BIOST 527 - FINAL PROJECT

KATIE WOOD

Proof for Metric Space

Let $\mathcal{B}$ denote the set of bike trips, with $x, y, z \in \mathcal{B}$ arbitrary. Then define $d : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ where

$$d(x,y) = \lambda \left[\min\left(\frac{|t_{1,x} - t_{1,y}|}{43200}, 2 - \frac{|t_{1,x} - t_{1,y}|}{43200} \right) \right] + \\ (1-\lambda) \left[(1-\gamma) \left[\frac{\text{dist}(s_{1,x}, s_{1,y}) + \text{dist}(s_{2,x}, s_{2,y})}{2 \max_{v,w \in \mathcal{B}, i \in \{1,2\}} \text{dist}(s_{i,v}, s_{i,w})} \right] + \gamma \left| \sin(\theta_{x,y}) \right| \right]$$

We summarize the meaning of the notation as follows:

$$\begin{aligned} \lambda, \gamma &:= \text{positive constants} \in [0,1] \text{ chosen by cross-validation} \\ t_{1,v} &:= \text{the start time of day (in seconds) of a trip} \\ 43200 &:= \text{the number of seconds in half a day} \\ s_{1,v} &:= \text{the start location of a trip} \\ s_{2,v} &:= \text{the end location of a trip} \\ \text{dist}(s_{i,v}, s_{i,w}) &:= \text{Great Circle distance between two locations} \\ \theta_{v,w} &:= \text{difference in initial compass bearings between two trips} \end{aligned}$$

Next, we investigate whether the ordered pair $(\mathcal{B}, d)$ satisfies the four metric space axioms.[1]

1. The distance from a point to itself is zero.

Proof: With $x \in \mathcal{B}$ arbitrary, we verify that

$$\begin{aligned} |t_{1,x} - t_{1,x}| &= 0 \\ \text{dist}(s_{i,x}, s_{i,x}) &= 0, \quad i = 1,2 \\ \theta_{x,x} &= 0 \quad \Rightarrow \quad |\sin(\theta_{x,x})| = 0 \end{aligned}$$

and thus $d(x,x) = 0$.

2. The distance between two distinct points is always positive.

Proof: As defined, $d(x,y) \geq 0$. Thus, we consider whether it is possible for $d(x,y) = 0$ when $x \neq y$. BWOC, assume this is indeed the case. Then

$$\begin{aligned} \min\left(\frac{|t_{1,x} - t_{1,y}|}{43200}, 2 - \frac{|t_{1,x} - t_{1,y}|}{43200} \right) &= 0 \\ \text{dist}(s_{1,x}, s_{1,y}) &= 0 \\ \text{dist}(s_{2,x}, s_{2,y}) &= 0 \\ |\sin(\theta_{x,y})| &= 0 \quad \Rightarrow \quad \theta_{x,y} = n\pi, \; n \in \mathbb{Z} \end{aligned}$$

[1] https://en.wikipedia.org/wiki/Metric_space

Then either $|t_{1,x} - t_{1,y}| = 0$ or $|t_{1,x} - t_{1,y}| = 2 \times 43200 = 0 \mod 86400$ (where there are 86400 seconds in a day). Thus x and y start at the same time of day, or $t_{1,x} = t_{1,y}$. Also, that the distance between the start locations and the distance between the end locations are both zero implies that x and y begin at the same location and end at the same location, or $s_{1,x} = s_{1,y}$ and $s_{2,x} = s_{2,y}$. However, this further implies that the initial bearing of trip x is the same as that of trip y. Thus, x and y are the same trip $\Rightarrow\Leftarrow$.

3. The distance from x to y is always the same as the distance from y to x.

Proof: We take as trivially true that the third axiom holds for the start times, start locations, and end locations of x and y. It remains to investigate whether $|\sin(\theta_{x,y})| = |\sin(\theta_{y,x})|$. Since sin is an odd function, and $\theta_{x,y} = -\theta_{y,x}$, we have $\sin(\theta_{x,y}) = -\sin(\theta_{y,x})$, so by virtue of the absolute values the equality indeed holds.

