

Oct 1

Last Lecture

1. Given a convex function $f: \mathbb{R} \rightarrow \mathbb{R}$ satisfying $f(1) = 0$, we define

$$D_f(p||q) := \mathbb{E}_q \left[f\left(\frac{p(x)}{q(x)}\right) \right] = \begin{cases} \sum q(w) f\left(\frac{p(w)}{q(w)}\right) & \text{discrete} \\ \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx & \text{continuous} \end{cases}$$

2. TV, KL, χ^2 , χ^q are instances of f -divergence

3. Rényi divergence D_α is a monotone function of χ^q

4. Hockey-stick divergence: $\mathbb{E}_r(p||q) := D_f(p||q)$ with

$$f(x) = \begin{cases} (x-r)_+ & r \geq 1 \\ (r-x)_+ & r < 1 \end{cases}$$

 $r \geq 0$

Wasserstein distance
is NOT f -diverg



$$\hookrightarrow E_r(p||q) = \sum a_{ij} \cdot \left(r - \frac{p_{ij}}{a_{ij}} \right)_+ \quad r \leq 1$$

$$\text{Properties of } E_r\text{-divergence} = \sum (r a_{ij} - p_{ij})_+$$

1. It inherits all the properties of D_f :

$$\cdot) E_r(p||q) \geq 0$$

$\hookrightarrow E_r(p||q)$ is jointly convex

$$\cdot) (p, q) \mapsto E_r(p||q) \text{ is convex}$$

$\cdot)$ conditioning increases E_r -divergence

$$\cdot) D_{PI}$$

2. If $p=q$, then $E_r(p||q)=0$. Converse is NOT true!

$$\text{If for some } r \geq 0 \quad E_r(p||q)=0 \quad \Rightarrow \quad p=q$$

Hockey-stick divergence is an asymmetric generalization of TV.

3. $E_r(p||q) = TV(p, q)$

4. For any p, q : $r \mapsto E_r(p||q)$ is non-increasing & convex for $r \geq 1$ & is non-decreasing & convex for $r \leq 1$.

proof. For $r \geq 1$: $E_r(p||q) = \sum (p(x) - r q(x))_+$ so monotonicity is immediate.

How to prove convexity: we need to prove: $r_t := t r_1 + \bar{t} r_2$ $t \in [0, 1]$

$E_{r_t}(p||q) \stackrel{\text{def of convexity}}{\leq} t E_{r_1}(p||q) + \bar{t} E_{r_2}(p||q)$

