Bounds on the expected entropy and KL-divergence of sampled multinomial distributions

Brandon C. Roy
bcroy@media.mit.edu

Original: May 18, 2011
Revised: June 6, 2011

Abstract

Information theoretic quantities calculated from a sampled multinomial distribution deviate from the true quantity as a function of the number of samples. In this report, bounds on the *expected* entropy and KL-divergence for a sampled distribution are derived. These bounds can be helpful in understanding the error between the empirical quantity and the true quantity for a distribution.

1 Expected entropy lower bound

Consider a multinomial distribution p with B bins, and estimates of p obtained by sampling. Let p_n be the distribution obtained by taking n samples from p. What is the *expected* entropy of p_n as a function of n?

We should expect that for $n = 1$ samples, all the probability mass will be in one particular bin of p_n and the entropy should be 0. As $n \to \infty$ we expect $p_n \to p$ and $H(p_n) \to H(p)$. This derivation explores this relationship.

We want to know the expected value of the entropy of p_n,

$$\begin{aligned} \mathrm{E}\left[H(p_n)\right] &= -\mathrm{E}\left[\sum_{i=1}^{B} p_{n,i} \log p_{n,i}\right] \\ &= -\sum_{i=1}^{B} \mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \end{aligned} \tag{1}$$

We now consider only the expected value term in (1). The maximum likelihood estimate

of p from n samples is $p_{n,i} = \frac{x_i}{n}$ for all $i = 1 \dots B$. Thus, for a particular bin i we have

$$\begin{aligned} \mathrm{E}\left[p_{n,i} \log p_{n,i}\right] &= \mathrm{E}\left[\frac{x_i}{n} \log \frac{x_i}{n}\right] \\ &= \sum_{k=0}^{n} \Pr(x_i = k) \frac{k}{n} \log \frac{k}{n} \end{aligned} \tag{2}$$

Assuming the samples are iid p, then the expected number of samples in bin i can be calculated as

$$\begin{aligned} &= \sum_{k=0}^{n} \binom{n}{k} p_i^k \left(1 - p_i\right)^{n-k} \frac{k}{n} \log \frac{k}{n} \\ &= \frac{1}{n} \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} \quad p_i^k \left(1 - p_i\right)^{n-k} k \log \frac{k}{n} \\ &= \frac{1}{n} \sum_{k=1}^{n} \frac{n!}{(n-k)!k!} \quad p_i^k \left(1 - p_i\right)^{n-k} k \log \frac{k}{n} \end{aligned}$$

Note that in this equation, p_i is the *true* probability of bin i in p rather than the estimated value. Also note that the last line above results from the fact that when $k = 0$, the summand is zero. Next we simplify $\frac{k}{k!}$, pull an n out of $n!$, and pull a p_i into the front of the sum obtaining

$$= p_i \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} \quad p_i^{k-1} \left(1 - p_i\right)^{n-k} \log \frac{k}{n}$$

Now, let $j = k - 1$, $m = n - 1$ and apply Jensen's inequality for the following derivation:

$$= p_i \sum_{j=0}^{m} \frac{m!}{(m-j)!j!} \quad p_i^j \left(1 - p_i\right)^{m-j} \log \frac{j+1}{m+1}$$

$$= p_i \sum_{j=0}^{m} \Pr(x_i = j) \log \frac{j+1}{m+1} \tag{3}$$

$$\leq p_i \log \sum_{j=0}^{m} \Pr(x_i = j) \frac{j+1}{m+1} \tag{4}$$

$$= p_i \log \frac{mp_i + 1}{m+1} \tag{5}$$

$$= p_i \log \frac{(n-1)p_i + 1}{n} \tag{6}$$

$$= p_i \log \left(p_i + \frac{1 - p_i}{n}\right) \tag{7}$$

We obtain (4) from (3) using Jensen's inequality for the concave function $\log(\cdot)$. Equation (5) is just the expected value of the function $j+1$ for the binomial distribution. Substituting n back in for $m + 1$ yields (6) which simplifies to (7).

Putting all this back together, we have

$$\mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \leq p_i \log \left(p_i + \frac{1 - p_i}{n}\right) \tag{8}$$

$$-\mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \geq -p_i \log \left(p_i + \frac{1 - p_i}{n}\right) \tag{9}$$

Recalling equation (1), and replacing the expectation with our lower bound we have

$$\mathrm{E}\left[H(p_n)\right] = \sum_{i=1}^{B} -\mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \tag{10}$$

$$\geq -\sum_{i=1}^{B} p_i \log \left(p_i + \frac{1 - p_i}{n}\right) \tag{11}$$

Note that intuitively, as $n \to \infty$, $\log \left(p_i + \frac{1-p_i}{n}\right) \to \log p_i$ in equation (11) yielding the true entropy. When $n = 1$, $\log \left(p_i + \frac{1-p_i}{n}\right) = \log(1) = 0$ implying $H(p_1) = 0$ as expected. Moreover, for each i, $-\log \left(p_i + \frac{1-p_i}{n}\right) < -\log p_i$ and therefore contributes a smaller factor to the total entropy, implying that for small n the expected entropy is smaller.

