An Invitation to Statistics in Wasserstein Space

Shen Ziliang 2021213218

Shanghai University of Finance and Economics

February 26, 2022

Outline

Content

Although the Wasserstein space $\mathscr{W}_p(\mathscr{X})$ is non-linear in terms of measures, it is linear in terms of maps. Indeed, if $\mu \in \mathscr{W}_p(\mathscr{X})$and $T_i : \mathscr{X} \to \mathscr{X}$ are such that $\|T_i\| \to L_p(\mu)$ then $(\alpha T_1 + \beta T_2)\#\mu \in \mathscr{W}_p(\mathscr{X})$ for all $\alpha, \beta \in \mathbb{R}$ We assume here that $\mathscr{X}$ is a Hilbert space and that $p = 2$; the results extend to any $p > 1$. Absolutely continuous measures are assumed to be so with respect to Lebesgue measure if $\mathscr{X} = \mathbb{R}^d$ and otherwise refer to Definition 1.6.4.

Definition 1.6.4:Gaussian Null Sets and Absolutely Continuous Measures

A subset $\mathscr{N} \subseteq \mathscr{X}$ is a Gaussian null set if whenever ν is a nondegenerate Gaussian measure, $\nu(\mathscr{N}) = 0$. A probability measure $\mu \in P(\mathscr{X})$ is absolutely continuous if μ vanishes on all Gaussian null sets.

Definition:Wasserstein space

Let X be a separable Banach space. The p-Wasserstein space on X is defined as:

$$\mathscr{W}_p(\mathscr{X}) = \{\mu \in P(\mathscr{X}) : \int_{\mathscr{X}} \|x\|^p \mathrm{d}\mu(x) < \infty\}, p \geq 1$$

We will sometimes abbreviate and write simply $\mathscr{W}_p$ instead of $\mathscr{W}_p(\mathscr{X})$.

Geodesics in $\mathscr{W}_2(\mathscr{X})$

Let $\gamma \in \mathscr{W}_2(\mathscr{X})$ be absolutely continuous and $\mu \in \mathscr{W}_2(\mathscr{X})$ arbitrary.From Sect.1.6.1(Theorem 1.6.2), we know that there exists a unique solution to the Monge–Kantorovich problem, and that solution is given by a transport map that we denote by t_γ^μ Recalling that $\mathbf{i} : \mathscr{X} \to \mathscr{X}$ is the identity map, we can define a curve:

$$\gamma_t = [\mathbf{i} + t(t_\gamma^\mu - \mathbf{i})]\#\mu, t \in [0, 1]$$

This curve is known as McCann's [93] interpolant. As hinted in the introduction to this section, it is constructed via classical linear interpolation of the transport maps t_γ^μ and the identity. Clearly $\gamma_1 = \gamma, \gamma_0 = \mu$.

Definition:Constant speed geodesics

A curve $\gamma : [0, 1] \to \mathscr{S}$ is a (constant speed) geodesic if

$$d(\gamma_t, \gamma_s) = d(\gamma_0, \gamma_1)(t - s), \forall 0 \leq s \leq t \leq 1$$

Geodesics in $\mathscr{W}_2(\mathscr{X})$

McCann's interpolant is a constant-speed geodesic in $\mathscr{W}_2(\mathscr{X})$, that is

$$dW_2(\gamma_t, \gamma_s) = W_2(\gamma_0, \gamma_1)(t-s), \forall 0 \leq s \leq t \leq 1$$

Tangent space of $\mathscr{W}_2(\mathscr{X})$ at μ

$$Tan_\mu = \overline{\{t(\mathbf{t}-\mathbf{i}) : \mathbf{t} = t_\mu^\nu, \exists \nu \in \mathscr{W}_2(\mathscr{X}), t > 0\}}^{L_2(\mu)}$$

Strictly speaking, Tan_μ is a subset of the space of functions $f: \mathscr{X}\,\mathscr{X}$ such that $\|f\| \in L^2(\mu)$ rather than $L^2(\mu)$ itself, as in Definition, but we will write $Tan_\mu \subseteq L^2(\mu)$ for simplicity.