LHS: $E_{r_t}(p||q) = \sum (p(x) - r_t q(x))_+$

$= \sum (p(x) - (t r_1 + \bar{t} r_2) q(x))_+$

$= \sum (t (p(x) - r_1 q(x)) + \bar{t} (p(x) - r_2 q(x)))_+$

Jensen's
ineq

$\leq t \sum (p(x) - r_1 q(x))_+ + \bar{t} \sum (p(x) - r_2 q(x))_+$

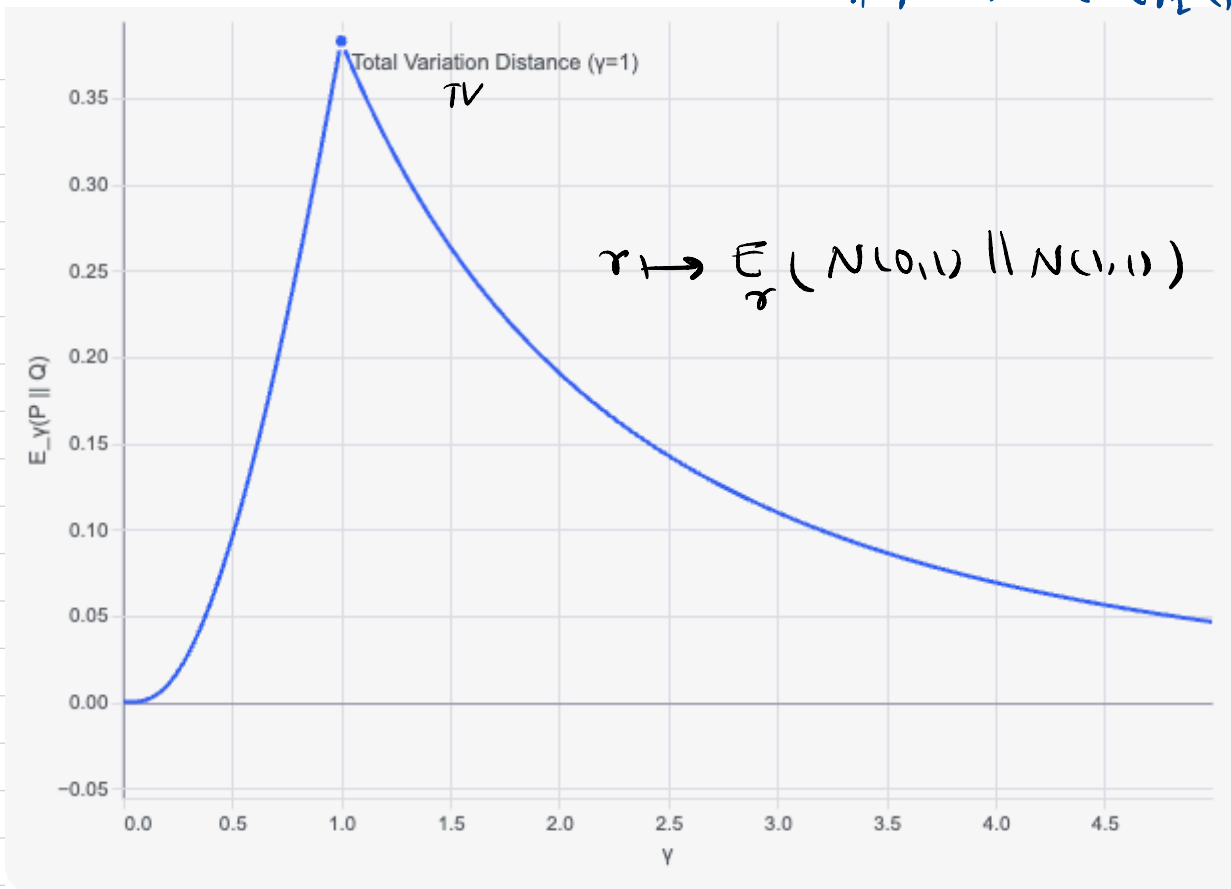
$\text{Pos}(x)_+ = (x)_+$

5. $\lim_{r \rightarrow 0} E_r(p||q) = 0$

$\lim_{r \rightarrow \infty} E_r(p||q) = 0$

$$= t \cdot E_{\tau_1}(p||q) + \bar{t} E_{\tau_2}(p||q)$$

□



6. Reciprocity: $E_r(p||q) = r E_{1/r}(q||p) \quad \forall r > 0$

proof. Suppose $r \geq 1$

$$E_r(p||q) = \sum (p(x) - r q(x))_+$$

$$E_{1/r}(q||p) = \sum (\frac{1}{r} p(x) - q(x))_+$$

RHS: $r E_{1/r}(q||p) = r \cdot \left[\sum (\frac{1}{r} p(x) - q(x))_+ \right]$

$$= \sum (p(x) - r q(x))_+$$

$$= E_r(p||q)$$

If $r < 1$:

$$E_r(p||q) = r \cdot E_{\frac{1}{r}}(q||p)$$

$\frac{1}{r} \geq 1$

From now on, we can restrict to $r \geq 1$

7. Variational expression:

Recall that:

$$TV(p, q) = \sup_A [p(A) - q(A)]$$

For any $r \geq 1$ & distributions p & q :

$$E_r(p||q) = \sup_A [p(A) - r q(A)]$$

proof: we need to prove:

