subcube $\mathcal{C}_i$ has side-lengths upper bounded by $n^{-1/d}$, and so $\operatorname{diam}\left(\mathcal{C}_i\right) \lesssim \sqrt{d} \cdot n^{-1/d}$. The second inequality just follows from $\sum_{i=1}^{n} k_i = n$. $\square$

2.1 Proof of Proposition 2.2

By the layered cake representation of an expectation, we have

$$\begin{aligned} \mathbb{E}\left[\operatorname{dist}\left(\boldsymbol{p}, \mathcal{P}\right)\right] &= \int_0^{\sqrt{d}} \Pr\left[\operatorname{dist}\left(\boldsymbol{p}, \mathcal{P}\right) \geq R\right] \, dR \\ &= \int_0^{\sqrt{d}} \Pr_{\boldsymbol{q} \sim \mathsf{Unif}[0,1]^d} \left[\operatorname{dist}(\boldsymbol{p}, \boldsymbol{q}) \geq R\right]^n \, dR. \qquad \text{(Using independence of } \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\text{)} \end{aligned}$$

Observe that there are constants $0 < c(d) < C(d) < 1$, depending only on d, such that the volume of the radius-R Euclidean ball around $\boldsymbol{p}$, intersected with $[0,1]^d$, has volume

$$c(d) \cdot R^d \leq \operatorname{Vol}\left(\mathcal{B}_2(\boldsymbol{p}, R) \cap [0,1]^d\right) \leq C(d) \cdot R^d, \qquad \forall 0 \leq R \leq \sqrt{d}.$$

Using this, we have

$$\Pr_{\boldsymbol{q} \sim \mathsf{Unif}[0,1]^d} \left[\operatorname{dist}(\boldsymbol{p}, \boldsymbol{q}) \geq R\right] \geq 1 - C(d) \cdot R^d$$

Letting $R_0 = \left(\frac{1}{C(d) \cdot n}\right)^{1/d}$, we obtain the lower bound

$$\mathbb{E}\left[\operatorname{dist}\left(\boldsymbol{p}, \mathcal{P}\right)\right] \geq R_0 \cdot \Pr_{\boldsymbol{q} \sim \mathsf{Unif}[0,1]^d} \left[\operatorname{dist}(\boldsymbol{p}, \boldsymbol{q}) \geq R_0\right]^n \geq \left(\frac{1}{C(d) \cdot n}\right)^{1/d} \cdot \left(1 - \frac{1}{n}\right)^n \gtrsim \frac{1}{n^{1/d}}.$$

For the upper bound, if we let $R_0 = \left(\frac{1}{c(d) \cdot n}\right)^{1/d}$ instead, and let $T = \left\lceil \sqrt{d}/R_0 \right\rceil$, then we have

$$\mathbb{E}\left[\operatorname{dist}\left(\boldsymbol{p}, \mathcal{P}\right)\right] \leq \sum_{t=0}^{T} \int_{tR_0}^{(t+1)R_0} \left(1 - c(d) \cdot R^d\right)^n \, dR \leq \sum_{t=0}^{\infty} R_0 \cdot e^{-t^d} = O_d(1) \cdot R_0 \lesssim \frac{1}{n^{1/d}}.$$

3 Concentration for OPT

In this section, we prove the concentration estimate stated in Theorem 1.2. As a first attempt, observe that

$$\mathsf{OPT}\left(\mathcal{P}\right) = \inf_{\text{Tours } \boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(m)} \in \mathcal{P}} \sum_{i=1}^{m-1} \left\|\boldsymbol{p}^{(i+1)} - \boldsymbol{p}^{(i)}\right\|_2,$$

viewed as a function of n-tuples of points, is $2\sqrt{d}$-Lipschitz with respect to Hamming distance on $\mathcal{X}^n$, where $\mathcal{X} = [0,1]^d$; this is an immediate consequence of the fact that the diameter of $[0,1]^d$ with respect to Euclidean distance is $\sqrt{d}$. Hence, McDiarmid's Inequality applies and we get

$$\Pr\left[|\mathsf{OPT} - \mathbb{E}\left[\mathsf{OPT}\right]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{8dn}\right), \qquad \forall t \geq 0.$$

This is nice since we get order-$\sqrt{n}$ deviation with probability at most some constant, say, 1%. This is pretty good for large d, but is still rather far from Theorem 1.2 for small d, especially when $d = 2$ and $\mathbb{E}\left[\mathsf{OPT}\right] \asymp \sqrt{n}$.

