Lec 3 : f-divergence

Yanjun Han

Defn. (f-divergence, Csiszár '63)

Let $f : (0, \infty) \to \mathbb{R}$ be convex with $f(1) = 0$. The f-divergence between two distributions P and Q on the same space is

$$D_f(P \| Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right].$$

Remark: 1. Some defn. additionally assumes that $f'(1) = 0$. This is WLOG: $f(x)$ and $f(x) + c(x-1)$ give the same f-divergence.

2. If $\frac{dP}{dQ} = 0$, define $f(0) := f(0+)$;

If $P \not\ll Q$, define $D_f(P \| Q) = \int_{q>0} q f\left(\frac{p}{q}\right) d\mu + f'(\infty) P(q=0)$, with $f'(\infty) := \lim_{x \to \infty} \frac{f(x)}{x}$.

Examples.

☆ 1: $f(x) = \frac{1}{2}|x-1|$: $D_f(P \| Q) = TV(P, Q) = \frac{1}{2} \int |dP - dQ|$

(total variation (TV) distance)

☆ 2: $f(x) = (\sqrt{x} - 1)^2$: $D_f(P \| Q) = H^2(P, Q) = \int (\sqrt{dP} - \sqrt{dQ})^2$

(squared Hellinger distance)

☆ 3: $f(x) = x \log x$: $D_f(P \| Q) = D_{KL}(P \| Q) = \int dP \log \frac{dP}{dQ}$

☆ 4: $f(x) = (x-1)^2$: $D_f(P \| Q) = \chi^2(P \| Q) = \int \frac{(dP - dQ)^2}{dQ}$

(χ^2 divergence)

5. $f(x) = \frac{1-x}{2(1+x)}$: $D_f(P \| Q) = LC(P, Q) = \frac{1}{2} \int \frac{(dP - dQ)^2}{dP + dQ}$

(Le Cam distance)

6. $f(x) = x \log x + (x+1) \log \frac{2}{x+1}$: $D_f(P \| Q) = JS(P, Q) = D_{KL}(P \| \frac{P+Q}{2}) + D_{KL}(Q \| \frac{P+Q}{2})$

(Jensen-Shannon divergence)

Basic properties. ① $D_f(P \| Q) \geq 0$

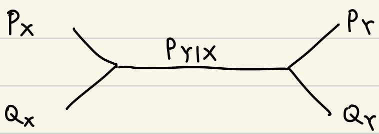
Pf. $D_f(P \| Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \geq f\left(\mathbb{E}_Q\left[\frac{dP}{dQ}\right]\right) = f(1) = 0$ ∎

② $(P, Q) \mapsto D_f(P \| Q)$ is jointly convex.

Pf. For convex f, the perspective transform $\mathbb{R}_+^2 \ni (x, y) \mapsto y f\left(\frac{x}{y}\right)$ is also convex.

Check Hessian: $\begin{bmatrix} \frac{1}{y} f''\left(\frac{x}{y}\right) & -\frac{x}{y^2} f''\left(\frac{x}{y}\right) \\ -\frac{x}{y^2} f''\left(\frac{x}{y}\right) & \frac{x^2}{y^3} f''\left(\frac{x}{y}\right) \end{bmatrix} \succeq 0.$ $\square$

③ Data processing inequality: $D_f(P_X \| Q_X) \geq D_f(P_Y \| Q_Y)$



Pf. Follow from joint convexity (similar to the KL proof) $\square$

<u>Why f-divergence? Binary hypothesis testing.</u>

Recall the simple hypothesis testing problem:

Null H_0: $X \sim P$

Alternative H_1: $X \sim Q$

Test: $T: \mathcal{X} \to \{0, 1\}$

Type I error: $P(T(X) = 1)$

Type II error: $Q(T(X) = 0)$

Thm.

$$\inf_T \Big(P(T(X) = 1) + Q(T(X) = 0) \Big) = 1 - TV(P, Q)$$

Pf. Easy to show $TV(P, Q) = \sup_A P(A) - Q(A)$

($\leq$) Take $T(X) = 1(X \notin A)$ for A attaining the supremum;

($\geq$) Take $A = \{T(X) = 0\}$. $\square$

Remark: ① $TV(P,Q)=0$: $P=Q$. totally indistinguishable

② $TV(P,Q)=1$: $P \perp Q$, perfectly distinguishable

③ $TV(P,Q)<1$: partially indistinguishable

(Important quantity for establishing minimax lower bounds later)

Why not just TV?

