Confidence Expectation and Convergence for k-Tails

Supporting materials on log completeness, August 2014

Available from http://smlab.cs.tau.ac.il/logcompleteness

Hila Cohen and Shahar Maoz

School of Computer Science
Tel Aviv University, Israel

In this document we compute the expected confidence of a log of size n. We then show how as n gets larger, the expected confidence approaches one, and the expected change in confidence following the addition of a new trace, approaches zero. We demonstrate these results on a simple model, and also compute the partial expectation with respect to confidence decrease for this model using various log sizes.

Let C_n be the computed confidence for k-Tails after n traces.

Theorem. $\lim_{n\to+\infty} \mathbb{E}[C_n] = 1$

Proof.

$$\begin{aligned}
\mathbb{E}[C_n] = \mathbb{E}\left[1 - \sum_{\{es|\hat{q}_{es}>0\}} (1-\hat{q}_{es})^n\right] &= 1 - \mathbb{E}\left[\sum_{\{es|\hat{q}_{es}>0\}} (1-\hat{q}_{es})^n\right] \\
&= 1 - \mathbb{E}\left[\sum_{es} (1-\delta(\hat{q}_{es},0))(1-\hat{q}_{es})^n\right] \\
&= 1 - \mathbb{E}\left[\sum_{es} (1-\hat{q}_{es})^n - \delta(\hat{q}_{es},0)(1-\hat{q}_{es})^n\right] \\
&= 1 - \sum_{es} \mathbb{E}\left[(1-\hat{q}_{es})^n - \delta(\hat{q}_{es},0)(1-\hat{q}_{es})^n\right] \\
&= 1 - \sum_{es} \mathbb{E}\left[(1-\hat{q}_{es})^n\right] - \mathbb{E}\left[\delta(\hat{q}_{es},0)(1-\hat{q}_{es})^n\right] \\
&= 1 - \sum_{es} \mathbb{E}\left[(1-\hat{q}_{es})^n\right] - \mathbb{P}\left[\hat{q}_{es}=0\right] \\
&= 1 - \sum_{es} \mathbb{E}\left[(1-\sum_{i=1}^{n}\frac{Y_i(es)}{n})^n\right] - (1-q_{es})^n
\end{aligned}$$

For a fixed es, denote $\sum_{i=1}^{n} Y_i(es) \equiv Z_n$, then $Z_n \sim Bin(n, q_{es})$.

$$
\begin{aligned}
1 - \sum_{es} \mathbb{E}\left[(1 - \frac{Z_n}{n})^n\right] - (1 - q_{es})^n &= 1 - \sum_{es}\left(\sum_{k=0}^{n}\left((1 - \frac{k}{n})^n \cdot \mathbb{P}\left[Z_n = k\right]\right) - (1 - q_{es})^n\right) \\
&\uparrow \\
&\text{law of the unconscious statistician} \\
&= 1 - \sum_{es}\left(\sum_{k=1}^{n}\left((1 - \frac{k}{n})^n \cdot \mathbb{P}\left[Z_n = k\right]\right) + \mathbb{P}\left[Z_n = 0\right] - (1 - q_{es})^n\right. \\
&= 1 - \sum_{es}\left(\sum_{k=1}^{n}\left((1 - \frac{k}{n})^n \cdot \mathbb{P}\left[Z_n = k\right]\right)\right) \\
&= 1 - \sum_{es}\left(\sum_{k=1}^{n}\left((1 - \frac{k}{n})^n \binom{n}{k} q_{es}^k (1 - q_{es})^{n-k}\right)\right)
\end{aligned}
$$

Let $\hat{q}_{es,n}$ be the computed $\hat{q}_{es}$ after n traces. Let $C_\Delta = C_{n+1} - C_n$.

$$
\begin{aligned}
\mathbb{E}\left[C_\Delta\right] &= \mathbb{E}\left[C_{n+1} - C_n\right] = \mathbb{E}\left[C_{n+1}\right] - \mathbb{E}\left[C_n\right] \\
&= 1 - \sum_{es} \mathbb{E}\left[(1 - \frac{Z_{n+1}}{n+1})^{n+1}\right] - (1 - q_{es,n+1})^{n+1} - 1 + \sum_{es} \mathbb{E}\left[(1 - \frac{Z_n}{n})^n\right] - (1 - q_{es,n})^n \\
&= \sum_{es} \mathbb{E}\left[(1 - \frac{Z_n}{n})^n\right] - (1 - q_{es,n})^n - \sum_{es} \mathbb{E}\left[(1 - \frac{Z_{n+1}}{n+1})^{n+1}\right] - (1 - q_{es,n+1})^{n+1} \\
&\leq \sum_{es} \mathbb{E}\left[(1 - \frac{Z_n}{n})^n\right] - (1 - q_{es,n})^n - \sum_{es}\left(1 - \frac{\mathbb{E}\left[Z_{n+1}\right]}{n+1}\right)^{n+1} - (1 - q_{es,n+1})^{n+1} \\
&\uparrow \\
&\text{Jensen's inequality} \\
&= \sum_{es} \mathbb{E}\left[(1 - \frac{Z}{n})^n\right] - (1 - q_{es,n})^n - \sum_{es}(1 - q_{es,n+1})^{n+1} - (1 - q_{es,n+1})^{n+1} \\
&= \sum_{es} \mathbb{E}\left[(1 - \frac{Z}{n})^n\right] - (1 - q_{es,n})^n \\
&= \sum_{es}\left(\sum_{k=1}^{n}\left((1 - \frac{k}{n})^n \cdot \mathbb{P}\left[Z_n = k\right]\right)\right) = 1 - \mathbb{E}\left[C_n\right]
\end{aligned}
$$