1- $\forall A$ we have: $p(A) - r q(A) \leq E_r(p||q)$

2- $\exists A^*$ such that: $p(A^*) - r q(A^*) = E_r(p||q)$

$$\begin{aligned} \downarrow \\ \sup_A [p(A) - r q(A)] &\geq p(A^*) - r q(A^*) \\ &= E_r(p||q) \end{aligned}$$

1: Pick any event A :

we throw away negative terms

$$p(A) - r Q(A) = \sum_{x \in A} p(x) - r \cdot Q(x) \leq \sum_{x \in A} (p(x) - r Q(x))_+$$

$$\leq \sum_x (p(x) - r Q(x))_+ = E_r(p \parallel Q). \quad (r \geq 1)$$

2- Define

$$A^* = \{x: p(x) \geq r \cdot Q(x)\}$$

$$p(A^*) - r Q(A^*) = \sum_{x \in A^*} p(x) - r Q(x) = \sum_{x \in A^*} (p(x) - r Q(x))_+$$

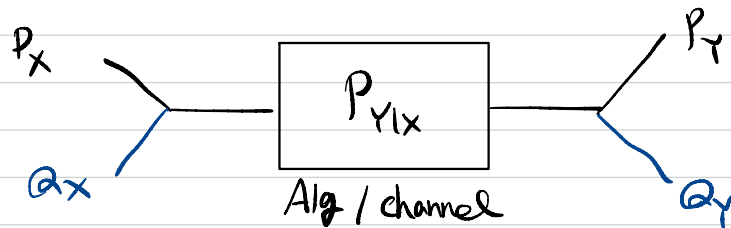
$$= \sum_{x \in A^*} (p(x) - r Q(x))_+ + \underbrace{\sum_{x \in A^{*c}} (p(x) - r Q(x))_+}_{=0}$$

$$= \sum_x (p(x) - r Q(x))_+ = E_r(p \parallel Q)$$

□

Strong DPI

Recall that all f -divergences satisfy DPI:



$$D_f(P_Y \| Q_Y) \leq D_f(P_X \| Q_X)$$

output divergence input divergence

we want to find $\eta \leq 1$ such that

$$D_f(P_Y \| Q_Y) \leq \eta D_f(P_X \| Q_X)$$

$$\eta \geq \frac{D_f(P_Y \| Q_Y)}{D_f(P_X \| Q_X)}$$

if $\exists \eta < 1$ then we
say that the algorithm
satisfies **Strong** DPI
for the f -divergence!

we want this inequality to hold for all P_X & Q_X .

$$\eta \geq \sup_{\substack{P_X, Q_X \\ D_f(P_X \| Q_X) \neq 0}} \frac{D_f(P_Y \| Q_Y)}{D_f(P_X \| Q_X)} =: \eta_f(P_{Y|X})$$

contraction coefficient
of $P_{Y|X}$ under
 f -divergence

SDPI :

$$D_f(P_Y \| Q_Y) \leq \eta_f(P_{Y|X}) D_f(P_X \| Q_X)$$

✓ P_X & Q_X

How to compute $\eta_{\text{TV}}(P_{Y|X})$? It is generally very challenging to compute because the feasible set (the set where P_X & Q_X takes value in) is very large. (infinite-dimensional)

Theorem (Dobrushin) : For any $P_{Y|X}$:
1950

$$\eta_{\text{TV}}(P_{Y|X}) := \sup_{P_X \neq Q_X} \frac{\text{TV}(P_Y, Q_Y)}{\text{TV}(P_X, Q_X)} = \max_{x_1, x_2} \text{TV}(P_{Y|X=x_1}, P_{Y|X=x_2})$$

contraction coefficient
of $P_{Y|X}$ under TV

Two-point characterization
of contraction of TV
coefficient

proof. we need to prove the following:

$$\max_{x_1, x_2} TV(P_{Y|X=x_1}, P_{Y|X=x_2}) \leq \eta_{TV}(P_{Y|X})$$

$\uparrow$

$$1- \forall x_1, x_2: TV(P_{Y|X=x_1}, P_{Y|X=x_2}) \leq \eta_{TV}(P_{Y|X})$$

$$2- \max_{x_1, x_2} TV(P_{Y|X=x_1}, P_{Y|X=x_2}) \geq \eta_{TV}(P_{Y|X})$$

$$1- \text{Begin from RHS: } \eta_{TV}(P_{Y|X}) = \sup_{P_X \neq Q_X} \frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)} \geq \frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)}$$

