Div. from Info. Theory

(1) Motivation from odds:

If we want to decide whether given data is from dist. μ or ν:

i) If $\exists A, B \in \mathcal{B}_{\mathbb{R}^d}$. s.t. $\mu(A) > 0$, $\nu(B) > 0$ and $\mu(B) = \nu(A) = 0$. Assume sample $\mathcal{X} = (X_1 \cdots X_n)$ is from μ of size n. So with proba $= 1 - (1-\mu(A))^n$ we can observe at least one sample is from A. $\Rightarrow$ Make decision that $\mathcal{X}$ is from dist. μ.

Rmk: $1 - (1-\mu(A))^n \xrightarrow{n \to \infty} 1$. So the conclusion can be reached exponentially fast.

ii) If $\mu \sim \nu$. (i.e. $\mu \ll \nu$ & $\nu \ll \mu$). Then: the decision become uncertained no matter how many samples we draw.

Thm. (Radon - Nikodym)

For $\mu, \nu \in \mathcal{M}_1^+(\mathbb{R}^d)$. $\nu \ll \mu$. Then $\exists Z(x) \in$

$L^1(d\mu)$. s.t. $Z(x) \geq 0$ μ-a.s. $dV(x) = Z(x) d\mu(x)$

<u>Pf</u>: Consider $L^2(\frac{1}{2}(\mu+V))$. We see:

$|L_\mu(g)| \leq 2 \|g\|_{L^2(\frac{1}{2}(\mu+V))}$. So $L_\mu(\cdot)$ is

BLO on $L^2(\frac{1}{2}(\mu+V)) \subset L^2(\mu) \cap L^2(V)$.

By Riesz Thm. $\exists \xi \in L^2(\frac{1}{2}(\mu+V))$. s.t.

$\int g \, d\mu = \int g \xi \, d\frac{1}{2}(\mu+V)$. $\forall g \in L^2(\frac{1}{2}(\mu+V))$

$\Rightarrow \int g(2-\xi) d\mu = \int g \xi \, dV$.

Set $A = [\xi \leq 0]$. We see that:

$0 \leq \mu(A) \leq \frac{1}{2} \int I_A (2-\xi) d\mu = \int I_A \xi \, dV$

$\leq 0 \Rightarrow \mu(A) = 0 = V(A)$.

So: $V(B) = \int I_B \cdot \xi(x)/\xi(x) \, dV$

$= \int I_B \frac{2-\xi}{\xi} d\mu$. $\Rightarrow Z = \frac{2-\xi}{\xi}$

Let $A = \{Z < 0\}$. Then:

$0 \leq V(A) = \int I_A Z \, d\mu \overset{(*)}{\leq} 0$. $\overset{(*)}{\Rightarrow} \mu(A) = 0$.

<u>Rmk</u>: $Z(x) = dV/d\mu\,(x)$ is odd of V v.s.

μ at observed data pt. x.

(2) <u>Kullback-Leibler Div.</u>:

Def: For $\mu, \nu \in \mathcal{M}_1^+(\mathbb{R}^d)$, $\nu \ll \mu$. KL div. is

$$d_{KL}(\mu \| \nu) = \int (\log \frac{d\mu}{d\nu}(x)) \cdot d\mu(x)$$

Rmk: i) If $\nu \not\ll \mu$. We set $d_{KL}(\mu\|\nu) = \infty$

ii) If $\mu \sim \nu$. We have $d_{KL}(\mu\|\nu)$

$$= \int -\log(\frac{d\nu(x)}{d\mu(x)})\, d\mu(x).$$

Lem. (Jensen inequi.)

$g : I \underset{\text{interval}}{\subseteq} \mathbb{R}^1 \to \mathbb{R}^1$. Convex. For X. $g(X) \in L^1$.

$\Rightarrow \mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$

Besides if g is strictly convex. and X isn't deterministic. $\Rightarrow \mathbb{E}(g(X)) > g(\mathbb{E}(X))$

Thm. For $\mu, \nu \in \mathcal{M}_1^+(\mathbb{R}^d)$. Then: $d_{KL}(\mu\|\nu) \geq 0$ and $d_{KL}(\mu\|\nu) = 0 \iff \mu = \nu$.

Rmk: d_{KL} is a divergence but not a metric on $\mathcal{M}_1^+(\mathbb{R}^d)$.

Claim: There's no metric $d(\cdot,\cdot)$ on $\mathcal{M}_1^+(\mathbb{R}^d)$. s.t. $d \sim d_{KL}$.

