A Minmax Redundancy and Probability Assignment

$\underline{x} \in \mathcal{X}^n \rightsquigarrow (X_1, \dots, X_n) = (x_1, \dots, x_n)$

$X_i \sim P$, $(X_1, \dots, X_n)$ are iid with common distribution P. } probabilistic modelling

→ Universal compression algorithms assume the generating dist. is unknown

→ Benchmark of performance: "Redundancy"

$\mathcal{C}$ be a prefix-free code that assigns a codeword of length $\ell(\underline{x}$

$$\boxed{r_n(\mathcal{C}, P^n) := \sum_{\underline{x} \in \mathcal{X}^n} P^n(\underline{x})\,\ell(\underline{x}) - H(P^n)} \rightsquigarrow$$

"redundancy of $\mathcal{C}$ for P^n"

Worst-case redundancy:

$\mathcal{P}(\mathcal{X})$

$$r_n(\mathcal{C}) := \max_{P \in \mathcal{P}(\mathcal{X})} r_n(\mathcal{C}, P^n)$$

→ This is our measure of for the prefix-free coc

Minmax redundancy

$$r_n^* := \min_{\mathcal{C}} r_n(\mathcal{C})$$

$$= \min_{\mathcal{C}} \max_{P \in \mathcal{P}(\mathcal{X})} \Big[\sum_{\underline{x}} P^n(\underline{x})\,\ell(\underline{x}$$

How large is r_n^*? And which codes $\mathcal{C}$ attain it?

Probability assignment

We want distribution $Q^{(n)}$ on $\mathcal{X}^n$ s.t. $\rightsquigarrow Q^{(n)}(\underline{x})$

$$R_n^* = \min_{Q^{(n)} \in \mathcal{P}(\mathcal{X}^n)} \max_{P \in \mathcal{P}(\mathcal{X})} D(P^n \| Q^{(n)})$$

(information rad

Lemma. $R_n^* \approx r_n^*$

P_n, $Q^{(n)}$.

$$\boxed{R_n^* \le r_n^* \le R_n^* + 1}$$

Proof. (Probability assignment ⟹ prefix-free code)

Given a prob. $Q^{(n)}$ on $\mathcal{X}^n$, let
$$\ell(\underline{x}) = \left\lceil \log \frac{1}{Q^{(n)}(\underline{x})} \right\rceil.$$

(Check: $\ell(\underline{x})$ satisfies Kraft's inequality) ✓

Therefore, for the prefix-free code associated with $(\ell(\underline{x}), \underline{x}$

we have
$$\mathbb{E}_P[\ell(\underline{X})] = \sum_{\underline{x}} P^n(\underline{x})\, \ell(\underline{x})$$
$$\le \sum_{\underline{x}} P^n(\underline{x}) \log \frac{1}{Q^{(n)}(\underline{x})} \cdot \frac{P^n(}{P^n}$$
$$= D(P^n \| Q^{(n)}) + nH(P) + 1$$
$$\Rightarrow \mathbb{E}_P[\ell(\underline{X})] - nH(P) \le D(P^n \| Q^{(n)}) + 1$$
$$\Rightarrow \max_{P \in \mathcal{P}(\mathcal{X})} \mathbb{E}_P[\ell(\underline{X})] - nH(P) \le \max_{P \in \mathcal{P}(\mathcal{X})} D(P^n \|$$
$$\Rightarrow \boxed{r_n^* \le R_n^* + 1}$$

(prefix-free codes ⟹ probability assignment)

Given a prefix-free code with codeword lengths $(\ell(\underline{x}), \underline{x} \in \mathcal{X}$

consider
$$Q^{(n)}(\underline{x}) = \frac{2^{-\ell(\underline{x})}}{\sum_{\underline{x}'} 2^{-\ell(\underline{x}')}}$$

