An Invitation to Statistics in Wasserstein Space

Shen Ziliang 2021213218

Shanghai University of Finance and Economics

January 29, 2022

Outline

Content

Complementary Slackness

$$\int_{\mathscr{X}\times\mathscr{Y}}[c(x,y)-\varphi(x)-\psi(y)]\mathrm{d}\pi(x,y)=0$$

which is in turn equivalent to $\varphi(x)+\psi(y)=c(x,y)$ π-almost surely. it has already been established that there exists an optimal transference plan π^*. Assuming that $C(\pi^*)=\int_{\mathscr{X}\times\mathscr{Y}}c(x,y)\mathrm{d}\pi^*(x,y))<\infty$ (otherwise all transference plans are optimal), a
pair$(\varphi,\psi)\in\Phi^c=\{\varphi(x)+\psi(y)\leq c(x,y)\}$is optimal if and only if
Assuming that $C(\pi^*)<\infty$ (otherwise all transference plans are optimal), a pair $(\varphi,\psi)\in\Phi^c$is optimal if and only if

$$\varphi(x)+\psi(y)=c(x,y),\pi^*-almost$$

Conversely, if (φ_0,ψ_0) is an optimal pair, then π is optimal if and only if it is concentrated on the set

$$\{(x,y):\varphi_0(x)+\psi_0(y)=c(x,y)\}$$

Unconstrained Dual Kantorovich Problem

Definiton : c- transform

$$\varphi^c(y) := inf_{x\in\mathscr{X}}[c(x,y)-\varphi(x)]$$

is the largest possible ψ satisfying $(\varphi,\psi)\in\Phi^c=\{\varphi(x)+\psi(y)\le c(x,y)\}$.
A function taking this form is called **c-concave**.

the dual problem can thus be formulated as **the unconstrained problem**:

$$sup_{\varphi\in L_1(\mu)}[\int_{\mathscr{X}}\varphi(x)\mathrm{d}\mu+\int_{\mathscr{Y}}\varphi^c(y)\mathrm{d}v]$$

One can apply this c-transform again and replace φ by

$$\varphi^{cc}(x)=(\varphi^c)^c(x)=inf_{y\in\mathscr{Y}}[c(x,y)-\varphi^c(y)]\ge\varphi(x)$$

An elementary calculation shows that in general $\varphi^{ccc}=\varphi^c$. If $\varphi^c=\psi$ and $\psi^c=\varphi$; in words, φ and ψ are **c-conjugate**.

Unconstrained Dual Kantorovich Problem

Proposition 1.8.1 (Existence of an Optimal Pair)

Let μ and v be probability measures on $\mathscr{X}$ and $\mathscr{Y}$ such that the independent coupling with respect to the nonnegative and lower semi-continuous cost function is finite: $\int_{\mathscr{X}\times\mathscr{Y}} c(x,y)\mathrm{d}\mu(x)\mathrm{d}v(y) < \infty$
Then there exists an optimal pair (φ, ψ) for the dual Kantorovich problem. Furthermore, the pair can be chosen in a way that μ-almost surely, $\varphi = \psi^c$ and v-almost surely, $\psi = \varphi^c$.

we see that if φ is optimal (for the unconstrained dual problem), then any optimal transference plan π^* must be concentrated on the set $\{(x,y) : \varphi(x) + \varphi^c(y) = c(x,y)\}$. If for μ-almost every x this equation defines y uniquely as a (measurable) function of x, then π^* is induced by a transport map.

The Kantorovich-Rubinstein Theorem

We have seen an example where c was the quadratic Euclidean distance. Here, we shall consider another useful case, where c is a metric. Assume that $\mathscr{X} = \mathscr{Y}$, denote their metric by d, and let $c(x, y) = d(x, y)$. If $\varphi = \psi^c$ is c-concave, then it is 1-Lipschitz. Indeed, by definition and the triangle inequality.

$$\varphi(z) \leq inf_{y \in \mathscr{Y}}[d(z, y) - \psi(y)] \leq \inf_{y \in \mathscr{Y}}[d(x, y) + d(x, z) - \psi(y)]$$

$$= \varphi(x) + d(x, z)$$

Interchanging x and z yields $|\varphi(x) - \varphi(z)| \leq d(x, z)$
the duality formula takes the form:

$$inf_{\varpi \in \Pi(\mu, v)} \int_{\mathscr{X}^2} d(x, y) \mathrm{d}\pi(x, y) = sup_{\|\varphi\|_{lip} \leq 1} \left| \int_{\mathscr{X}} \varphi \mathrm{d}\mu - \int_{\mathscr{X}} \varphi \mathrm{d}v \right|$$

$$\|\varphi\|_{lip} = sup_{x \neq y} \frac{\varphi(x) - \varphi(y)}{d(x, y)}$$

Kantorovich-Rubinstein theorem

Under assumption above, Kantorovich problem is equivalent to

$$sup\{\int_{\mathscr{X}} \varphi \mathrm{d}(\mu - v) : \varphi \in \bigcup L^1(d|\mu - v|), \|\varphi\|_{lip} \leq 1\}$$

Morevoer, it does not change the value of the supremum above to impose the additional condition that φ be bounded.

