To: *C. M. Caves*
From: *C. M. Caves*
Subject: **Laws of large numbers and typical-sequence theorems**
2000 December 11

Most of the following is based on Chapters 3 and 12 of *Elements of Information Theory* by Thomas M. Cover and Joy A. Thomas.

Chebyshev's inequality.

$$P(|\mathbf{x}| \ge a) \le \frac{\langle |\mathbf{x}|^2\rangle}{a^2} \quad \text{where} \quad \mathbf{x} = (x_1, \ldots, x_L)$$

Proof.

$$P(|\mathbf{x}| \ge a) = \int_{|\mathbf{x}|\ge a} d^L x\, p(\mathbf{x}) \le \int_{|\mathbf{x}|\ge a} d^L x\, \frac{|\mathbf{x}|^2}{a^2} p(\mathbf{x}) \le \frac{1}{a^2}\int d^L x\, |\mathbf{x}|^2 p(\mathbf{x}) = \frac{\langle |\mathbf{x}|^2\rangle}{a^2}$$

■

Corollary.

$$P\big(|\mathbf{x} - \langle\mathbf{x}\rangle| \ge a\big) \le \frac{\langle |\mathbf{x} - \langle\mathbf{x}\rangle|^2\rangle}{a^2}$$

i.i.d.'s. Let $\mathbf{x} = (x_1, \ldots, x_N)$ be a sequence of N draws from a probability distribution $\mathbf{p} = (p_1, \ldots, p_L)$ for L alternatives. Each sequence has a vector of occurrence numbers $\mathbf{n}_\mathbf{x} = (n_1, \ldots, n_L) = \mathbf{n}$ and an associated vector of frequencies $\mathbf{f} = (f_1, \ldots, f_L)$, where $f_j = n_j/N$. Sequences with the same occurrence numbers (frequencies) make up a *type*. The probability of sequence $\mathbf{x}$ is

$$P(\mathbf{x}) = p_1^{n_1} \cdots p_L^{n_L} \ ,$$

and the probability of type $\mathbf{f}$ is

$$P(\mathbf{n}) = \frac{N!}{n_1! \cdots n_L!} p_1^{n_1} \cdots p_L^{n_L} \ .$$

It is easy to calculate means and second moments for the occurrence numbers and frequencies:

$$\begin{aligned}
\langle n_j \rangle &= \sum_{\mathbf{n}} n_j P(\mathbf{n}) \\
&= \left(p_j \frac{\partial}{\partial p_j} \sum_{\mathbf{n}} \frac{N!}{n_1! \cdots n_L!} p_1^{n_1} \cdots p_L^{n_L} \right) \Bigg|_{p_1+\cdots+p_L=1} \\
&= \left(p_j \frac{\partial}{\partial p_j} (p_1 + \cdots + p_L)^N \right) \Bigg|_{p_1+\cdots+p_L=1} \\
&= N p_j \ , \\
\langle n_j n_k \rangle &= \sum_{\mathbf{n}} n_j n_k P(\mathbf{n}) \\
&= \left(p_j \frac{\partial}{\partial p_j} p_k \frac{\partial}{\partial p_k} \sum_{\mathbf{n}} \frac{N!}{n_1! \cdots n_L!} p_1^{n_1} \cdots p_L^{n_L} \right) \Bigg|_{p_1+\cdots+p_L=1} \\
&= \left(p_j \frac{\partial}{\partial p_j} p_k \frac{\partial}{\partial p_k} (p_1 + \cdots + p_L)^N \right) \Bigg|_{p_1+\cdots+p_L=1} \\
&= N(N-1) p_j p_k - N p_j \delta_{jk} \ ,
\end{aligned}$$

One thus obtains the correlation matrix of the occurrence numbers,

$$\langle \Delta n_j \Delta n_k \rangle = \langle n_j n_k \rangle - \langle n_j \rangle \langle n_k \rangle = N(p_j \delta_{jk} - p_j p_k) \ ,$$

and the means and correlation matrix of the frequencies,

$$\langle f_j \rangle = p_j \ , \quad \langle \Delta f_j \Delta f_k \rangle = \frac{(p_j \delta_{jk} - p_j p_k)}{N} \ .$$