Then,
$$D(P^n \| Q^{(n)}) = \sum_{\underline{x}} P^n(\underline{x}) \log \frac{1}{Q^{(n)}(\underline{x})} - nH(P$$
$$= \sum_{\underline{x}} P^n(\underline{x}) \log 2^{\ell(\underline{x})} + \log \sum_{\underline{x}'} 2$$

$- n H(P)$

$$\leq \mathbb{E}_{P^n}[\ell(\underline{X})] - n H(P),$$

for every $P \in \mathcal{P}(\mathcal{X})$. Therefore,

$$\min_{Q^{(n)} \in \mathcal{P}(\mathcal{X}^n)} \max_{P \in \mathcal{P}(\mathcal{X})} D(P^n \| Q^{(n)}) \leq \max_{P \in \mathcal{P}(\mathcal{X})} \mathbb{E}_{P^n}[$$

for every prefix-free code.

$$\Rightarrow \boxed{R_n^* \leq r_n^*}$$ ■

Universal prefix-free code design $\approx$ minmax probability as

[B] Compression using word frequencies

$\mathcal{X} = \{0,1\}$

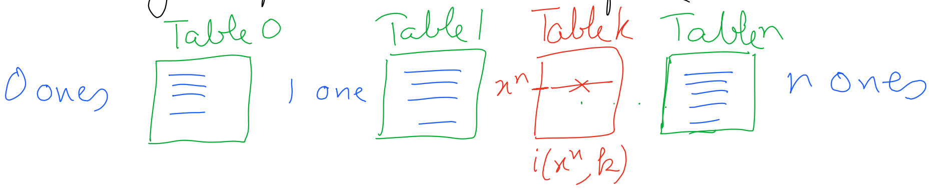
Our scheme: Input: $x^n \in \{0,1\}^n$

Output: A binary codeword $c(\underline{x}) \in \{0,1\}^*$

1) Count the # of 1s in x^n. Denote it by k.

2) Let $i(x^n, k)$ be the index of the sequence x^n among all sequen k 1s.

(3) $c(\underline{x}) \equiv$ binary representation of $(k, i(x^n, k))$



What is the number bits used to represent a sequence x^n ?

→ k takes $\log(n+1)$ bits to represent

→ $i(x^n, k)$ can take $\binom{n}{k}$ values, and so, needs $\log \binom{n}{k}$ bi

Therefore,

$$\ell(\underline{x}) = \log(n+1) + \log \binom{n}{k(\underline{x})},$$

where $k(\underline{x}) = $ # 1s in x.

Important fact: $\binom{n}{k} \approx 2^{n h(k/n)}$,

where $h(t) = t \log \frac{1}{t} + (1-t) \log \frac{1}{(1-t)}$.

More formally,

$$\binom{n}{k} \le 2^{n h\left(\frac{k}{n}\right)}$$

$$\Rightarrow \quad \ell(x) \le \log(n+1) + nh\left(\frac{k(x)}{n}\right)$$

Thus,

$$\mathbb{E}_{P^n}[\ell(X^n)] \le \log(n+1) + n\mathbb{E}_{P^n}\left[h\left(\frac{k(X}{n}\right.\right.$$

(by Jensen's ineq., since h is a concave function)

$$\le \log(n+1) + nh\left(\frac{\mathbb{E}_{P^n}[k(X^n}{n}\right.$$

$$\mathbb{E}_{P^n}[k(X^n)] = \mathbb{E}_{P^n}\left[\sum_{t=1}^{n} X_t\right] = np$$

$$\Rightarrow \quad \mathbb{E}_{P^n}[\ell(X^n)] \le \log(n+1) + n\underbrace{h(\overset{P(1)}{p})}_{= H(P)}$$

$$\Rightarrow \quad \mathbb{E}_{P^n}[\ell(X^n)] - nH(P) \le \log(n+1)$$

for every $P \in \mathcal{P}(\{0,1\})$

$$\Rightarrow \quad \max_{P \in \mathcal{P}(\mathcal{X}^n)} \mathbb{E}_{P^n}[\ell(X^n)] - nH(P) \le \log(n$$

Remarks 1) $\log(n+1)$ extra cost for "universality" is negligible in comparison with the optimal avg. length $nH(P)$.