① $TV(P,Q)$ can be hard to compute

② TV does not tensorize; e.g. $TV(P^{\otimes n}, Q^{\otimes n}) \le n\,TV(P,Q)$ is the best possible inequality in general, but is often loose.

Example. How large is $TV(\mathrm{Bern}(\frac{1}{2})^{\otimes n}, \mathrm{Bern}(\frac{1}{2}+\delta)^{\otimes n})$?

Using $TV(P^{\otimes n}, Q^{\otimes n}) \le n\,TV(P,Q)$: $n\delta$ upper bound

Using Pinsker's inequality: $TV(P^{\otimes n}, Q^{\otimes n}) \le \sqrt{\frac{1}{2} D_{KL}(P^{\otimes n} \| Q^{\otimes n})}$

$$= \sqrt{\frac{n}{2} D_{KL}(P\|Q)} = O(\sqrt{n}\,\delta)!$$

Popular f-divergences that tensorize:

① H^2 :
$$1-\frac{1}{2}H^2\Big(\prod_i P_i, \prod_i Q_i\Big) = \prod_i \Big(1-\frac{1}{2}H^2(P_i, Q_i)\Big)$$

② KL:
$$D_{KL}\Big(\prod_i P_i \,\Big\|\, \prod_i Q_i\Big) = \sum_i D_{KL}(P_i \| Q_i)$$

③ χ^2 :
$$\chi^2\Big(\prod_i P_i \,\Big\|\, \prod_i Q_i\Big) + 1 = \prod_i \big(\chi^2(P_i \| Q_i) + 1\big).$$

Remark (optional): All of them follow from the tensorization of Rényi divergences

i.e. $D_\lambda(\prod_i P_i \| \prod_i Q_i) = \sum_i D_\lambda(P_i \| Q_i)$, with

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda-1}\log \mathbb{E}_Q\Big[\Big(\frac{dP}{dQ}\Big)^\lambda\Big].$$

For $\lambda = \frac{1}{2}, 1, 2$, D_λ corresponds to H^2, KL and χ^2.

Similarities and differences between f-divergences

Locally χ^2-like: when $f''(1)$ exists and $P \approx Q$:

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$
$$\approx \mathbb{E}_Q\Big[\underbrace{f(1)}_{=0} + f'(1)\underbrace{\left(\frac{dP}{dQ}-1\right)}_{\mathbb{E}_Q[\cdot]=0} + \frac{f''(1)}{2}\left(\frac{dP}{dQ}-1\right)^2\Big]$$
$$= \frac{f''(1)}{2}\chi^2(P\|Q).$$

In parametric models: Fisher information: if $(P_\theta)_{\theta\in\Theta}$ is a "regular" parametric model with $\theta \in \mathbb{R}^d$, then for $h \in \mathbb{R}^d$ and $t \approx 0$:

$$\chi^2(P_{\theta+th}\|P_\theta) = \int \frac{(f_{\theta+th}-f_\theta)^2}{f_\theta}\mu(dx) \qquad \left(\text{assume } \frac{dP_\theta}{d\mu} = f\right)$$
$$\approx t^2 h^T \int \frac{(\dot f_\theta)^2}{f_\theta}\mu(dx)\, h \qquad \left(\dot f_\theta(x) = \frac{\partial f}{\partial\theta}(x)\right)$$
$$=: t^2 h^T I(\theta) h,$$

where $I(\theta) \in \mathbb{R}^{d\times d}$ is the Fisher information:

$$I(\theta) = \int \frac{(\dot f_\theta)^2}{f_\theta} d\mu = \mathbb{E}\left[(\nabla_\theta \log f_\theta(X))(\nabla_\theta \log f_\theta(X))^T\right]$$
$$= \mathbb{E}\left[-\nabla_\theta^2 \log f_\theta(X)\right].$$

f-divergence as "average statistical information".

