Presented at the 38th *Allerton Conference on Communication, Control and Computing*, Monticello, IL, 2000. In *Proceedings of the* 38th *Allerton Conference on Communication, Control and Computing*, UIUC Press: Champaign-Urbana, IL, 2001.

Knowing When to Stop

Brendan J. Frey
Department of Computer Science
University of Waterloo
http://www.cs.toronto.edu/~frey

Abstract

In the last decade, simulation has played an important role in demonstrating the excellent performance of iterative decoders. Although techniques for theoretical analysis are now emerging, even the best theoretical papers include simulation results to justify assumptions made in the analysis. A question that arises when running simulations is: "When should the simulation be stopped?" Although the answer to this question has been well-studied in the statistics literature, coding researchers continue to present simulation results without giving readers a sense of how close the results are to the truth. We report the results of an uncontrolled (or maybe, "out of control") experiment carried out on coding researchers during the presentation of this paper at the Allerton conference. The experiment justifies this paper, which reviews some properties of estimators and some simple methods for computing confidence intervals for simulation results.

1 An uncontrolled experiment on coding researchers

Suppose we simulate M blocks in an attempt to estimate the true probability of word error, p_w, of a coding system. Let $e_1, \ldots, e_M$ be indicator variables for word errors. That is $e_i = 1$ indicates that the ith word was erroneous, whereas $e_i = 0$ indicates that the ith word was error-free. A commonly used estimate of p_w is

$$\bar{p}_w = (1/M) \sum_{i=1}^{M} e_i. \tag{1}$$

During the presentation of this paper at the 2000 Allerton conference, the audience was shown the a series of plots from a simulation, showing $\bar{p}_w$ as a function of M for increasing values of M. After being shown each plot, members of the audience were asked to make rough guesses at the true value p_w.

Figs. 1a to 1f show these plots. After seeing only Fig. 1a, the majority of the participants guessed that $p_w < 1 \times 10^{-3}$. After seeing Fig. 1b, the majority guessed that $1 \times 10^{-3} < p_w < 1.8 \times 10^{-3}$. After seeing Fig. 1c, the majority guessed that