Weak law of large numbers

$$\langle | \, \mathbf{f} - \mathbf{p} |^2 \rangle \le 1/N$$

Proof.

$$\langle | \, \mathbf{f} - \mathbf{p} |^2 \rangle = \sum_{j=1}^{L} (\Delta f_j)^2 = \sum_{j=1}^{L} \frac{p_j(1-p_j)}{N} = \frac{1 - \sum_{j=1}^{L} p_j^2}{N} \le 1/N$$

■

Weak law of large numbers. A second version.. For any $\delta, \epsilon > 0$, there exists an N_0 such that for all $N \geq N_0$,

$$P\big(|f_j - p_j| < \epsilon,\ j = 1, \ldots, L\big) \geq P\big(|\,\mathbf{f} - \mathbf{p}| < \epsilon\big) \geq 1 - \delta\ .$$

Proof. Start with

$$P\big(|\,\mathbf{f} - \mathbf{p}| < \epsilon\big) = 1 - P\big(|\,\mathbf{f} - \mathbf{p}| \geq \epsilon\big) \geq 1 - \frac{\langle |\,\mathbf{f} - \langle \mathbf{p} \rangle|^2 \rangle}{\epsilon^2} \geq 1 - \frac{1}{N\epsilon^2}\ .$$

Given δ and ϵ, choose $N_0 \geq 1/\delta\epsilon^2$. ■

The strong law of large numbers is a much stronger statement that sequences whose frequencies limit to the probabilities have probability 1 in the infinite limit.

Typical sequences. The set of typical sequences of length N, denoted $\mathrm{TYP}_\epsilon(N)$, is defined by

$$\mathrm{TYP}_\epsilon(N) \equiv \big\{\mathbf{x}\ \big|\ 2^{-N[H(\mathbf{p})+\epsilon]} < P(\mathbf{x}) < 2^{-N[H(\mathbf{p})-\epsilon]}\big\} = \big\{\mathbf{x}\ \big|\ |{-\log P(\mathbf{x})/N} - H(\mathbf{p})| < \epsilon\big\}\ ,$$

where

$$H(\mathbf{p}) = -\sum_{j=1}^{L} p_j \log p_j$$

is the entropy of the distribution $\mathbf{p}$. Notice that

$$\frac{-\log P(\mathbf{x})}{N} = -\sum_{j=1}^{L} f_j \log p_j\ ,$$

so

$$\left\langle \frac{-\log P(\mathbf{x})}{N} \right\rangle = H(\mathbf{p})$$

and

$$\frac{-\log P(\mathbf{x})}{N} - H(\mathbf{p}) = -\sum_{j=1}^{L} \Delta f_j \log p_j\ .$$

Typical-sequence theorem. For any $\delta, \epsilon > 0$, there exists an N_0 such that

1. For all $N \geq N_0$,
$$P\big(\mathrm{TYP}_\epsilon(N)\big) \geq 1 - \delta\ ;$$
2. For all N the number of sequences in $\mathrm{TYP}_\epsilon(N)$ satisfies
$$\big|\mathrm{TYP}_\epsilon(N)\big| < 2^{N[H(\mathbf{p})+\epsilon]}\ .$$

3. For all $N \geq N_0$,
$$\big|\mathrm{TYP}_\epsilon(N)\big| > (1-\delta)2^{N[H(\mathbf{p})-\epsilon]} \ .$$

Proof.

1. Since
$$\left\langle \left| \frac{-\log P(\mathbf{x})}{N} - H(\mathbf{p}) \right|^2 \right\rangle = \left\langle \left(\sum_{j=1}^{L} \Delta f_j \log p_j \right)^2 \right\rangle$$
$$= \sum_{j,k} \langle \Delta f_j \Delta f_k \rangle \log p_j \log p_k$$
$$= \sum_{j,k} \frac{(p_j \delta_{jk} - p_j p_k)}{N} \log p_j \log p_k$$
$$= \frac{1}{N} \left(\sum_{j=1}^{L} p_j (\log p_j)^2 - H^2 \right) ,$$

we have that
$$P\big(\mathrm{TYP}_\epsilon(N)\big) = 1 - P\big(| -\log P(\mathbf{x})/N - H(\mathbf{p})| \geq \epsilon\big)$$
$$\geq 1 - \frac{\big\langle | -\log P(\mathbf{x})/N - H(\mathbf{p})|^2 \big\rangle}{\epsilon^2}$$
$$= 1 - \frac{1}{N\epsilon^2} \left(\sum_{j=1}^{L} p_j (\log p_j)^2 - H^2 \right) .$$

