Modifizierte Wassersteindistanz

Um die Wasserstein Distanz für nicht gleichgroße Knotenmengen zu erweitern, betrachten wir folgendes für X und Y zwei Punktmengen oberhalb der Diagonalen $\Delta = \{(x,x) \in \mathbb{R}^2 | x \in \mathbb{R}\}$, mit $|X| < |Y|$:

$$W(X,Y) := \inf_{\varphi: X \hookrightarrow Y} \left(\sum_{x \in X} \|x - \varphi(x)\| + \sum_{y \in Y \setminus im(\varphi)} \gamma(y) \right)$$

wobei φ injektiv sein soll, und man $\gamma : \mathbb{R}^2 \to \mathbb{R}$ z.B. definieren als die "Lebensdauer" des Punktes, also $\gamma(x,y) = y - x$. Alternativ auch der minimale Abstand von dem Punkt (x,y) zu der Diagonalen Δ. Allerdings ist dieser Abstand einfach $\frac{y-x}{\sqrt{2}}$, also ein Skalar von dem anderen Abstand. Dieser Abstand $W(X,Y)$ lässt sich interpretieren als die normale Wasserstein Distanz, wobei alle Punkte in der Zielmenge, die nicht im Bild von φ enthalten sind, getroffen werden von einem Punkt auf der Diagonalen, die keinen Mehrwert für die persistente Homologie haben.

Diese Alternative respektiert die Wichtigkeit jedes einzelnen Punktes, insbesondere den "lange lebenden". Die Wasserstein Distanz so zu erweitern ergibt natürlich nur im Zusammenhang mit persistenter Homologie Sinn. Eine andere Methode, die ich aber schon verworfen habe, ist erst auf der Menge Y ein $|X|$-Clustering C zu finden, und dann über die Mappings $X \to C$ zu optimieren. Hierbei verliert man aber wichtige Details, wenn zwei Punkte die "lange leben", und zeitgleich geboren werden z.B. zu einem Cluster werden. Dadurch verliert man eine wichtige Information über einen Erzeuger der ersten Homologie.

Effiziente Berechnung

Das ganze lässt sich polynomiell lösen. Hierbei reduziert man das Problem auf eine min-cost-flow-Instanz. Gegeben X mit $|X| = k$ und Y mit $|Y| = n > k$. Dann definieren wir einen Graphen $G = (V,E)$, wobei $V = X \cup \{s,t,h\}$. s,t und h sind hierbei Hilfsknoten. Die Menge der Kanten besteht aus der Vereinigung von $\{(s,a) | a \in X \cup \{h\}\}$, $\{(a,b) | a \in X \cup \{h\}, b \in Y\}$ und $\{(b,t) | b \in Y\}$. Die Kapazitäten $u : E \to \mathbb{R}$ definieren wir hierbei als $n - k$ für alle Kanten der Form (h,y) mit $y \in Y$ und die Kante (s,h), und 1 für alle anderen Kanten. Die Kosten für Kanten der Form (x,y) mit $x \in X$ und $y \in Y$ sei $\|x - y\|$. Die Kosten für Kanten der Form (h,y) mit $y \in Y$ sei der oben beschriebene Abstand $\gamma(y)$. Alle anderen Kosten sind 0. Und Flow-Variablen $b(s) = -b(t) = n$, $b(v) = 0$ für alle anderen v. Gesucht wird jetzt ein min-cost-flow. Dieser ist optimal lösbar in polynomieller Zeit durch z.B. einen Algorithmus wie Cycle-Cancling. Weil die Kapazitäten alle ganzzahlig sind, ist die Lösung auch ganzzahlig, und somit können wir einfach gucken welche Kanten ausgewählt wurden, welches dann die optimale Abbildung φ bestimmt. Sogar genauer, $W(X,Y)$ ist dann genau gleich dem Gewicht der optimalen Lösung. Hier beispielhaft der Graph G ohne Kosten und Kapazitäten für $k = 2$ und $n = 4$.