4. The triangle inequality holds.

Proof: Since the Great Circle distance is the shortest distance between two points when travelling on the surface of a sphere, and all bike trips are routes along Earth's surface, we have

$$\begin{aligned}\text{dist}(s_{1,x}, s_{1,z}) &\leq \text{dist}(s_{1,x}, s_{1,y}) + \text{dist}(s_{1,y}, s_{1,z})\\ \text{dist}(s_{2,x}, s_{2,z}) &\leq \text{dist}(s_{2,x}, s_{2,y}) + \text{dist}(s_{2,y}, s_{2,z})\end{aligned}$$

Next, with $t_{1,x} < t_{1,z}$ (WLOG), if $t_{1,y} \in [t_{1,x}, t_{1,z}]$, the temporal distance from x to z equals the sum of the temporal distances from x to y and from y to z. If $t_{1,y} \notin [t_{1,x}, t_{1,z}]$, WLOG $t_{1,y} < t_{1,x}$, the temporal distance from x to z equals itself plus twice the sum of the temporal distance from x to y. Thus,

$$\begin{aligned}\min\left(\frac{|t_{1,x} - t_{1,z}|}{43200}, 2 - \frac{|t_{1,x} - t_{1,z}|}{43200}\right) \leq \min\left(\frac{|t_{1,x} - t_{1,y}|}{43200}, 2 - \frac{|t_{1,x} - t_{1,y}|}{43200}\right) +\\ \min\left(\frac{|t_{1,y} - t_{1,z}|}{43200}, 2 - \frac{|t_{1,y} - t_{1,z}|}{43200}\right)\end{aligned}$$

Next, we investigate whether

$$|\sin(\theta_{x,z})| \leq |\sin(\theta_{x,y})| + |\sin(\theta_{y,z})|$$

Since $\theta_{x,z} = \theta_{x,y} + \theta_{y,z}$, we in particular investigate whether

$$\begin{aligned}&|\sin(\theta_{x,z})| \leq |\sin(\theta_{x,z} - \theta_{y,z})| + |\sin(\theta_{y,z})|\\ \text{or}\qquad &|\sin\alpha| \leq |\sin(\alpha - \beta)| + |\sin\beta|\end{aligned}$$

Where for ease of notation we have let $\alpha = \theta_{x,z}$ and $\beta = \theta_{y,z}$. First, we note that the function $f(x) = |\sin(x)|$ is π-periodic in x. Second, we note that the right-hand side $f(\alpha - \beta) + f(\beta)$ is a translation of the left-hand side, $f(\alpha)$, namely, a translation right β and up $f(\beta)$. Thus, using the π-periodicity, $f(\alpha) = f(\alpha - \beta) + f(\beta)$ whenever $\beta = \alpha + \pi m, m \in \mathbb{Z}$. Let $g(\alpha, \beta) = f(\alpha - \beta) + f(\beta)$ for additional ease of notation. Next, we note that

$$\frac{\partial}{\partial\alpha}|\sin\alpha| = \begin{cases}\cos\alpha, & \alpha \in \cup_{m=-\infty}^{\infty}[2m\pi, (2m+1)\pi]\\ -\cos\alpha, & \alpha \in \cup_{m=-\infty}^{\infty}[(2m-1), 2m\pi]\end{cases}$$

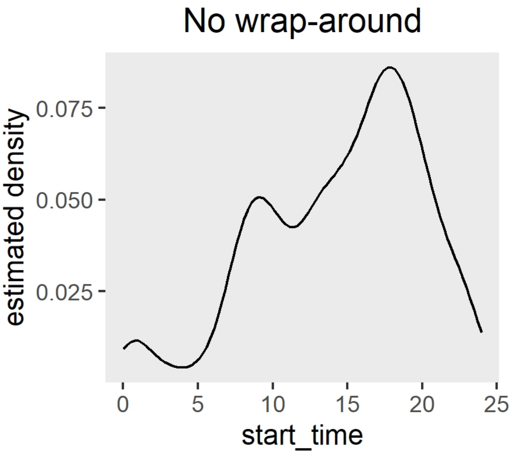
$$\frac{\partial}{\partial\alpha}\left(|\sin(\alpha - \beta)| + |\sin\beta|\right) = \begin{cases}\cos(\alpha - \beta), & \alpha \in \cup_{m=-\infty}^{\infty}[\beta + 2m\pi, \beta + (2m+1)\pi]\\ -\cos(\alpha - \beta), & \alpha \in \cup_{m=-\infty}^{\infty}[\beta + (2m-1), \beta + 2m\pi]\end{cases}$$