$$0 \leq \lim_{n \to +\infty} P\left[Z_n = k\right] = \lim_{n \to +\infty} \binom{n}{k} q_{es}^k (1 - q_{es})^{n-k} = \lim_{n \to +\infty} \left(\frac{q_{es}}{1 - q_{es}}\right)^k \binom{n}{k} (1 - q_{es})^n$$
$$\leq const(q_{es}, k) \lim_{n \to +\infty} n^k (1 - q_{es})^n = 0$$

Finally, we have

$$\lim_{n \to +\infty} \mathbb{E}\left[C_\Delta\right] = \lim_{n \to +\infty} \sum_{es} \left(\sum_{k=1}^{n} \left(\underbrace{(1 - \frac{k}{n})^n}_{\downarrow\, e^{-k}} \cdot \mathbb{P}\left[Z_n = k\right] \right) \right) = 0$$

□

Example Computation

We give an example computation for the open-read-close model shown in Figure 1. In this example, the confidence is computed with regard to the k-Tails algorithm with $k = 2$. The model's 2-directly-follows probabilities are as follows:
$q_{\texttt{close;read}} = 0$
$q_{\texttt{close;open}} = 0.5$
$q_{\texttt{open;read}} = 1$
$q_{\texttt{open;open}} = 0$
$q_{\texttt{open;close}} = 0$
$q_{\texttt{read;open}} = 0$
$q_{\texttt{read;read}} = 0.67$
$q_{\texttt{read;close}} = 1$
$q_{\texttt{close;close}} = 0$

Thus, the expected confidence for a log of size n is

$$\mathbb{E}\left[C_n\right] = 1 - \left(\sum_{k=1}^{n} (1 - \frac{k}{n})^n \binom{n}{k} 0.67^k (1 - 0.67)^{n-k} + \sum_{k=1}^{n} (1 - \frac{k}{n})^n \binom{n}{k} 0.5^k (1 - 0.5)^{n-k} \right) \tag{1}$$

Table 1 shows the confidence expectation results for $n \leq 10$.

	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$	$n=10$
$\mathbb{E}[C_n]$	1	.764	.793	.845	.891	.926	.951	.967	.978	.986
$\mathbb{E}[C_\Delta]$	-	-.236	+.029	+.052	+.046	+.035	+.024	+.017	+.011	+.007

Table 1: Example of expected confidence for a log of size n. Computation is done according to Eq. 1.

One might be interested in a partial expectation with respect to confidence decrease, namely computing the expectation over the cases in which the confidence drops after the addition of a new trace.

$$\begin{aligned}\mathbb{E}^{-}\left[C_\Delta\right] \equiv \mathbb{E}\left[C_\Delta | C_\Delta < 0\right]\mathbb{P}\left[C_\Delta < 0\right] &= \mathbb{E}\left[C_{n+1} - C_n | C_{n+1} < C_n\right]\mathbb{P}\left[C_{n+1} < C_n\right] \\ &= \sum_{\substack{(x-y)<0 \\ x,y\in\mathbb{R}}} (x-y)\mathbb{P}\left[C_{n+1}=x, C_n=y\right]\end{aligned}$$

$\mathbb{P}\left[C_{n+1}=x, C_n=y\right]$ is the joint probability mass function of the confidence after n traces (C_n), and the confidence after $n+1$ traces (C_{n+1}). Since C_n is a discrete random variable, in general, y has at most $(n+1)^{\Sigma^k}$ possible values for which $\mathbb{P}\left[C_n=y\right] > 0$: $C_n(\hat{q}_{es_1,n}, \hat{q}_{es_2,n}, \ldots, \hat{q}_{es_{\Sigma^k},n})$, where each $\hat{q}_{es,n}$ has $n+1$ possible values.
For C_{n+1}, on the other hand, x has at most 2^{Σ^k} possible values for which $\mathbb{P}\left[C_{n+1}=x | C_n=y\right] > 0$: $C_{n+1}(\hat{q}_{es_1,n+1}, \hat{q}_{es_2,n+1}, \ldots, \hat{q}_{es_{\Sigma^k},n+1})$, where each $\hat{q}_{es,n+1}$ has two possible values representing whether es appeared in the $n+1$'th trace or not.
The random variables C_{n+1} and C_n are clearly dependant, and their joint mass function is model-specific.

