Statistical Aspects of Wasserstein Distances

Victor M. Panaretos*and Yoav Zemel†

June 8, 2018

【数学】Wasserstein Distance

张楚珩
清华大学 交叉信息院博士在读

今天来学一个数学知识，Wasserstein Distance。如果听说过WGAN的话，里面的W就是代表Wasserstein。

参考资料

Panaretos, Victor M., and Yoav Zemel. "Statistical aspects of Wasserstein distances." Annual Review of Statistics and Its Application (2018).

Optimal Transport and Wasserstein Distance (slides)

特色

常见的有很多衡量概率分布差异的度量方式，比如total variation（TRPO推导里面有用到），还有经常被用到的KL散度。相比于这些度量方式，Wasserstein距离有如下一些好处。

- 能够很自然地度量离散分布和连续分布之间的距离；
- 不仅给出了距离的度量，而且给出如何把一个分布变换为另一分布的方案；
- 能够连续地把一个分布变换为另一个分布，在此同时，能够保持分布自身的几何形态特征；

过程

1. 其他距离度量的缺陷

首先注意到KL散度不是距离度量，它不满足对称性。常见的距离度量有

$$\text{Total Variation}: \sup_A |P(A)-Q(A)| = \frac{1}{2}\int |p-q|$$

$$\text{Hellinger}: \sqrt{\int (\sqrt{p}-\sqrt{q})^2}$$

$$L_2: \int (p-q)^2$$

$$\chi^2: \int \frac{(p-q)^2}{q}.$$

- 这些距离度量没法衡量离散分布和连续分布之间的距离：假设 p 是均匀分布 $U[0,1]$ 的概率密度，q 是离散均匀分布 $\{0,1/N,\cdots,1\}$ 的概率密度。其total variation等于1，即完全不相似，但是凭感觉上来说，它们两个是很相似的。在Wasserstein距离度量下，它们的距离为 $1/N$ ，这看起来就比较合理了。
- 这些距离都忽略了概率分布之间的几何特性：它们几乎都有一个共同的特征，那就是都是对应点的概率密度函数相比较，这会忽略其几何特性。比如图1中，左边的分布应该离中间的分布更近，而中间的离右边的更远，但是其他度量无法反应这个特性，但Wasserstein距离可以。

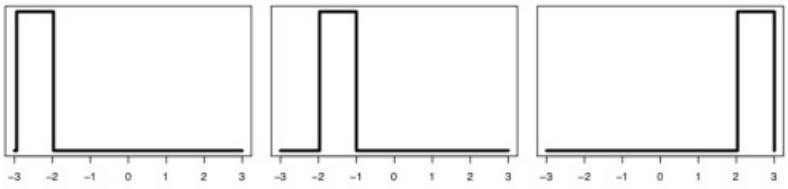


Figure 1: *Three densities p_1, p_2, p_3. Each pair has the same distance in L_1, L_2, Hellinger etc. But in Wasserstein distance, p_1 and p_2 are close.*

图1

- 基于Wasserstein距离可以找出Wasserstein平均（Wasserstein barycenter），相比于欧式平均（Eulidean average）来说，它更能够描述其形态特征，如图2所示。

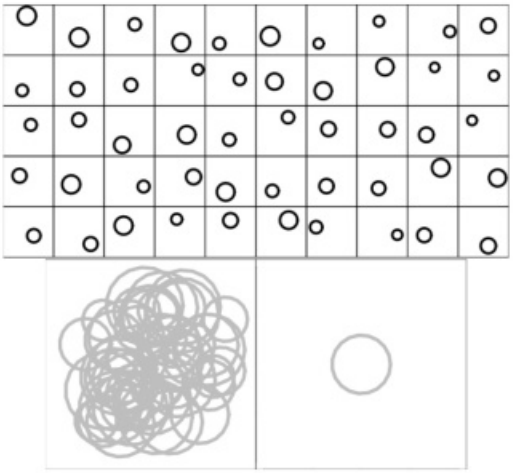


Figure 2: *Top: Some random circles. Bottom left: Euclidean average of the circles. Bottom right: Wasserstein barycenter.*