3.1 Refining the Bound: Proof of Theorem 1.2

The basic idea is to consider again the Doob martingale given by $Y_k = \mathbb{E}\left[\mathsf{OPT} \mid \mathcal{P}_{\leq k}\right]$ for $k = 0, \ldots, n$, where we write $\mathcal{P}_{\leq k} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\}$; we also define $\mathcal{P}_{>k} = \mathcal{P} \setminus \mathcal{P}_{\leq k}$, and $\mathcal{P}_{-k} = \mathcal{P} \setminus \{\boldsymbol{p}_k\}$. Rather than applying McDiarmid's Inequality, which uses a uniform bound on the Lipschitzness of OPT, we will combine Azuma–Hoeffding with a more refined bound on the almost sure boundedness of the increments $Y_k - Y_{k-1}$.

We will need the following geometric result.

Lemma 3.1. *Let $\mathcal{P} \subseteq [0,1]^d, \boldsymbol{p} \in [0,1]^d$ be arbitrary. Then*

$$\mathsf{OPT}\left(\mathcal{P}\right) \leq \mathsf{OPT}\left(\mathcal{P} \cup \{\boldsymbol{p}\}\right) \leq \mathsf{OPT}\left(\mathcal{P}\right) + 2 \cdot \operatorname{dist}\left(\boldsymbol{p}, \mathcal{P}\right)$$

Proof. The first inequality is immediate. For the second, we can build a tour for $\mathcal{P} \cup \{\boldsymbol{p}\}$ by taking an optimal tour for $\mathcal{P}$ and appending the moves $\boldsymbol{q} \to \boldsymbol{p} \to \boldsymbol{q}$, where $\boldsymbol{q} \in \mathcal{P}$ minimizes $\|\boldsymbol{p} - \boldsymbol{q}\|_2$. This yields a tour with cost $\mathsf{OPT}(\mathcal{P}) + 2 \cdot \operatorname{dist}\left(\boldsymbol{p}, \mathcal{P}\right)$. □

Let us use it to bound the increments and deduce the desired concentration estimate.

Corollary 3.2. *For every k, $|Y_k - Y_{k-1}| \leq \min\left\{2\sqrt{d}, \frac{O_d(1)}{(n-k)^{1/d}}\right\}$ almost surely.*

Proof. Arbitrarily fix the first k points $\mathcal{P}_{\leq k} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\} \subseteq [0,1]^d$. Our goal is to show that

$$\left|\mathbb{E}\left[\mathsf{OPT} \mid \mathcal{P}_{\leq k}\right] - \mathbb{E}\left[\mathsf{OPT} \mid \mathcal{P}_{\leq k-1}\right]\right| \leq \min\left\{2\sqrt{d}, \frac{O_d(1)}{(n-k)^{1/d}}\right\}.$$

The first bound is immediate from the diameter of $[0,1]^d$. For the second bound, observe that we may perfectly couple the random choices of the remaining points $\mathcal{P}_{>k} = \{\boldsymbol{p}_{k+1}, \ldots, \boldsymbol{p}_n\}$ to obtain the upper bound