(2) Recall that we can find prob. assignment using this schem

$$R_n^* \le r_n^* \le \log(n+1)$$

iid $Q^{(n)}$ can only give $O(n)$ bounds

(3) The analysis above can be improved to get $\frac{1}{2}\log(n+1)$.

Extension to an arbitrary alphabet $\mathcal{X}$: Types (Method of Types, Csiszár-Körner b

Definition (Type of a sequence) The type of a sequence $\underline{x}$ is a pmf denoted $P_{\underline{x}}$, given by

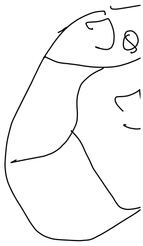
$$P_{\underline{x}}(a) = \frac{\overbrace{N(a|\underline{x})}^{\text{\# of times a appea } (x_1,\dots,x_n)}}{n}, \quad a \in \mathcal{X}.$$

The set of all sequences of a given type Q is called the type type class Q, denoted $\mathcal{T}_Q^{(n)}$.

The set of all types is denoted by $T^{(n)}$.

Fact: All sequences $\underline{x} \in \mathcal{T}_Q$ have equal probabilities under any iid

Proof.
$$\begin{aligned}
P^n(\underline{x}) &= \prod_{t=1}^{n} P(x_t) \\
&= \prod_{a\in\mathcal{X}} P(a)^{N(a|\underline{x})} \\
&= 2^{\sum_{a\in\mathcal{X}} \underbrace{N(a|\underline{x})}_{\to n Q(a)} \log P(a)} \\
&= 2^{-\sum_{a\in\mathcal{X}} N(a|\underline{x}) \log \frac{1}{P(a)} \cdot \frac{Q(a)}{Q(a)}} \\
&= 2^{-n\left(\sum_{a\in\mathcal{X}} Q(a) \log \frac{Q(a)}{P(a)} + H(Q)\right)} \\
&= 2^{-n\left(D(Q\|P) + H(Q)\right)}
\end{aligned}$$
■



Lemma (Type counting lemma) For a finite alphabet $\mathcal{X}$,

$$\boxed{|T^{(n)}| \le (n+1)^{|\mathcal{X}|-1}}.$$

Lemma (Type class cardinality lemma) For every $Q \in T^{(n)}$,

$$\boxed{\frac{2^{nH(Q)}}{(n+1)^{|\mathcal{X}|-1}} \le |\mathcal{T}_Q| \le 2^{nH(Q)}}$$

$$\left(\frac{1}{n}\log|\mathcal{T}_Q| \simeq H(Q)\right)$$

<u>Proof</u>

* $1 \geq Q^n(\mathcal{T}_Q) = \sum_{\underline{x}\in\mathcal{T}_Q} Q^n(\underline{x}) = |\mathcal{T}_Q|\, 2^{-nH(Q)}$

$\Rightarrow \quad |\mathcal{T}_Q| \leq 2^{nH(Q)}$.

* $$\boxed{\max_{P\in\mathcal{T}^{(n)}} Q^n(\mathcal{T}_P) = Q^n(\mathcal{T}_Q)}$$ will show this

Proof.