In binary hypothesis testing, if $\mathbb{P}(H_0) = \pi \in (0,1)$, then the Bayes error is

$$B_\pi(P,Q) = \inf_T \left(\pi P(T(X)=1) + (1-\pi)Q(T(X)=0)\right)$$
$$= \int (\pi\, dP \wedge (1-\pi)\, dQ) \qquad (x\wedge y := \min\{x,y\})$$

The <u>statistical information</u> is the difference between "a priori" and "a posteriori" Bayes losses:

$$I_\pi(P, Q) = \pi \wedge (1-\pi) - B_\pi(P, Q),$$

which is a f-divergence with $f_\pi(t) = \pi \wedge (1-\pi) - (\pi t) \wedge (1-\pi)$.

<u>Thm (Liese & Vajda '06)</u>. For any f-divergence, $\exists$ a measure Γ_f on $(0,1)$ s.t.

$$D_f(P\|Q) = \int_0^1 I_\pi(P, Q)\,\Gamma_f(d\pi). \quad \forall P, Q.$$

Remark: every f-divergence is an "average" statistical information, with different weights on π.

<u>Pf</u>. $f(1) = 0$, and WLOG assume $f'(1) = 0$. Then

$$f(t) = \int_1^t (t-x) f''(dx) \qquad \text{(For } f \in C^2,\ f''(dx) = f''(x)\,dx\text{;}$$

$$\overset{\text{check}}{=} \int_0^1 (x - t\wedge x) f''(dx) \qquad \text{in general, any convex function gives}$$

$$+ \int_1^\infty (t - t\wedge x) f''(dx). \qquad \text{rise to a "measure" } f''(dx)\text{)}$$

Define $\tilde{f}(t) = \int_0^1 (x - t\wedge x) f''(dx) + \int_1^\infty (1 - t\wedge x) f''(dx)$, then

$$\mathbb{E}_Q\left[(f-\tilde{f})\left(\tfrac{dP}{dQ}\right)\right] = \mathbb{E}_Q\left[\int_1^\infty \left(\tfrac{dP}{dQ} - 1\right) f''(dx)\right] = 0.$$

On the other hand,

$$1\wedge x - t\wedge x = (1+x)\left(\tfrac{1}{1+x} \wedge \tfrac{x}{1+x} - \tfrac{t}{1+x} \wedge \tfrac{x}{1+x}\right) = (1+x) f_{\frac{1}{1+x}}(t).$$

So

$$\int_0^\infty (1+x) I_{\frac{1}{1+x}}(P, Q) f''(dx) = \mathbb{E}_Q\left[\int_0^\infty (1+x) f_{\frac{1}{1+x}}\left(\tfrac{dP}{dQ}\right) f''(dx)\right]$$

$$= \mathbb{E}_Q\left[\tilde{f}\left(\tfrac{dP}{dQ}\right)\right] = \mathbb{E}_Q\left[f\left(\tfrac{dP}{dQ}\right)\right] = D_f(P\|Q),$$

and $\Gamma_f(\pi)$ is the pushforward measure of $(1+x) f''(dx)$ by the map

$$x \in (0, \infty) \longmapsto \frac{1}{1+x} \in (0, 1)$$

$\blacksquare$

Different guarantees on contiguity

Def (contiguity) $\{P_n\}$ is contiguous w.r.t. $\{Q_n\}$ (written as $\{P_n\} \triangleleft \{Q_n\}$)
if $Q_n(A_n) \to 0$ implies $P_n(A_n) \to 0$.

Clearly, $TV(P_n, Q_n) \to 0$ implies $\{P_n\} \triangleleft \{Q_n\}$.