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathsf{OPT} \mid \mathcal{P}_{\leq k}\right] - \mathbb{E}\left[\mathsf{OPT} \mid \mathcal{P}_{\leq k-1}\right]\right| \\
&\quad\leq \sup_{\boldsymbol{p}'_{k+1} \in [0,1]^d} \mathbb{E}_{\boldsymbol{p}_{k+1}, \ldots, \boldsymbol{p}_n}\left[\left|\mathsf{OPT}\left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k, \ldots, \boldsymbol{p}_n\right) - \mathsf{OPT}\left(\boldsymbol{p}_1, \ldots, \boldsymbol{p}'_k, \ldots, \boldsymbol{p}_n\right)\right|\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Triangle Inequality)} \\
&\quad\leq 2 \cdot \sup_{\boldsymbol{p}'_{k+1} \in [0,1]^d} \mathbb{E}_{\boldsymbol{p}_{k+1}, \ldots, \boldsymbol{p}_n}\left[\operatorname{dist}\left(\boldsymbol{p}_k, \mathcal{P}_{-k}\right) + \operatorname{dist}\left(\boldsymbol{p}'_k, \mathcal{P}_{-k}\right)\right] \qquad \text{(Lemma 3.1)} \\
&\quad\leq 2 \cdot \sup_{\boldsymbol{p}'_{k+1} \in [0,1]^d} \mathbb{E}_{\boldsymbol{p}_{k+1}, \ldots, \boldsymbol{p}_n}\left[\operatorname{dist}\left(\boldsymbol{p}_k, \mathcal{P}_{>k}\right) + \operatorname{dist}\left(\boldsymbol{p}'_k, \mathcal{P}_{>k}\right)\right] \\
&\quad\leq \frac{O_d(1)}{(n-k)^{1/d}}, \qquad \text{(Proposition 2.2)}
\end{aligned}$$

□

To complete the proof of Theorem 1.2, we let $c_k = \min\left\{2\sqrt{d}, \frac{O_d(1)}{(n-k)^{1/d}}\right\}$ for $k = 1, \ldots, n$ by Corollary 3.2. Observe that

$$C(n,d) = \sum_{k=1}^{n} c_k^2 \leq O_d(1) \sum_{k=1}^{n-1} \frac{1}{(n-k)^{2/d}} \leq O_d(1) \cdot \int_1^n \frac{1}{x^{2/d}}\, dx \leq \begin{cases} O\left(\log n\right), & \text{if } d = 2 \\ O_d\left(n^{1-\frac{2}{d}}\right), & \text{if } d > 2 \end{cases}.$$

Invoking Azuma–Hoeffding then concludes the proof.

References

[Aro98] Sanjeev Arora. “Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems”. In: *J. ACM* 45.5 (Sept. 1998), pp. 753–782. ISSN: 0004-5411. DOI: 10.1145/290179.290180 (cit. on p. 1).

[BHH59] Jillian Beardwood, J. H. Halton, and J. M. Hammersley. “The shortest path through many points”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 55.4 (1959), pp. 299–327. DOI: 10.1017/S0305004100034095 (cit. on p. 1).

[Mit99] Joseph S. B. Mitchell. “Guillotine Subdivisions Approximate Polygonal Subdivisions: A Simple Polynomial-Time Approximation Scheme for Geometric TSP, k-MST, and Related Problems”. In: *SIAM Journal on Computing* 28.4 (1999), pp. 1298–1309. DOI: 10.1137/S0097539796309764 (cit. on p. 1).

[Rhe91] Wansoo T. Rhee. “On the Fluctuations of the Stochastic Traveling Salesperson Problem”. In: *Mathematics of Operations Research* 16.3 (1991), pp. 482–489. ISSN: 0364765X, 15265471 (cit. on p. 1).

[Rhe92] Wansoo T. Rhee. “On the Travelling Salesperson Problem in Many Dimensions”. In: *Random Structures & Algorithms* 3.3 (1992), pp. 227–233 (cit. on p. 1).

[RT87] Wansoo T. Rhee and Michel Talagrand. “Martingale Inequalities and NP-Complete Problems”. In: *Mathematics of Operations Research* 12.1 (1987), pp. 177–181. ISSN: 0364765X, 15265471 (cit. on p. 1).

[RT89] WanSoo T. Rhee and Michel Talagrand. “A Sharp Deviation Inequality for the Stochastic Traveling Salesman Problem”. In: *The Annals of Probability* 17.1 (1989), pp. 1–8 (cit. on p. 1).