$$\frac{Q^n(\mathcal{T}_P)}{Q^n(\mathcal{T}_Q)} = \frac{|\mathcal{T}_P|}{|\mathcal{T}_Q|}\cdot\frac{\prod_a Q(a)^{nP(a)}}{\prod_a Q(a)^{nQ(a)}}$$

$$= \frac{|\mathcal{T}_P|}{|\mathcal{T}_Q|}\cdot\prod_a Q(a)^{n(P(a)-Q(a))}$$

Show this

$$\boxed{\frac{k!}{l!} \leq k^{k-l}}$$

$$= \frac{n!}{\prod_a (nP(a))!}\cdot\frac{\prod_a (nQ(a))!}{n!}\quad \prod_a Q(a)^{n(P(a)}$$

$$= \prod_a \frac{(nQ(a))!}{(nP(a))!}\, Q(a)^{n(P(a)-Q(a))}$$

$$\leq \prod_a nQ(a)^{n\overbrace{(Q(a)-P(a)+P(a)-Q(}^{0}}$$

$$= 1 \quad \square$$

$\rightarrow \quad 1 = \sum_{P\in\mathcal{T}^{(n)}} Q^n(\mathcal{T}_P) \leq |\mathcal{T}^{(n)}|\; Q^n(\mathcal{T}_Q|$

$\leq (n+1)^{|\mathcal{X}|-1}\cdot 2^{-nH(Q)}\, |\mathcal{T}_Q|$

$$\Rightarrow \boxed{|\mathcal{T}_Q| \geq \frac{2^{nH(Q)}}{(n+1)^{|\mathcal{X}|-1}}}$$

<u>General scheme</u> Given a sequence $\underline{x}\in\mathcal{X}^n$,

(1) Find type Q of $\underline{x}$

(2) Find the index $i(\underline{x};Q)$ of $\underline{x}$ in the list of all sequences of t

[3] Store binary representation of $(Q, i(\underline{x}; Q))$

$\leq \log (n+1)^{|\mathcal{X}|-1}$ bits $\quad$ $\log 2^{n}$

$$\Rightarrow \ell(\underline{x}) \leq (|\mathcal{X}|-1)\log(n+1) + nH(Q)$$

Therefore,

$$\mathbb{E}_{P^n}\left[\ell(X^n)\right] \leq (|\mathcal{X}|-1)\log(n+1) + n\,\mathbb{E}_{P^n}\left[H(P_{X^n})\right]$$

(by Jensen's ineq. since $H(P)$ is concave in P)

$$\leq (|\mathcal{X}|-1)\log(n+1) + nH\Big(\underbrace{\mathbb{E}_{P^n}[P_{X^n}]}_{}\Big) \quad \rightarrow P$$

$\mathbb{E}_{P^n}\{P_{X^n}(a)\} = \frac{n}{n}$

$$= (|\mathcal{X}|-1)\log(n+1) + nH(P)$$

Thus, $\quad \mathbb{E}_{P^n}\left[\ell(X^n)\right] - nH(P) \leq (|\mathcal{X}|-1)\log(n+1)$