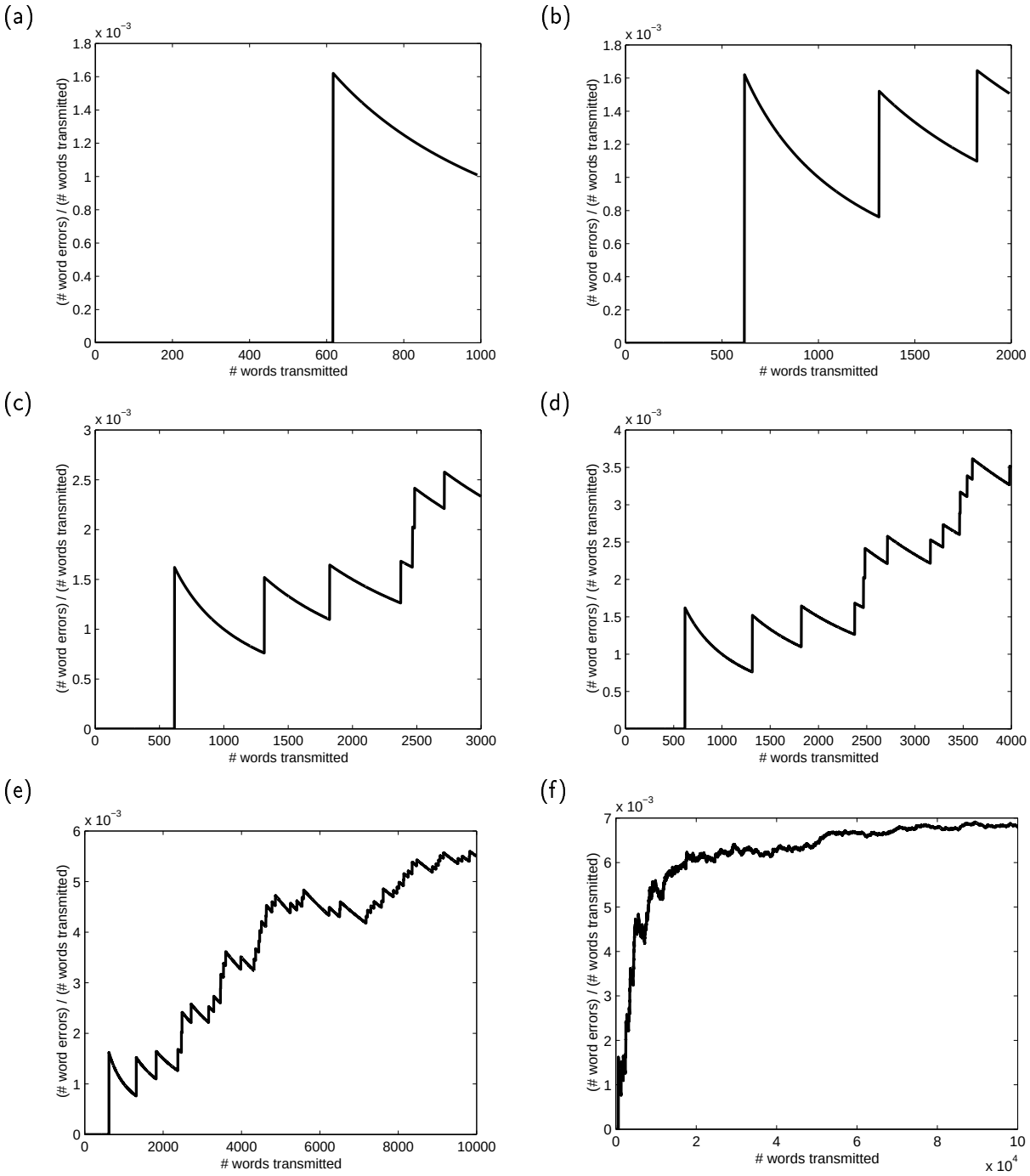

Figure 1: (a) to (f) show the progression of the common estimate of word error rate. In this case, plots (a) to (d) give overly optimistic results. NOTE THE CHANGE IN SCALES.

$1.5\times10^{-3} < p_w < 3\times10^{-3}$. After seeing Fig. 1d, the majority guessed that $p_w < 4\times10^{-3}$. Eventually, it was clear that all of these estimates were overly optimistic and that $p_w \approx 7 \times 10^{-3}$ (see Fig. 1f).

Of course, the audience was partly playing into the game of over-optimism. Nonetheless, results *are* frequently reported for simulations where the number of error events puts the results in the category of Figs. 1a to 1d, as opposed to the preferable category of Fig. 1f.

2 Problems with frequentist estimates

2.1 Bias

When evaluating error rates, we usually use a logarithmic scale. So, instead of having a good estimate of p_w, we actually want a good estimate of $\log p_w$. Usually, $\log \bar{p}_w$ is used as an estimate of $\log p_w$.

A useful question to ask of an estimate is: "Is the estimate overly optimistic or overly pessimistic, on average?" The bias of an estimate is

$$BIAS = \mathrm{E}[estimate] - true\ \ value, \tag{2}$$

where E[] is an expectation over the distribution of estimates produced by different random experiments.