Given δ and ϵ, choose
$$N_0 \geq \frac{1}{\delta\epsilon^2} \left(\sum_{j=1}^{L} p_j (\log p_j)^2 - H^2 \right) .$$

Then, for all $N \geq N_0$, we have
$$P\big(\mathrm{TYP}_\epsilon(N)\big) \geq 1 - \frac{1}{N_0\epsilon^2} \left(\sum_{j=1}^{L} p_j (\log p_j)^2 - H^2 \right) \geq 1 - \delta \ .$$

2.
$$1 \geq P\big(\mathrm{TYP}_\epsilon(N)\big)$$
$$= \sum_{\mathbf{x} \in \mathrm{TYP}_\epsilon(N)} P(\mathbf{x})$$
$$> \sum_{\mathbf{x} \in \mathrm{TYP}_\epsilon(N)} 2^{-N[H(\mathbf{p})+\epsilon]}$$
$$= \big|\mathrm{TYP}_\epsilon(N)\big| 2^{-N[H(\mathbf{p})+\epsilon]}$$

3. Choosing N_0 as in 1, we have for all $N \geq N_0$,

$$
\begin{aligned}
1-\delta &\leq P\big(\mathrm{TYP}_\epsilon(N)\big) \\
&= \sum_{\mathbf{x}\in \mathrm{TYP}_\epsilon(N)} P(\mathbf{x}) \\
&< \sum_{\mathbf{x}\in \mathrm{TYP}_\epsilon(N)} 2^{-N[H(\mathbf{p})-\epsilon]} \\
&= \big|\mathrm{TYP}_\epsilon(N)\big| 2^{-N[H(\mathbf{p})-\epsilon]} \;.
\end{aligned}
$$

■

Type classes. As noted above, sequences $\mathbf{x}$ with the same occurrence numbers $\mathbf{n}$—i.e., with the same frequencies $\mathbf{f}$—make up a type. Formally, we define the type $T_{\mathbf{f}}(N)$ to be the set

$$T_{\mathbf{f}}(N) \equiv \{\mathbf{x}|\mathbf{n}_{\mathbf{x}} = N\mathbf{f}\} \;.$$

The number of sequences in the type $T_{\mathbf{f}}(N)$ is given by a multinomial coefficient:

$$\big|T_{\mathbf{f}}(N)\big| = \frac{N!}{(Nf_1)!\cdots(Nf_L)!} \;.$$

The probability for any sequence in $T_{\mathbf{f}}(N)$ is given by

$$P(\mathbf{x}) = p_1^{Nf_1}\cdots p_L^{Nf_L} = 2^{N[f_1\log p_1+\cdots+f_L\log p_L]} = 2^{-N[H(\mathbf{f})+H(\mathbf{f}||\mathbf{p})]} \;,$$

where

$$
\begin{aligned}
H(\mathbf{f}||\,\mathbf{p}) &\equiv \sum_{j=1}^{L} f_j \log\left(\frac{f_j}{p_j}\right) \\
&= \sum_{j=1}^{L} f_j \log f_j - \sum_{j=1}^{L} f_j \log p_j \\
&= -H(\mathbf{f}) - \sum_{j=1}^{L} f_j \log p_j \\
&= -H(\mathbf{f}) - \frac{\log P(\mathbf{x})}{N}
\end{aligned}
$$

is the *relative entropy* of the distributions $\mathbf{f}$ and $\mathbf{p}$.