图2

- Wasserstein距离不仅告诉两个分布之间的距离，而且能够告诉我们它们具体如何不一样，即如何从一个分布转化为另一个分布。如图3所示，Wasserstein能够告诉我们每一份probability density的转移方案。

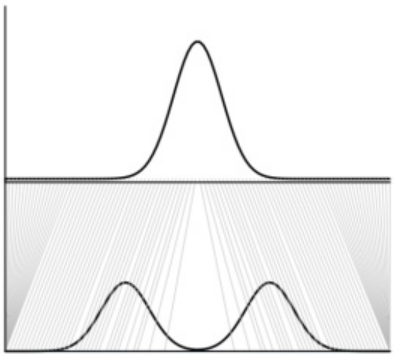


Figure 3: *Two densities p and q and the optimal transport map to that morphs p into q.*

图3

- 这个转化过程还可以做成一个连续的过程，可以把A分布连续转化为B分布，并且这个转化过程是能够保持其几何特征的，如下面两图所示。

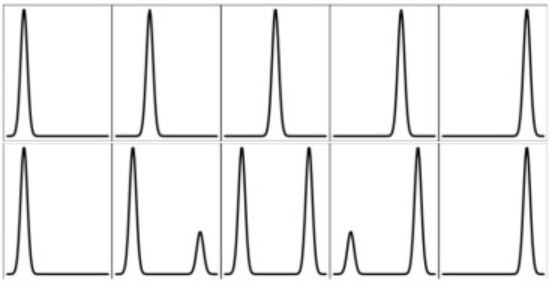


Figure 5: *Top row: Geodesic path from P_0 to P_1. Bottom row: Euclidean path from P_0 to P_1.*

Figure 6: *Morphing one image into another using the Wasserstein geodesic. Image credit: Bauer, Joshi and Modin 2015.*

图4

2. Wasserstein距离定义

Wasserstein距离的起源是optimal transport problem，把概率分布想象成一堆石子，如何移动一堆石子，通过最小的累积移动距离把它堆成另外一个目标形状，这就是optimal transport所关心的问题。

首先，要能完成这个操作，先要确保本来的这一堆石子的总质量要和目标石子堆总质量一样；考虑到概率分布的归一化条件，这一点是自然被满足的。

其次，我们暂时假设石子都是很小的，无限可分的；毕竟如果一个大石块要求堆成两座山仅仅通过移动肯定没法做到。（后面会有另外的定义方式来把大石块“劈开”）

假设地面上 $\mathcal{X}=\mathbb{R}^2$ 堆了一些石子，石子的分布我们用 $\mu:\mathcal{X}\to\mathbb{R}$ 来表示，采取这样的表示方法对于地面上的任意一块面积 $A\subseteq\mathcal{X}$ ， $\mu(A)$ 表示这块面积上放置了质量为多少的石子。同样的我们可以定义目标石子堆的分布 ν 。定义一个输运方案 $T:\mathcal{X}\to\mathcal{X}$ 把现有的石子堆变成目标石子堆。 $T(A)=B$ 表示把原来放在A处的石子都运到B处放好，类似地可以定义反函数 $T^{-1}(B)=A$ 。该输运方案成立需要满足 $\nu(B)=\mu(T^{-1}(B)),\forall B\subseteq\mathcal{X}$ ，即任意位置的石子通过输运过后都刚好满足分布 μ 的要求。这也可以写为 $T\#\mu=\nu$ 。

由此，两堆石子之间的距离可以被定义成把一堆石子挪动成另外一堆所需要的最小输运成本

$$W_p(\mu,\nu)=\left(\inf_{T:T\#\mu=\nu}\int_{\mathcal{X}}\|x-T(x)\|^p\mu(x)dx\right)^{1/p}$$

遇到石子不可分的情况上述的定义就不再保证能存在可行的输运方案了，比如 $\mathcal{X}=\mathbb{R}$ ， $\mu(x)=\delta(x)$ ， $\nu(x)=\frac{1}{2}\delta(x-1)+\frac{1}{2}\delta(x+1)$ ，这样位于 $x=0$ 的质量为 1 的“大石头”又不能劈开，那应该运到哪个位置呢？下面引入一个新的定义来解决这个问题。