The Space $\mathscr{L}_p(\mu)$

Let $\mathscr{X}$ be a Banach space and μ a Borel measure on $\mathscr{X}$. Then the space $\mathscr{L}_p(\mu)$ is the space of measurable functions $f: \mathscr{X} \to \mathscr{X}$ such that

$$\|f\|_{\mathscr{L}_p(\mu)} = \left(\int_{\mathscr{X}} \|f(x)\|_{\mathscr{X}}^p d\mu(x)\right)^{1/p} < \infty$$

The Tangent Bundle

Although not obvious from the definition, this is a linear space. The reason is that, in $\mathbb{R}^d$, Lipschitz functions are dense in $L_2(\mu)$, and for t Lipschitz the negative of a tangent element

$$-t(\mathbf{t}-\mathbf{i}) = s(\mathbf{s}-\mathbf{i}), s > t\|\mathbf{t}\|_{Lip}, \mathbf{s} = \mathbf{i} + \frac{t}{s}(\mathbf{i}-\mathbf{t})$$

lies in the tangent space, since $\mathbf{s}$ can be seen to belong to the subgradient of a convex function by definition of $\mathbf{s}$.This also shows that Tan_μ can be seen to be the $L_2(\mu)$-closure of all gradients of C_c^∞ functions. The tangent space definition is valid for arbitrary measures in $\mathscr{W}_2(\mathscr{X})$. The exponential map at $\gamma \in \mathscr{W}_2(\mathscr{X})$ is the restriction to Tan_γ of the mapping that sends $\mathbf{r} \in L_2(\gamma)$ to $[\mathbf{r}+\mathbf{i}]\#\gamma \in \mathscr{W}_2(\mathscr{X})$.

The Log Map and The Exponential Map

The Exponential Map

$exp_\gamma : Tan_\gamma \to \mathscr{W}_2(\mathscr{X})$ takes the form

$$exp_\gamma(t(\mathbf{t}-\mathbf{i})) = exp_\gamma([t\mathbf{t}+(1-t)\mathbf{i}]-\mathbf{i}) = [t\mathbf{t}+(1-t)\mathbf{i}]\#\gamma$$

Thus, when γ is absolutely continuous, exp exp_γ is surjective, as can be seen from its right inverse, the log map

The Log Map

$log_\gamma : \mathscr{W}_2(\mathscr{X}) \to Tan_\gamma$ takes the form

$$log_\gamma(\mu) = t_\gamma^\mu - \mathbf{i}$$

because convex combinations of optimal maps are optimal maps as well. In particular, McCann's interpolant $[t\mathbf{t}+(1-t)\mathbf{i}]\#\gamma$ is mapped bijectively to the line segment $t(t_\gamma^\mu - \mathbf{i}) \in Tan_\gamma$ through the log map.

Other Definition McCann's interpolant

McCann's interpolant can also be defined as

$$[tp_2 + (1-t)p_1]\#\pi, p_1(x,y) = x, p_2(x,y) = y$$

where $p_1, p_2 : \mathscr{X}^2 \to \mathscr{X}$ are projections and π is any optimal transport plan between γ and μ. This is defined for arbitrary measures $\gamma, \mu \in \mathscr{W}_2(\mathscr{X})$,and reduces to the previous definition if γ is absolutely continuous.

Content

Curvature

Let $\gamma, \mu, \nu \in \mathscr{W}_2(\mathscr{X})$ be absolutely continuous measures.

$$W_2^2(\mu,\nu) \leq \int_{\mathscr{X}} \|t_\gamma^\mu(x) - t_\gamma^\nu(x)\|^2 d\gamma(x) = \|log_\gamma(\mu) - log_\gamma(\nu)\|^2$$

In other words, the distance between μ and ν is smaller in $\mathscr{W}_2(\mathscr{X})$ than the distance between the corresponding vectors $log_\gamma(\mu)$ and $log_\gamma(\nu)$ in the tangent space Tan_γ .In the terminology of differential geometry, this means that the Wasserstein space has nonnegative sectional curvature at any absolutely continuous γ. It is instructive to see when equality holds. As $t_\nu^\gamma = (t_\gamma^\nu)^{-1}$, a change of variables gives

$$W_2^2(\mu,\nu) \leq \int_{\mathscr{X}} \|t_\gamma^\mu(t_\nu^\gamma(x)) - x\|^2 d\nu(x)$$

Since the map $t_\gamma^\mu \circ t_\nu^\gamma$ pushes forward ν to μ,equality holds if and only if $t_\gamma^\mu \circ t_\nu^\gamma = t_\nu^\mu$. This motivates the following definition.

