| **Data Science Through a Geometric Lens** | Fall 2019 |
|---|---|

# Homework 2

*Due: Wednesday 10/30/19, 5pm*

- **Solving 3 of the following 5 problems** will lead to full credit. You may attempt more than 3 problems, but the grading will be based on the 3 problems with the highest scores.
- Email the solutions to both the instructor and TA (emails listed on the course website).
- You may work in groups of size 1-3. If you do, please hand-in a single assignment with everyone's names on it. It is strongly encouraged to type up the solutions in Latex.
- If the question asks to prove something, you must write out a formal mathematical proof.
- If the question involves analyzing an algorithm, you must formally explain the time and/or space usage, along with the approximation guarantees (when applicable).
- When you are asked to prove a bound, it suffices to prove it up to multiplicative constants, i.e., using $O(\cdot)$, $\Theta(\cdot)$, or $\Omega(\cdot)$ notation. No need to optimize (multiplicative) constants!
- You may use other resources, but you must cite them. If you use any external sources, you still must provide a complete and self-contained proof/result for the homework solution.

## 1 Problem 1: Tales of different norms

(a) Prove that the following two relationships hold for any vector $x \in \mathbb{R}^n$:

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty \qquad \text{and} \qquad \|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2.$$

(b) Provide example vectors that satisfy each of the above four inequalities with an equality.

## 2 Problem 2: Bourgain for $\ell_1$

The strategy of Bourgain's embedding from Lecture 6 also works for $\ell_1$. Prove there is an embedding from any $n$-point metric space $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ to $k$-dimensional $\ell_1$ with $k = O(\log^2 n)$, which we denote as the map $f : \mathcal{X} \to \mathbb{R}^k$, where $f$ satisfies

$$d(x, y) \leq \|f(x) - f(y)\|_1 \leq c \log(n) \cdot d(x, y),$$

for any $x, y \in \mathcal{X}$, for some constant $c > 0$. Note that in the proof it suffices to show that

$$c' m \cdot d(x, y) \leq \|f(x) - f(y)\|_1 \leq m \log(n) \cdot d(x, y),$$

for a constant $c' > 0$ because you can divide the resulting vectors by $c'm$ to get the desired bound on the distortion.

# 3 Problem 3: Another view of Frechet

(a) Prove that Frechet's embedding from Lecture 6 provides an isometric embedding from an $n$-point metric space into $n$-dimensional $\ell_\infty$, which means that the distances are preserved exactly: $d(x, y) = \|f(x) - f(y)\|_\infty$.

(b) Improve the embedding from part (a) to only using $n - 1$ dimensions instead of $n$.

# 4 Problem 4: Good embeddings may or may not be possible

(a) Let $C_n$ denote the cycle graph: the vertices are $\{1, 2, \ldots, n\}$, and there are $n$ total edges, connecting $i$ and $i+1$ for $i \in \{1, 2, \ldots, n-1\}$, and also connecting $n$ and 1. The shortest path metric $d(i, j)$ on $C_n$ is the length of the shortest path in $C_n$ between vertices $i, j \in \{1, 2, \ldots, n\}$.

Show that any embedding of the shortest path metric on $C_n$ into $\mathbb{R}$ has distortion $\Omega(n)$. In other words, if $f : \{1, 2, \ldots, n\} \to \mathbb{R}$ satisfies $|f(i) - f(j)| \geq d(i, j)$ for all $1 \leq i, j \leq n$, then there must be some pair $i', j'$ with $|f(i') - f(j')| \geq cn \cdot d(i', j')$ for a constant $c > 0$ (where $c$ does not depend on $n$).

*Hint:* Consider three vertices on the cycle separated by distances roughly $n/3$.

(b) A tree metric $(X, d)$ is the shortest path metric on the vertices of a connected tree (that is, $d(x, y)$ is the length of the shortest path between vertices $x$ and $y$).

Letting $n = |X| \geq 2$ be the number of nodes in the tree, prove that a tree metric can be embedded with distortion 1 into $(n-1)$-dimensional $\ell_1$. In other words, show that there exists a mapping $f : X \to \mathbb{R}^{n-1}$ such that $d(x, y) = \|f(x) - f(y)\|_1$ for every $x, y \in X$.

*Hint:* Use induction on $n$ with base case $n = 2$.

# 5 Problem 5: Implementing Dimensionality Reduction

Implement and test the Johnson-Lindenstrauss dimensionality reduction method from Lecture 5.

(a) Find a dataset of $n \geq 100$ points, either randomly generated or from a public repository (e.g., UCI, ScikitLearn, etc).

(b) Provide results (in a table or plot, clearly labeled) for the distortion of the projected points versus the original points, as you increase the dimensionality of the embedded points.

(c) For the same dataset and parameter settings, replace the normal distribution with $\pm 1$ random variables. How does the embedding change (better, worse, different, ...)?