下面的这个方案就是认为每个位置的石子都能够分开并且按照比例输运到不同的位置，先写出其定义

$$W_p(\mu,\nu)=\left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{X}}\|x-y\|^p d\gamma(x,y)\right)^{1/p}$$

其中 γ 是一个联合概率分布，称coupling，它要求其边缘分布刚好是 μ 和 ν ，即 $\gamma(A\times\mathcal{X})=\mu(A)$ ， $\gamma(\mathcal{X}\times B)=\nu(B)$ ；如下图所示，它能够表示分布 μ 上的某个位置的质量被拆开（红竖线），然后再按照权重被分配给目标分布（红箭头）。

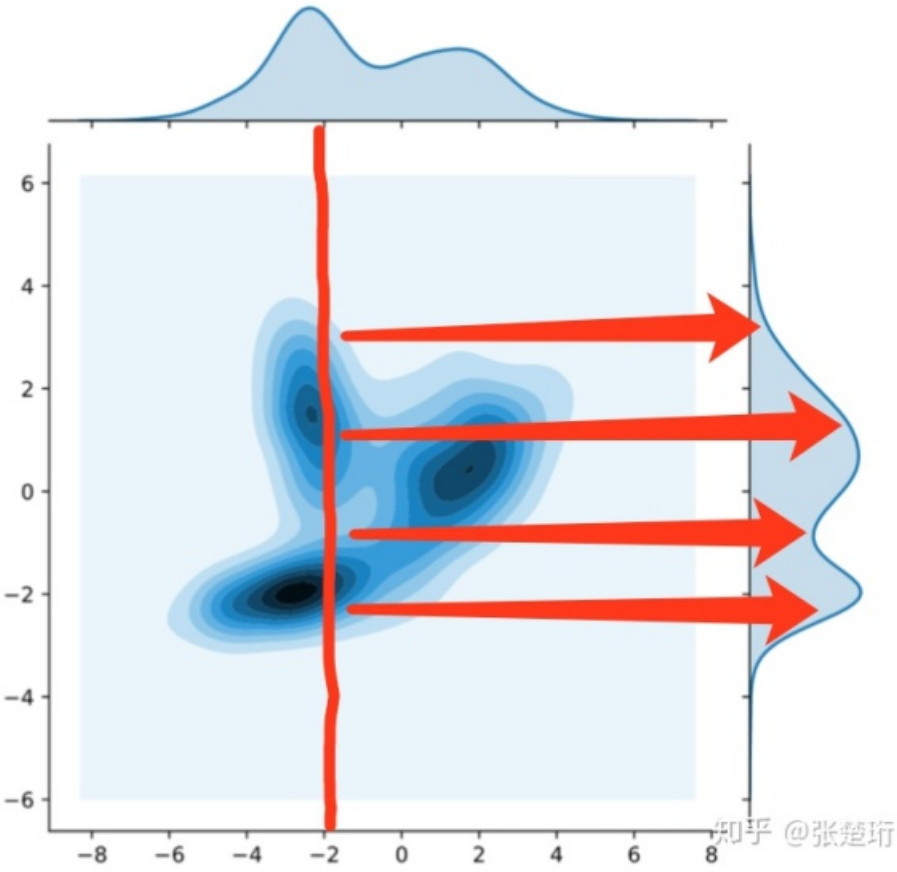


图5

要注意到，图中画的这种联合概率分布是很难实际成为真实的optimal coupling的，一般来讲optimal coupling都会比较稀疏。排除少数不可分的情况，大多数的optimal coupling都是稀疏的，即 $\gamma(A \times B) = \mu(A \cap T^{-1}(B))$ 。

3. 对偶形式

最小化输运成本也可以写成对偶形式

$$\sup_{\phi,\psi}\Big\{\mathbb{E}\phi(X)+\mathbb{E}\psi(Y)\Big\},\qquad \text{subject to}\quad \phi(x)+\psi(y)\le\|x-y\|^p$$

注意到可以利用后面的约束关系推到类似前面定义的形式，它有弱对偶（weak duality）性质，即这个式子的最大值不超过Wasserstein距离定义中的最小值。