In comparison, $KL(P_n \| Q_n) \le C$ already establishes contiguity, as

$$P_n(A_n) \log \frac{P_n(A_n)}{eQ_n(A_n)} \le KL(P_n \| Q_n) \le C \quad \text{(see Lec 2)}$$

$\chi^2(P_n \| Q_n) \le C$ leads to an even stronger guarantee:

$$\frac{(P_n(A_n) - Q_n(A_n))^2}{Q_n(A_n)(1-Q_n(A_n))} \overset{\text{DPI}}{\le} \chi^2(P_n \| Q_n) \le C$$

$$\Rightarrow \quad P_n(A_n) \le Q_n(A_n) + \sqrt{C \cdot Q_n(A_n)}.$$

Therefore, different f-divergences have different powers in establishing contiguity results, due to different growth of $f(t)$ as $t \to \infty$. In this context, a popular choice is to upper bound $\chi^2(P_n \| Q_n)$, known as the "second moment method" in random graph theory & property testing (Lec 8).

Dual representations of f-divergence.

Similar to KL, f-divergences also admit dual representations.

Def (convex conjugate): for a convex function f on $\mathbb{R}$, its convex conjugate is defined as

$$f^*(y) = \sup_x \left(xy - f(x)\right).$$

Properties: ① f^* is convex;
② $f^{**} = f$;
③ <u>Young's inequality</u>: $f(x) + f^*(y) \geq xy$.

The following result is then immediate:

<u>Thm.</u> $$D_f(P\|Q) = \sup_{g:\, \mathbb{E}_Q[f^*\circ g]<\infty} \mathbb{E}_P\, g - \mathbb{E}_Q[f^*\circ g].$$

<u>Pf.</u> $$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \mathbb{E}_Q\left[\sup_y\, y\frac{dP}{dQ} - f^*(y)\right]$$

$$= \sup_{g:\, \mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g] - \mathbb{E}_Q[f^*\circ g]. \qquad \blacksquare$$

<u>Example 1 (TV)</u>: When $f(x) = \frac{1}{2}|x-1|$, $f^*(y) = \begin{cases} y & \text{if } |y| \leq \frac{1}{2} \\ +\infty & \text{if } |y| > \frac{1}{2} \end{cases}$. so

$$TV(P, Q) = \sup_{\|g\|_\infty \leq \frac{1}{2}} \mathbb{E}_P\, g - \mathbb{E}_Q\, g = \frac{1}{2}\sup_{\|g\|_\infty\leq 1} \left|\mathbb{E}_P\, g - \mathbb{E}_Q\, g\right|.$$

<u>Example 2 (KL)</u>, When $f(x) = x\log x$, $f^*(y) = e^{y-1}$, so

$$D_{KL}(P\|Q) = \sup_g \mathbb{E}_P\, g - \mathbb{E}_Q\, e^{g-1} = \sup_g \mathbb{E}_P\, g - (\mathbb{E}_Q e^g - 1).$$

As $\mathbb{E}_Q e^g - 1 \geq \log \mathbb{E}_Q e^g$, this is weaker than Donsker-Varadhan.
A way to recover Donsker-Varadhan is

$$D_{KL}(P\|Q) = \sup_g \sup_{a\in\mathbb{R}} \mathbb{E}_P[g+a] - \mathbb{E}_Q\, e^{g+a-1}$$

$$= \sup_g \left(\mathbb{E}_P[g] - \underbrace{\inf_{a\in\mathbb{R}} \left(\mathbb{E}_Q\, e^{g+a-1} - a\right)}_{= \log \mathbb{E}_Q e^g,\ \text{by taking } a = 1 - \log \mathbb{E}_Q e^g.}\right)$$

Example 3 (χ^2): When $f(x) = (x-1)^2$, $f^*(y) = y + \frac{y^2}{4}$, so

$$\chi^2(P\|Q) = \sup_g \mathbb{E}_P[g] - \mathbb{E}_Q\left[g + \frac{g^2}{4}\right]$$

$$= \sup_g \sup_{\lambda, c \in \mathbb{R}} \mathbb{E}_P[\lambda(g+c)] - \mathbb{E}_Q\left[\lambda(g+c) + \frac{\lambda^2(g+c)^2}{4}\right]$$

$$= \sup_g \frac{(\mathbb{E}_P g - \mathbb{E}_Q g)^2}{\mathrm{Var}_Q\, g}.$$

Corollary (Hammersley-Chapman-Robbins (HCR) lower bound)