It turns out that $\log \bar{p}_w$ is a *downward* biased estimate of $\log p_w$, which makes it overly optimistic. Using Jensen's inequality, we have

$$BIAS = \mathrm{E}[\log \bar{p}_w] - \log p_w \leq \log \mathrm{E}[(1/M)\sum_{i=1}^{M} e_i] - \log p_w = \log\big((1/M)\sum_{i=1}^{M} p_w\big) - \log p_w = 0. \tag{3}$$

Another common approach is to *wait* until L errors are observed. Suppose we choose L ahead of time and then measure runlengths $r_1, \ldots, r_L$. Runlength j is the number of words transmitted between the last error and error j, *including* the jth erroneous word. This gives the following estimate of p_w:

$$\hat{p}_w = L/(\sum_{j=1}^{L} r_j).$$

Using $\log \hat{p}_w$ as an estimate of $\log p_w$, it turns out that $\log \hat{p}_w$ as an *upward* biased estimate of $\log p_w$, which makes it overly pessimistic. Again, from Jensen's inequality, we have

$$BIAS = \mathrm{E}[-\log\big((\sum_{j=1}^{L} r_j)/L\big)] - \log p_w \geq -\log\big((\sum_{j=1}^{L} \mathrm{E}[r_j])/L\big) - \log p_w = 0.$$

2.2 Variance

The variance of $\bar{p}_w$ is easy to estimate:

$$VAR = \mathrm{VAR}[(1/M)\sum_{i=1}^{M} e_i] = \mathrm{VAR}[e_i]/M = p_w(1-p_w)/M. \approx \bar{p}_w(1-\bar{p}_w)/M. \tag{4}$$

However, how can we use an estimate of the variance to produce a confidence interval for $\bar{p}_w$ or $\log \bar{p}_w$? For Gaussian random variables, the variance directly provides, *e.g.*, 95confidence intervals. However, $\bar{p}_w$ is clearly not Gaussian (since its bounded) and $\log \bar{p}_w$ will not generally be Gaussian.

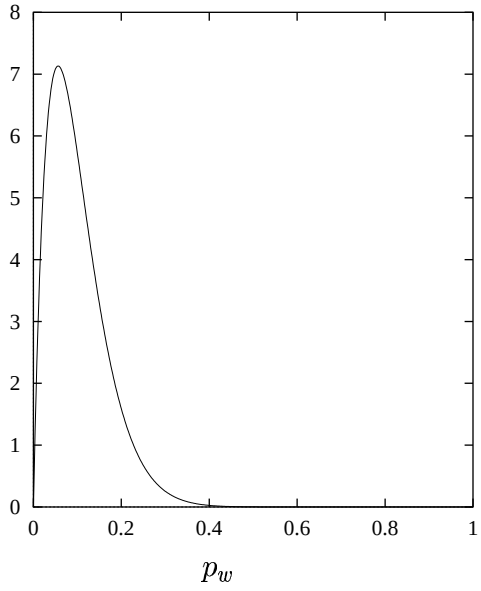

Figure 2: Posterior distribution over the probability of word error p_w, after observing 1 error in 17 words, assuming a uniform prior.

3 Bayesian confidence intervals

Bayesian techniques allow us to specify a prior $p(parameter)$ over the unknown parameter (be it p_w or $\log p_w$) and then include the likelihood $p(data|parameter)$ to obtain the posterior, $p(parameter|data)$. We can then compute whatever confidence intervals we like from the posterior, or just present the posterior as the result.

Beta model for probability of word error

Suppose we observe a total of n_w word errors in M transmitted word. The likelihood is

$$P(n_w|p_w) = p_w^{n_w}(1-p_w)^{M-n_w}. \tag{5}$$

For analytic convenience, we now choose a *conjugate prior*, which has the same form as the likelihood. The conjugate prior for the binomial likelihood is the Beta prior,

$$P(p_w) \propto p_w^{\alpha_w-1}(1-p_w)^{\beta_w-1}. \tag{6}$$

Finally, we obtain the posterior,

$$P(p_w|n_w) \propto p_w^{n_w+\alpha_w-1}(1-p_w)^{M-n_w+\beta_w-1}. \tag{7}$$

For example, suppose $M = 18$ and we observe $n_w = 1$. The frequentist estimate is $\bar{p}_e \approx 0.06$ with $2\sqrt{\mathrm{VAR}} \approx 0.1$. Assuming a uniform prior, $\alpha = \beta = 1$, we obtain the posterior, $P(p_w|n_w) \propto p_w(1-p_w)^{17}$. This distribution is shown in Fig. 2.