It is easy to show that

$$\begin{aligned}
H(\mathbf{f}||\,\mathbf{p}) &= -\sum_{j=1}^{L} f_j \log\left(\frac{p_j}{f_j}\right) \\
&\ge \frac{1}{\ln 2}\sum_{j=1}^{L} f_j \left(1-\frac{p_j}{f_j}\right) \\
&= \frac{1}{\ln 2}\sum_{j=1}^{L}(f_j - p_j) \\
&= 0\ ,
\end{aligned}$$

with equality holding if and only if $\mathbf{f}=\mathbf{p}$. (Here we use $-\log x = -\ln x/\ln 2 \ge (1-x)/\ln 2$, with equality if and only if $x=1$.) When $\mathbf{f}=\mathbf{p}+\Delta\mathbf{f}$ is close to $\mathbf{p}$, the relative entropy becomes the Wootters distance:

$$H(\mathbf{f}||\,\mathbf{p}) = \frac{1}{2\ln 2}\sum_{j=1}^{L}\frac{(\Delta f_j)^2}{p_j}\ .$$

Notice that the difference between the entropies of $\mathbf{f}$ and $\mathbf{p}$ has two parts, one the relative entropy and the other the difference that defines typical subsets:

$$H(\mathbf{f}) - H(\mathbf{p}) = -H(\mathbf{f}||\,\mathbf{p}) - \sum_{j=1}^{L}\Delta f_j \log p_j = -H(\mathbf{f}||\,\mathbf{p}) + \left(\frac{-\log P(\mathbf{x})}{N} - H(\mathbf{p})\right)\ .$$

The probability of the type $T_{\mathbf{f}}(N)$ can be expressed as

$$P\big(T_{\mathbf{f}}(N)\big) = P(\mathbf{n}) = \big|T_{\mathbf{f}}(N)\big| p_1^{Nf_1}\cdots p_L^{Nf_L} = \big|T_{\mathbf{f}}(N)\big| 2^{-N[H(\mathbf{f})+H(\mathbf{f}||\mathbf{p})]}\ .$$

Notice that the probability is bounded by

$$P\big(T_{\mathbf{f}}(N)\big) \le \big|T_{\mathbf{f}}(N)\big| 2^{-NH(\mathbf{f})}\ ,$$

with equality holding if and only if $\mathbf{p}=\mathbf{f}$.

Number of types. Let $\mathcal{P}_N$ be the set of types for sequences of length N. The possible occurrence numbers are in one-to-one correspondence with binary strings of the form

$$\underbrace{0\ldots0}_{n_1 0's}1\underbrace{0\ldots0}_{n_2 0's}1 \quad \ldots \quad 1\underbrace{0\ldots0}_{n_L 0's},$$

where the 1's are used to separate substrings of 0's whose lengths give the occurrence numbers n_j. These binary strings have $L-1$ 1's and total length $N+L-1$. The number of such binary strings—and, hence, the number of types—is

$$|\mathcal{P}_N| = \frac{(N+L-1)!}{N!(L-1)!} .$$

The number of types is the same as the number of states for a Bose-Einstein system of N particles occupying L single-particle states, and the argument leading to $\mathcal{P}_N$ is the standard one for determining the number Bose-Einstein states.

The number of types can be bounded by

$$|\mathcal{P}_N| \le (N+1)^{L-1} \le (N+1)^L .$$

Proof. The second inequality follows directly from noting that a type is specified by L occurrence numbers, each of which can take on $N+1$ values. The first inequality, which provides a better bound, comes from

$$\begin{aligned}
|\mathcal{P}_N| &= \frac{N+L-1}{L-1}\frac{N+L-2}{L-2}\cdots\frac{N+2}{2}\frac{N+1}{1} \\
&= \prod_{k=1}^{L-1}\frac{N+k}{k} \\
&\le \prod_{k=1}^{L-1}\frac{kN+k}{k} \\
&= \prod_{k=1}^{L-1}(N+1) \\
&= (N+1)^{L-1} .
\end{aligned}$$

■

Bounds on sizes of type classes.

$$\frac{1}{(N+1)^{L-1}}2^{NH(\mathbf{f})} \le \frac{1}{|\mathcal{P}_N|}2^{NH(\mathbf{f})} \le \big|T_{\mathbf{f}}(N)\big| \le 2^{NH(\mathbf{f})}$$

Proof. The rightmost inequality is easy:

$$1 \ge P\big(T_{\mathbf{p}}(N)\big) = \big|T_{\mathbf{p}}(N)\big|2^{-NH(\mathbf{p})} .$$