Pf: It's because d_{KL} isn't sym:

$d_{KL}(\mathcal{U}[0,d_m] \| \mathcal{U}[0,1]) \xrightarrow{m\to\infty} 0. \quad (d_m \downarrow 1)$

But $d_{KL}(\mathcal{U}[0,1] \| \mathcal{U}[0,d_m]) \equiv \infty$

Since $d\mathcal{U}[0,d_m] / d\mathcal{U}[0,1] = \frac{1}{d_m}$ and

$\mathcal{U}[0,d_m] \not\ll \mathcal{U}[0,1]$.

<u>Pf</u>: $d_{KL}(\mu \| \nu) = \int \left(\log \frac{d\mu}{d\nu}\right) \frac{d\mu}{d\nu}\, d\nu$

$f(t) = t\log t$ is strictly convex on $\mathbb{R}^{>0}$.

$\Rightarrow \mathbb{E}_\nu(f(X)) \geq f(\mathbb{E}_\nu(X)) = 0$.

with "=" holds if $X = \frac{d\mu}{d\nu} \equiv \text{const}$.

(But $\mu(\mathbb{R}^d) = \text{const} \cdot \nu(\mathbb{R}^d) = 1 \Rightarrow c = 1$)

<u>Def</u>: For $\mu, \nu \in \mathcal{M}_1^+(\mathbb{R}^d)$, $\mu \sim \nu$. Jensen-Shannon div. is $d_{JS}(\mu \| \nu) = \frac{1}{2}\left(d_{KL}\left(\mu \| \frac{\mu+\nu}{2}\right) + d_{KL}\left(\nu \| \frac{\mu+\nu}{2}\right)\right)$

<u>Rmk</u>: d_{JS} is also a div. which's of importance in GANs.

Comparison of Topo.

Next, we want to prove:

$$d_{KL} \Rightarrow d_{TV} \Rightarrow d_{RD} \underset{set}{\overset{cpt}{\Rightarrow}} \underbrace{d_{W,(s)} \Rightarrow \text{weak topo}}_{\text{equi in cpt set.}}$$

$$d_{JS} \Longleftrightarrow$$

<u>Rmk</u>: These relation can help us transit some property (e.g. measurability) from one topo. to another topo.

<u>Prop</u>. (Gibbs' variational principle)

$X : (\mathbb{R}^k, \mathcal{B}_{\mathbb{R}^k}, \mu) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ with CGF $\Psi_{X,\mu}(\alpha)$

Then: $\forall \alpha > 0$. Set $\mathbb{E}_\nu(X) = -\infty$. if $\mathbb{E}_\nu(X_-) = -\infty$.

$$\Psi_{X,\mu}(\alpha) = \sup_{\nu \in M_1(\mathbb{R}^k)} (\alpha \mathbb{E}_\nu(X) - d_{KL}(\nu \| \mu))$$

<u>Rmk</u>: Note that we don't restrict on L^1-r.v. on condition.

<u>Pf</u>: Set $X_n = X \wedge n$. And define:

$$d\mu_{\alpha,n}(x) = e^{\alpha X_n} d\mu(x) / \mathbb{E}_\mu(e^{\alpha X_n})$$

(It's well-def since X_n is upper-bdd)

For $d_{KL}(\nu \| \mu) < \infty$. (since $d_{KL}(\nu\|\nu) = 0$ and RHS take supremum ≥ 0)

$\Psi_{X_n,\mu}(\alpha) - d_{KL}(\nu \| \mu_{\alpha,n}) =$

$\Psi_{X_n,\mu}(\alpha) + \int \log(d\mu_{\alpha,n}/d\nu)\, d\nu =$

$\Psi_{X_n,\mu}(\alpha) + \int [\log(d\mu_{\alpha,n}/d\mu) + \log(d\mu/d\nu)]\, d\nu$

$= \Psi_{X_n,\mu}(\alpha) + \mathbb{E}_\nu(\alpha X_n) - \Psi_{X_n,\mu}(\alpha) - d_{KL}(\nu\|\mu).$

$= \mathbb{E}_\nu(\alpha X_n) - d_{KL}(\nu\|\mu).$

Take $\sup_{\nu \in \mathcal{M}_1^+(\Omega)}$ on both sides:

LHS $= \Psi_{X_n,\mu}(\alpha)$ by: $\nu = \mu_{\alpha,n}$ is optimal.