From this distribution, we can compute whatever confidence intervals we wish to compute. Notice that the confidence intervals will be within the bounds for p_w.

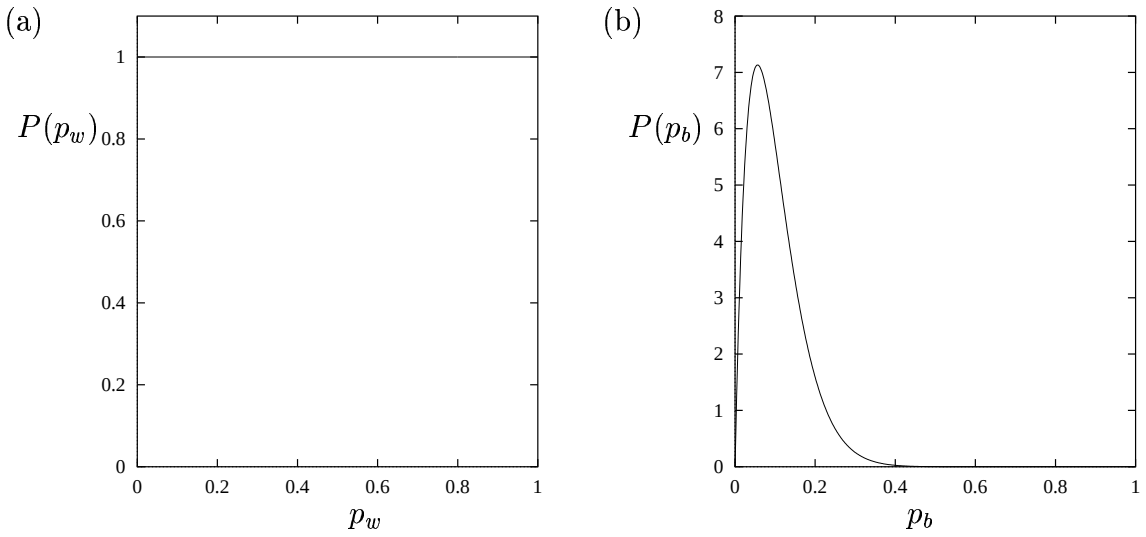

Figure 3: (a) The prior distribution over the probability of word error p_w. (b) The prior distribution over the probability of bit error p_b within erroneous words. This distribution is designed so that its median is equal to the probability of bit error for uncoded transmission.

3.1 Hierarchical Beta model for probability of bit error

In cases where confidence intervals for the probability of bit error are desired, we can use a Beta model for the probability of word error – as described above – and a second Beta model for the probability of bit error within erroneous words.

The error model contains two unknowns: the probability p_w of word error, and the probability p_b of bit error *within erroneous words*. This is a rather crude approximation, since in practice we expect there to be more than one failure mode, *i.e.*, there ought to be several p_b's corresponding to different failure modes.

Let M be the number of words transmitted and let n_w be the number of measured word errors. Let K be the number of information bits per word, and let n_b be the *total* number of bit errors measured while transmitting all M blocks. From the Bayesian perspective, before observing n_w, n_b, we place a prior distribution $P(p_w, p_b)$ on the unknown probabilities. After observing n_w, n_b, we draw conclusions (*e.g.*, compute a confidence interval) from the posterior distribution $P(p_w, p_b | n_w, n_b)$, where

$$P(p_w, p_b | n_w, n_b) \propto P(p_w, p_b) P(n_w, n_b | p_w, p_b). \tag{8}$$

In this equation, the constant of proportionality does not depend on p_w or p_b.