Equation (11) can also be written as

$$\begin{aligned}
&\geq -\sum_{i=1}^{B} p_i \log\left(p_i\left(1+\frac{1-p_i}{np_i}\right)\right) \\
&= -\sum_{i=1}^{B} p_i \log p_i - \sum_{i=1}^{B} p_i \log\left(1+\frac{1-p_i}{np_i}\right) \\
&= H(p) - \sum_{i=1}^{B} p_i \log\left(1+\frac{1-p_i}{np_i}\right) \qquad (12)
\end{aligned}$$

Equation (12) is useful because it isolates the component that varies with n. If $c_i = \frac{1-p_i}{p_i} \leq n$ then $x_i = c_i/n \leq 1$ and the Taylor expansion to $\log(1+x)$ can be applied. The Taylor series of $\log(1+x)$ for $-1 < x \leq 1$ is $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$ and so applying for $x = c_i/n$ we have

$$\log(1+c_i/n) = \frac{c_i}{n} - \frac{c_i^2}{2n^2} + \frac{c_i^3}{3n^3} - \frac{c_i^4}{4n^4} + \ldots$$

Plugging this into the summation in (12) yields the first line below, and in the second line we apply the fact that $p_i c_i^k = (1-p_i)c_i^{k-1}$ to get

$$\begin{aligned}
&= -\sum_{i=1}^{B} p_i \left(\frac{c_i}{n} - \frac{c_i^2}{2n^2} + \frac{c_i^3}{3n^3} - \frac{c_i^4}{4n^4} + \ldots\right) \\
&= -\sum_{i=1}^{B} \frac{1-p_i}{n} + \sum_{i=1}^{B} \frac{(1-p_i)c_i}{2n^2} - \sum_{i=1}^{B} \frac{(1-p_i)c_i^2}{3n^3} \cdots \\
&\approx -\frac{B-1}{n} \qquad (13)
\end{aligned}$$

with the approximation improving as n increases. Also note that this preserves the bound, since including the first odd number of terms of the Taylor series (ie. 1,3,5,... terms) is always greater than the whole series. Therefore, we can write an alternative bound on the expected entropy as

$$\mathrm{E}\left[H(p_n)\right] \geq H(p) - \frac{B-1}{n} \qquad (14)$$

This shows that the lower bound approaches the true entropy with $\frac{1}{n}$, a property that will come into play later for the KL-divergence.

We performed some simulations to obtain the expected entropy as a function of n and compared this to the lower bound obtained using this derivation. Interestingly, this lower bound seems to be a good approximation to $H(p_n)$, and for $n = 1$ and $n \to \infty$ it seems that equality holds. These simulations are shown in figure 1.