现在来考虑 $p=2$ 的形式，考虑到

$$\|x-y\|^2=\|x\|^2+\|y\|^2-2\langle x,y\rangle$$

做替换 $\varphi(x)=\|x\|^2-\phi(x)$ ， $\Psi(y)=\|y\|^2-\psi(y)$ ，可以得到如下形式

$$\inf_{\varphi,\Psi}\Big\{\mathbb{E}\varphi(X)+\mathbb{E}\Psi(Y)\Big\},\qquad \text{subject to}\quad \varphi(x)+\Psi(y)\ge\langle x,y\rangle\,.$$

现在考虑把 $\Phi(y)$ 替换掉，这样就只有一个变量，在最优情况下有 $\Phi=\varphi^*$ 以及 $\varphi=\varphi^{**}$ ，其中

$$\varphi^*(y)=\sup_{x\in\mathcal{X}}\{\langle x,y\rangle-\varphi(x)\},$$

是 $\varphi(x)$ 的 Fenchel conjugate function。这样

- 优化问题就转变为了unconstrained问题， $\inf_{\varphi}\mathbb{E}[\varphi(X)+\varphi^*(Y)]$ ；
- $\varphi=\varphi^{**}$ 表明 φ 是一个凸函数；
- $y=\nabla\varphi(x)$ ，注意到这刚好是输运函数 $y=T(x)$ 的形式，即输运函数为 φ 的梯度；

另外注意到，Wasserstein距离不仅可以对两个分布来计算，也可以对从两个分布里面采样得到的样本来计算，如下示意图。

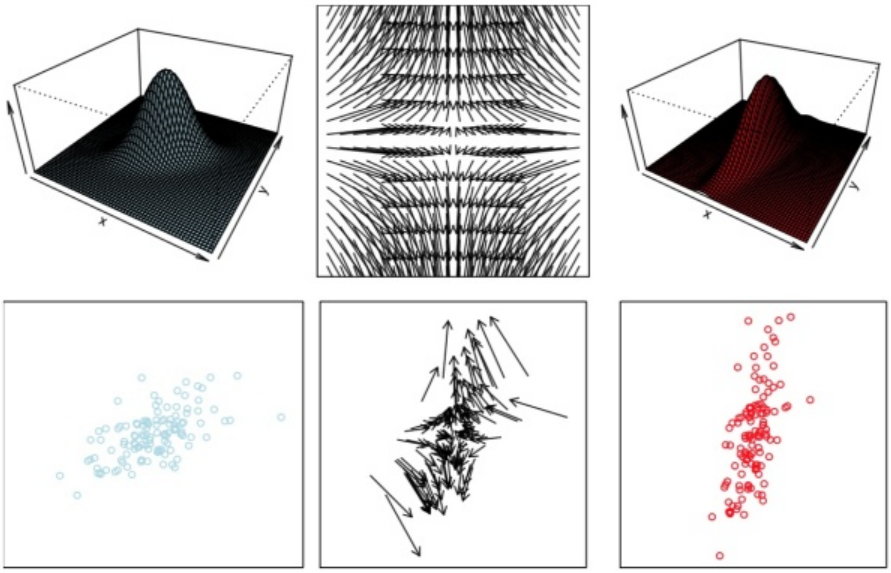


Figure 1: Illustration of the "analytic" and "probabilistic" definitions. The top row of plots shows the densities of two Gaussian probability measures μ (on the left, in blue) and ν (on the right, in red), and the optimal deterministic map T (in the middle) that deforms μ into ν, i.e., $T\#\mu = \nu$. The map is plotted in the form of the vector field $T(x) - x$, where each arrow indicates the source and destination of the mass being transported. Reversing the direction of the arrows would produce the inverse map, optimally deforming the measure ν to obtain μ. The bottom row features two independent random samples $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \mu$ (on the left, in blue) and $Y_1, \ldots, Y_N \overset{\text{i.i.d.}}{\sim} \nu$ (on the right, in red), for $N = 120$. The sample $\{X_i\}_{i=1}^N$ was constructed by sampling μ directly. The sample $\{Y_i\}_{i=1}^N$ was constructed by applying the optimal map T to the sample $\{X_i\}_{i=1}^N$, i.e. $Y_i = T(X_i)$. The plot in the middle illustrates how the sample $\{X_i\}_{i=1}^N$ is re-arranged in order to produce the sample $\{Y_i\}_{i=1}^N$, by plotting the vectors $T(X_i) - X_i$. The optimality of T can be understood in terms of minimising the average squared length of these arrows. In all plots, the x and y axes range from -3 to 3.

