Homework 3: Geodesics, Distance, and Metric Embedding

Due March 31, 2021

This is the third homework assignment for 6.838. Check the course website for additional materials and the late policy. You may work on assignments in groups, but every student must submit their own write up; please note your collaborators, if any, on your write up. **Submit your code as** `6838-hw3-<yourkerberos>.zip` **and writeup as** `6838-hw3-<yourkerberos>.pdf`**, where** `<yourkerberos>` **is replaced with your MIT Kerberos ID.**

Problem 1 (25 points)**.** In this problem, you will examine the relationships between curvature, geodesics, and the *parallel transport* operator on a surface. Let $\gamma : [0, L] \to M$ be an arc length–parametrized curve on an orientable surface $M \subset \mathbb{R}^3$. Let $\mathbf{v}$ be a tangent vector field defined along the curve—that is, a smooth assignment of tangent vectors $\mathbf{v}(s) \in T_{\gamma(s)}M$. The tangential component of the derivative $\mathbf{v}'(s)$ tells us how $\mathbf{v}$ is varying *intrinsically*, i.e., "as seen from inside the surface." If

$$\mathbf{0} = \text{Proj}_{T_{\gamma(s)}M}\mathbf{v}'(s) = \mathbf{v}'(s) - \mathbf{n}(\gamma(s))[\mathbf{n}(\gamma(s)) \cdot \mathbf{v}'(s)], \tag{1}$$

we say that $\mathbf{v}$ is *parallel*.

(a) Show that the unit tangent vector field of a geodesic is parallel.

(b) Suppose that $\mathbf{u}$ and $\mathbf{v}$ are parallel fields along γ. Show that $\mathbf{u} \cdot \mathbf{v}$ and $\|\mathbf{u}\|_2$ are constant.

Equation (1) is a first-order ODE whose solution is unique given an initial condition $\mathbf{v}(0)$. As such, we can define the *parallel transport operator* P_γ by $P_\gamma \mathbf{V} = \mathbf{v}(L)$, where $\mathbf{v}$ is the unique parallel field along γ with $\mathbf{v}(0) = \mathbf{V}$.

(c) Use (b) to argue that parallel transport around a closed loop (known as *holonomy*) amounts to a rotation in the tangent plane.

(d) Let $\mathbf{v}$ be parallel along γ. Let $\theta(s)$ be the angle from $\gamma'(s)$ to $\mathbf{v}(s)$, measured counterclockwise about the surface normal $\mathbf{n}$. Show that

$$\theta'(s) = -\kappa_g$$

where (recall from lecture) κ_g is the *geodesic curvature* of γ, defined by projection of the second derivative of γ into the tangent plane of the surface:

$$\text{Proj}_{T_{\gamma(s)}M}\gamma''(s) = \kappa_g(\mathbf{n} \times \gamma'(s)).$$

(e) A geodesic polygon is a polygon formed from geodesic segments. Using parts (a) and (b), show that parallel transporting a vector around a geodesic polygon rotates it by the angle $\sum_i (\pi - \alpha_i)$, where α_i are the interior angles of the polygon.

Note: If $\gamma_1 : [0, L_1] \to M$ and $\gamma_2 : [0, L_2] \to M$ are two successive geodesics in the polygon with $\gamma_1(L_1) = \gamma_2(0) = p$, the exterior angle between the two curves at p is defined to be the angle measured counterclockwise between their tangent vectors $\gamma_1'(L_1)$ and $\gamma_2'(0)$.

(f) More generally, let γ be a smooth closed curve on a surface M bounding a disk-shaped region U. Using part (d) and the Gauss-Bonnet theorem, show that the parallel transport P_γ rotates vectors by an angle

$$\theta_\gamma = \int_U K \, dA,$$

where K is the Gaussian curvature.

(g) Drawing on your intuition from the previous parts, define a notion of parallel transport between the tangent planes of adjacent triangles in a triangle mesh. Show a relationship between your definition and a notion of Gaussian curvature introduced in lecture.

Problem 2 (25 points)**.** In this problem, you will show the existence of low-distortion embeddings of arbitrary finite metric spaces, a version of *Bourgain's Theorem*. Recall the definition of metric distortion.

> **Definition.** *Let (X, d_X) and (Y, d_Y) be metric spaces and $f : X \to Y$ a map between them. The* distortion *is the smallest positive ρ such that for some c and all pairs $x_1, x_2 \in X$,*
>
> $$c \, d_X(x_1, x_2) \leq d_Y(f(x_1), f(x_2)) \leq \rho \, c \, d_X(x_1, x_2).$$

The distortion measures how tightly f preserves distances, modulo uniform scaling. Or, in other words, it measures how well f preserves distance ratios. ρ is always at least 1—a distortion of 1 indicates distances are exactly preserved up to uniform scale.

The result we will prove is the following. **Note:** we use big-O, big-Ω, and big-Θ notation to simplify the statements. If you haven't seen this notation in a while, it might be worth looking it up to recall the definitions.