Fix $P_X = \delta_{x_1}$ distribution on $\{x_1, x_2\}$ for a fixed (P_X, Q_X)

$$P_X(x_1) = 1 \quad \text{for a fixed } x_1$$

$Q_X = \delta_{x_2}$ distribution on $\{x_2\}$ where $x_1 \neq x_2$

$$Q_X(x_2) = 1$$

Therefore

$$\eta_{TV}(P_{Y|X}) \geq \frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)} \quad (*)$$



note that $TV(P_X, Q_X) = \frac{1}{2} \sum_x |P_X(x) - Q_X(x)| = \frac{1}{2} [|P_X(x_1) - Q_X(x_1)| + |P_X(x_2) - Q_X(x_2)|]$

$$= \frac{1}{2} [1 + 1] = 1$$

when $P_X = \delta_{x_1}$ then $P_Y(y) = \sum_x P_X(x) \cdot P_{Y|X}(y|x)$

$$= P_X(x_1) \cdot P_{Y|X}(y|x_1)$$
$$= 1 \cdot P_{Y|X}(y|x_1)$$

$$\Rightarrow P_Y = P_{Y|X=x_1}$$

similarly:

$$Q_X = \delta_{x_2} \text{ then } Q_Y = P_{Y|X=x_2}$$

$$\} \Rightarrow TV(P_Y, Q_Y) = TV(P_{Y|X=x_1}, P_{Y|X=x_2})$$

From (*), we have: $\eta_{TV}(P_{Y|X}) \geq \frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)} = TV(P_{Y|X=a_1}, P_{Y|X=a_2})$

2. Fix two distributions P_X & Q_X & define:

$$A = \{x: P_X(x) \geq Q_X(x)\}$$

Now, consider the corresponding output dist P_Y & Q_Y .

For any event B , we can write:

$$\begin{aligned} P_Y(B) - Q_Y(B) &= \sum_y P_Y(y) - Q_Y(y) \\ &= \sum_x [P_{Y|X}(B|x) P_X(x) - P_{Y|X}(B|x) \cdot Q_X(x)] \end{aligned}$$

$$= \sum_x P_{Y|X}(B|x) \cdot [p_x(u) - q_x(u)]$$

$$= \sum_{x \in A} P_{Y|X}(B|x) \cdot [p_x(u) - q_x(u)]$$

$$+ \sum_{x \in A^c} P_{Y|X}(B|x) \cdot [p_x(u) - q_x(u)]$$

$$= \sum_{x \in A} P_{Y|X}(B|x) \cdot [p_x(u) - q_x(u)]$$

$$- \sum_{x \in A^c} P_{Y|X}(B|x) \cdot [q_x(u) - p_x(u)]$$

$$= \text{TV}(p_x, q_x) \left[\sum_{x \in A} P_{Y|X}(B|x) \cdot [p_x(u) - q_x(u)] \cdot \frac{1}{\text{TV}(p_x, q_x)} \right. \\ \left. - \sum_{x \in A^c} P_{Y|X}(B|x) \cdot [q_x(u) - p_x(u)] \cdot \frac{1}{\text{TV}(p_x, q_x)} \right]$$

$\uparrow$ $U(x)$

$$u(x) := \frac{p_x(x) - q_x(x)}{TV(p_x, q_x)}$$

on A

$$\begin{aligned} \sum_{x \in A} u(x) &= \sum_{x \in A} \frac{p_x(x) - q_x(x)}{TV(p_x, q_x)} \\ &= 1 \quad \text{why?} \end{aligned}$$