In a parametric family $(P_\theta)_{\theta \in \mathbb{R}}$, if an estimator $\hat{\theta}$ is unbiased, then

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'} \| P_\theta)}.$$

In particular, by taking $\theta' \to \theta$, it recovers the Cramér-Rao bound

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

Example 4 (JS): When $f(x) = x\log x + (x+1)\log\frac{2}{x+1}$, $f^*(y) = \begin{cases} -\log(2-e^y), & y < \log 2 \\ +\infty, & y \geq \log 2 \end{cases}$

$$JS(P, Q) = \sup_{g \leq \log 2} \mathbb{E}_P\, g + \mathbb{E}_Q[\log(2-e^g)]$$

$$\overset{h = \frac{e^g}{2}}{=} \sup_{0<h<1} \mathbb{E}_P[\log h] + \mathbb{E}_Q[\log(1-h)] + \log 2.$$

So generative adversarial networks (GAN) aim to minimize

$$\min_{\underset{\text{generator}}{G}} JS(\underset{\text{data distribution}}{P}, P_{G(\underset{\text{noise}}{Z})}) = \min_G \sup_{\underset{\text{discriminator}}{D}} \mathbb{E}_{X\sim P}[\log D(X)] + \mathbb{E}_{Z\sim N}[\log(1-D(G(Z)))]$$

Joint range: given two f-divergences, how to prove inequalities between them?

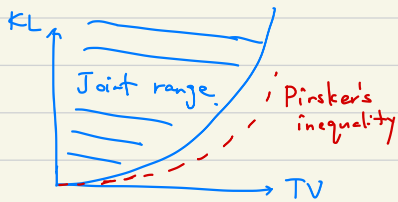
(For example, is there a general paradigm to prove Pinsker's inequality
$$2\,TV(P,Q)^2 \le D_{KL}(P\|Q)\ ?$$
)

Def (Joint range): Fix two f-divergences $D_f(P\|Q)$ and $D_g(P\|Q)$.

Define: $\mathcal{R} = \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ general prob. measures}\}$

$\mathcal{R}_k = \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ prob. measures on } [k]\}$.

Example (TV vs. KL):



$$\boxed{\textbf{Thm (Harremoës–Vajda '11)} \quad \mathcal{R} = \mathrm{conv}(\mathcal{R}_2) = \mathcal{R}_4.}$$

Implication: to establish inequalities between D_f and D_g, suffices to prove them for $P = (p, 1-p)$ and $Q = (q, 1-q)$!

Pf (of a simpler case $P \ll Q$)

① $\mathcal{R} \subseteq \mathrm{conv}(\mathcal{R}_2)$: Fix any point $(D_f(P\|Q), D_g(P\|Q)) \in \mathcal{R}$.

Then $L = \frac{dP}{dQ}$ is a RV in $[0,\infty)$ with $\mathbb{E}_Q[L] = 1$, and
$$(D_f(P\|Q), D_g(P\|Q)) = (\mathbb{E}_Q[f(L)], \mathbb{E}_Q[g(L)]).$$

Next consider the set C of all prob. measures on $[0,\infty)$ with mean 1.

For $\mu \in C$, we associate a point $(\mathbb{E}_\mu f(L), \mathbb{E}_\mu g(L)) \in \mathbb{R}^2$.

Clearly C is convex, and

extremal points of C = {distributions with mean 1 and support size ≤ 2}.

(i.e. all points x that cannot be expressed as $x = \lambda y + (1-\lambda) z$ with $y, z \in C$, $\lambda \in (0,1)$)

In fact, if A_1, A_2, A_3 form a partition of $[0, \infty)$, and

$$\mu = \lambda_1 \mu_1 + \lambda_2 \mu_2 + \lambda_3 \mu_3 , \quad \lambda_i > 0, \quad \mathrm{supp}(\mu_i) \subseteq A_i .$$

Then the probability and mean constraints only require

$$\begin{cases} \lambda_1 + \lambda_2 + \lambda_3 = 1, \\ \lambda_1 m(\mu_1) + \lambda_2 m(\mu_2) + \lambda_3 m(\mu_3) = 1, \end{cases}$$

which is a line containing $(\lambda_1, \lambda_2, \lambda_3)$. So μ cannot be an extremal point.

