Section 1

Definition and Basic Facts
For $p = (p_1, \ldots, p_m)$ and $q = (q_1, \ldots, q_m)$, let

$$D(p\|q) = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}$$

$0 \log \frac{0}{0} = 0$, $p \log \frac{p}{0} = \infty$

The relative entropy of pair of rv’s, is the relative entropy of their distributions.

Names: Entropy of $p$ relative to $q$, relative entropy, information divergence, Kullback-Leibler (KL) divergence/distance

Many different interpretations

Main interpretation: the information we gained about $X$, if we originally thought $X \sim q$ and now we learned $X \sim p$
Numerical Example

\[ D(p \parallel q) = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} \]

- \( p = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, 0), \quad q = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}) \)

- \( D(p \parallel q) = \frac{1}{4} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{4} + \frac{1}{4} \log \frac{1}{8} + 0 \log 0 = \frac{1}{4} \cdot (-1) + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 1 = \frac{1}{2} \)

- \( D(q \parallel p) = \frac{1}{2} \log \frac{1}{4} + \frac{1}{4} \log \frac{1}{2} + \frac{1}{8} \log \frac{1}{4} + \frac{1}{8} \log \frac{1}{0} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot (-1) + \frac{1}{8} \cdot (-1) + \infty = \infty \)
Supporting the interpretation

- $X$ rv over $[m]$
- $H(X)$ — measure for amount of information we do not have about $X$
- $\log m - H(X)$ — measure for information we do have about $X$
  (just by knowing its distribution)
- Example $X = (X_1, X_2) \sim (\frac{1}{2}, 0, 0, \frac{1}{2})$ over $\{00, 01, 10, 11\}$
- $H(X) = 1$, $\log m - H(X) = 2 - 1 = 1$
- Indeed, we know $X_1 \oplus X_2$

\[
H(\sim [m]) - H(p_1, \ldots, p_m) = \log m - H(p_1, \ldots, p_m)
\]
\[
= \log m + \sum_i p_i \log p_i = \sum_i p_i (\log p_i - \log \frac{1}{m})
\]
\[
= \sum_i p_i \log \frac{p_i}{\frac{1}{m}} = D(p\|\sim [m])
\]
- $D(X\|\sim [m])$ — measures the information we gained about $X$, if we originally thought it is $\sim [m]$ and now we learned it is $\sim p$
Supporting the interpretation, cont.

- (generally) $D(p\|q) \neq H(q) - H(p)$
- $H(q) - H(p)$ is not a good measure for information change
- Example: $q = (0.01, 0.99)$ and $p = (0.99, 0.01)$
- We were almost sure that $X = 1$ but learned that $X$ is almost surely $0$
- But $H(q) - H(p) = 0$
- Also, $H(q) - H(p)$ might be negative

- We understand $D(p\|q)$ as the information we gained about $X$, if we originally thought it is $\sim q$ and now we learned it is $\sim p$
Changing distribution

- What does it mean: originally thought $X \sim q$ and now we learned $X \sim p$?

- How can a distribution change?

- Typically, this happens by learning additional information

- $q_i = \Pr[X = i]$ and $p_i = \Pr[X = i|E]$

- Example $X \sim (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)$; someone saw $X$ and tells us that $X \leq 2$

- The distribution changes to $X \sim (\frac{2}{3}, \frac{1}{3}, 0, 0)$

- Another example

<table>
<thead>
<tr>
<th>$X \backslash Y$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
</tbody>
</table>

- $Y \sim (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, 0)$, but

- $Y \sim (\frac{1}{2}, \frac{1}{2}, 0, 0)$ conditioned on $X = 0$

- $Y \sim (\frac{1}{2}, 0, \frac{1}{2}, 0)$ conditioned on $X = 1$

- Generally, a distribution can change if we condition on event $E$
Additional properties

- $0 \log \frac{0}{0} = 0$, $p \log \frac{p}{0} = \infty$ for $p > 0$
- $\exists i$ s.t. $p_i > 0$ and $q_i = 0$, then $D(p\|q) = \infty$
- If originally $\Pr[X = i] = 0$, then it cannot be more than 0 after we learned something.
- Hence, it make sense to think of it as infinite amount of information learnt
- Alternatively, we can define $D(p\|q)$ only for distribution with $q_i = 0 \implies p_i = 0$
  (recall that $\Pr[X = i] = 0 \implies \Pr[X = i|E] = 0$, for any event $E$
- If $p_i$ is large and $q_i$ is small, then $D(p\|q)$ is large
- $D(p\|q) \geq 0$, with equality iff $p = q$ (hw)
Example