图6

4. 从一个分布渐变到另一个分布（Geodesics）

找一个分布的路径 $c(t), t \in [0,1]$ 从 μ 到 ν，其中 $c(0) = \mu$， $c(1) = \nu$，如果对于任意的 t，$W_p(c(0), c(t)) + W_p(c(t), c(1)) = W_p(c(0), c(1))$，那么这条路径就是最短路径。举例来说就是本文的图4。

5. 求多个分布的平均（Barycenters）

对于多个分布 $P_j, j \in [N]$，其Wasserstein Barycenter就是 $\arg\min_P \sum_{j=1}^{N} W_p(P, P_j)$。连续情况下就是这样。

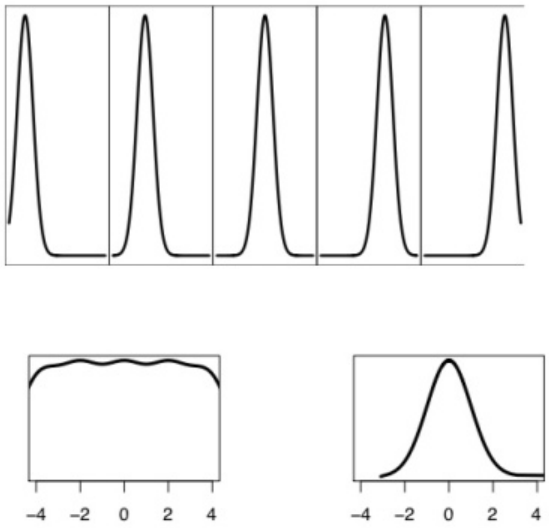


Figure 7: *Top: Five distributions. Bottom left: Euclidean average of the distributions. Bottom right: Wasserstein barycenter.*

离散情况下就是

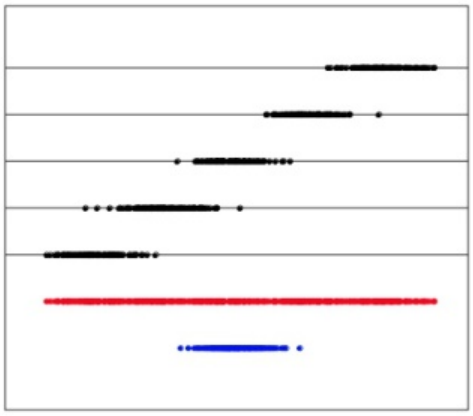


Figure 8: *The top five lines show five, one-dimensional datasets. The red points the what happens if we simple average the give empirical distributions. The blue dots show the Wasserstein barycenter which, in this case, can be obtained simply by averaging the order statistics.*