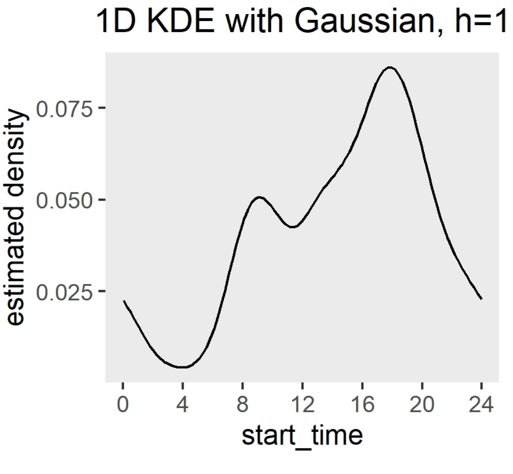
We have shown above that $f'(\alpha - \beta) = g_\alpha(\alpha, \beta)$. Also, $f'(\alpha)$ is monotone decreasing over each interval $\cup_{m=-\infty}^{\infty}[m\pi, (m+1)\pi]$ (with jumps at the intervals' endpoints). Consider now a single period of $f(\alpha)$, and restrict β to the same period as α. For $\alpha > \beta$, we then have $f'(\alpha) < f'(\alpha - \beta) = g_\alpha(\alpha, \beta)$. Since f and g agree at $\alpha = \beta$, and g is always increasing faster than f for $\alpha > \beta$, it follows that $g > f$ for $\alpha > \beta$. For $\alpha < \beta$, consider instead $\alpha' = -\alpha$ and $\beta' = -\beta$, such that $\alpha' > \beta'$. Then by the result we just proved, $g > f$ for $\alpha' > \beta'$, which implies that $g > f$ for $\alpha < \beta$. By periodicity in α and in β, these results hold for all α and β. Therefore, the desired bound $f \leq g$ holds.

Combining all of these results together by multiplying by the appropriate constants, the triangle inequality holds for d. $\square$

Implementing Kernel Methods for d

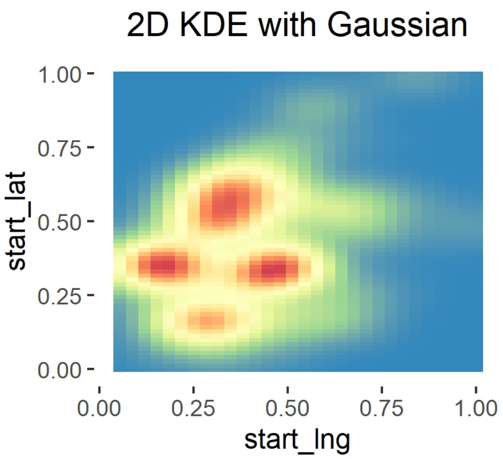
We begin by implementing a 1D Gaussian kernel for the time component of our custom distance metric. The principal aim here is to show a continuous transition from before midnight to after midnight:

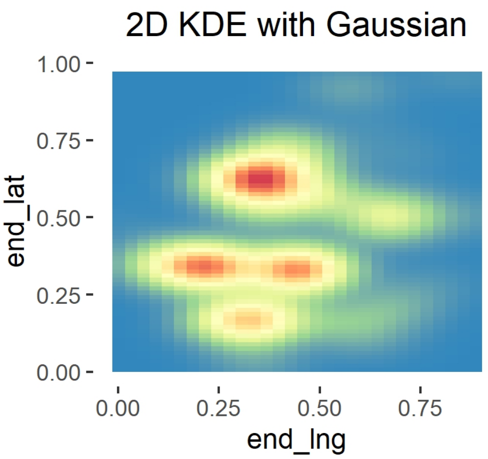




Indeed, we show above right that our custom distance metric “repairs” the discontinuity at midnight.

