Supplementary Document

Proof of Remark 3.3. It is shown in page 73 of (Villani, 2003) that given two probability measures $\alpha, \beta \in \mathcal{P}(\mathbb{R})$, we have

$$d_{W,1}(\alpha,\beta) = \int_{-\infty}^{\infty} |F(t) - G(t)| dt,$$

where F and G are the cumulative distribution functions of α and β respectively. Now let us also use F and G to represent the cumulative distribution functions of $m_\alpha^{(\varepsilon)}(x)$ and $m_\alpha^{(\varepsilon)}(x')$ respectively. Then we have explicitly

$$F(t) = \begin{cases} 0, & t \le x-\varepsilon \\ \frac{\alpha((x-\varepsilon,t])}{\alpha((x-\varepsilon,x+\varepsilon))}, & x-\varepsilon \le t < x+\varepsilon \\ 1, & x+\varepsilon \le t \end{cases}$$

and

$$G(t) = \begin{cases} 0, & t \le x'-\varepsilon \\ \frac{\alpha\left((x'-\varepsilon,t]\right)}{\alpha((x'-\varepsilon,x'+\varepsilon))}, & x'-\varepsilon \le t < x'+\varepsilon \cdot \\ 1, & x'+\varepsilon \le t \end{cases}$$

Now suppose ε is small enough such that $x+\varepsilon \le x'-\varepsilon$, then we have $F = G$ on $(-\infty, x-\varepsilon] \cup [x'+\varepsilon, \infty)$ and thus

$$\begin{aligned}
&\int_{-\infty}^{\infty} |F(t) - G(t)| dt \\
=&\left(\int_{-\infty}^{x-\varepsilon} + \int_{x-\varepsilon}^{x'+\varepsilon} + \int_{x'+\varepsilon}^{\infty}\right) |F(t) - G(t)| dt \\
=&\int_{x-\varepsilon}^{x'+\varepsilon} |F(t) - G(t)| dt \\
=&\int_{x-\varepsilon}^{x'-\varepsilon} F(t) dt + \int_{x'-\varepsilon}^{x'+\varepsilon} |1 - G(t)| dt \\
=&x' - x + \underbrace{\int_{x-\varepsilon}^{x+\varepsilon} \frac{\alpha\left((x-\varepsilon,t]\right)}{\alpha\left((x-\varepsilon,x+\varepsilon)\right)} dt}_{H(x)} \\
&- \underbrace{\int_{x'-\varepsilon}^{x'+\varepsilon} \frac{\alpha\left((x'-\varepsilon,t]\right)}{\alpha\left((x'-\varepsilon,x'+\varepsilon)\right)} dt}_{H(x')}
\end{aligned}$$

Now recalling that α has density function f, and using its Taylor expansion, we have

$$\begin{aligned}
H(x) &= \frac{\int_{x-\varepsilon}^{x+\varepsilon}\int_{x-\varepsilon}^{t} f(s)\, ds\, dt}{\int_{x-\varepsilon}^{x+\varepsilon} f(s) ds} = \frac{\int_{-\varepsilon}^{\varepsilon}\int_{-\varepsilon}^{t} f(x+s)\, ds\, dt}{\int_{-\varepsilon}^{\varepsilon} f(x+s) ds} \\
&= \frac{\int_{-\varepsilon}^{\varepsilon}\int_{-\varepsilon}^{t} \left(f(x) + f'(x)s + \frac{f''(x)}{2}s^2 + O(s^3)\right) ds\, dt}{\int_{-\varepsilon}^{\varepsilon} \left(f(x) + f'(x)s + \frac{f''(x)}{2}s^2 + O(s^3)\right) ds} \\
&= \frac{2\varepsilon^2 f(x) + \frac{2\varepsilon^3}{3} f'(x) + \frac{\varepsilon^4}{3} f''(x) + O(\varepsilon^5)}{2\varepsilon f(x) + \frac{\varepsilon^3}{3} f''(x) + O(\varepsilon^4)} \\
&= \varepsilon - \frac{f'(x)}{3f(x)}\varepsilon^2 + O(\varepsilon^3).
\end{aligned}$$

Similarly,

$$H(x') = \varepsilon - \frac{f'(x')}{3f(x')}\varepsilon^2 + O(\varepsilon^3).$$

Therefore

$$\begin{aligned}
d_\alpha^{(\varepsilon)}(x,x') &= d_{W,1}\left(m_\alpha^{(\varepsilon)}(x), m_\alpha^{(\varepsilon)}(x')\right) \\
&= x' - x + H(x) - H(x') \\
&= d_\alpha^{(\varepsilon)}(x,x') \\
&= x' - x + \frac{1}{3}\left[\frac{f'(x')}{f(x')} - \frac{f'(x)}{f(x)}\right]\varepsilon^2 + O(\varepsilon^3).
\end{aligned}$$

□

Details about Grassmann manifolds data

In this subsection we describe the setting of our experiments on the data described in (Cetingul & Vidal, 2009).