Compatibility

Definition :Compatible Measures

A collection of absolutely continuous measures $\mathscr{C} \subseteq \mathscr{W}_2(\mathscr{X})$ is compatible if for all $\gamma, \mu, \nu \in \mathscr{C}$ we have $t_\gamma^\mu \circ t_\nu^\gamma = t_\nu^\mu$.(in $L_2(\nu)$).

Remark

The absolute continuity is not necessary and was introduced for notational simplicity. A more general definition that applies to general measures is the following: every finite subcollection of $\mathscr{C}$ admits an optimal multicoupling whose relevant projections are simultaneously pairwise optimal.

A collection of two (absolutely continuous) measures is always compatible.

Example:Absolutely continuous measures in $\mathbb{R}$

if $\mathscr{X} = \mathbb{R}$, then the entire collection of absolutely continuous (or even just continuous) measures is compatible. This is because of the simple geometry of convex functions in $\mathbb{R}$: gradients of convex functions are nondecreasing, and this property is stable under composition. In a more probabilistic way of thinking, one can always push-forward μ to ν via the uniform distribution $Leb|_{[0,1]}$ (see Sect.1.5). Letting F_μ^{-1} and F_ν^{-1} denote the quantile functions, we have seen that

$$W_2^2(\mu,\nu) = \|F_\mu^{-1} - F_\nu^{-1}\|_{L_2(0,1)}$$

In other words, $\mu \to F_\mu^{-1}$ is an isometry from $\mathscr{W}_2(\mathbb{R})$ to the subset of $L_2(0,1)$ formed by (equivalence classes of) left-continuous nondecreasing functions on (0,1).

If $\gamma = Leb|[0,1]$, then $F_\mu^{-1} = t_\gamma^\mu$ for all μ, so we can write the above equality as

$$W_2^2(\mu,\nu) = \|F_\mu^{-1} - F_\nu^{-1}\|_{L_2(0,1)} = \|log_\gamma(\mu) - log_\gamma(\nu)\|^2$$

so that if $\mathscr{X} = \mathbb{R}$, the Wasserstein space is essentially flat (has zero sectional curvature). The importance of compatibility can be seen as mimicking the simple one- dimensional case in terms of a Hilbert space embedding. Let $\mathscr{C} \subseteq \mathscr{W}_2(\mathscr{X})$ be compatible and fix $\gamma \in \mathscr{C}$. Then for all $\mu, \nu \in \mathscr{C}$

$$W_2^2(\mu,\nu) = \int_{\mathscr{X}} \|t_\gamma^\mu(x) - t_\gamma^\nu(x)\|^2 d\gamma(x) = \|log_\gamma(\mu) - log_\gamma(\nu)\|^2_{L^2(\gamma)}$$

Consequently, once again, $\mu \to t_\gamma^\mu$ is an isometric embedding of $\mathscr{C}$ into $L_2(\gamma)$.

Example:Gaussian compatible measures

The Gaussian case presented in Sect.1.6.3 is helpful in shedding light on the structure imposed by the compatibility condition. Let $\gamma \in \mathscr{W}_2(\mathbb{R}^d)$ be a standard Gaussian distribution with identity covariance matrix. Let Σ_μ denote the covariance matrix of a measure $\mu \in \mathscr{W}_2(\mathbb{R}^d)$ When μ and ν are centred nondegenerate Gaussian measures,

$$t_\gamma^\mu = \Sigma_\mu^{1/2}; t_\gamma^\nu = \Sigma_\nu^{1/2}; t_\mu^\nu = \Sigma_\mu^{-1/2}[\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2}]^{1/2}\Sigma_\mu^{-1/2}$$