> **Theorem** (Bourgain)**.** *For any finite metric space (X, d) of size $|X| = n$, there is an embedding $f : X \to \mathbb{R}^m$ with $m \in \Theta(\log^2 n)$ such that*
>
> $$|f_i(x) - f_i(y)| \leq d(x, y) \tag{2}$$
> $$\|f(x) - f(y)\|_1 \in \Omega((\log n) d(x, y)). \tag{3}$$
>
> *Moreover, the distortion is $O(\log n)$ for any ℓ_p metric on $\mathbb{R}^m$, $p \geq 1$.*

(a) Using Hölder's inequality, show that the ℓ_1 results (2) and (3) imply the conclusion for all p.

(b) The map f can be built up from the distances themselves. Show that if S_i is any nonempty subset of X and

$$f_i(x) := d(x, S_i) := \min_{y \in S_i} d(x, y),$$

then (2) holds. *Hint:* apply the triangle inequality.

(c) The construction of the sets S_i and the proof of (3) rely on a *probabilistic* argument. The basic idea is to sample enough subsets S_i to "separate" any given pair of points x and y.

Suppose a subset $S \subset X$ is sampled as follows: $\mathbb{P}(x \in S) = 1/k$ independently for each $x \in X$, where $k \geq 2$. Let B be a fixed subset of X. Show that

$$|B| \geq \alpha k \implies \mathbb{P}(B \cap S \neq \varnothing) \geq 1 - e^{-\alpha} > 0$$
$$|B| \leq \alpha k \implies \mathbb{P}(B \cap S = \varnothing) \geq 4^{-\alpha} > 0.$$

(d) For each $i \in \{1, \ldots, \log_2 n\}$, sample a subset S_i as follows: $\mathbb{P}(x \in S_i) = 2^{-i}$ independently for each $x \in X$. Defining $f_i(x) = d(x, S_i)$ as above, show that for a fixed pair $x, y \in X$,

$$\mathbb{E}\left[\sum_i |f_i(x) - f_i(y)|\right] \in \Omega(d(x,y)),$$

where the expectation is taken over the sets S_i.

Hint: Define sequences of suitable neighborhoods around x and y, and use part (c) to analyze the probability that S_i distinguishes a pair of x– and y–neighborhoods.

(e) Using the following concentration inequality, show that for $m \in \Theta(\log^4 n)$, you can find an f with

$$\|f(x) - f(y)\|_1 \in \Omega((\log^3 n) d(x,y)) \tag{4}$$

for all pairs $x, y \in X$. Note that (4), together with (2), implies $\log n$ distortion in ℓ_1.

> **Lemma** (Hoeffding)**.** *Let $z_1, \ldots, z_r$ be independent random variables such that for each i, $z_i \in [0, M]$ almost surely. Define $z = (1/r)\sum_i z_i$, and let $\mu = \mathbb{E}[z]$. Then*
>
> $$\mathbb{P}(\mu - z \geq t) \leq e^{-2rt^2/M^2}.$$

Hint: for each $j \in \{1, \ldots, r\}$, sample a sequence of $\log n$ subsets S_{ij} as in part (d). You need to show that for the right choice of r, the embedding simultaneously separates *all* pairs of points with $\Omega(1)$ probability.

(f) **Extra Credit** (*5 points*): By a more careful analysis of your construction in parts (d) and (e), you can prove the original statement of Bourgain's Theorem, with $m \in \Theta(\log^2 n)$. *Hint:* consider binary events $\chi_{ij} \in \{0, 1\}$ indicating whether S_{ij} separates your neighborhoods from part (d). Then apply Hoeffding's inequality.

Problem 3 (25 points)**.** In this problem, you will prove a version of the famous *Johnson-Lindenstrauss Lemma*, which states that n points in Euclidean space can be embedded with low distortion into $\mathbb{R}^{O(\log n)}$ with the ℓ_2 metric—note that this is $\log n$ better than Bourgain's result for arbitrary metrics. More precisely,

> **Lemma** (Johnson-Lindenstrauss)**.** *Let $X \subset \mathbb{R}^d$ consisting of n points. For any $\epsilon \in (0, 1)$ there exists a map $f : X \to \mathbb{R}^m$ with $m \in O(\epsilon^{-2} \log n)$ and such that*
>
> $$(1 - \epsilon)\|x - y\|_2^2 \leq \|f(x) - f(y)\|_2^2 \leq (1 + \epsilon)\|x - y\|_2^2.$$

(a) Constructing a suitable f by hand is difficult, but as in Problem 2, it turns out we can find one via sampling.

> **Lemma** (Distributional Johnson-Lindenstrauss). *Fix a dimension $d > 1$ and $\epsilon, \delta \in (0, 1/2)$. Then for $m \in O(\epsilon^{-2} \log(1/\delta))$, there exists a distribution over matrices $\Pi \in \mathbb{R}^{m \times d}$ such that for any $z \in S^{d-1} \subset \mathbb{R}^d$,*
>
> $$\mathbb{P}(|\|\Pi z\|_2^2 - 1| > \epsilon) < \delta.$$

Suppose such a distribution is given. Show that for a fixed set $X \subset \mathbb{R}^d$, if you sample Π enough times, you will eventually find a function $f(x) = \Pi x$ satisfying the JL Lemma.