Description of the Grassmannian as a Metric Space

The *Grassmannian* $\mathcal{G}_{k,m-k}$ is a set consisting of all k-dimensional subspaces of $\mathbb{R}^m$. In particular, $\mathcal{G}_{1,m}$ denotes the projective space $\mathbb{R}\mathrm{P}^m$. To induce a natural metric on this set, we will consider an alternative description as follows. Suppose $w \in \mathcal{G}_{k,m-k}$ is a k-dimensional subspace of $\mathbb{R}^m$, then we can choose k orthonormal vectors in w and form a $m \times k$ matrix W. Consider the $m \times m$ matrix $P = WW^\top$, then it is easy to show that P is invariant under orthonormal transformations of W. Hence we have an equivalent definition of Grassmannian which regards it as a submanifold in $\mathbb{R}^{m\times m}$ (Hüper et al., 2010):

$$\mathrm{G}_{k,m} := \{P \in \mathbb{R}^{m\times m} |\, P^\top = P, P^2 = P, \mathrm{tr}(P) = k\}.$$

This is the manifold of rank k symmetric projection operators of $\mathbb{R}^m$. We will use the equivalence between $\mathcal{G}_{k,m-k}$ and $\mathrm{G}_{k,m}$. Since $\mathrm{G}_{k,m}$ is a manifold embedded in $\mathbb{R}^{m\times m}$, we can endow it with the restriction of the Euclidean metric. For any two points $P, Q \in \mathrm{G}_{k,m-k}$, we can write $P = XX^\top, Q = YY^\top$ for some $m \times k$ matrices X and

Y with orthogonal column vectors. The intrinsic distance between P, Q is given by

$$d_I^2(P,Q) = 2\,\mathrm{tr}\big(\arccos^2\big(\sqrt{Y^\top XX^\top Y}\big)\big).$$

The extrinsic distance (Euclidean distance) between them is

$$\begin{aligned} d_E^2(P,Q) &= \|P-Q\|_F^2 = 2k - 2\,\mathrm{tr}(PQ) \\ &= 2k - 2\,\mathrm{tr}(Y^\top XX^\top Y). \end{aligned}$$

Following Formula (9) in (Cetingul & Vidal, 2009), in the remainder we always use the metric $d := \frac{d_E}{\sqrt{2}}$ as the metric on the Grassmannian. We call this metric as the **manifold distance.**

To obtain an explicit description of the smooth structure and tangent spaces of the Grassmannian, we can adopt the following alternative perspective. We can regard $\mathcal{G}_{k,m-k}$ as the quotient manifold $\mathcal{V}_{k,m}/\mathrm{GL}_k(\mathbb{R})$, where $\mathcal{V}_{k,m} = \{X \in \mathbb{R}^{m\times k} | \,\mathrm{rank}(X) = k\}$ is the *noncompact Stiefel manifold* and $\mathrm{GL}_k(\mathbb{R})$ is the group of $k \times k$ invertible real matrices (Absil et al., 2004). $\mathcal{V}_{k,m}$ is an open subset of $\mathbb{R}^{m\times k}$, so the tangent space of $\mathcal{V}_{k,m}$ at $X \in \mathcal{V}_{k,m}$ is trivial, i.e. $T_X\mathcal{V}_{k,m} = \mathbb{R}^{m\times k}$. As a quotient manifold, we can regard the tangent space of $\mathcal{G}_{k,m-k}$ as a quotient of $\mathbb{R}^{m\times k}$. Hence we can represent the tangent vectors at $[X] \in \mathcal{G}_{k,m-k}$ by $m \times k$-matrices (sometimes we also use X to denote an element of $\mathcal{G}_{k,m-k}$). In fact, let $\pi : \mathcal{V}_{k,m} \to \mathcal{V}_{k,m}/\mathrm{GL}_k(\mathbb{R}) = \mathcal{G}_{k,m-k}$ be the canonical projection and $X \in \mathcal{V}_{k,m}$, then we can identify $T_{[X]}\mathcal{G}_{k,m-k}$ with a subspace H_X of $T_X\mathcal{V}_{k,m}$, where