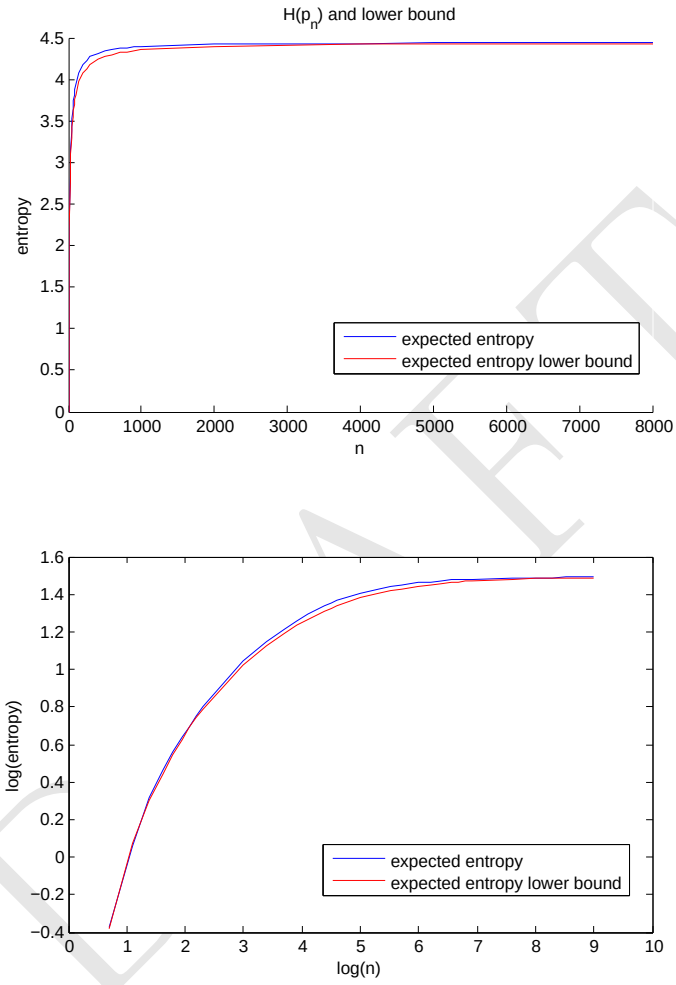

Figure 1: Simulations of the expected entropy for various sample sizes n and the lower bound.

1.1 Expected entropy upper bound

The lower bound of the expected entropy converges to the true entropy of the distribution as $n \to \infty$. This is not surprising, as $p_n \to p$. Nevertheless, we can also find an upper bound on the expected entropy to better understand how it varies with n.

Starting with equation (2), we have

$$\begin{aligned}
\mathrm{E}\left[p_{n,i} \log p_{n,i}\right] &= \sum_{k=0}^{n} \Pr(x_i = k) \frac{k}{n} \log \frac{k}{n} \\
&= \sum_{k=0}^{n} \Pr(x_i = k) \frac{k}{n} \log \frac{k \cdot \Pr(x_i = k)}{n \cdot \Pr(x_i = k)}
\end{aligned}$$

Let $a_k = \Pr(x = k)\frac{k}{n}$ and $b_k = \Pr(x = k)$. Then rewriting, and applying the log-sum inequality yields

$$\begin{aligned}
&= \sum_{k=0}^{n} a_k \log \frac{a_k}{b_k} \\
&\geq \left(\sum_{k=0}^{n} a_k\right) \frac{\sum_{k=0}^{n} a_k}{\sum_{k=0}^{n} b_k} \\
&= p_i \log p_i
\end{aligned}$$

since $\sum_{k=0}^{n} a_k = p$ and $\sum_{k=0}^{n} b_k = 1$. Therefore, $\mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \geq p_i \log p_i$ or equivalently, $-\mathrm{E}\left[p_{n,i} \log p_{n,i}\right] \leq -p_i \log p_i$. Putting this bound on equation (2) back in for equation (1) gives

$$\begin{aligned}
\mathrm{E}\left[H(p_n)\right] &\leq -\sum_{i=1}^{B} p_i \log p_i \\
&= H(p)
\end{aligned}$$

In other words, the expected entropy of the sampled distribution obtained after n samples is upper bounded by the entropy of the true distribution.

It would be nice to find a tighter upper bound on the expected entropy, namely, one that varies with n. One crude way to show that the upper bound increases with n is to find the maximum entropy p_n for each n. The maximum entropy p_n would be one where each sample lands in a new bin, and for $n \leq B$ we have $p_i = \log 1/n$. However, this is not a very satisfying upper bound. It may be more fruitful to focus on probabilistic bounds, that may instead take the variance into account.

2 Expected cross entropy

The cross entropy between q and p, here denoted as $H(q,p) = -\sum_i q_i \log p_i$, can be thought of as the cost in bits of encoding q using a code for p. Suppose we have q_n – the distribution obtained by taking n samples from q. Then what is the expected cross entropy $H(q_n, p)$?

Similar to the expected entropy calculation, we seek

$$\begin{aligned}
\mathrm{E}\left[H(q_n,p)\right] &= -\mathrm{E}\left[\sum_{i=1}^{B} q_{n,i} \log p_i\right] \\
&= -\sum_{i=1}^{B} \mathrm{E}\left[q_{n,i} \log p_i\right] \\
&= -\sum_{i=1}^{B} \mathrm{E}\left[\frac{x_i}{n} \log p_i\right]
\end{aligned}$$

The expected value $E[x_i]$ above is just the expected count for bin i under the true probability distribution q. Thus, $E[x_i] = nq_i$ and

$$\begin{aligned}
-\sum_{i=1}^{B} \frac{1}{n} \log p_i \, \mathrm{E}\left[x_i\right] &= -\sum_{i=1}^{B} q_i \log p_i \\
&= H(q,p) \qquad (15)
\end{aligned}$$

So the expected cross entropy $\mathrm{E}\left[H(q_n,p)\right]$ is just the true cross entropy $H(q,p)$. Note that $H(p,p) = H(p)$, and in the special case where $q_n = p_n$ we have that $\mathrm{E}\left[H(p_n,p)\right] = H(p)$.