so that γ, μ, and ν are compatible if and only if

$$t_\mu^\nu = t_\gamma^\nu \circ t_\mu^\gamma = \Sigma_\nu^{1/2}\Sigma_\mu^{-1/2}$$

equivalently, Σ_μ and Σ_ν commute.
We see that a collection $\mathscr{C}$ of Gaussian measures on $\mathbb{R}^d$ that includes the standard Gaussian distribution is compatible if and only if all the covariance matrices of the measures in $\mathscr{C}$ are *simultaneously diagonalisable*.

Gaussian compatible measures

there exists an orthogonal matrix U such that $D_\mu = U\Sigma_\mu U^T$ is diagonal for all $\mu \in \mathscr{C}$.In that case, formula

$$W_2^2(\mu,\nu) = tr(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}) = tr(\Sigma_\mu + \Sigma_\nu - 2\Sigma_\mu^{1/2}\Sigma_\nu^{1/2})$$

$$= tr(D_\mu + D_\nu - 2D_\mu^{1/2}D_\nu^{1/2}) = \sum_{i=1}^{d}(\sqrt{\alpha_i} - \sqrt{\beta_i})^2, \alpha_i = [D_\mu]_{ii}, \beta_i = [D_\nu]_{ii}$$

and identifying the (nonnegative) number $a \in \mathbb{R}$, with the map $x \to ax$ on $\mathbb{R}$, the optimal maps take the "orthogonal separable" form

$$t_\mu^\nu = \Sigma_\nu^{1/2}\Sigma_\mu^{-1/2} = UD_\nu^{1/2}D_\mu^{1/2}U^T = U \circ (\sqrt{\beta_1/\alpha_1}, ..., \sqrt{\beta_d/\alpha_d}) \circ U^T$$

In other words, up to an orthogonal change of coordinates, the optimal maps take the form of d nondecreasing real-valued functions. This is yet another crystallisation of the one-dimensional-like structure of compatible measures.

With the intuition of the Gaussian case at our disposal, we can discuss a more general case. Suppose that the optimal maps are continuously differentiable. Then differentiating the equation $t_\mu^\nu = t_\gamma^\nu \circ t_\mu^\gamma$ gives

$$\nabla t_\mu^\nu(x) = \nabla t_\gamma^\nu(t_\mu^\gamma(x)) \nabla t_\mu^\gamma(x)$$

Since optimal maps are gradients of convex functions, their derivatives must be symmetric and positive semidefinite matrices.A product of such matrices stays symmetric if and only if they commute, so in this differentiable setting, compatibility is equivalent to commutativity of the matrices $\nabla t_\gamma^\nu(t_\mu^\gamma(x))$ and $\nabla t_\mu^\gamma(x)$ for μ-almost all x. In the Gaussian case, the optimal maps are linear functions, so x does not appear in the matrices.

Radial transformations

Consider the collection of functions $\mathbf{t} : \mathbb{R}^d \to \mathbb{R}^d$ of the form $\mathbf{t}(x) = xG(\|x\|)$ with $G : \mathbb{R}_+ \to \mathbb{R}$ differentiable. Then a straightforward calculation shows that

$$\nabla \mathbf{t}(x) = G(\|x\|)\mathbf{I} + [G'(\|x\|)/\|x\|]xx^T$$

Since both $\mathbf{I}$ and xx^T are positive semidefinite, the above matrix is so if both G and G' are nonnegative. If $\mathbf{s}(x) = xH(\|x\|)$ is a function of the same form, then $\mathbf{s}(\mathbf{t}(x)) = xG(\|x\|)H(\|x(\|x\|)\|)$ which belongs to that family of functions (since G is nonnegative). Clearly

$$\nabla \mathbf{s}(\mathbf{t}(x)) = H(\|x\| G(\|x\|))\mathbf{I} + [G(\|x\|)H'(\|x\|)/\|x\|]xx^T$$

commutes with $\nabla \mathbf{t}(x)$, since both matrices are of the form $a\mathbf{I} + bxx^T$ with a, b scalars (that depend on x).