$$\Rightarrow \boxed{r_n^{*} \leq (|\mathcal{X}|-1)\log n}$$

C Arithmetic Code: An online scheme

$Q^{(n)}(x^n)$ pmf over sequences of length n from $\mathcal{X}$

→ We want a prefix-free code with codeword lengths $\lceil \log$

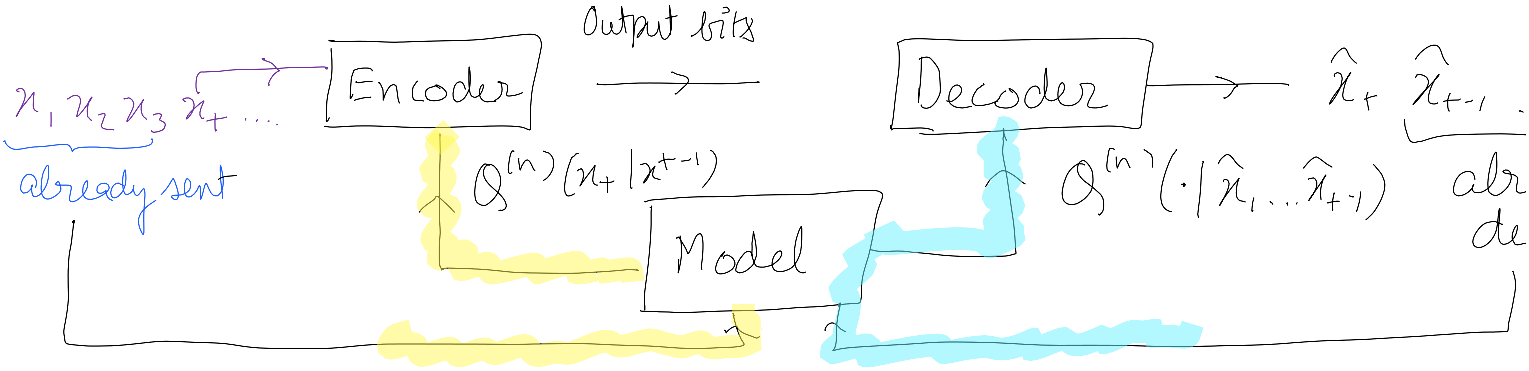
- a scheme that uses $Q^{(n)}(x_i | x^{i-1})$ to encode symbol after it has encoded $(x_1, \ldots, x_{i-1})$ in the previous slots
- "Streaming implementation", can operate symbol-by-sy
- at the decoder side upon decoding $\hat{x}$ $\hat{x}_i$ (

at the decoder side, upon decoding $\hat{x}_1, \ldots, \hat{x}_{i-1}$ (

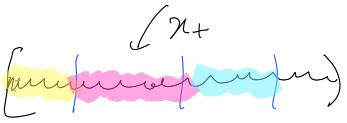
you can decode $\hat{x}_i$ using $Q^{(n)}(x_i | x^{i-1})$

FIFO schemes : "First-in

At time t


Output bits
Encoder
Decoder
$x_1\, x_2\, x_3\, x_t \ldots$
already sent
$Q^{(n)}(x_t | x^{t-1})$
Model
$Q^{(n)}(\cdot | \hat{x}_1 \ldots \hat{x}_{t-1})$
$\hat{x}_t \;\; \hat{x}_{t-1}$
alr
de

Arithmetic code (interval representation) Maintain an inte I_{t-1}



(1) Initial step : $I_0 = [0, 1)$

(2) At time t, (a) Divide I_{t-1} into sub-intervals $\{I_{t-1,x}, x \in \mathcal{X}\}$

($|I| \equiv$ length of interval I)

s.t. $$\frac{|I_{t-1,x}|}{|I_{t-1}|} = Q^{(n)}(X_t = x | X^{t-1} = x^{t-1})$$

(b) Let $I_t = I_{t-1, x_t}$

(3) When you stop, you have an interval I_n of length

$$|I_n| = |I_1| \cdot \frac{|I_2|}{|I_1|} \cdots \frac{|I_n|}{|I_{n-1}|} = Q^{(n)}(x_1)\, Q^{(n)}(x_2 | x_1) .$$

$$= Q^{(n)}(x_1, \ldots, x_n)$$

How do we represent these final intervals using a prefix-free code?

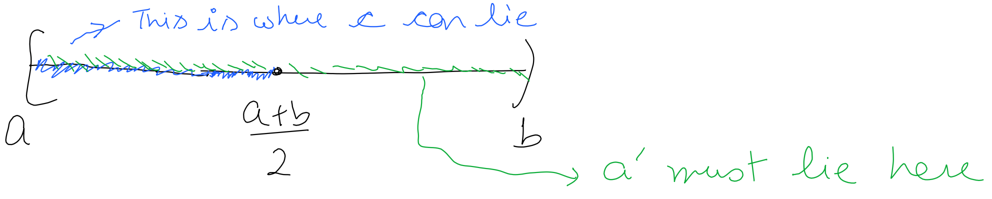
Shannon-Fano-Elias code

Represent the interval $I = [a, b)$ using the $\boxed{l = \lceil -\log(b-a) \rceil + 1}$ most significant bits in the binary representation of $\boxed{\frac{a+b}{2}}$

This code is prefix free

Suppose $\frac{a+b}{2} = 0.r_1 r_2 \ldots$

Let $c = 0.r_1 r_2 \ldots r_l$



Since $\ell = \lceil -\log(b-a) \rceil + 1$, $2^{-\ell} \le \frac{b-a}{2}$.

