Beginning of 02/14/2025

More on Euclidean TSP

Some results that we recall from last lecture:
Let $X_1, \cdots, X_n$ be drawn i.i.d. from $[0,1]^2$. We let $f(X_1, \cdots, X_n)$ denote the optimal TSP length; we showed last time that $\mathbb{E}[f] = \Omega(\sqrt{n})$, and our remaining goal is to bound the tail probabilities $\mathbb{P}(|f - \mathbb{E}f| \geqslant \epsilon)$.
We constructed a sequence of martingales via:

- $Z_0 = \mathbb{E}[f(X_1, \cdots, X_n)]$, which is completely deterministic with value $\Omega(\sqrt{n})$.
- $Z_1 = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1]$, a non-deterministic value which varies over values of X_1.
- $Z_2 = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, X_2]$, and likewise.
- $Z_n = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_n] = f(X_1, \cdots, X_n)$, the TSP function itself.

We observe that

$$\mathbb{E}[Z_1] = \mathbb{E}_{x_1}[\mathbb{E}[f(X_1, \cdots, X_n) \mid X_1]] = Z_0,$$

and likewise for the rest. This shows $\{Z_i\}$ forms a martingale.
In order to appeal to Azuma's inequality we want to compute almost sure bounds on pairwise difference $|Z_k - Z_{k-1}|$.
We shoed that

$$\begin{aligned} C_k = |Z_k - Z_{k-1}| &\leqslant \max_{X_k, \hat{X}_k} \mathbb{E}_S[f(X_1, \cdots, X_{k-1}, X_k, \underbrace{X_{k+1}, \cdots, X_n}_{S}) \mid X_1, \cdots, X_{k-1}] \\ &\quad - \mathbb{E}[f(X_1, \cdots, X_{k-1}, \hat{X}_k, X_{k+1}, \cdots, X_n) \mid X_1, \cdots, X_{k-1}] \\ &\leqslant 2\mathbb{E}_S[d(X_k, S) + d(\hat{X}_k, S)] \end{aligned}$$

where $d(x, S)$ is the set distance from S to x.
Note S is just $n-k$ random points drawn i.i.d. in $[0,1]^2$. Fix any x drawn from $[0,1]^2$. Consider the ball centered at x with radius $r = q/\sqrt{\pi(n-k)}$ for some q. If no points in S lies in $B(x,r)$ then $d(x,S) \geqslant q/\sqrt{\pi(n-k)}$. This happens with probability

$$\mathbb{P}(d(x, S \geqslant q/\sqrt{\pi(n-k)}) \leqslant \left(1 - \frac{q^2}{\pi(n-k)}\pi\right)^{n-k} \leqslant \exp(-q^2).$$

Then splitting $\mathbb{E}(d(x,S))$ into tail sums we obtain

$$\begin{aligned} \mathbb{E}[d(x,S)) &\leqslant \sum_{q \geqslant 1} \mathbb{P}\left(d(x,S) \geqslant q/\sqrt{\pi(n-k)}\right) \cdot \frac{q}{\sqrt{\pi}(n-k)} \\ &\leqslant \sum_{q \geqslant 1} e^{-q^2} \frac{q}{\sqrt{\pi(n-k)}} = o\left(\frac{1}{\sqrt{n-k}}\right). \end{aligned}$$

Therefore $C_k = o(1/\sqrt{n-k})$, and in particular

$$\sum_k C_k^2 = o\left(\sum_{k=1}^{n-1} \frac{1}{n-k}\right) = o(\log n).$$

Choose $\lambda = h\eta(\log n)$ and let $\mu = \mathbb{E}f$. Then Azuma's inequality implies

$$\mathbb{P}\left[|f(X_1, \cdots, X_n) - \mu| \geqslant \log n\right] \leqslant \exp(-\lambda^2 / \sum C_k^2) = \exp(-\log^2 n / \log n) = \frac{1}{n}.$$

This shows that **w.h.p. the optimal TSP length is highly concentrated around $\Omega(\sqrt{n})$**.

Part II: Karp's Algorithm

Theorem

There exists a polynomial-time algorithm such that for all Euclidean TSP instances on $[0,1]^2$,

$$\mathrm{ALG} \leqslant \mathrm{OPT} + \mathcal{O}\left(\sqrt{\frac{n \log\log n}{\log n}}\right).$$

Observe the error term is $o(\sqrt{n})$, as $\log\log n / \log n$ vanishes as $n \to \infty$.

This result is useless for certain rare instances where the optimal TSP length is small: for example if all X_i's align along a diagonal, for the error term then becomes large in comparison. However, using our previous concentration bound, most instances have high cost, so w.h.p. this shows the algorithm is $1 + o(1)$ approximate.

Proof. Let $s = \log n / \log\log n$. Consider a collection of n points in $[0,1]^2$. We partition the space into different horizontal stripes of width 1, each containing $\sqrt{ns}$ points. The total number of horizontal stripes is $\sqrt{n/s}$.

Then, for *each* individual horizontal stripe, partition them using vertical lines such that each box has s points. That is, each horizontal stripe now has $\sqrt{n/s}$ boxes.

Step 1. For each box, we brute force compute an optimal tour for the s points contained. Each brute force takes factorial time w.r.t. s, and with some algebra one can show that $s! = \mathcal{O}(n)$, so indeed the algorithm runs in polynomail time.

Step 2. Connect the individual tours using serpent line, i.e., $(0,0) \to (0,1) \to \cdots \to (0, \sqrt{n/s}) \to (1, \sqrt{n/s}) \to \cdots \to (1,0) \to (2,0) \to \cdots$. We just connect each optimal individual box subtour to this serpent line.

The total distance traveled by the serpent line itself is $2 + \sqrt{n/s}$ (2 for vertical distance traveled and 1 for each of the $\sqrt{n/s}$ rows). The total extra distance induced by connecting serpent lines to a subtours and from subtours back to the line is $2\sqrt{ns}$ since we have $\sqrt{ns}$ boxes, each with length no more than 2.

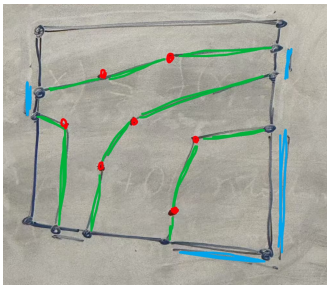

Figure 1: Suppose the red vertices are those in the box. Green edges = edges that appear in OPT (truncated for those that extend beyond this box). We add blue edges.

Finally, we consider the cost related to the subtours. Clearly the OPT tour restricted to a box is likely different

from the locally optimal subtour. We bridge this gap by introducing Eulerian paths to an augmented local subtour. See figure.

Immediately we see that every vertex in this augmented graph has degree 2 or 4, so there exists a Eulerian path on this graph. Therefore,

$$\begin{aligned}\text{Karp restricted to box} &\leqslant \text{length of Eulerian path} \leqslant \text{green} + 2 \cdot \text{blue} \\ &\leqslant \text{OPT restricted to box} + 2 \cdot \text{ perimeter of box.}\end{aligned}$$

Therefore,

$$\text{Karp} = \sum_{\text{boxes}} \text{Karp restricted to box} \leqslant \sum_{\text{boxes}} \text{OPT restricted to box} + 2 \sum_{\text{boxes}} \text{perimeter} = \text{OPT} + 8\sqrt{n/s}.$$

All additional costs are $\mathcal{O}(\sqrt{n/s})$ so our proof is complete. □