Now by Choquet-Bishop-de Leeuw, any $\mu \in C$ can be written as a convex combination of extremal points of C, i.e. $R \subseteq \mathrm{conv}(R_2)$.

Thm (Choquet-Bishop-de Leeuw): if C is a metrizable convex compact subset of a locally convex topological vector space, then $C = \mathrm{conv}(\mathrm{extremal}(C))$.

② $\mathrm{conv}(R_2) \subseteq R_4$: by Carathéodory theorem below, any point of $\mathrm{conv}(R_2) \subseteq \mathbb{R}^2$ (which is connected) can be written as a convex combination of 2 points of R_2, which belongs to R_4.

Thm (Carathéodory): Let $S \subseteq \mathbb{R}^d$ and $x \in \mathrm{conv}(S)$. Then there exists $S' = \{x_1, \cdots, x_k\}$ s.t. $x \in \mathrm{conv}(S')$, with

① $k \le d+1$ in general;

② $k \le d$ if S has at most d connected components. ✍

Examples of inequalities:

① TV vs. H^2: $\quad \frac{H^2}{2} \le \mathrm{TV} \le \sqrt{H^2\left(1 - \frac{H^2}{4}\right)}$ (also the joint range)

② TV vs. KL: $\quad \mathrm{TV}^2 \le \frac{1}{2} \mathrm{KL}$

$\qquad \mathrm{TV} \le 1 - \frac{1}{2} \exp(-\mathrm{KL})$

③ KL vs. χ^2: $\quad \mathrm{KL} \le \log(1 + \chi^2)$ (also the joint range)

Special topic: chain rule for H^2

> **Thm (Jayram '09)** For all P_{X^n}, Q_{X^n}:
> $$H^2(P_{X^n}, Q_{X^n}) \le C \sum_{i=1}^{n} \mathbb{E}_P\left[H^2(P_{X_i|X^{i-1}}, Q_{X_i|X^{i-1}})\right],$$
> with $C = \prod_{i=1}^{\infty} \frac{1}{1-2^{-i}} \approx 3.46$.

The proof is surprisingly combinatorial. First, it suffices to prove the case $n = 2^k$: for general $2^{k-1} < n \le 2^k$, can consider $P_{2^k} = P_{X^n} \otimes P_0^{2^k-n}$, $Q_{2^k} = Q_{X^n} \otimes P_0^{2^k-n}$. The proof uses several properties of H^2.

Lemma 1 (L^2 geometry). For arbitrary distributions $P_0, \cdots, P_m$:
$$\frac{1}{m}\sum_{1\le i<j\le m} H^2(P_i, P_j) \le \sum_{i=1}^{m} H^2(P_i, P_0).$$

Pf. This result holds for all L^2 distance:
$$\frac{1}{m}\sum_{1\le i<j\le m} \| P_i - P_j\|^2 \le \sum_{i=1}^{m} \| P_i - P_0\|^2.$$

In fact,
$$\begin{aligned}
2\cdot \text{LHS} &= \frac{1}{m}\sum_{i,j=1}^{m} \| P_i - P_j\|^2 \\
&= \frac{1}{m}\sum_{i,j=1}^{m} \| P_i - P_0 - (P_j - P_0)\|^2 \\
&= \frac{1}{m}\sum_{i,j=1}^{m} \left(\| P_i - P_0\|^2 + \| P_j - P_0\|^2 - 2\langle P_i - P_0, P_j - P_0\rangle\right) \\
&= 2\cdot \text{RHS} - \frac{2}{m}\left\| \sum_{i=1}^{m} (P_i - P_0)\right\|^2 \le 2\cdot \text{RHS}.
\end{aligned}$$

Finally, note that
$$H^2(P, Q) = \int (\sqrt{dP} - \sqrt{dQ})^2$$
is indeed an L^2 distance. $\square$

Now for $A \subseteq [n]$, define interpolations

$$p^A = \prod_{i=1}^{n} \left(P_{X_i|X^{i-1}}\right)^{\mathbb{1}(i \notin A)} \left(Q_{X_i|X^{i-1}}\right)^{\mathbb{1}(i \in A)}.$$

Then $p^{\emptyset} = P_{X^n}$, $p^{[n]} = Q_{X^n}$.