- \( q = (q_1, \ldots, q_m) \) with \( \sum_{i=1}^{n} q_i = 2^{-k} \) (i.e., \( n < m \))
- \( p_i = \begin{cases} q_i/2^{-k}, & 1 \leq i \leq n \\ 0, & \text{otherwise.} \end{cases} \)
- \( p = (p_1, \ldots, p_m) \) — the distribution of \( q \) conditioned on the event \( i \in [n] \)
- \( D(p\|q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{n} p_i \log 2^k = \sum_{i=1}^{n} p_i k = k \)
- We gained \( k \) bits of information
- Example: \( \sum_{i=1}^{n} q_i = \frac{1}{2} \), and we were told that \( i \leq n \) or \( i > n \), we got one bit of information
Section 2

Axiomatic Derivation
Axiomatic derivation

Let \( \tilde{D} \) is a continuous and symmetric (wrt each distribution) function such that

1. \( \tilde{D}(p\parallel m) = \log m - H(p) \)

2. \( \tilde{D}((p_1, \ldots, p_m)\parallel(q_1, \ldots, q_m)) = \tilde{D}((p_1, \ldots, p_{m-1}, \alpha p_m, (1 - \alpha)p_m)\parallel(q_1, \ldots, q_{m-1}, \alpha q_m, (1 - \alpha)q_m)) \), for any \( \alpha \in [0, 1] \)

then \( \tilde{D} = D \).

Interpretation

Proof: Let \( p \) and \( q \) be distributions over \([m]\), and assume \( q_i \in \mathbb{Q} \setminus \{0\} \).

\[ \begin{align*}
\tilde{D}(p\parallel q) &= \tilde{D}((\alpha_{1,1}p_1, \ldots, \alpha_{1,k_1}p_1, \ldots, \alpha_{m,1}p_m, \ldots, \alpha_{m,k_m}p_m)\parallel\left(\alpha_{1,1}q_1, \ldots, \alpha_{1,k_1}q_1, \ldots, \alpha_{m,1}q_m, \ldots, \alpha_{m,k_m}q_m\right)), \text{ for } \sum_j \alpha_{i,j} = 1 \text{ and } \alpha_{i,j} \geq 0 \\
\text{Taking } \alpha_i \text{'s s.t. } \alpha_{i,1} = \alpha_{i,2} \ldots, \alpha_{i,k_i} = \alpha_i \text{ and } \alpha_i q_i = \frac{1}{M}, \text{ it follows that } \\
\tilde{D}(p\parallel q) &= \log M - H((\alpha_{1,1}p_1, \ldots, \alpha_{1,k_1}p_1, \ldots, \alpha_{m,1}p_m, \ldots, \alpha_{m,k_m}p_m)) \\
&= \sum_i p_i \log M + \sum_i p_i \log \alpha_i p_i = \sum_i p_i (\log M + \log \frac{p_i}{q_i M}) = \sum_i p_i \log \frac{p_i}{q_i}.
\end{align*} \]

\[ \begin{align*}
\text{Zeros and non-rational } q_i \text{'s are dealt by continuity}
\end{align*} \]
Section 3

Relation to Mutual Information
Mutual information as expected relative entropy

Claim 1

\[ E_{Y \leftarrow Y}[D(X|Y=y||X)] = I(X;Y). \]

Proof:

- Let \( X \sim (q_1, \ldots, q_m) \) over \([m]\), and \( Y \) be rv over \( \{0,1\} \) (to keep it simple)
- \( (X|Y=j) \sim p_j = (p_{j,1}, \ldots, p_{j,m}) \), \( \quad p_{j,i} = \Pr[X = i|Y = j] \)

\[
E_Y[D(p_Y||q)] = \Pr[Y = 0] \cdot D(p_{0,1}, \ldots, p_{0,m}||q_1, \ldots, q_m) \\
+ \Pr[Y = 1] \cdot D(p_{1,1}, \ldots, p_{1,m}||q_1, \ldots, q_m) \\
= \Pr[Y = 0] \cdot \sum_i p_{0,i} \log \frac{p_{0,i}}{q_i} + \Pr[Y = 1] \cdot \sum_i p_{1,i} \log \frac{p_{1,i}}{q_i} \\
= \Pr[Y = 0] \cdot \sum_i p_{0,i} \log p_{0,i} + \Pr[Y = 1] \cdot \sum_i p_{1,i} \log p_{1,i} \\
- \Pr[Y = 0] \cdot \sum_i p_{0,i} \log q_i - \Pr[Y = 1] \cdot \sum_i p_{1,i} \log q_i \\
= -H(X|Y) - \sum_i (\Pr[Y = 0] \cdot p_{0,i} + \Pr[Y = 1] \cdot p_{1,i}) \log q_i \\
= -H(X|Y) + H(X) = I(X;Y). \]
Equivalent definition for mutual information

Claim 2

Let $(X, Y) \sim p$, then $I(X; Y) = D(p\|p_Xp_Y) = D((X, Y)\| (X \otimes Y))$.

Proof:

\[
D(p\|p_Xp_Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p_X(x)p_Y(y)}
\]

\[
= \sum_{x,y} p(x, y) \log \frac{p_X|Y(x|y)}{p_X(x)}
\]

\[
= -\sum_{x,y} p(x, y) \log p_X(x) + \sum_{x,y} p(x, y) \log p_X|Y(x|y)
\]

\[
= H(X) + \sum_y p_Y(y) \sum_x p_X|Y(x|y) \log p_X|Y(x|y)
\]

\[
= H(X) - H(X|Y) = I(X; Y).
\]

We will later relate the above two claims.
Section 4

Relation to Data Compression
Wrong code

**Theorem 3**

Let \( p \) and \( q \) be distributions over \([m]\), and let \( C \) be code with

\[
\ell(i) = |C(i)| = \left\lceil \log \frac{1}{q_i} \right\rceil.
\]

Then

\[
H(p) + D(p\|q) \leq E_{i \leftarrow p} [\ell(i)] \leq H(p) + D(p\|q) + 1
\]

- Recall that \( H(q) \leq E_{i \leftarrow q} [\ell(i)] \leq H(q) + 1 \).
- Proof of upperbound (lowerbound is proved similarly)

\[
E_{i \leftarrow p} [\ell(i)] = \sum_i p_i \left\lceil \log \frac{1}{q_i} \right\rceil < \sum_i p_i (\log \frac{1}{q_i} + 1)
\]

\[
= 1 + \sum_i p_i (\log \frac{p_i}{q_i} + \frac{1}{p_i}) = 1 + \sum_i p_i (\log \frac{p_i}{q_i}) + \sum_i p_i (\log \frac{1}{p_i})
\]

\[
= 1 + D(p\|q) + H(p)
\]

- Can there be a (close) to optimal code for \( q \) that is better for \( p \)? HW
Section 5

Conditional Relative Entropy
Conditional relative entropy

For dist. $p$ over $\mathcal{X} \times \mathcal{Y}$, let $p_X$ and $p_{Y|X}$ be its marginal and conditional dist.

**Definition 4**

For two distributions $p$ and $q$ over $\mathcal{X} \times \mathcal{Y}$:

$$D(p_{Y|X} \parallel q_{Y|X}) := \sum_{x \in \mathcal{X}} p_X(x) \cdot \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)}$$

- $D(p_{Y|X} \parallel q_{Y|X}) = E(x,y) \sim p(x,y) \left[ \log \frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)} \right]$.
- Let $(X_p, Y_p) \sim p$ and $(X_q, Y_q) \sim q$, then
  $$D(p_{Y|X} \parallel q_{Y|X}) = E_{x \leftarrow X_p} \left[ D(Y_p|x_p=x \parallel Y_q|x_q=x) \right]$$