To prove the left-hand inequalities, we first need to show that $P\big(T_{\mathbf{f}}(N)\big) \le P\big(T_{\mathbf{p}}(N)\big)$. We proceed by noting that

$$\frac{P\big(T_{\mathbf{p}}(N)\big)}{P\big(T_{\mathbf{f}}(N)\big)} = \frac{\big|T_{\mathbf{p}}(N)\big|p_1^{Np_1}\cdots p_L^{Np_L}}{\big|T_{\mathbf{f}}(N)\big|p_1^{Nf_1}\cdots p_L^{Nf_L}} = \frac{(Nf_1)!\cdots(Nf_L)!}{(Np_1)!\cdots(Np_L)!}p_1^{N(p_1-f_1)}\cdots p_L^{N(p_L-f_L)} .$$

The factorials can be bounded by $m!/n! \ge n^{m-n}$, which is easily proved by separately considering $m \ge n$ and $m < n$. We find

$$\begin{aligned}
\frac{P\big(T_{\mathbf{p}}(N)\big)}{P\big(T_{\mathbf{f}}(N)\big)} &\ge (Np_1)^{N(f_1-p_1)}\cdots(Np_L)^{N(f_L-p_L)}p_1^{N(p_1-f_1)}\cdots p_L^{N(p_L-f_L)} \\
&= N^{N(f_1-p_1+\cdots+f_L-p_L)} \\
&= N^{N(1-1)} \\
&= 1 .
\end{aligned}$$

Now we can write

$$1 = \sum_{\mathbf{f}} P\big(T_{\mathbf{f}}(N)\big) \le \sum_{\mathbf{f}} P\big(T_{\mathbf{p}}(N)\big) = |\mathcal{P}_N|P\big(T_{\mathbf{p}}(N)\big) = |\mathcal{P}_N|\big|T_{\mathbf{p}}(N)\big|2^{-NH(\mathbf{p})} .$$

■

Bounds on probabilities of type classes.

$$\frac{1}{(N+1)^{L-1}}2^{-NH(\mathbf{f}||\,\mathbf{p})} \le \frac{1}{|\mathcal{P}_N|}2^{-NH(\mathbf{f}||\,\mathbf{p})} \le P\big(T_{\mathbf{f}}(N)\big) \le 2^{-NH(\mathbf{f}||\,\mathbf{p})}$$

Proof. The bounds follow immediately from applying the bounds on the sizes of type classes to

$$P\big(T_{\mathbf{f}}(N)\big) = \big|T_{\mathbf{f}}(N)\big|2^{-N[H(\mathbf{f})+H(\mathbf{f}||\,\mathbf{p})]} .$$

■

Another set of typical sequences. We can define another kind of set of typical sequences of length N:

$$\mathrm{TYP}'_\epsilon(N) \equiv \big\{\mathbf{x} \bigm| H(\mathbf{f}_\mathbf{x}||\,\mathbf{p}) < \epsilon\big\} = \bigcup_{H(\mathbf{f}||\,\mathbf{p})<\epsilon} T_\mathbf{f}(N)\ .$$

Typical-sequence theorem. For any $\delta, \epsilon > 0$, there exists an N_0 such that for all $N \geq N_0$,

$$P\big(\mathrm{TYP}'_\epsilon(N)\big) \geq 1 - \delta\ .$$

Proof. We first note that

$$\begin{aligned}
1 - P\big(\mathrm{TYP}'_\epsilon(N)\big) &= \sum_{\{\mathbf{f}|H(\mathbf{f}||\,\mathbf{p})\geq\epsilon\}} P\big(T_\mathbf{f}(N)\big) \\
&\leq \sum_{\{\mathbf{f}|H(\mathbf{f}||\,\mathbf{p})\geq\epsilon\}} 2^{-NH(\mathbf{f}||\,\mathbf{p})} \\
&\leq \sum_{\{\mathbf{f}|H(\mathbf{f}||\,\mathbf{p})\geq\epsilon\}} 2^{-N\epsilon} \\
&\leq \sum_{\mathbf{f}} 2^{-N\epsilon} \\
&= |\mathcal{P}_N|\, 2^{-N\epsilon} \\
&\leq (N+1)^{L-1} 2^{-N\epsilon} \\
&= 2^{-N\big(\epsilon + [(L-1)\log(N+1)]/N\big)}\ .
\end{aligned}$$