Suppose $y_1 \dots y_m$ has $x_1 \dots x_\ell$ as its prefix. Consider $a' = 0.y_1 \dots y_m$. $(c \le a' \le c + 2^{-\ell})$

$\Rightarrow$ Both c and a' must lie in I.

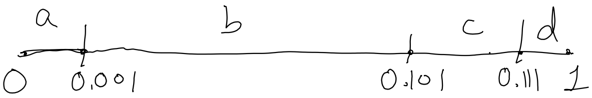
$\Rightarrow$ The code is prefix-free.

→ The average length of the resulting code is $\le H(P) + 1 + 1 = H($

Example: $\mathcal{X} = \{a, b, c, d\}$, $n = 4$, $Q^{(n)}(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} P(x_i)$

where $P(a) = \frac{1}{8}$, $P(b) = \frac{1}{2}$, $P(c)$

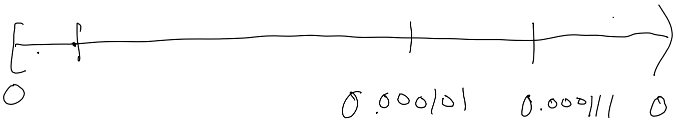
1.



Encode the sequence a c b

$I_1 = [0, 0.001)$

2.



$I_2 = [0.000101, 0.000111)$

3. $I_3 = [0.00010101, 0.000101111)$

4. $I_4 = [\underset{a}{0.0001010101}, \underset{b}{0.00010101111})$

$\hookrightarrow \frac{a+b}{2} = 0.0001010110$

$b - a = 2^{-9} \Rightarrow \ell = 11$

$\Rightarrow$ The codeword corresponding to this interval is $\boxed{0001010}$

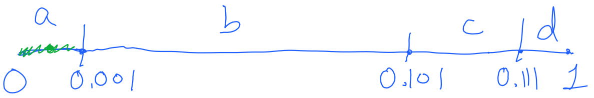
Decoder of arithmetic code

→ Decoder just "inverts" the encoding procedure.

We illustrate using the example above.

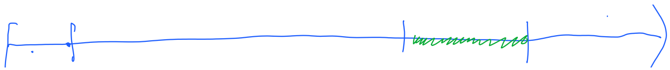
Given 000101011 00

1.



$\Rightarrow x_1 = a$

2.



0 $\quad 0.000101 \quad 0.000111 \quad 0.001$

$\Rightarrow x_2 = c.$

$x_3 = b, \; x_4 = b.$

In conclusion, we have a "practical" algorithm that attains average length $\leq H(Q^{(n)}) + 2$ and requires the model to provide $Q^{(n)}(\cdot \mid x^{t-1})$ at time t.

→ If $Q^{(n)}$ attains R_n^*, we get a scheme with regret less than

$\boxed{D}$ Online probability assignment and a universal scheme

input: $Q^{(n)}$ on $\mathcal{X}^n$ $\quad \left(Q^{(n)}(x_t \mid x^{t-1}) \right)$

output: a code with $l(\underline{x}) = \left\lceil \log \frac{1}{Q^{(n)}(\underline{x})} \right\rceil + 1$