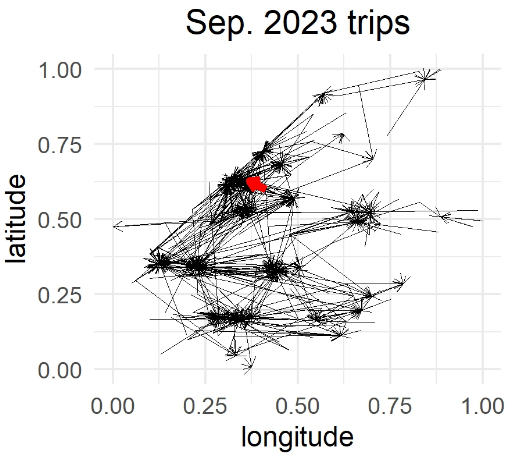
Next, we explore the spatial components of our data. These data are 4-dimensional, since they contain start & end latitudes and start & end longitudes for every trip. To simplify the graph, we first visualize only start latitude & longitude. We apply a standard 2D Gaussian kernel density estimation (not using d), with smoothing parameters $h_1 = 0.08$ and $h_2 = 0.05$.

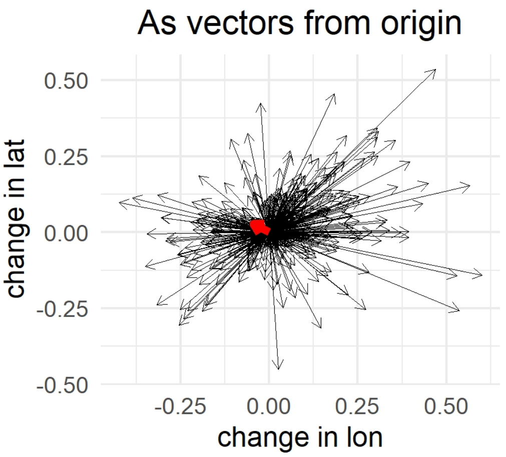




Above left, four prominent start locations appear in red. These correspond to particular areas of Manhattan from which CitiBike users frequently departed on trips (during the time this sample was collected: September 2023). Above right, we show where this same set of trips ended. We find that there are also four prominent end locations, though we cannot tell from where any particular trip originated.

Next, we display the set of trips from September 2023[2] in a different manner: as arrows that trace each trip from tail to tip. We highlight one arbitrary trip in red, chosen to be near one of the four "hot spots" found above.