- Numerical example:
  $$p = \begin{bmatrix} \mathcal{X} & \mathcal{Y} \\ 0 & 1 & 0.1 & 0.9 \\ 1 & 0.8 & 0.2 \end{bmatrix} \quad q = \begin{bmatrix} \mathcal{X} & \mathcal{Y} \\ 0 & 1 & 0.8 & 0.2 \\ 1 & 0.5 & 0.5 \end{bmatrix}$$
  $$D(p_{Y|X} \parallel q_{Y|X}) = \frac{1}{4} \cdot D((1/2, 1/2) \parallel (1/3, 2/3)) + \frac{3}{4} \cdot D((1/3, 2/3) \parallel (4/5, 1/5))$$
  $$\quad = \ldots$$
Chain rule

Claim 5

For any two distributions $p$ and $q$ over $\mathcal{X} \times \mathcal{Y}$, it holds that

$$D(p\|q) = D(p_{\mathcal{X}}\|q_{\mathcal{X}}) + D(p_{\mathcal{Y}|\mathcal{X}}\|q_{\mathcal{Y}|\mathcal{X}})$$

Proof:

$$D(p\|q) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{q(x,y)}$$

$$= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p_{\mathcal{X}}(x)p_{\mathcal{Y}|\mathcal{X}}(y|x)}{q_{\mathcal{X}}(x)q_{\mathcal{Y}|\mathcal{X}}(y|x)}$$

$$= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p_{\mathcal{X}}(x)}{q_{\mathcal{X}}(x)} + \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p_{\mathcal{Y}|\mathcal{X}}(y|x)}{q_{\mathcal{Y}|\mathcal{X}}(y|x)}$$

$$= D(p_{\mathcal{X}}\|q_{\mathcal{X}}) + D(p_{\mathcal{Y}|\mathcal{X}}\|q_{\mathcal{Y}|\mathcal{X}}) \square$$

Hence, for $(X, Y) \sim p$:

$$I(X, Y) = D(p\|p_X p_Y) = D(p_X\|p_X) + \mathbb{E}_{x \leftarrow X} \left[ D(p_{Y|X=x}\|p_Y) \right]$$

$$= \mathbb{E}_{x \leftarrow X} \left[ D(p_{Y|X=x}, p_Y) \right].$$
Section 6

Data-processing inequality
Data-processing inequality

Claim 6
For any rv’s $X$ and $Y$ and function $f$, it holds that $D(f(X)\|f(Y)) \leq D(X\|Y)$.

- Analogues to $H(X) \geq H(f(X))$

Proof:
- $D(X, f(X)\| Y, f(Y)) = D(X\| Y)$
- $D(X, f(X)\| Y, f(Y)) = D(f(X)\|f(Y)) + \mathbb{E}_{z \leftarrow f(X)} \left[ D(X\|f(X)=z \| Y\|f(Y)=z) \right] \geq D(f(X)\|f(Y))$

Hence, $D(f(X)\|f(Y)) \leq D(X\|Y)$. 
Section 7

Relation to Statistical Distance
Relation to statistical distance

- \( D(p\|q) \) is used many times to measure the distance from \( p \) to \( q \)
- It is not a distance in the mathematical sense: \( D(p\|q) \neq D(q\|p) \) and no triangle inequality
- However,

Theorem 7 (Pinsker inequality)

\[
SD(p, q) \leq \sqrt{\frac{\ln 2}{2} \cdot D(p\|q)}
\]

- Corollary: For rv \( X \) over \( [m] \) with \( H(X) \geq \log m - \varepsilon \), it holds that
  \[
  SD(X, \sim [m]) \leq \sqrt{\frac{\ln 2}{2} \cdot (\log m - H(X))} = \sqrt{\frac{\ln 2}{2} \cdot \varepsilon}
  \]
- Other direction is incorrect: \( SD(p, q) \) might be small but \( D(p\|q) = \infty \)
- Does \( SD(p, \sim [m]) \) being small imply \( D(p\|\sim [m]) = \log m - H(p) \) is small?