Therefore, given a probability assignment $Q^{(n)}(\underline{x})$, arithmetic code gives a scheme with average length

$$\leq \sum_{\underline{x}} P^n(\underline{x}) \log \frac{1}{Q^{(n)}(\underline{x})} + 2$$

$$= \boxed{D(P^n \| Q^{(n)})} + nH(P) + 2$$

↳ we want to choose a $Q^{(n)}$ that minimi

$$\max_{P \in \mathcal{P}(\mathcal{X})} D(P^n \| Q^{(n)})$$

→ Further, we should be able to efficiently compute $Q^{(n)}(x_t \mid \underline{\underline{x^{t-1}}})$

(ideally, we should not be required ↵ to remember the entire sequence x^{t-1})

$\boxed{\text{Restrict to } \mathcal{X} = \{0, 1\}}$ We use a "Bayesian heuristic" and assum the sequence x^n was generated from iid $\mathrm{Ber}(p)$, where $p \sim \mathrm{unif}\,[$

(i.e., $Q^{(n)}(\underline{x})$ is given by first generating $p \sim \mathrm{unif}\,[0,1]$ and then taking n indep. samples from $\mathrm{Ber}(p)$).

$$Q\,(X_{t+1}=1 \mid X_1=x_1, X_2=x_2, \dots, X_t=x_t) = \frac{Q\,(X_{t+1}=1,\ X_j=x_j,\ }{Q^{(n)}(X_j=x_j,\ 1\le}$$

$$= \frac{\int_0^1 p^{k+1}(1-p)^{t-k}\,dp}{\int_0^1 p^{k}(1-p)^{t-k}\,dp}, \quad \text{where } k = \sum_{j=1}^{t} x_j .$$

Note that

$$\int_0^1 p^m (1-p)^n\,dp = \frac{n}{m+1}\int_0^1 p^{m+1}(1-p)^{n-1}\,dp$$

$$\vdots$$

$$= \frac{n}{m+1}\cdot\frac{(n-1)}{m+2}\cdot\ \cdots\ \frac{1}{m+n}\int_0^1 p^{m+n}\,dp$$

$$= \frac{n!\,m!}{(n+m)!}\ \frac{1}{(n+m+1)}$$

Thus, $Q^{(n)}(X_{t+1}=1 \mid X^t = x^t) = \boxed{\frac{k+1}{t+2}} = \frac{\underbrace{k+1}_{}}{\underbrace{(k+1)}_{\hookleftarrow} + \underbrace{(t-k+1)}_{\hookrightarrow \#\text{ of 0s in } x^t}}$ → empirical esti after adding each

of ~~1s~~ 1s in x^t

This probability assignment is called the $\boxed{\text{add-1 estimate}}$ → L

In general, we can consider the add-α estimate, which corr ~~to~~ to a different prior on p.

The add-$1/2$ is known to ha the "best" performance, and it to Jeffrey's prior ($\pi(p) \propto$

→ Consider a sequence x^n with k ones (and $(n-k)$ zeros).

e.g.

$$Q\,(0010) = Q(0).\ Q(0|0)\ Q(1|00)\ Q(0|001)$$

$$= \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{1}{4}\cdot\frac{3}{5}$$

$$Q\,(0100) = \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{2}{4}\cdot\frac{3}{5}$$

In general, $Q^{(n)}(x^n) = \frac{(1.2\ldots k)(1.2\ldots(n-k))}{2.3\ldots(n+1)} = \frac{1}{(n+1)}\cdot\frac{1}{\binom{n}{k}}$, where k i # of ones in x^n.

$$D(P^n \| Q^{(n)}) = \sum_{\underline{x}} P^n(\underline{x}) \log \frac{1}{Q^{(n)}(\underline{x})} - n H(P)$$

$$= \mathbb{E}_{P^n}\left[\log (n+1)\binom{n}{k}\right] - n H(P)$$

(random) # of ones in $X^n \sim$ iid P

$$\leq \log(n+1) + n\,\mathbb{E}_{P^n}\left[h\left(\frac{k}{n}\right)\right] - n H(P)$$

$$\leq \log(n+1) .$$

$\blacksquare$