The function $(N+1)^{L-1}2^{-N\epsilon}$ is equal to $2^{L-1-\epsilon} \geq 1$ at $N = 1$, increases to a maximum ≥ 1 at $N = N_c = -1 + (L-1)/\epsilon \ln 2$, and then decreases for $N > N_c$. Choose $N_0 > N_c$ to satisfy

$$\delta = (N_0+1)^{L-1} 2^{-N_0\epsilon}\ .$$

Then for all $N \geq N_0$, we have

$$P\big(\mathrm{TYP}'_\epsilon(N)\big) \geq 1 - (N+1)^{L-1}2^{-N\epsilon} \geq 1 - (N_0+1)^{L-1}2^{-N_0\epsilon} = 1 - \delta\ .$$

■

Csiszár-Körner typical sequences. For a maximum entropy H_0, define the Csiszár-Körner set of typical sequences of length N:

$$\mathrm{CK}_{H_0}(N) \equiv \big\{\mathbf{x} \bigm| H(\mathbf{f}_\mathbf{x}) \le H_0\big\} = \bigcup_{H(\mathbf{f}) \le H_0} T_\mathbf{f}(N) \ .$$

Csiszár-Körner typical-sequence theorem. For any $\delta, \epsilon > 0$, there exists an N_0 such that for all $N \ge N_0$,

1. For any $\mathbf{p}$ such that $H(\mathbf{p}) < H_0$, $P\big(\mathrm{CK}_{H_0}(N)\big) \ge 1 - \delta$,
2. $\big|\mathrm{CK}_{H_0}(N)\big| < 2^{N(H_0+\epsilon)}$.

Proof. We need two properties:

$$\begin{aligned}
1 - P\big(\mathrm{CK}_{H_0}(N)\big) &= \sum_{\{\mathbf{f}|H(\mathbf{f})>H_0\}} P\big(T_\mathbf{f}(N)\big) \\
&\le \sum_{\{\mathbf{f}|H(\mathbf{f})>H_0\}} 2^{-NH(\mathbf{f}||\,\mathbf{p})} \\
&\le \sum_{\{\mathbf{f}|H(\mathbf{f})>H_0\}} 2^{-NH^*_{H_0,\mathbf{p}}} \\
&\le \sum_{\mathbf{f}} 2^{-NH^*_{H_0,\mathbf{p}}} \\
&= |\mathcal{P}_N|\, 2^{-NH^*_{H_0,\mathbf{p}}} \\
&\le (N+1)^{L-1} 2^{-NH^*_{H_0,\mathbf{p}}} \ ,
\end{aligned}$$

where

$$H^*_{H_0,\mathbf{p}} \equiv \inf_{\{\mathbf{f}|H(\mathbf{f})>H_0\}} H(\mathbf{f}||\,\mathbf{p}) \ ,$$

and

$$\begin{aligned}
\big|\mathrm{CK}_{H_0}(N)\big| &= \sum_{\{\mathbf{f}|H(\mathbf{f})\le H_0\}} \big|T_\mathbf{f}(N)\big| \\
&\le \sum_{\{\mathbf{f}|H(\mathbf{f})\le H_0\}} 2^{NH(\mathbf{f})} \\
&\le \sum_{\{\mathbf{f}|H(\mathbf{f})\le H_0\}} 2^{NH_0} \\
&\le \sum_{\mathbf{f}} 2^{NH_0} \\
&= |\mathcal{P}_N|\, 2^{NH_0} \\
&\le (N+1)^{L-1} 2^{NH_0} \\
&= 2^{N\left(H_0 + [(L-1)\log(N+1)]/N\right)} \ .
\end{aligned}$$

If $\mathbf{p}$ is such that $H(\mathbf{p}) \geq H_0$, then $H^*_{H_0,\mathbf{p}} = 0$, and there is no bound on the probability $1 - P\big(\mathrm{CK}_{H_0}(N)\big)$. In contrast, if $\mathbf{p}$ is such that $H(\mathbf{p}) < H_0$, then $H^*_{H_0,\mathbf{p}} > 0$. Then, choosing $N_0 = \max(N_1, N_2)$, where N_1 and N_2 are defined by

$$\epsilon = (N_1 + 1)^{L-1} 2^{-N_1 H^*_{H_0,\mathbf{p}}}$$
$$\delta = \frac{(L-1)\log(N_2 + 1)}{N_2} \ ,$$

we have the two results for all $N \geq N_0$. ■