Next, we take $\sup\limits_n$ on both sides:

By MCT, $\Psi_{X_n,\mu}(\alpha) \uparrow \Psi_{X,\mu}(\alpha)$ with

$$\mathbb{E}_\nu(\alpha X_n) \uparrow \mathbb{E}_\nu(\alpha X).$$

(It's consistent with $\mathbb{E}_\nu(X_-) = -\infty$ since $(X_n)_- = X_- \Rightarrow \mathbb{E}_\nu(X_n) = -\infty$)

<u>Lem.</u> $\sup\limits_x \sup\limits_y f(x,y) \overset{a)}{=} \sup\limits_y \sup\limits_x f(x,y) \overset{b)}{=} \sup\limits_{x,y} f(x,y)$

<u>Pf</u>: Note $\sup\limits_y f(x,y) \geq f(x,y)$. $\Rightarrow$ a): $\geq$ ✓

And by sym. $\Rightarrow$ a) : $\le$ holds as well.

For b). Note: $\exists (x_k, y_k)$. st.

$f(x_k, y_k) \to \sup_{x,y} f(x,y)$. So we have:

$\sup_x \sup_y f(x,y) \geq \sup_y f(x_k, y) \geq f(x_k, y_k)$

Also $f(x,y) \le \sup_{x,y} f(x,y) \Rightarrow$ "$\le$" holds.

$\Rightarrow \Psi_{X,\mu}(\tau) = \sup_{n} \sup_{v} (\mathbb{E}_v(\tau X_n) - \lambda_{KL}(v \| \mu_n))$

$= \sup_v \sup_n \mathbb{E}_v(\tau X_n) - \lambda_{KL}(v \| \mu_n)$

$= \sup_{v \in M_+^1(\mathbb{R}^d)} \mathbb{E}_v(\tau X) - \lambda_{KL}(v \| \mu)$.

Thm. (Bobkov - Götze)

$\mu \in M_+^1(\mathbb{R}^d)$. δ is metric on $\mathbb{R}^d$. St. $x \in \mathbb{R}^d$ $\mapsto \delta(x,y)$ is Borel - measurable. $\forall y \in \mathbb{R}^d$. For

i) For $X \sim \mu$, $g \in Lip(\delta)$. $Y_g = g(X) - \mathbb{E}_\mu(g(X))$ is subgaussian with var. proxy $\sigma^2 > 0$ that doesn't depend on g.

ii) $\mathcal{W}_{1,(\delta)}(v \| \mu) \le (2\sigma^2 \lambda_{KL}(v \| \mu))^{\frac{1}{2}}$. $\forall v \in M_+^1$.

Then: We have i) $\Leftrightarrow$ ii).

Pf: i) $\Leftrightarrow$ $\Psi_{Y_g, \mu}(\tau) \le \frac{1}{2}\sigma^2\tau^2$. $\forall g \in Lip(\delta)$. $\forall \tau \geq 0$

$\iff \sup_{g \in Lip(\delta)} \sup_{\alpha \geq 0} \sup_{\nu \in \mathcal{M}_1^+} (E_\nu(\alpha Y_g) - \lambda KL(\nu \| \mu) - \frac{1}{2}\sigma^2\alpha^2)$

≤ 0 from variational principle.

$$LHS \overset{Y_g = \cdots}{=} \sup_\nu \sup_\alpha \sup_g \{\alpha(\int g\, d\nu - \int g\, d\mu) - \frac{\sigma^2}{2}\alpha^2 - \lambda KL(\nu \| \mu)\}.$$

$$\overset{\sup_{g \in Lip(\delta)}}{=} \sup_\nu \sup_\alpha (\alpha\, \lambda_{W_1(\delta)}(\nu \| \mu) - \frac{1}{2}\sigma^2\alpha^2 - \lambda KL(\nu \| \mu))$$

$$\overset{\alpha = \square}{=} \sup_\nu \{(2\sigma^2)^{-1} d_{W_1(\delta)}(\nu \| \mu)^2 - \lambda KL(\nu \| \mu)\}$$

So: $LHS \leq 0 \iff$ ii) holds.

