$\mathcal{J}$ - Divergence

We want to learn about the entire dist. μ rather expected value. Because any value comes with risk and in order to quantify the value w.r.t the dist. Approxi. knowledge of dist. μ is model:

Next, we will consider $\mathcal{J}$-divergence.

Rmk: For $\mathcal{J}$ larger, $d_{\mathcal{J}}$ obtains more guarantees for approxi. accuracy.

(1) Topo & Metric:

Def: i) V, W, n.v.s. $L: (V, \|\cdot\|_V) \to (W, \|\cdot\|_W)$ is strongly conti. if it's BLO.

ii) $(V', \|\cdot\|'_V)$ denote dual space of $(V, \|\cdot\|_V)$.

iii) $\|g\|_\infty := \sup_{x \in \mathbb{R}^d} |g(x)|$. Set $L_B^\infty(\mathbb{R}^d) := \{ g \text{ measurable} \mid \|g\|_\infty < \infty \}$.

Rmk: $\mathcal{M}_1^+(\mathbb{R}^d)$ can be embedded into $(L_B^\infty)^*$ by Set $L_\nu(g) \overset{\Delta}{=} \int g\, d\nu$ for $\nu \in \mathcal{M}_+^1(\mathbb{R}^d)$. $|L_\nu(g)| \leq \|g\|_\infty$.

iv) $\mathcal{M}(\mathbb{R}^d)$ is finite signed measure on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$. It can be embedded in $(L_B^\infty)^*$ as well: for $\gamma = \alpha\mu + \beta\nu$,

$L_\gamma(g) = \alpha L_\mu(g) + \beta L_\nu(g)$

$= \alpha \int g\, d\mu + \beta \int g\, d\nu = \int g\, d\gamma$.

v) Equip $\mathcal{M}(\mathbb{R}^d)$ with norm $\|\gamma\|_{TV} := \sup\{L_\gamma(g) \mid g \in L_B^\infty(\mathbb{R}^d), \|g\|_\infty \leq 1\}$.

Rmk: By Jordan decomp, $\gamma = \gamma^+ - \gamma^-$, s.t. $\gamma^+ \perp \gamma^-$. So $\exists A^\pm \in \mathcal{B}_{\mathbb{R}^d}$ s.t. $g = I_{A^+} - I_{A^-}$ attains its sup.

vi) $\mathcal{G}_{TV} := L_B^\infty(\mathbb{R}^d)$. $d_{TV}(\mu, \nu) = k_{\mathcal{G}_{TV}}(\mu, \nu)$ $= \|\mu - \nu\|_{TV}$ for $\mu, \nu \in \mathcal{M}_+^1(\mathbb{R}^d)$.

Rmk: i) Set $\mathcal{G}_{KO} := C_B(\mathbb{R}^d)$. And we restrict on it. We can obtain

Radon-Norm: $\|\cdot\|_{RD}$. Which's weaker than $\|\cdot\|_{TV}$. Also we can also consider $\mathcal{G}_c := C_c(\mathbb{R}^d)$.

ii) We can also introduce $d_{TV}(\mu, \nu) = 2\sup_{A \in \mathcal{B}_{\mathbb{R}^d}} |\mu(A) - \nu(A)|$. i.e. set $\mathcal{G} = \{2 I_A \mid A \in \mathcal{B}_{\mathbb{R}^d}\}$.

<u>Lem.</u> $X_n \sim \mu_n$. $X \sim \mu$. $\mathbb{R}^d$-r.v. Then: We have
$$X_n \xrightarrow{d} X \iff \mu_n \to \mu \text{ in } C_c(\mathbb{R}^d)'.$$

<u>Pf</u>: prove that $\mu_n \to \mu$ in $C_c(\mathbb{R}^d)' \iff \mu_n \to \mu$ in $C_B(\mathbb{R}^d)'$.

For $g \in C_B(\mathbb{R}^d)$. $B = \|g\|_\infty$. Choose K s.t. $\mu(B_K(0)^c) \leq \varepsilon/B$.

φ is bump func. s.t. $\varphi = 1$ on $B_K(0)$. $\varphi \geq 0$. $\|\varphi\|_\infty = 1$. $\varphi \in C_c^\infty(\mathbb{R}^d)$.

$$\int g\, d\mu_n = \int \varphi g\, d\mu_n + \int (1-\varphi) g\, d\mu_n$$

$$\left|\int (1-\varphi) g\, d\mu_n - \int (1-\varphi) g\, d\mu\right| \leq B\left(1 - \int \varphi\, d\mu_n\right) + B\left(1 - \int \varphi\, d\mu\right) \xrightarrow{n\to\infty} 2B\left(1 - \int \varphi\, d\mu\right) \leq B \cdot \frac{\varepsilon}{B} = \varepsilon \to 0.$$

Cor. For $\mathcal{J} \subseteq \mu_1^+(\mathbb{R}^n)$ to be the space of p.m.'s $\ll dx$ on $\mathbb{R}^n$. i.e. $\mu(dx) = f(x)dx$. Then: $\forall \mu, \nu \in \mathcal{J}$, we have:

$d_{TV}(\mu,\nu) = d_{C_c}(\mu,\nu) = d_{\infty}(\mu,\nu)$.