HW
Proving Thm 7, Boolean case

- Let \( p = (\alpha, 1 - \alpha) \) and \( q = (\beta, 1 - \beta) \) and assume \( \alpha \geq \beta \)
- \( \text{SD}(p, q) = \alpha - \beta \)
- We will show that
  \[
  D(p\|q) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1-\alpha}{1-\beta} \geq \frac{4}{2 \ln 2} (\alpha - \beta)^2 = \frac{2}{\ln 2} \text{SD}(p, q)^2
  \]
  
- Let \( g(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y} - \frac{4}{2 \ln 2} (x - y)^2 \)
  
  \[
  \frac{\partial g(x, y)}{\partial y} = -\frac{x}{y \ln 2} + \frac{1-x}{(1-y) \ln 2} - \frac{4}{2 \ln 2} 2(y - x)
  \]
  
  \[
  = \frac{y-x}{y(1-y) \ln 2} - \frac{4}{\ln 2} (y - x)
  \]
  
  \[
  = \frac{(y-x)}{\ln 2} (\frac{1}{y(1-y)} - 4)
  \]

- Since \( y(1-y) \leq \frac{1}{4} \), \( \frac{\partial g(x,y)}{\partial y} \leq 0 \) for \( y < x \).

- Since \( g(x, x) = 0 \), \( g(x, y) \geq 0 \) for \( y < x \). □
Proving Thm 7, general case

- Let $\mathcal{U} = \text{Supp}(p) \cup \text{Supp}(q)$
- Let $S = \{u \in \mathcal{U} : p(u) > q(u)\}$
- $\text{SD}(p, q) = \Pr_p [S] - \Pr_q [S]$ (by homework)
- Let $P \sim p$, and let the indicator $\hat{P}$ be 1 iff $P \in S$.
- Let $Q \sim q$, and let the indicator $\hat{Q}$ be 1 iff $Q \in S$.
- $\text{SD}(\hat{P}, \hat{Q}) = \Pr [P \in S] - \Pr [Q \in S] = \text{SD}(p, q)$

\[
D(p \parallel q) \geq D(\hat{P} \parallel \hat{Q}) \quad \text{(data-processing inequality)}
\]
\[
\geq \frac{2}{\ln 2} \cdot \text{SD}(\hat{P}, \hat{Q})^2 \quad \text{(the Boolean case)}
\]
\[
= \frac{2}{\ln 2} \cdot \text{SD}(p, q)^2. \square
\]
Extreme events

- Assume $SD(P, Q) \leq \delta$, then $\forall E: \Pr_Q[E] \geq \Pr_P[E] - \delta$.
- Assume $D(P\|Q) \leq \delta$, then $\Pr_Q[E] \geq \Pr_P[E] - \Theta(\sqrt{\delta})$.

  but if $\Pr_P[E] = 1$, then $\Pr_Q[E] \geq 1 - \Theta(\delta)$.

Proof: let $\Pr_Q[E] = 1 - \varepsilon$, then

$$\delta \geq D(1\|\Pr_Q[E]) = D(1\|1 - \varepsilon) = \log(1/1 - \varepsilon) = \Theta(\log e^\varepsilon) = \Theta(\varepsilon).$$

- Tight bounds
- KL divergence is a better mean for bounding extreme events
Section 8

Conditioned Distributions
Main theorem

**Theorem 8**

Let $X_1, \ldots, X_k$ be iid over $\mathcal{U}$, and let $Y = (Y_1, \ldots, Y_k)$ be rv over $\mathcal{U}^k$. Then

$$D(Y\|X_1, \ldots, X_k) = \sum_{j=1}^k D(Y_j\|X_j) + D(Y\|\bigotimes_{i=1}^k Y_i) \geq \sum_{j=1}^k D(Y_j\|X_j).$$

For rv $Z$, let $Z(z) = \Pr[Z = z]$.

We prove for $k = 2$, general case follows similar lines. Let $X = (X_1, X_2)$

$$D(Y\|X) = \sum_{y \in \mathcal{U}^2} Y(y) \log \frac{Y(y)}{X(y)} = \sum_{y = (y_1, y_2)} Y(y) \log \frac{Y_1(y_1) Y_2(y_2)}{X_1(y_1) X_2(y_2)} Y(y) Y_1(y_1) Y_2(y_2)$$

$$= \sum_{y = (y_1, y_2)} Y(y) \log \frac{Y_1(y_1)}{X_1(y_1)} + \sum_{y = (y_1, y_2)} Y(y) \log \frac{Y_2(y_2)}{X_2(y_2)}$$

$$+ \sum_{y = (y_1, y_2)} Y(y) \log \frac{Y(y)}{Y_1(y_1) Y_2(y_2)}$$

$$= D(Y_1\|X_1) + D(Y_2\|X_2) + D(Y\|Y_1 \otimes Y_2).$$
Conditioning distributions, relative entropy case