(b) One simple choice of Π is as follows: let $\Pi_{ij} = m^{-1/2}\sigma_{ij}$, where σ_{ij} are independent Rademachers. (A *Rademacher* random variable σ is uniformly distributed over $\{-1, 1\}$.) Show that multiplication by Π preserves norms in expectation, i.e., $\mathbb{E}\|\Pi z\|_2^2 = 1$ for $z \in S^{d-1}$.

(c) Using the following tail bound, show that the random variable Π defined in part (b) fulfills the Distributional JL lemma:

> **Lemma** (Hanson-Wright). *Let $\sigma \in \mathbb{R}^n$ be a vector of independent Rademachers and $A \in \mathbb{R}^{n \times n}$. Then for some constants K, C,*
>
> $$\mathbb{P}(|\sigma^\top A\sigma - \mathbb{E}[\sigma^\top A\sigma]| > \lambda) \leq K\left(e^{-C\lambda^2/\|A\|_F^2} + e^{-C\lambda/\|A\|}\right), \tag{5}$$
>
> *where $\|A\|$ and $\|A\|_F$ denote the operator and Frobenius norms, respectively.*

Hint: The quantity $\|\Pi z\|_2^2 - 1$ is quadratic in the σ_{ij}. Find a symmetric matrix $A \in \mathbb{R}^{md \times md}$ depending only on z such that $\|\Pi z\|_2^2 = \sigma^\top A\sigma$, where $\sigma \in \mathbb{R}^{md}$ consists of the σ_{ij} rearranged into a vector. Then bound the norms of A. Finally, use the right-hand side of (5) to choose a suitable m.

(d) **Extra Credit** (*5 points*): A faster DJL transform is defined as follows:

$$\Pi = \frac{1}{\sqrt{m}} SHD,$$

where

- D is a diagonal matrix of Rademacher variables;
- H is a discrete Fourier transform, which can be computed via FFT in $O(d \log d)$ time; and
- $S \in \mathbb{R}^{m \times d}$ is a sampling matrix with each row uniformly distributed over the standard basis of $\mathbb{R}^d$.

Test the DJL property: sample a fixed set of n random points on S^{d-1} ($d = 10000, n \in [100, 1000]$ are reasonable choices). Then, for a variety of embedding dimensions m, sample 100–1000 values of Π and plot a histogram of the resulting distortion values. What happens to this distribution as you increase m?

Note: you may find the MATLAB functions `fft` and `randi` (or their Julia equivalents `FFTW.fft` and `rand`) useful.

Problem 4 (25 points)**.** In this problem, you will implement a dimensionality reduction algorithm and examine its behavior.

(a) *Maximum Variance Unfolding (MVU)* assigns new positions $\boldsymbol{y}_i \in \mathbb{R}^d$ to a set of n points $\mathbf{x}_i \in \mathbb{R}^D$ by solving (a relaxation of) the following optimization problem:

$$\begin{aligned} \max_{\mathbf{y}} \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \\ \text{subject to } \|\mathbf{y}_i - \mathbf{y}_j\| = \|\mathbf{x}_i - \mathbf{x}_j\| \quad (i,j) \in E \\ \sum_i \mathbf{y}_i = 0, \end{aligned} \tag{6}$$

where E is the edge set of a graph in which each point $\mathbf{x}_i$ and its k nearest neighbors in $\mathbb{R}^D$ form a $(k+1)$-clique.

Show that when $d = n$, (6) is equivalent to the following *semidefinite program*:

$$\begin{aligned} \max_{Y} \operatorname{trace} Y \\ \text{subject to } Y_{ii} + Y_{jj} - 2Y_{ij} = X_{ii} + X_{jj} - 2X_{ij} \quad (i,j) \in E \\ \mathbf{1}^\top Y \mathbf{1} = 0 \\ Y \succeq 0, \end{aligned} \tag{7}$$

where X is a symmetric positive definite matrix with entries $X_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$. How can you recover an embedding $\mathbf{y}$ from Y?

(b) What goes wrong in the formulations (6) and (7) if the graph is disconnected? How can you address this in practice?

(c) If the solution Y to (7) has rank $d < n$, show that the corresponding embedding $\mathbf{y}$ solves the original problem (6) for d. When might you expect this to happen in practice?

(d) Implement the SDP (7) in the file `mvu.m` or `mvu.jl`. You are encouraged to use CVX or `Convex.jl` with a solver such as Mosek or SCS. You can visualize your embedding by running the `runMVU` function. Try unfolding the provided "Swiss roll" dataset. What happens when you vary the number of nearest neighbors k? What about if you use the raw k nearest neighbor graph rather than the "union of $(k+1)$-cliques" graph? What about replacing the edgewise constraint by an inequality?

(e) **Extra Credit** (*5 points*): Try to find or create a dataset that breaks the algorithm, i.e., where MVU is unable to recover an embedding that matches the "intrinsic dimensionality" of the data. What goes wrong?