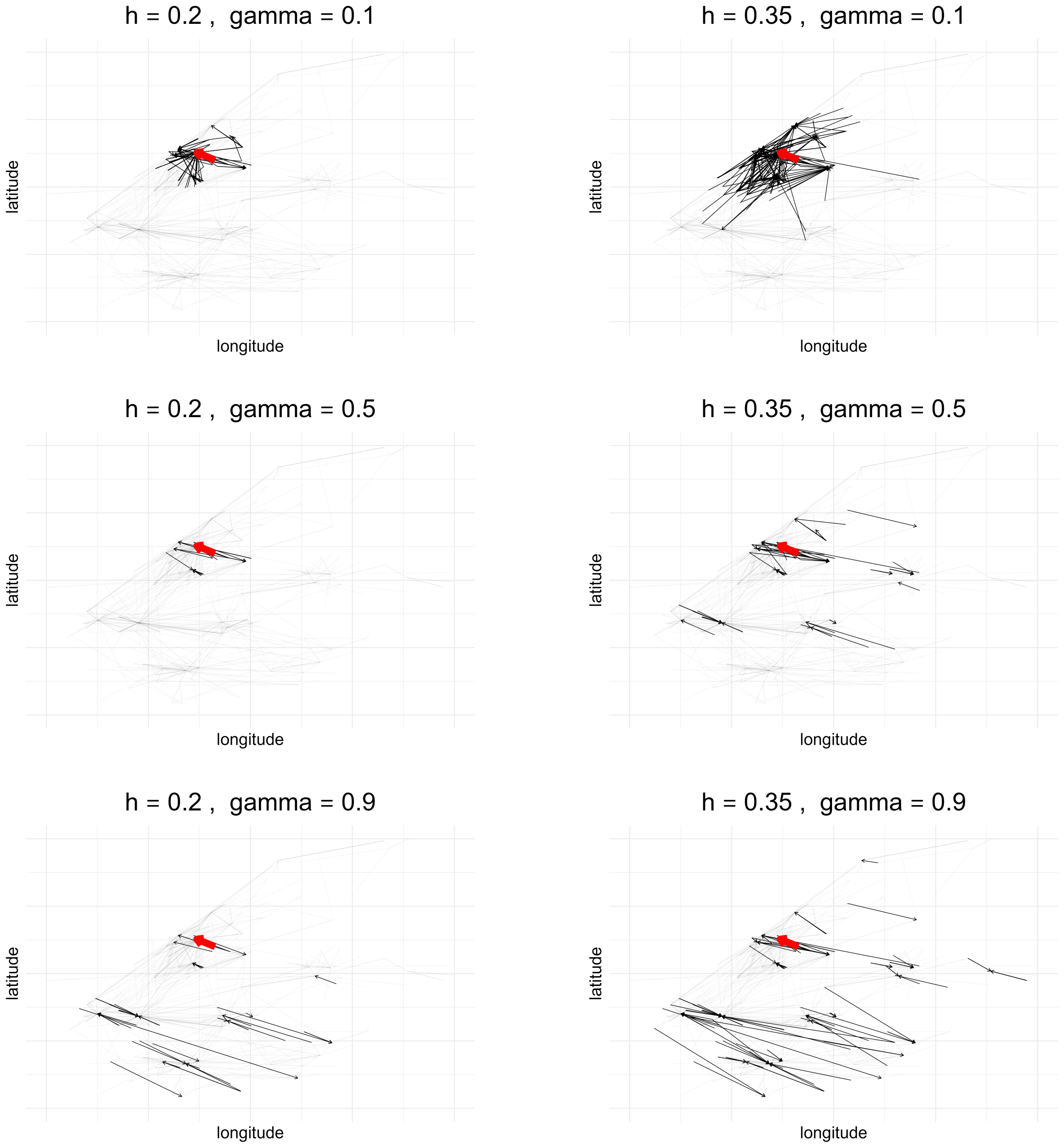




Above left, we show the trip arrows plotted in their true positions in a normalized latitude-longitude plan-view of Manhattan. Above right, meanwhile, shows the trip arrows as vectors emanating from the origin, the purpose of which is to highlight the role θ, the angle between two trips, will play[3]. The challenge, which we have addressed in section **1**, is to come up with a suitable metric that can measure the distance between trip vectors (4D objects) in a meaningful way.

[2] Actually, a small sample of size 1k (out of $\approx$ 500k)

[3] Note that in this implementation, I have simplified my calculations by using a planar approximation, rather than the true spherical geometry of the Earth.



Above, we showcase our custom distance metric in six plots. Trips less than a distance h from the red trip are black, while other trips are greyed out. We switch from a Gaussian kernel to an indicator kernel for two reasons: (i) the indicator is easier to visualize, (ii) the indicator could lead to sparse data structures, which would expedite an otherwise costly computation[4]. The six plots show how the hyper-parameter h tunes the "window size", with a larger h value permitting a larger set of trips to be considered close to the trip highlighted in red[5]. Meanwhile, γ tunes the importance of the angular penalty, with smaller values of γ preferring trips whose start and end locations are close, and larger values of γ preferring trips whose angles are close.

[4] We leave the task of computing the kernel density estimate on a dense grid of 4-dimensional points, likely necessitating the use of sparse data structures, to a future project.

[5] The parameter behavior is normalized so that $h = 1$ permits all trips, $h = 0$ permits no trips, etc.