We let p_w and p_b be independent Beta-distributed random variables under the prior: $P(p_w, p_b) = P(p_w)P(p_b)$, where

$$P(p_w) \propto p_w^{\alpha_w - 1}(1 - p_w)^{\beta_w - 1}, \quad \text{and} \quad P(p_b) \propto p_b^{\alpha_b - 1}(1 - p_b)^{\beta_b - 1}. \tag{9}$$

In frequentist terms, α_w and β_w have the effect of shrinking our measurements toward a word error rate of $\alpha_w/(\alpha_w + \beta_w)$, where the influence of this shrinkage grows with $\alpha_w + \beta_w$. Typically, we choose $\alpha_w = \beta_w = 1$, which gives a uniform prior over p_w as shown in Fig. 3a.

As for the prior over p_b, it should be chosen while keeping in mind the behavior of the decoder. If the main source of bit errors is a failure to decode, and if we believe that for failures the decoder will produce a probability of bit error that is roughly equal to the probability p_u of bit error for uncoded transmission, then the prior should place weight on $p_b = p_u$. In this case, we choose $\alpha_b = 2$ and $\beta_b = 1/p_u$, which ensures that the mode of the prior occurs at p_u and that the prior is relatively broad. For example, for $E_b/N_0 = 1$ dB we have $p_u = 0.0563$, and so We choose $\alpha_b = 2$ and $\beta_b = 1/0.0563 = 17.76$, giving the prior distribution for p_b shown in Fig. 3b.

It is straightforward to show that the likelihood is

$$P(n_w, n_b|p_w, p_b) = P(n_w, n_b|p_w, p_b) \propto p_w^{n_w}(1-p_w)^{M-n_w} p_b^{n_b}(1-p_b)^{n_w K - n_b}. \tag{10}$$

This distribution is the product of a binomial distribution for the number of word errors and a binomial distribution for the number of bit errors. Combining this likelihood with the prior, we obtain the posterior,

$$P(p_w, p_b|n_w, n_b) \propto p_w^{\alpha_w - 1 + n_w}(1-p_w)^{\beta_w - 1 + M - n_w} p_b^{\alpha_b - 1 + n_b}(1-p_b)^{\beta_b - 1 + n_w K - n_b}, \tag{11}$$

which is just the product of a Beta distribution over p_w and a separate Beta distribution over p_b.

Of course, we are actually interested in the posterior distribution $P(p_w p_b|n_w, n_b)$ over the total probability of a bit error $p_w p_b$. We use Monte Carlo to obtain a sample from $P(p_w p_b|n_w, n_b)$. First, we draw p_w — p_b pairs from the posterior in (11) and then we take the product of p_w and p_b in each pair. This sample is sorted in ascending order, and the value of $p_w p_b$ occurring half-way through the sorted list is taken as an estimate of the median of $P(p_w p_b|n_w, n_b)$. Similarly, the values of $p_w p_b$ occurring 2.5% and 97.5% through the sorted list are taken as the 95% confidence interval.

In one of our experiments, we simulated the transmission of $M = 332$ blocks using a block length of $N = 131,072$. We measured $n_w = 14$ and $n_b = 34,225$. Using the prior presented above, a sample of 1000 points from the posterior over p_w and p_b was obtained and is shown in Fig. 4a. As described above, for $\gamma = 0.025$, 0.5 and 0.975, we found the values for p_γ such that $\hat{p}(p_w p_b < p_\gamma|n_w, n_b) = \gamma$, where $\hat{p}$ is the sample distribution. The corresponding three curves of the form $p_w p_b = p_\gamma$ are shown in Fig. 4a, and the corresponding values of p_γ give a median of 1.7×10^{-3} and a 95% confidence interval of $(9.9 \times 10^{-4}, 2.6 \times 10^{-3})$.

Clearly, in this case it is the values for p_w that determine the p_γ's for these curves, whereas the values for p_b are well-determined by the measurements. We could have assumed that p_b took on its measured value instead of sampling from the posterior.