Content

Cost Functions

We now return to the Euclidean case $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$ and explore the structure of c-transforms.

definition: c-superdifferential

we can still apply the idea that $\varphi(x) + \psi(y) = c(x, y)$ if and only if the infimum (definition of c-transform) is attained at x. So that $\varphi(x) + \psi(y) = c(x, y)$ if and only if:

$$\varphi(z) - \varphi(x) \leq c(z, y) - c(x, y), z \in \mathscr{X}$$

we call the collection of y's satisfying the above in equality the c-superdifferential of ϕ at x, and we denote it by $\partial^c \varphi$.

The following result generalises Theorem 1.6.2(Optimal Solution of Quadratic Cost in Euclidean Spaces) to other powers $p > 1$ of the Euclidean norm.

Theorem 1.8.2 (Strictly Convex Costs on $\mathbb{R}^d$)

Let $c(x, y) = h(x - y)$ with $h(v) = \|v\|^p/p$ for some $p > 1$ and let μ and v be probability measures on $\mathbb{R}^d$ with finite p-th moments such that μ is absolutely continuous with respect to Lebesgue measure. Then the solution to the Kantorovich problem with cost function c is unique and induced from a transport map T. Furthermore, there exists an optimal pair (φ, φ^c) of the dual problem, with φ c-concave. The solutions are related by

$$T(x) = x - \nabla\varphi(x)\nabla\varphi(x)^{1/(p-1)-1}(\mu - a.s.).$$

Proof: (Assuming v has CompactSupport).

1. **Step 1: φ is c-superdifferentiable.** Let π^* be an optimal coupling. By duality arguments, π is concentrated on the set of (x, y) such that $y \in \partial^c \varphi(x)$. Consequently, for μ-almost any x, the c-superdifferential of φ at x is nonempty.
2. **Step 2: φ is differentiable.** Here, we impose the additional condition that v is compactly supported. Then can be taken as a c-transform on the compact support of v . Since h is locally Lipschitz (it is C^1 because $p > 1$) this implies that φ is locally Lipschitz. Hence, it is differentiable Lebesgue almost surely, and consequently μ-almost surely.
3. **Conclusion.**

For μ-almost every x there exists $y \in \partial^c\varphi(x)$ and a gradient $u = \nabla\varphi(x)$. In particular, u is a subgradient of φ:

$$\varphi(z) - \varphi(x) \geq \langle u, z - x\rangle + o(\|z - x\|).$$

Here and more generally, $o(\|z - x\|)$ denotes a function $r(z)$ (defined in a neighbourhood of x) such that $r(z)/\|z - x\| \to 0$ as $z \to x$. (If φ were convex, then we could take $r \equiv 0$, so the definition for convex functions is equivalent, and then the inequality holds globally and not only locally.) But $y \in \partial^c\varphi(x)$ means that as $z \to x$,

$$\begin{aligned} h(z - y) - h(x - y) = c(z, y) - c(x, y) &\geq \varphi(z) - \varphi(x) \\ &\geq \langle u, z - x\rangle + o(\|z - x\|). \end{aligned}$$

In other words, u is a subgradient of h at $x - y$. But h is differentiable with gradient $h(v) = v\|v\|^{p-2}$ (zero if $v = 0$). We obtain $\nabla\varphi(x) = u = \nabla h(x - y)$ and since the gradient of h is invertible, we conclude

$$y = T(x) := x - (\nabla h)^{-1}[\nabla\varphi(x)],$$

which defines y as a (measurable) function of x. Hence, the optimal transference plan π is unique and induced from the transport map T.

1.It holds for a larger class of functions h, those that are strictly convex on $\mathbb{R}^d$ (this yields that ∇h is invertible) Furthermore, if h is sufficiently smooth, namely $h \in C^{(1,1)}$ locally (it is if $p \geq 2$, but not if $p \in (1, 2)$), then μ does not need to be absolutely continuous;
2.It is also noteworthy that for strictly concave cost functions (e.g., $p \in (0, 1)$), the situation is similar when the supports of μ and v are disjoint. The reason is that h may fail to be differentiable at 0, but it only needs to be differentiated at $x - y$ with $x \in supp\mu$ and $y \in suppv$. If the supports are not disjoint,

Content

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 $\mathscr{W}_p(\mathscr{X})$.