**Theorem 9**

Let $X_1, \ldots, X_k$ be iid over $\mathcal{X}$, let $X = (X_1, \ldots, X_k)$ and let $W$ be an event (i.e., Boolean rv). Then

$$\sum_{j=1}^{k} D((X_j|W)\|X_j) \leq D((X|W)\|X) \leq \log \frac{1}{\Pr[W]}.$$
Theorem 10

Let $X_1, \ldots, X_k$ be iid over $\mathcal{X}$ and let $W$ be an event. Then

$$\sum_{j=1}^{k} \text{SD}((X_j|W), X_j)^2 \leq \log \frac{1}{\Pr[W]}.$$ 

Proof: follows by Thm 7, and Thm 8. □

Using $(\sum_{j=1}^{k} a_j)^2 \leq k \cdot \sum_{j=1}^{k} a_j^2$, it follows that

Corollary 11

$$\sum_{j=1}^{k} \text{SD}((X_j|W), X_j) \leq \sqrt{k \log \left( \frac{1}{\Pr[W]} \right)}, \text{ and}$$

$$E_{j \leftarrow k} \text{SD}((X_j|W), X_j) \leq \sqrt{\frac{1}{k} \log \left( \frac{1}{\Pr[W]} \right)}$$

Extraction
Numerical example

- Let $X = (X_1, \ldots, X_{40}) \leftarrow \{0, 1\}^{40}$ and let $f: \{0, 1\}^{40} \rightarrow \{0, 1\}$ be such that $\Pr[f(X) = 0] = 2^{-10}$.

- $E_{j \leftarrow [40]} \text{SD}((X_j|f(X)=0), \sim \{0, 1\}) \leq \sqrt{\frac{1}{40} \cdot 10} = \frac{1}{2}$

- Typical bits are not too biased, even when conditioning on a very unlikely event.
Section 9

Appendix
Extension

Theorem 12

Let $X = (X_1, \ldots, X_k)$, $T$ and $V$ be rv’s over $\mathcal{X}^k$, $T$ and $V$ respectively. Let $W$ be an event and assume that the $X_i$’s are iid conditioned on $T$. Then

$$\sum_{j=1}^k D((TVX_j)\| (TV)|_W X_j'(T)) \leq \log \frac{1}{\Pr[W]} + \log |\text{Supp}(V|_W)|,$$

where $X_j'(t)$ is distributed according to $X_j|_{T=t}$.

Interpretation.
Proving Thm 12

Let \( X = (X_1, \ldots, X_k) \), \( T \) and \( V \) be rv’s over \( \mathcal{X}^k \), \( T \) and \( V \) respectively, such that \( X_i \)'s are iid conditioned on \( T \). Let \( W \) be an event and let \( X_j'(t) \) be distributed according to the distribution of \( X_j|T=t \).

\[
\sum_{j=1}^{k} D((TVX_j)|W \parallel (TV)|W X_j'(T))
\]
\[
= \mathbb{E}_{(t,v) \leftarrow (TV)|W} \left[ \sum_{j=1}^{k} D(X_j|W,V=v,T=t \parallel (X_j|T=t)) \right]
\]
\[
= \mathbb{E}_{(t,v) \leftarrow (TV)|W} \left[ \sum_{j=1}^{k} D((X_j|W,V=v)|T=t \parallel (X_j|T = t)) \right]
\]
\[
\leq \mathbb{E}_{(t,v) \leftarrow (TV)|W} \left[ \log \frac{1}{\Pr[W \land V = v | T = t]} \right]
\]
\[
\leq \log \mathbb{E}_{(t,v) \leftarrow (TV)|W} \left[ \frac{1}{\Pr[W \land V = v | T = t]} \right]
\]
\[
= \log \sum_{(t,v) \in \text{Supp}((TV)|W)} \frac{\Pr[T = t]}{\Pr[W]} \leq \log \frac{||\text{Supp}(V|W)||}{\Pr[W]}.
\]

\[\square\]