3 Expected KL-divergence upper bound

The KL-divergence between distributions q and p is written as $D(q||p) = \sum_i q_i \log \frac{q_i}{p_i}$. This can be rewritten as

$$\begin{aligned}
D(q||p) &= \sum_i q_i \log q_i - \sum_i q_i \log p_i \qquad (16) \\
&= H(q,p) - H(q) \qquad (17)
\end{aligned}$$

For all q and p of the same dimension, $D(q||p) \geq 0$ with equality iff $q = p$. The KL-divergence can be thought of as the additional bits required to encode q using a code for p rather than the code for q. What is the expected KL-divergence of a sampled distribution p_n to the true distribution p? Using the results from the previous sections, we have

$$\begin{aligned}
\mathrm{E}\left[D(p_n||p)\right] &= \mathrm{E}\left[H(p_n, p) - H(p_n)\right] \\
&= \mathrm{E}\left[H(p_n, p)\right] - \mathrm{E}\left[H(p_n)\right] \\
&= H(p) - \mathrm{E}\left[H(p_n)\right] \\
&\leq H(p) + \sum_{i=1}^{B} p_i \log\left(p_i + \frac{1 - p_i}{n}\right) \qquad (18)
\end{aligned}$$

If we instead write the KL-divergence bound using equation (14) we have

$$\begin{aligned}
\mathrm{E}\left[D(p_n||p)\right] &\leq H(p) - H(p) + \frac{B-1}{n} \\
&= \frac{B-1}{n} \qquad (19)
\end{aligned}$$

For comparing a sampled distribution q_n against p, we have

$$\begin{aligned}
\mathrm{E}\left[D(q_n||p)\right] &\leq H(q, p) - H(q) + \frac{B-1}{n} \\
&= D(q||p) + \frac{B-1}{n}
\end{aligned}$$

In other words, for a given number of samples n, we expect the sampled KL-divergence to be within a certain range of the true KL-divergence, depending on n.

We tested this upper bound by taking n samples of p to obtain p_n, computing $D(p_n||p)$, and repeating this many times for each n. The average of $D(p_n||p)$ is an estimate of the expected KL-divergence for n. We then computed the upper bound using equation (18) and plotted this as well. The simulation results are shown in figure 2.

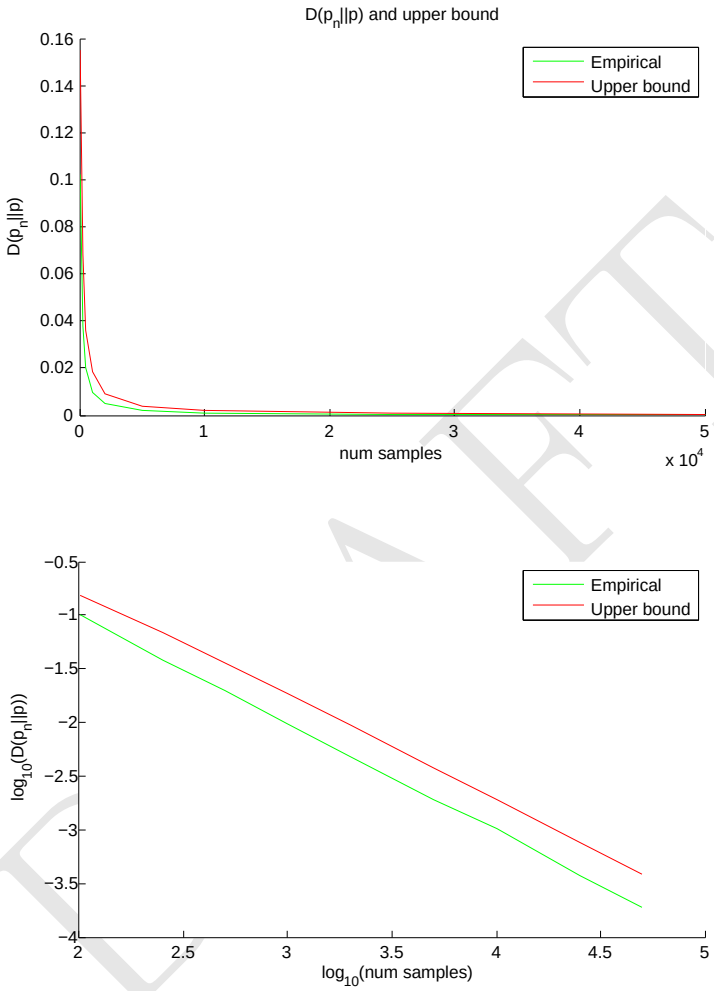

Figure 2: KL-divergence for p_n sampled from p for various n, and the upper bound on the expected value.