Table 2 shows the computation of the expected confidence drop for $n=1$. The confidence is expected to drop by 0.236 with the addition of a new trace. The columns of the table correspond to the possible values of $\hat{q}_{\texttt{<close,open>},1}$ and $\hat{q}_{\texttt{<read,read>},1}$ in a single trace. E.g., column (1,0) refers to `<close,open>` appearing in the trace and `<read,read>` not appearing in the trace. The rows of the table correspond to the possible values of $\hat{q}_{\texttt{close,open},2}$ and $\hat{q}_{\texttt{read,read},2}$ in two traces. E.g., row (2,1) refers to `<close,open>` appearing in both traces and `<read,read>` appearing in one of the traces.
Each cell of the table $((i_1,i_2),(j_1,j_2))$ contains the multivariate joint probability $\mathbb{P}\left[\hat{q}_{\texttt{<close,open>},1}=j_1, \hat{q}_{\texttt{<read,read>},1}=j_2, \hat{q}_{\texttt{close,open},2}=i_1, \hat{q}_{\texttt{read,read},2}=i_2\right]$, and the calculated decrease in confidence $C_2\left(\hat{q}_{\texttt{close,open},2}, \hat{q}_{\texttt{read,read},2}\right) - C_1\left(\hat{q}_{\texttt{<close,open>},1}, \hat{q}_{\texttt{<read,read>},1}\right)$. The decrease in confidence is calculated only for positive joint probabilities.
The expected decrease is then calculated from the table cells by multiplying the joint probability and the actual confidence difference for each cell, and summing these values across all cells. Note that for $n=1$, since the confidence can only decrease or stay the same, none of the table cells have a positive confidence difference. For $n>1$, the corresponding table contains cells with positive, negative and zero confidence differences. In order to compute

	(0,0)	(0,1)	(1,0)	(1,1)	
(0,0)	0.063 (0.00)	0.000	0.000	0.000	0.000
(0,1)	0.063 (-0.25)	0.063 (-0.25)	0.000	0.000	-0.031
(0,2)	0.000	0.063 (0.00)	0.000	0.000	0.000
(1,0)	0.021 (-0.25)	0.000	0.021 (-0.25)	0.000	-0.010
(1,1)	0.104 (-0.50)	0.021 (-0.50)	0.021 (-0.50)	0.104 (-0.50)	-0.125
(1,2)	0.000	0.104 (-0.25)	0.000	0.104 (-0.25)	-0.052
(2,0)	0.000	0.000	0.007 (0.00)	0.000	0.000
(2,1)	0.000	0.000	0.035 (-0.25)	0.035 (-0.25)	-0.017
(2,2)	0.000	0.000	0.000	0.174 (0.00)	0.000
					-0.236

Table 2: Example of expected confidence decrease for a log with a single trace ($n = 1$). With the addition of a new trace, the confidence is expected to drop by 0.236.

	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$	$n = 8$
$\mathbb{E}^{-}[C_{\Delta}]$	-0.236	-0.099	-0.064	-0.044	-0.030	-0.020	-0.014	-0.009

Table 3: Example of partial expectation with respect to confidence decrease for a log of size n.

the expected confidence decrease, only cells with negative confidence difference should be included in the summation.

Table 3 shows the partial expectation with respect to confidence decrease for a log of size n. For example, assume we have a log of 5 traces and we extend it by an additional trace. According to Table 1, without further assumptions, the confidence is expected to increase by 0.035. However, if we assume that the confidence is going to decrease, then it is expected to decrease by 0.03. Note that in this example, the expected decrease goes to zero very quickly, and already at the 8th trace the confidence is likely to decrease by only 0.009.

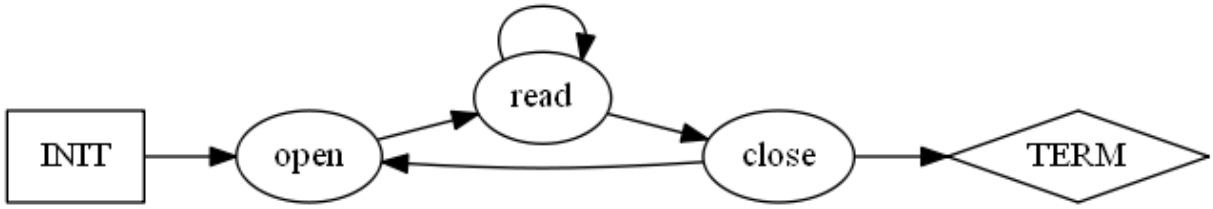

Figure 1: A simple example model used to demonstrate confidence expectation computation.