$$H_X = \{X^\perp K : \, K \in \mathbb{R}^{(m-k)\times k}\},$$

and $X^\perp$ is an $m \times (m-k)$ matrix, denoting an orthogonal complement of X . Equivalently, we can define $H_X = \{\Delta \in \mathbb{R}^{m\times k} : \, X^\top \times \Delta = 0\}$.

In practice, we represent an element w of $\mathcal{G}_{k,m-k}$ by an $m \times k$ matrix X, such that the columns of X constitute an orthonormal basis of w. For a tangent vector $\Delta \in T_{[X]}\mathcal{G}_{k,m-k}$, the exponential map is of the form

$$\exp_{[X]}(\Delta) = [XV\cos(\Sigma) + U\sin(\Sigma)]V^\top,$$

where $U\Sigma V^\top$ is the singular value decomposition of Δ and the right hand side is a representative of a k-dimensional subspace.

Mean Shift on the Grassman Manifold and Experimental Setup

With the help of the explicit formula of exponential map, one is able to carry out mean shift method on Grassmann manifolds. In (Cetingul & Vidal, 2009), an extrinsic mean shift method was proposed as follows: given a set of points $\mathcal{X} = \{X_n\}_{n=1}^N \subset \mathcal{G}_{k,m-k}$, and some kernel-related function ψ with bandwidth ε, we update the points by

$$X^{(i+1)} = \exp_{X^{(i)}}\big(m(X^{(i)};\varepsilon)\big),$$

where $m(X^{(i)};\varepsilon)$ is defined as

$$-\frac{\sum_{n=1}^N \nabla_{X^{(i)}} d^2(X^{(i)}, X_n)\psi\big(d^2(X^{(i)}, X_n);\varepsilon\big)}{\sum_{n=1}^N \psi\big(d^2(X^{(i)}, X_n);\varepsilon\big)}.$$

Notice that $\nabla_X d^2(X, X_n)$ on Grassmannian is nothing but

$$\nabla_X d^2(X, X_n) = -2(I_m - XX^\top)X_nX_n^\top X.$$

In our experiments, we carry out the mean shift with respect to Gaussian kernels as comparison. More precisely, we take $\psi(x^2;\varepsilon) = C_\varepsilon \exp(-\frac{x^2}{(\frac{2}{3}\varepsilon)^2})$, where C_ε is the normalization coefficient.

We also carry out the mean shift with respect to the truncation kernels in a slightly different formula from (Cetingul & Vidal, 2009): we first pull back the points on the manifold to the tangent space, do the mean shift method on tangent space (this is an Euclidean space), and then map the points back to the manifold by the exponential map, whose explicit formula is given above. To pull back the points on the manifold to the tangent space, we will need an explicit formula of log, the inverse of exponential map (Subbarao & Meer, 2009):

$$\forall X, Y \in \mathcal{G}_{k,m-k}, \quad \log_X(Y) = U\arcsin(S)V^\top,$$

where, $USD^\top = Y - XX^\top Y$ and $VCD^\top = X^\top Y$ is the generalized SVD with $C^\top C + S^\top S = I$. A precise procedure for the mean shift variant is then described as follows:

$$X_k^{(i+1)} = \exp_{X_k^{(i)}}\big(m(X_k^{(i)};\varepsilon)\big), \quad \forall k = 1, 2, \cdots, N$$

where

$$m(X_k^{(i)};\varepsilon) = \frac{\sum_{X_n^{(i)} \in \mathrm{B}\left(X_k^{(i)},\varepsilon\right)} \log_{X_k^{(i)}}\left(X_n^{(i)}\right)}{\#\mathrm{B}\left(X_k^{(i)},\varepsilon\right)},$$

where as mentioned bove, the ball is determined by the metric d.

In practice, we have another parameter K besides ε. This means that ε neighborhoods are trimmed whenever their cardinality exceeds K: in such cases only the K closest points to the center of the ball are considered. This has the effect of limiting the size of the OT problems one has to solve in practice to $K \times K$ (as the size the of the distance matrix involved). In our experiments we fixed this parameter to the value 200.