$\hookrightarrow v(x)$

$$v(x) := \frac{q_x(x) - p_x(x)}{TV(p_x, q_x)}$$

on A^c

$$\begin{aligned} \sum_{x \in A^c} v(x) &= \sum_{x \in A^c} \frac{q_x(x) - p_x(x)}{TV(p_x, q_x)} \\ &= 1 \quad \text{why?} \end{aligned}$$

continue
from above
 $\downarrow$

$$= TV(p_x, q_x) \left[\sum_{x \in A} p_{Y|X}(B|x) u(x) - \sum_{x \in A^c} p_{Y|X}(B|x) v(x) \right]$$

$$= TV(p_x, q_x) \left[\sum_{x \in A} p_{Y|X}(B|x) u(x) - \sum_{x' \in A^c} p_{Y|X}(B|x') v(x') \right]$$

$$= TV(P_X, Q_X) \left[\sum_{x' \in A^c} \sum_{x \in A} P_{Y|X}(B|x) \cdot U(x) \cdot V(x') - \sum_{x' \in A^c} \sum_{x \in A} P_{Y|X}(B|x') \cdot V(x') \cdot U(x) \right]$$

$$= TV(P_X, Q_X) \sum_{x' \in A^c} \sum_{x \in A} U(x) \cdot V(x') \cdot \underbrace{[P_{Y|X}(B|x) - P_{Y|X}(B|x')]}_{\leq \max_{x, x'} [P_{Y|X}(B|x) - P_{Y|X}(B|x')]}$$

$$\leq TV(P_X, Q_X) \max_{x, x'} [P_{Y|X}(B|x) - P_{Y|X}(B|x')] \underbrace{\sum_{x' \in A^c} \sum_{x \in A} U(x) \cdot V(x')}_{=1}$$

$$= TV(P_X, Q_X) \cdot \max_{x, x'} [P_{Y|X}(B|x) - P_{Y|X}(B|x')]$$

$$\begin{aligned}
 P_Y(B) - Q_Y(B) &\leq TV(P_X, Q_X) \cdot \max_{x, x^-} \left[\underbrace{P_{Y|X}(B|x) - P_{Y|X}(B|x^-)}_{\leq \sup_B [P_{Y|X}(B|x) - P_{Y|X}(B|x^-)]} \right] \\
 &= TV(P_{Y|X=x}, P_{Y|X=x^-})
 \end{aligned}$$

$$P_Y(B) - Q_Y(B) \leq TV(P_X, Q_X) \cdot \max_{x, x^-} TV(P_{Y|X=x}, P_{Y|X=x^-})$$

Taking Supremum from both sides over B :

$$TV(P_Y, Q_Y) \leq TV(P_X, Q_X) \cdot \max_{x, x^-} TV(P_{Y|X=x}, P_{Y|X=x^-})$$

$$\Rightarrow \max_{x, x^-} TV(P_{Y|X=x}, P_{Y|X=x^-}) \geq$$

$$\frac{TV(P_Y, Q_Y)}{TV(P_X, Q_X)}$$

By taking supremum over $P_X \neq Q_X$:

$$\max_{x, x'} TV(P_{Y|X=x}, P_{Y|X=x'}) \geq \eta_{TV}(P_{Y|X})$$



For many years (from 50), TV was the only f-divergence whose contraction coefficient was easy to compute!

Despite many applications of other contraction coefficients, (such as η_{KL} , η_{x^2}), they weren't easily computable.

— η_{x^2} is closely related to Rényi maximal correlation & spectral methods in graph theory.

- η_{KL} is "dissipation of information in a network"

a simple closed-form maximization

Recently, $\sqrt{\cdot}$ for contraction coefficient under E_r -divergence was proved.

$$\eta_r(P_{Y|X}) := \sup_{\substack{P_X, Q_X \\ E_r(P_X || Q_X) \neq 0}} \frac{E_r(P_Y || Q_Y)}{E_r(P_X || Q_X)} \quad r \geq 1$$

Theorem.

For any $r \geq 1$ & any $P_{Y|X}$:

$$\eta_r(P_{Y|X}) = \max_{x_1, x_2} E_r(P_{Y|X=x_1} || P_{Y|X=x_2})$$

this is the generalization of Dobrushin's two-point characterization of η_{TV} .

proof is in HW1.