6. Wasserstein距离的数值计算

Wasserstein距离是很难计算的，这成为了一个限制其应用的难点。目前能显示计算出来的只有两种情况，一种是维度 $d=1$ 的情况，另一种是高斯分布。

$d=1$ 的情况下，

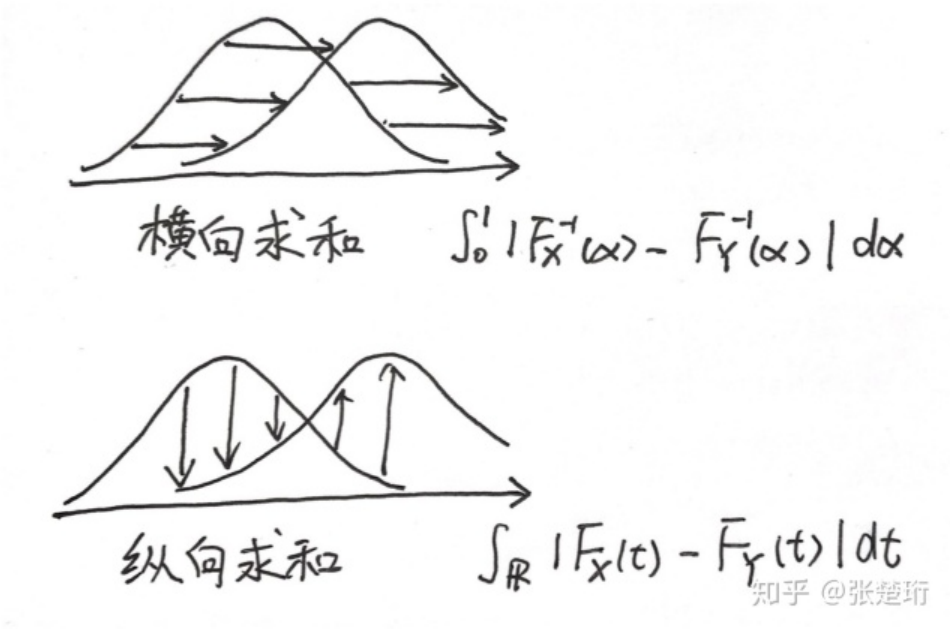
$$W_p(X,Y) = \|F_X^{-1} - F_Y^{-1}\|_p = \left(\int_0^1 |F_X^{-1}(\alpha) - F_Y^{-1}(\alpha)|^p \, \mathrm{d}\alpha \right)^{1/p}, \qquad \mathbf{t}_X^Y = F_Y^{-1} \circ F_X,$$

这里的 $\mathbf{t}_X^Y = \mathbf{t}_\mu^\nu = T$ 就是前面一直提到的输运函数。这种情况下能求解也很好理解，毕竟如果只有一个维度，那么输运方案就是顺着初始分布和目标分布上各个“石子”的序关系一一对应就可以了。

特殊地， $d=1, p=1$ 的情况下，还可以简化为

$$W_1(X,Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)| \, \mathrm{d}t.$$

这相比于上面的式子相当于是把前面的横向的求和换成了纵向的求和。



高斯分布的情况下有

$$W_2^2(X,Y) = \|m_1 - m_2\|^2 + \mathrm{tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}],$$
$$\mathbf{t}_X^Y(x) = m_2 + \Sigma_1^{-1/2}[\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}]^{1/2}\Sigma_1^{-1/2}(x - m_1),$$

关于Fenchel conjugate可以参考我导师的课件 sealzhang.tk/assets/fil...

下图说明了原函数与对偶函数的关系，注意到图中不同的颜色是一一对应的，具体见课件。

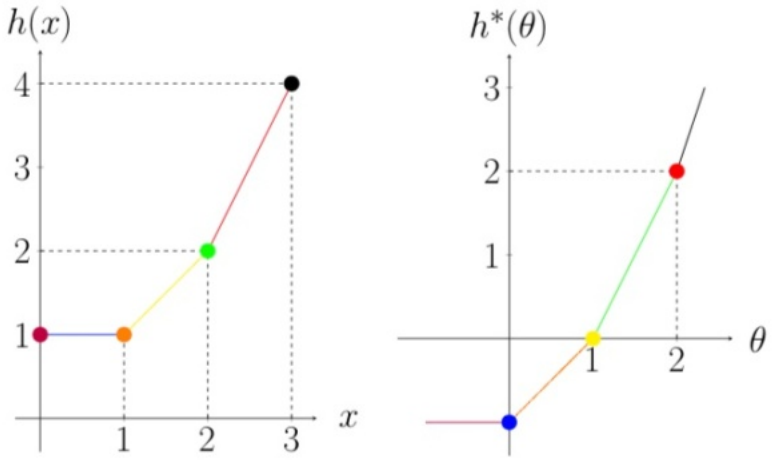


Figure 1: Conjugate Function Encodes the Tangent Planes

$$h(x) = \begin{cases} 1, & 0 \le x \le 1 \\ x, & 1 < x \le 2 \\ 2x-2, & 2 < x \le 3 \\ +\infty, & \text{otherwise} \end{cases} \qquad h^*(\theta) = \begin{cases} -1, & x \le 0 \\ x-1, & 0 < x \le 1 \\ 2x-2, & 1 < x \le 2 \\ 3x-4, & x > 2 \end{cases}$$

编辑于 2019-08-20

数学