The p-Wasserstein distance between μ and v is defined as the minimal total transportation cost between μ and v in the Kantorovich problem with respect to the cost function $c_p(x, y) = \|x - y\|^p$:$W_P(\mu, v) = (inf_{\pi \in \Pi(\mu,v) C_p(\pi)})^{1/p} = (inf_{\pi \in \Pi(\mu,v)} \int_{\mathscr{X} \times \mathscr{X}} \|x_1 - x_2\|^p \mathrm{d}\pi(x_1, x_2))^{1/p}$
The Wasserstein distance between μ and v is finite when both measures are in $\mathscr{W}_p(\mathscr{X})$, because

$$\|x_1 - x_2\|^p \leq 2^p \|x_1\| + 2^p \|x_2\|$$

Topological Properties

Theorem 2.2.1 (Convergence in Wasserstein Space)

Let $\mu, \mu_n \in \mathscr{W}_p(X)$. Then the following are equivalent:

1. $W_p(\mu, \mu_n) \to 0$ as $n \to \infty$;
2. $\mu_n \to \mu$ weakly and $\int_{\mathscr{X}} \|x\|^p \mathrm{d}\mu_n(x) \to \int_{\mathscr{X}} \|x\|^p \mathrm{d}\mu(x)$
3. $\mu_n \to \mu$ weakly and
$$sup_n \int_{\{x:\|x\|>R\}} \|x\|^p \mathrm{d}\mu_n(x) \to 0, R \to \infty$$
4. for any $C > 0$ and any continuous $f : \mathscr{X} \to \mathbf{R}$ such that $|f(x)| \leq C(1 + \|x\|^p)$ for all x,
$$\int_{\mathscr{X}} f(x)\mathrm{d}\mu_n(x) \to \int_{\mathscr{X}} f(x)\mathrm{d}\mu(x)$$
5. $\mu_n \to \mu$ weakly and there exists $v \in \mathscr{W}_p(X)$ such that $W_p(\mu_n, v) \to W_p(\mu, v)$.

Topological Properties:Convergence, Compact Subsets

Corollary 2.2.2

Let $K \subseteq \mathscr{X}$ be a bounded set and suppose that $_n(K) = 1$ for all $n \geq 1$. Then $W_p(\mu_n, \mu) \to 0$ if and only if $\mu_n \to \mu$ weakly.

Proposition 2.2.3 (Compact Sets in $\mathscr{W}_p$)

A weakly tight set $\mathscr{K} \subseteq \mathscr{W}_p$ is Wasserstein tight (has a compact closure in $\mathscr{W}_p$) if and only if

$$sup_{\mu \in \mathscr{K}} \int_{\{x:\|x\|>R\}} \|x\|^p \mathrm{d}\mu(x) \to 0, R \to \infty$$

Moreover, condition above is equivalent to the existence of a monotonically divergent function $g : \mathbf{R}_+ \to \mathbf{R}_+$ such that

$$sup_{\mu \in \mathscr{K}} \int_{\mathscr{X}} \|x\|^p g(\|x\|) \mathrm{d}\mu(x) \to 0$$

Convergence, Compact Subsets

Remark 2.2.4 For any sequence (μ_n) in $\mathscr{W}_p$ (tight or not) there exists a monotonically divergent g with $\int_{\mathscr{X}} \|x\|^p g(\|x\|) \mathrm{d}\mu_n(x) < \infty$ for all n.

Corollary 2.2.5 (Measures with Common Support)

Let $K \subseteq \mathscr{X}$ be a compact set. Then

$$\mathscr{K} = \mathscr{W}_p(K) = \{\mu \in P(\mathscr{X}) : \mu(K) = 1\} \subseteq \mathscr{W}_p(\mathscr{X})$$

is compact.

Definition:Uniform Absolute Continuity(consequence of ui)

$\forall \epsilon \exists \delta \forall n \forall BroelA \subseteq \mathscr{X}$, $\mu_n(A) \leq \delta$ then:

$$\int_A \|x\|^p \mathrm{d}\mu_n(x) \leq \epsilon$$

Dense Subsets and Completeness

Proposition 2.2.6 (Empirical Measures in $\mathscr{W}_p$)

For any $\mu \in P(\mathscr{X})$and the corresponding sequence of empirical measures μ_n ,$W_p(\mu_n, \mu) \to 0$ almost surely if and only if $\mu \in {}_p(\mathscr{X})$.

Theorem 2.2.7 (Dense Subsets of $\mathscr{W}$)

The following collections of measures are dense in $\mathscr{W}_p(\mathscr{X})$:

1. finitely supported measures with rational weights;
2. compactly supported measures;
3. finitely supported measures with rational weights on a dense subset $A \subseteq X$;
4. if $\mathscr{X} = \mathbb{R}^d$,the collection of absolutely continuous and compactly supported measures;
5. if $\mathscr{X} = \mathbb{R}^d$,the collection of absolutely continuous measures with strictly positive and bounded analytic densities.

Dense Subsets and Completeness

In particular, $\mathscr{W}_p$ is separable (the third set is countable as $\mathscr{X}$ is separable).

Proposition 2.2.8 (Completeness)

The Wasserstein space $\mathscr{W}_p(\mathscr{X})$ is complete.