Remark 2: This theorem gives $\eta_r(P_{Y|X})$ for $r \geq 1$.

What about $\eta_r(P_{Y|X})$ for $r < 1$?

For $r < 1$:

$$\eta_r(P_{Y|X}) = \sup_{P_X, Q_X} \frac{E_r(P_Y \| Q_Y)}{E_r(P_X \| Q_X)} = \sup_{P_X, Q_X} \frac{r \cdot E_{Yr}(Q_Y \| P_Y)}{r \cdot E_{Yr}(Q_X \| P_X)} = \eta_{Yr}(P_{Y|X})$$

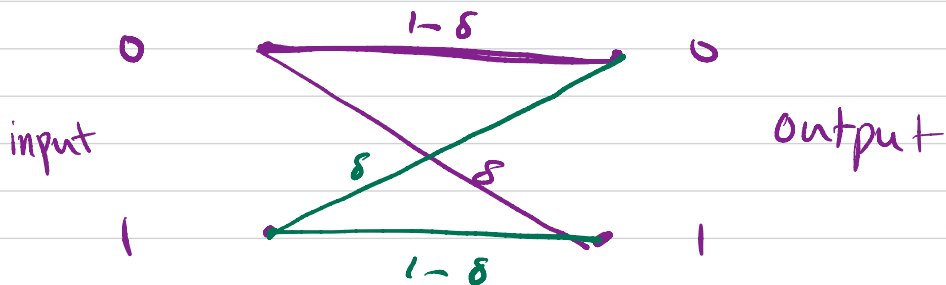
$\Rightarrow$ For $r < 1$:

$$\eta_r(P_{Y|X}) = \eta_{Yr}(P_{Y|X}) = \max_{x_1, x_2} E_{Yr}(P_{Y|X=x_1} \| P_{Y|X=x_2})$$

Example: consider channel $P_{Y|X}$ with binary input:

$$P_{Y|X=0} = \text{Bernoulli}(\delta)$$

$$P_{Y|X=1} = \text{Bernoulli}(1-\delta)$$



compute $\eta_r(P_{Y|X})$.
for $\delta \geq 1$

According to the theorem:

$$\eta_r(P_{Y|X}) = \max_{x_1, x_2} \mathbb{E}_\delta(P_{Y|X=x_1} \| P_{Y|X=x_2})$$

$$\begin{aligned}
 &= \max \left\{ \mathbb{E}_r(P_{Y|X=0} \| P_{Y|X=1}), \mathbb{E}_r(P_{Y|X=1} \| P_{Y|X=0}) \right\} \\
 &= \max \left\{ \mathbb{E}_Y(\text{Bernoulli}(\delta) \| \text{Bernoulli}(1-\delta)) \right. \\
 &\quad \left. , \mathbb{E}_Y(\text{Bernoulli}(1-\delta) \| \text{Bernoulli}(\delta)) \right\}
 \end{aligned}$$

$$\delta < 1/2$$

$$\downarrow \\
 = \max \left\{ (\bar{\delta} - r\delta)_+, (\bar{\delta} - r\delta)_+ \right\}$$

$$= (\bar{\delta} - r\delta)_+$$

$$\eta_{TV}(P_{Y|X}) = (\bar{\delta} - \delta)_+ = 1 - 2\delta$$

$$\eta_r(P_{Y|X}) = \frac{1}{r} \cdot (r\bar{\delta} - \delta)_+$$

for $r < 1$

Optional:

joint Range of f -divergence

Recall Pinsker's inequality:

$$2TV^2(p, q) \leq D(p||q) \quad \forall p, q$$

$$\text{If } D(p||q) \leq b \Rightarrow TV(p, q) \leq \sqrt{b/2}$$

$$\text{If } TV(p, q) \geq a \Rightarrow D(p||q) \geq 2a^2$$

Generally, these two lead to the following problems:

$$\begin{array}{l} \sup_{\substack{p, q \\ D(p||q) \leq b}} TV(p, q) \leq \sqrt{b/2} \end{array}$$