Pf: For the latter part. we set

φ_n is bump func. on $B_n(0)$.

and $\|\varphi_n\|_\infty \le 1$

And use DCT. We can get:

$\int \varphi_n g\, d\mu \to \int g\, d\mu$.

Rmk: Note that C_B, L_B^∞ are not separable. So it may lead to some measurability problem when considering SLAs.

(2) Wasserstein Metric:

Def: For δ metric on $\mathbb{R}^k$. s.t. $\delta(\cdot, y) \in \mathcal{B}_{\mathbb{R}^k, \mathbb{R}^1}$. $\forall y$.

1) $Lip(\delta) = \{ f : \mathbb{R}^n \to \mathbb{R}^1,\ f$ has Lip-const $= 1\}$. $Lip_y(\delta) := Lip(\delta) \cap \{ f \mid g(y) = 0\}$

Rmk: $g = x \in Lip_0^{1.1}$. $\mu \sim Cauchy(1)$, $L_\mu(g) = \infty$

So $L_\mu(g)$ doesn't have to exist.

ii) $\mathcal{M}^\delta(\mathbb{R}^d) := \{ \nu \in \mathcal{M}(\mathbb{R}^d) \mid \exists y \in \mathbb{R}^d$. s.t. $\int \delta(x,y)\, d|\nu|(x) < \infty \}$.

iii) Wasserstein 1-norm $\|\mu - \nu\|_{W_1(\delta)} = \sup\{ |L_{\mu-\nu}(g)| \mid g \in Lip_y(\delta)\}$ for some $y \in \mathbb{R}^d$ and $\mu, \nu \in \mathcal{M}^\delta(\mathbb{R}^d)$.

iv) Set $\mathcal{M}_1^{\delta,+}(\mathbb{R}^d) = \mathcal{M}_1^+(\mathbb{R}^d) \cap \mathcal{M}^\delta(\mathbb{R}^d)$.

Lip metric $d_{W_1(\delta)}(\mu,\nu) := \|\mu - \nu\|_{W_1(\delta)}$ is defined on $\mathcal{M}_1^{\delta,+}(\mathbb{R}^d)$.

Lem. For $\mu, \nu \in \mathcal{M}^\delta(\mathbb{R}^d)$. We have:

i) $\|\mu\|_{W_1(\delta)} < \infty$.

ii) If $\mu(\mathbb{R}^d) = \nu(\mathbb{R}^d)$. Then we can replace $Lip_y(\delta)$ by $Lip(\delta)$. i.e.

$\|\mu - \nu\|_{W_1(\delta)} = \sup_{Lip(\delta)} |L_{\mu-\nu}(g)|$

iii) $\|\cdot\|_{W_1(\delta)}$ is a semi-norm on $\mathcal{M}^\delta(\mathbb{R}^d)$.

iv) If δ-topo isn't weaker than $|\cdot|$ - topo. $^{(*)}$ ($x_n \xrightarrow{\delta} x \Rightarrow x_n \xrightarrow{|\cdot|} x$). Then:

(*) is initial topo.

$d_{W,(\delta)}(\mu,\nu) = 0 \iff \mu = \nu$ for $\mu,\nu \in \mathcal{M}_1^{s.t}(\mathbb{R}^d)$. So it's metric on $\mathcal{M}_1^{s.t}(\mathbb{R}^d)$.

<u>Pf</u>: i) Note $g(x) \overset{Lip}{\le} |g(y)| + \delta(x,y) = \delta(x,y)$

ii) Let $g^*(x) = g(x) - g(y)$. Then =

$L_{\mu-\nu}(g) = L_{\mu-\nu}(g^*)$ from and.

iii) $\|\cdot\|_{W,(\delta)}$ is Supremum of seminorm.

iv) $d_{W,(\delta)}(\mu,\nu) = 0 \overset{ii)}{\Rightarrow} L_\mu(g) = L_\nu(g)$

for $\forall g$ is Lip-conti.

Next, we prove $\mu(A) = \nu(A)$ for $\forall A$ closed in $\mathbb{R}^d$.

Set $g_{\varepsilon,A}(x) = (1 - \frac{1}{\varepsilon}\delta(x,A)) \vee 0$.

$\Rightarrow g_{\varepsilon,A}$ is Lip-conti. with const. $= \frac{1}{\varepsilon}$. And $g_{\varepsilon,A} \downarrow I_A$ because

$\delta(x,A) = 0 \overset{conv.}{\Rightarrow} \inf_y |x-y| = 0$.

Apply DCT on $\int g_{\varepsilon,A}\, d\mu = \int g_{\varepsilon,A}\, d\nu$.