In another experiment, we simulated the transmission of $M = 10,216$ blocks. In this case, we measured $n_w = 0$ and $n_b = 0$. Using naive methods, we might conclude that the bit error rate is 0 and that there isn't any variation in this value. A less naive method would be to assume that the *next* block would be in error and pick a value for n_b.

The Bayesian technique gives the sample from the posterior shown in Fig. 4b. In this case, the values of both p_w and p_b play a role in determining the p_γ's for the three curves. The median is 5.1×10^{-6} and the confidence interval is $(1.6 \times 10^{-7}, 4.8 \times 10^{-5})$.

Summary

We illustrated that just "watching" simulations can lead a researcher astray when trying to obtain an accurate estimate of word error rate or bit error rate. Frequentist estimates

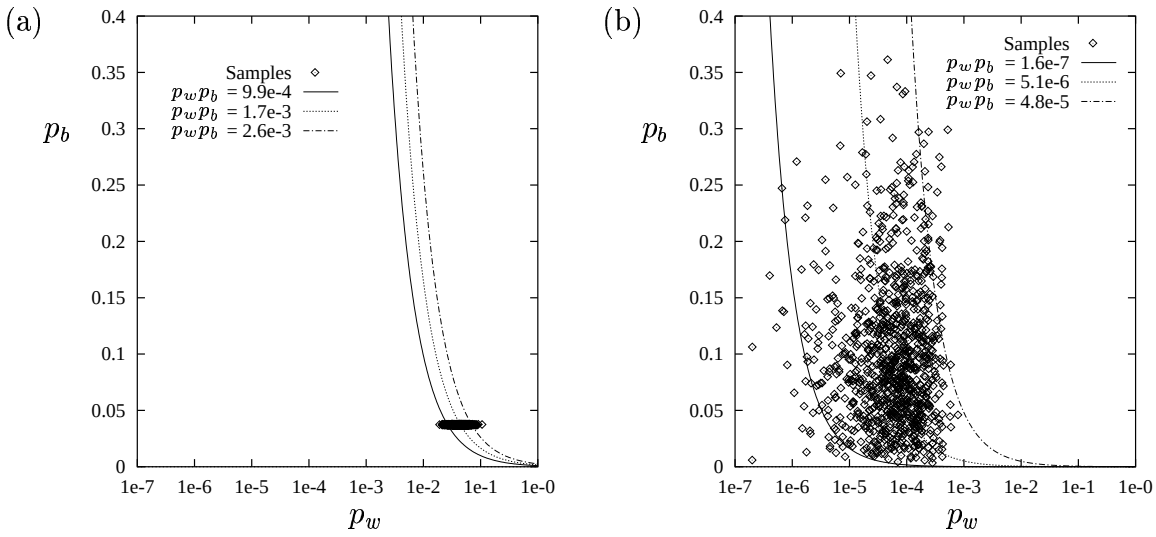

Figure 4: (a) A 1000-point sample from $P(p_w, p_b|n_w, n_b)$ for $M = 332$, $n_w = 14$, $K = 65,536$ and $n_b = 34,225$, for the prior described in the main text. (b) A 1000-point sample from $P(p_w, p_b|n_w, n_b)$ for $M = 10,216$, $n_w = 0$, $K = 65,536$ and $n_b = 0$, for the same prior.

can be biased upward or biased downward and there is little control over this bias. Further, frequentist confidence intervals may not make sense for bounded parameters.

Bayesian techniques combine a user-specified prior over the unknown parameter with the observation likelihood, producing a posterior distribution over the parameter of interest. This posterior can be used to compute a confidence interval. Conjugate priors can be used to simplify analysis, but if there is a good reason to use another type of prior, Monte Carlo methods can be used to obtain a sample from the posterior. Inferences can then be made using the sample.

References

Your undergraduate textbook on probability and statistics. Read it, and put confidence intervals on your plots from now on!