$$\inf_{p, q} D(p \| q) \geq 2q^2$$

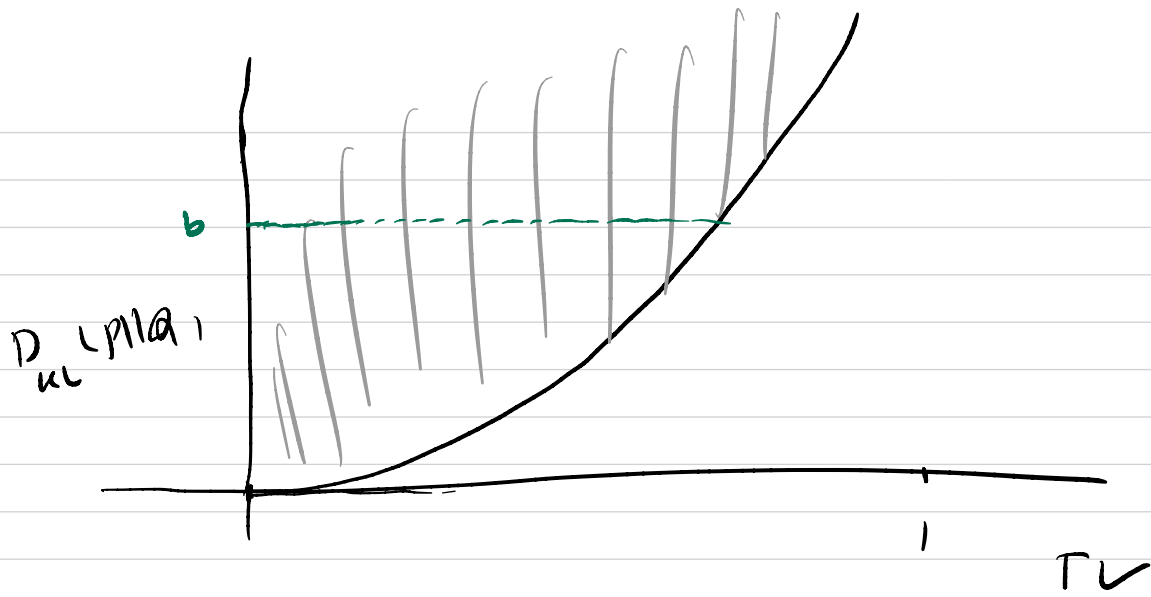
$$TV(p, q) \geq a$$

Both these two optimization problems can be cast as boundary of the following set:

$$R := \left\{ (TV(p, q), D(p \| q)) : p, q \right\}$$

Two-dimensional subset of $\mathbb{R}^2$

convex.



$b \mapsto \sup_{p, q} TV(p, q)$ is the lower-boundary of R
 $D(p|q) \leq b$

Theorem: R can be characterized by Only considering
(Harrémoes/Vajda) Bernoulli distributions.

$$R^{\text{binary}} := \left\{ \left(TV(\text{Bernoulli}(p), \text{Bernoulli}(q)), D(\text{Bernoulli}(p) \parallel \text{Bernoulli}(q)) \right) \mid 0 \leq p, q \leq 1 \right\}$$

$$R = \text{convex hull}(R^{\text{binary}})$$

* the same thing holds for any other pair
of f -divergence!

$$\sup_{P \in \mathcal{Q}} D_f(P \| Q)$$

$$P \in \mathcal{Q}$$

$$D_g(P \| Q) \leq a$$

$$\inf_{P \in \mathcal{Q}} D_g(P \| Q)$$

$$P \in \mathcal{Q}$$

$$D_f(P \| Q) \geq b$$

↙ joint range of f-divergence

$$R = \{ (D_f(P \| Q), D_g(P \| Q)) : P \in \mathcal{Q} \}$$

Above supremum corresponds to the lower bound of R

& infimum corresponds to the upper bound of R .

To characterize $\mathcal{R}$, you need to consider Only Bernoulli distributions.