<u>Lem</u>. $\delta(x,y) = |x-y|$ and $B \subset \mathbb{R}^d$ bdd. Let

$\mathcal{M}_1^{s.t}(B) = \{\mu \in \mathcal{M}_1^{s.t}(\mathbb{R}^d) \mid supp(\mu) \subset B\}$.

Let $r_B = \inf_{y\in B} \sup_{x\in B} \delta(x,y)$. Then: We have,

$d_{W_1(\delta)}(\mu,\nu) \le r_B\, d_{RD}(\mu,\nu)$ for $\forall \mu,\nu$

$\in \mathcal{M}_1^{\delta,+}(B)$.

<u>Pf</u>: $\exists y_\varepsilon \in B$. s.t. $\sup_{x\in B} \delta(x,y_\varepsilon) \le \varepsilon + r_B$.

So: $\forall g \in Lip_{y_\varepsilon}(\delta)$. $\Rightarrow g \le r_B+\varepsilon$ on B.

Let $\bar{g} = g(x) I_B + (r_B+\varepsilon) I_{B^c}$.

$\Rightarrow L_{\mu-\nu}(g) = L_{\mu-\nu}(\bar{g})$.

$= (r_B+\varepsilon)\, L_{\mu-\nu}\left(\frac{\bar{g}}{r_B+\varepsilon}\right)$

$\le (r_B+\varepsilon)\, \|\mu-\nu\|_{RD}$.

<u>Cor.</u> Under condition above. we have:

i) If B is cpt. then. $\|\cdot\|_{W_1(\delta)}$ is

weaker than $\|\cdot\|_{RD}$ or $\|\cdot\|_{TV}$.

ii) $\|\cdot\|_{W_1(\delta)}$ - topo = weak convergence

(3) <u>Metrics w.r.t density / para.</u>:

① For $d\mu = f_\mu dx$. $d\nu = f_\nu dx$. We set:

$d_p(\mu,\nu) = \|f_\mu - f_\nu\|_{L^p}$ on $\mathcal{J} = \{\mu\in\mathcal{M}_1^+ \mid \mu \ll dx\}$

Rmk: For $p=1$. Note $\mathrm{sgn}(f_1-f_2)\in L^\infty_{\mathcal{B}}$.

So it corresponds to $\|\cdot\|_{TV}=\|\cdot\|_{K_0}$.

For $p>1$. They're not comparable.

② For $\Theta\subset\mathbb{R}^d$. Consider map $\mu_{(\cdot)}:\Theta\to\mathcal{M}_1^+(\mathbb{R}^d)$.

is injective. And for $\nu,\nu'\in\mathrm{Im}(\mu)$. Set

$d_\Theta(\nu,\nu')=|\theta-\theta'|$. where $\nu=\mu_\theta$. $\nu'=\mu_{\theta'}$.

Rmk: It can develop para. stat. approach

$d(\hat{\mu}_n\|\mu)=d(\mu_{\hat{\theta}_n}\|\mu_\theta)\le C|\hat{\theta}_n-\theta|\to 0$.

Lem. $d(\cdot\|\cdot)$ is divergence on $\mathcal{M}_1^+(\mathbb{R}^d)$ and assume $\mu_{(\cdot)}:\Theta\to\mathcal{M}_1^+(\mathbb{R}^k)$ is injective and conti. w.r.t. $d_0=|\cdot|$. Then:

i) d_Θ - topo is stronger than d-topo.

ii) If Θ is cpt. Then: $d_\Theta\sim d$ on $\mathrm{Im}(\mu_{(\cdot)})$

Pf: i) $\theta_n\xrightarrow{|\cdot|}\theta_0\overset{\text{conti.}}{\Rightarrow}d(\mu_{\theta_0}\|\mu_{\theta_n})\to 0$.

ii) If $d(\mu_{\theta_0}\|\mu_{\theta_n})\to 0$. $|\theta_n-\theta_0|\not\to 0$

$\Rightarrow\exists\varepsilon>0,\ \theta_{n_k}.\ \text{s.t.}\ (\theta_{n_k})\subset\Theta_\varepsilon=\Theta/B_\varepsilon(\theta_0)$

Note Θ_ε is also cpt.

So $\delta = \inf_{\theta \in \Theta_\varepsilon} d(\mu_{\theta_0} \| \mu_\theta) > 0$ follows

from the conti. condition of $\mu(\cdot)$

(Otherwise it will contradict with

that "$\mu(\cdot)$ is injective" and d

is separating. i.e. $\exists \theta_* \neq \theta_0$. $\mu_{\theta_0} = \mu_{\theta^*}$)

$\Rightarrow d(\mu_{\theta_0} \| \mu_{\theta_{n_k}}) \geq \delta > 0$.

which contradict with $\mu_{\theta_k} \xrightarrow{d} \mu_{\theta_0}$.