__eg.__ Consider $\delta = |\cdot|$. $B \subseteq \mathbb{R}^d$. bdd domain with radius r_B. i.e. $\exists\, y_\varepsilon \in B$, s.t. $\sup_{x \in B} |y_\varepsilon - x| \leq r_B + \varepsilon$. Next, restrict $\nu, \mu \in \mathcal{M}_1^+(B)$:

First note Thm above works when we replace $Lip(\delta)$ by $Lip_y(\delta)$ and $Y_g = Y_{g^*}$ if $g - g^* = \text{const}$. So let $y = y_\varepsilon$. Then:

$|g(x)| \leq r_B + \varepsilon$, ν-a.s. $\forall \nu \in \mathcal{M}_1^+(B)$. $g \in Lip_{y_\varepsilon}(\delta)$.

$\Rightarrow$ By Hoeffding inequ.: $\sigma^2 = (r_B + \varepsilon)^2$.

So: $d_{W_1(|\cdot|)}(\nu \| \mu) \leq (r_B + \varepsilon)(2\lambda KL(\nu \| \mu))^{\frac{1}{2}}$

$\xrightarrow{\varepsilon \to 0} r_B (2\lambda KL(\nu \| \mu))^{\frac{1}{2}}$.

Note the Bobkov & Götze Thm above can be applied in any metric δ on $\mathbb{R}^d$.

<u>Def</u>: Metric of the discrete type $\delta_d(x,y)$ is:

$\delta_d(x,y) = I\{x \neq y\}.\ \forall x,y \in \mathbb{R}^d.$

<u>Rmk</u>: It's real metric. (check triangle inequi.)

<u>Lem</u>. i) $\mathcal{M}_1^{+,\delta_d}(\mathbb{R}^d) = \mathcal{M}_1^+(\mathbb{R}^d)$.

ii) $Lip(\delta_d) = \{g \in L^\infty(\mathbb{R}^d) \mid \sup_{\mathbb{R}^d} g - \inf_{\mathbb{R}^d} g \leq 1\}$.

iii) $d_{TV}(\mu,\nu) = 2\, d_{W_1(\delta_d)}(\mu,\nu)$ on $\mathcal{M}_1^+(\mathbb{R}^d)$.

<u>Pf</u>: i) Since $\{x\} \in \mathcal{B}_{\mathbb{R}^d} \Rightarrow y \mapsto \delta_d(y,x)$ is measurable for $\forall x$.

And note: $\int \delta_d(x,y)\, d\nu = 1 - \nu(\{x\}) \leq 1$.

ii) "$\leq$": $\sup g - \inf g = \sup_{x,y} |g(x) - g(y)|$

$\leq \delta_d(x,y) \leq 1$.

"$\supseteq$": $\sup_{x,y} |g(x) - g(y)| \leq$

$(\sup g - \inf g)\, \delta_d(x,y) \leq \delta_d(x,y)$.

iii) Set $g^*(x) = g(x) - \frac{1}{2}(\sup g + \inf g)$ for

$g \in Lip(\delta_d) \overset{ii)}{\Rightarrow} \|g^*\|_\infty \leq \frac{1}{2}.\ g^* \in Lip(\delta_d)$

Note $d_{TV}(\mu,\nu) = 2 \sup_{g\in L^\infty, \|g\|_\infty \le \frac{1}{2}} |\int g\,d\mu - \int g\,d\nu|$

$\le 2 \sup_{g \in Lip(\delta_d)} |\int g\,d\mu - \int g\,d\nu|$.

$= 2\, d_{W_1(\delta_d)}(\mu,\nu)$

$= 2 \sup_{g\in Lip(\delta_d)} |\int g^*\,d\mu - \int g^*\,d\nu|$.

$\le 2 \sup_{g^*\in L^\infty, \|g^*\|_\infty \le \frac{1}{2}} |\int g^*\,d\mu - \int g^*\,d\nu| = d_{TV}(\mu,\nu)$.

Thm. (Pinsker's inequality)

$d_{TV}(\mu,\nu) \le (2\, d_{KL}(\nu\|\mu))^{\frac{1}{2}}$ on $M_1^+(\mathcal{X})$.

Cor. KL-topo is stronger than TV-topo and Radon-topo.

Pf: By Lem ii) above. $\forall g \in Lip(\delta_d)$. $\overset{\text{Doeffling}}{\Rightarrow}$

Y_g is subgaussian with var. proxy $1/4$.

So: $d_{TV}(\mu,\nu) \overset{\text{Lem ii)}}{=} 2\, d_{W_1(\delta_d)}(\mu,\nu)$

$\overset{\text{Thm}}{\le} 2\,(2\cdot \frac{1}{4}\cdot d_{KL}(\mu,\nu))^{\frac{1}{2}}$.