Radial transformations

In order to be able to change the base measure γ, we need to check that the inverses belong to the family. But if $y = \mathbf{t}(x)$, then $x = ay$ for some scalar a that solves the equation

$$aG(a\|y\|) = 1$$

Such a is guaranteed to be unique if $a \longmapsto aG(a)$ is strictly increasing and it will exist (for y in the range of $\mathbf{t}$) if it is continuous. As a matter of fact, since the eigenvalues of $\nabla\mathbf{t}(x)$ are $G(a)$ and

$$G(a) + G'(a)a = (aG(a))', a = \|x\|$$

the condition that $a \longmapsto aG(a)$ is strictly increasing is sufficient (this is weaker than G itself increasing). Finally, differentiability of G is not required, so it is enough if G is continuous and $aG(a)$ is strictly increasing.

Separable variables

Consider the collection of functions $\mathbf{t} : \mathbb{R}^d \to \mathbb{R}^d$ of the form.

$$\mathbf{t}(x_1, ..., x_d) = (T_1(x), ..., T_d(x)), T_i : \mathbb{R} \to \mathbb{R}, 2.8)$$

with T_i continuous and strictly increasing. This is a generalisation of the compatible Gaussian case discussed above in which all the T_i's were linear. Here, it is obvious that elements in this family are optimal maps and that the family is closed under inverses and composition, so that compatibility follows immediately.
This family is characterised by measures having a *common dependence structure*.

Separable variables

copula

we say that $C : [0,1]^d \to [0,1]$ is a copula if C is (the restriction of) a distribution function of a random vector having uniform margins.

In other words, if there is a random vector $V = (V_1, ..., V_d)$ with $\mathbb{P}(V_i \leq a) = a$ for all $a \in [0,1]$ and all $j = 1, ..., d$, and

$$\mathbb{P}(V_1 < v_1, ..., V_d < v_d) = C(v_1, ..., v_d), v_i \in [0,1]$$

To any d-dimensional probability measure μ ,one can assign a copula $C = C_\mu$ in terms of the distribution function G of μ and its marginals G_j as

$$G(a_1, ..., a_d) = \mu((\infty, a_1] \times ... \times (\infty, a_d]) = C(G_1(a_1), ..., G_d(a_d)).$$

If each G_j is surjective on $(0,1)$, which is equivalent to it being continuous, then this equation defines C uniquely on $(0,1)^d$,and consequently on $[0,1]^d$.If some marginal G_j is not continuous, then uniqueness is lost, but C still exists.The connection of copulae to compatibility becomes clear in the following lemma.

Separable variables

Lemma 2.3.3 (Compatibility and Copulae)

The copulae associated with absolutely continuous measures $\mu, \nu \in \mathscr{W}_2(\mathbb{R}^d)$ are equal if and only if t_μ^ν takes the separable form (2.8).

Composition with linear functions

If $\phi : \mathbb{R}^d \to \mathbb{R}$ is convex with gradient $\mathbf{t}$ and A is a dd matrix,then the gradient of the convex function $x \longmapsto \phi(Ax)$ at x is $\mathbf{t}_A = A^T\mathbf{t}(Ax)$. Suppose ψ is another convex function with gradient $\mathbf{s}$ and that compatibility holds,i.e., $\nabla\mathbf{s}(\mathbf{t}(x))$ commutes with $\nabla\mathbf{t}(x)$ for all x. Then in order for

$$\nabla\mathbf{s}_A(\mathbf{t}_A(x)) = A^T\nabla\mathbf{s}(AA^T\mathbf{t}(Ax))A, \nabla\mathbf{t}_A(x) = A^T\nabla\mathbf{t}(Ax)A$$

to commute, it suffices that $AA^T = \mathbf{I}$, i.e., that A be orthogonal. Consequently, if $\{\mathbf{t}\#\mu_{t\in T}\}$ are compatible, then so are $\{\mathbf{t}_U\#\mu_{t\in T}\}$for any orthogonal matrix U.