Lemma 2 (cut-paste property) Let $a, b, c, d \in \{0,1\}^n$ be the indicators of sets $A, B, C, D \subseteq [n]$. If $a+b = c+d$, then $H^2(p^A, p^B) = H^2(p^C, p^D)$.

Pf.

$$H^2(p^A, p^B) = 2 - 2\int \sqrt{p^A p^B}$$

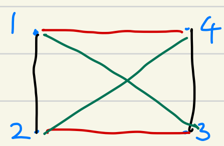
$$= 2 - 2 \int \sqrt{\prod_{i=1}^{n} P_{X_i|X^{i-1}}^{1-a_i+1-b_i} Q_{X_i|X^{i-1}}^{a_i+b_i}}$$

$$= 2 - 2 \int \sqrt{\prod_{i=1}^{n} P_{X_i|X^{i-1}}^{1-c_i+1-d_i} Q_{X_i|X^{i-1}}^{c_i+d_i}} = H^2(p^C, p^D) \quad \square$$

Lemma 3 (1-factorization of cliques) For even n, the complete graph K_n can be decomposed into $(n-1)$ edge-disjoint perfect matchings.

(i.e. round-robin tournaments)

Example of $n=4$.



A geometric construction:

Put node 1 in the center of a regular polygon with $(n-1)$ vertices. Use color i for $(1, i)$ and all edges perpendicular to $(1, i)$.

<u>Completing the proof</u>. For $n = 2^k$, prove by induction on $m = 0, 1, \cdots, k$ that for any partition $A_1, \cdots, A_{2^m}$ of $[n]$ (each of size 2^{k-m}):

$$\sum_{i=1}^{2^m} H^2(P^{A_i}, P^\phi) \geq c_m \cdot H^2(P^{[n]}, P^\phi),$$

with $c_m = \prod_{i=1}^{m} (1 - 2^{-i})$.

<u>Base $m = 0$</u>: trivial.

<u>Induction from $m-1$ to m</u>:

$$\sum_{i=1}^{2^m} H^2(P^{A_i}, P^\phi) \overset{\text{Lemma 1}}{\geq} \frac{1}{2^m} \sum_{1 \leq s < t \leq 2^m} H^2(P^{A_s}, P^{A_t})$$

$$\overset{\text{Lemma 2}}{=} \frac{1}{2^m} \sum_{1 \leq s < t \leq 2^m} H^2(P^{A_s \cup A_t}, P^\phi)$$

$$\overset{\text{Lemma 3}}{=} \frac{1}{2^m} \sum_{a=1}^{2^m - 1} \sum_{(s,t) \in E_a} H^2(P^{A_s \cup A_t}, P^\phi),$$

where each E_a is a perfect matching of K_{2^m}. By induction hypothesis,

$$\sum_{i=1}^{2^m} H^2(P^{A_i}, P^\phi) \geq \frac{2^m - 1}{2^m} c_{m-1} H^2(P^{[n]}, P^\phi) = c_m H^2(P^{[n]}, P^\phi).$$

<u>Conclusion</u>: choosing $m = k$ yields

$$H^2(P^{[n]}, P^\phi) \leq \frac{1}{c_k} \sum_{i=1}^{n} H^2(P^{\{i\}}, P^\phi)$$

$$= \frac{1}{c_k} \sum_{i=1}^{n} \mathbb{E}_P [H^2(P_{X_i | X^{i-1}}, Q_{X_i | X^{i-1}})].$$