Calculation of the Hessian for the Log-Log Frequency Distribution

Jason D. M. Rennie
jrennie@gmail.com

June 2, 2005*

Abstract

We calculate the diagonal entries of the Hessian for our log-log frequency models.

1 Introduction

We have discussed two parameterizations of the log-log term frequency model: (1) a model with a bias parameter for each word [2, 4], and (2) a model with an exponent parameter for each word [3]. Here we simply calculate additional derivatives for the two parameterizations.

2 Frequency Count Distribution

We calculate diagonal entries of the Hessian for the length conditional log-log term frequency count distribution.

2.1 Bias Per Word

Define a and $\{b_i\}$ as parameters of the distribution. Let x_{ij} be the frequency cound of word i in document j. Let $l_j = \sum_i x_{ij}$ be the length of document j. Assuming independence of word frequencies and documents, and conditioning on the length of each document, the data negative log-likelihood is

$$J = \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ \log \left(\sum_{x=0}^{l_j} (x + b_i)^a \right) - a \log(x_{ij} + b_i) \right\}, \tag{1}$$

*Updated July 6, 2005

The partial derivatives are

$$\frac{\partial J}{\partial a} = \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b_i)^a \log(x+b_i)}{\sum_{x=0}^{l_j} (x+b_i)^a} - \log(x_{ij}+b_i) \right\} \tag{2}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ E_{P_{ij}}[\log(x+b_i)] - E_{\hat{P}_{ij}}[\log(x+b_i)] \right\}, \tag{3}$$

$$\frac{\partial J}{\partial b_i} = \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b_i)^a \frac{a}{x+b_i}}{\sum_{x=0}^{l_j} (x+b_i)^a} - \frac{a}{x_{ij}+b_i} \right\} \tag{4}$$

$$= \sum_{j=1}^{n} \left\{ E_{P_{ij}} \left[\frac{a}{x+b_i} \right] - E_{\hat{P}_{ij}} \left[\frac{a}{x+b_i} \right] \right\}, \tag{5}$$

The diagonal elements of the Hessian are

$$\frac{\partial^2 J}{\partial a \partial a} = \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b_i)^a \log^2(x+b_i)}{\sum_{x=0}^{l_j} (x+b_i)^a} - \left(\frac{\sum_{x=0}^{l_j} (x+b_i)^a \log(x+b_i)}{\sum_{x=0}^{l_j} (x+b_i)^a} \right)^2 \right\} \tag{6}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ E_{P_{ij}} \left[\log^2(x+b_i) \right] - E_{P_{ij}}[\log(x+b_i)]^2 \right\}, \tag{7}$$

$$\frac{\partial^2 J}{\partial b_i \partial b_i} = \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b_i)^a \frac{a(a-1)}{(x+b_i)^2}}{\sum_{x=0}^{l_j} (x+b_i)^a} - \left(\frac{\sum_{x=0}^{l_j} (x+b_i)^a \frac{a}{x+b_i}}{\sum_{x=0}^{l_j} (x+b_i)^a} \right)^2 + \frac{a}{(x_{ij}+b_i)^2} \right\} \tag{8}$$

$$= \sum_{j=1}^{n} \left\{ E_{P_{ij}} \left[\frac{a(a-1)}{(x+b_i)^2} \right] - E_{P_{ij}} \left[\frac{a}{x+b_i} \right]^2 + E_{\hat{P}_{ij}} \left[\frac{a}{(x+b_i)^2} \right] \right\}. \tag{9}$$

Note that the a element is a variance; it is guaranteed to be non-negative. The $\{b_i\}$ elements may be negative if $a < 0$.

2.2 Exponent Per Word

Define $\{a_i\}$ and b as parameters of the distribution. Let x_{ij} be the frequency cound of word i in document j. Let $l_j = \sum_i x_{ij}$ be the length of document j. Assuming independence of word frequencies and documents, and conditioning on the length of each document, the data negative log-likelihood is

$$J = \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ \log \left(\sum_{x=0}^{l_j} (x+b)^{a_i} \right) - a_i \log(x_{ij}+b) \right\}, \tag{10}$$

The partial derivatives are

$$\frac{\partial J}{\partial a_i} = \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b)^{a_i} \log(x+b)}{\sum_{x=0}^{l_j} (x+b)^{a_i}} - \log(x_{ij}+b) \right\} \tag{11}$$

$$= \sum_{j=1}^{n} \left\{ E_{P_{ij}}[\log(x+b)] - E_{\hat{P}_{ij}}[\log(x+b)] \right\}, \tag{12}$$

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b)^{a_i} \frac{a_i}{x+b}}{\sum_{x=0}^{l_j} (x+b)^{a_i}} - \frac{a_i}{x_{ij}+b} \right\} \tag{13}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ E_{P_{ij}}\left[\frac{a_i}{x+b}\right] - E_{\hat{P}_{ij}}\left[\frac{a_i}{x+b}\right] \right\}, \tag{14}$$

The diagonal elements of the Hessian are

$$\frac{\partial^2 J}{\partial a_i \partial a_i} = \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b)^{a_i} \log^2(x+b)}{\sum_{x=0}^{l_j} (x+b)^{a_i}} - \left(\frac{\sum_{x=0}^{l_j} (x+b)^{a_i} \log(x+b)}{\sum_{x=0}^{l_j} (x+b)^{a_i}} \right)^2 \right\} \tag{15}$$

$$= \sum_{j=1}^{n} \left\{ E_{P_{ij}}\left[\log^2(x+b)\right] - E_{P_{ij}}[\log(x+b)]^2 \right\}, \tag{16}$$

$$\frac{\partial^2 J}{\partial b \partial b} = \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ \frac{\sum_{x=0}^{l_j} (x+b)^{a_i} \frac{a_i(a_i-1)}{(x+b)^2}}{\sum_{x=0}^{l_j} (x+b)^{a_i}} - \left(\frac{\sum_{x=0}^{l_j} (x+b)^{a_i} \frac{a_i}{x+b}}{\sum_{x=0}^{l_j} (x+b)^{a_i}} \right)^2 + \frac{a_i}{(x_{ij}+b)^2} \right\} \tag{17}$$

$$= \sum_{i=1}^{d} \sum_{j=1}^{n} \left\{ E_{P_{ij}}\left[\frac{a_i(a_i-1)}{(x+b)^2}\right] - E_{P_{ij}}\left[\frac{a_i}{x+b}\right]^2 + E_{\hat{P}_{ij}}\left[\frac{a_i}{(x+b)^2}\right] \right\}. \tag{18}$$

Note that the $\{a_i\}$ Hessian elements are each a variance; they are each guaranteed to be non-negative. We show in [1] that if b is held fixed, the optimization surface is convex. The b element may be negative if some of the $\{a_i\}$ are negative (which is the usual case).

3 Frequency Rate Distribution

3.1 Bias Per Word

Define $f_i(x,l) = \left(\frac{x}{l} + b_i\right)^a$. Then the data negative log-likelihood for a set of documents is

$$J = \sum_{i,j} \Big(\log[f_i(0,l_j) - f_i(l_j+1,l_j)] - \log[f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)] \Big), \tag{19}$$

where l_j denotes the length of document j and x_{ij} denotes the frequency of word i in document j. First we calculate derivatives of f_i with respect to the parameters:

$$\frac{\partial f_i(x,l)}{\partial a} = f_i(x,l)\log\left(\frac{x}{l}+b_i\right), \tag{20}$$

$$\frac{\partial f_i(x,l)}{\partial b_i} = f_i(x,l)\frac{a}{\frac{x}{l}+b_i}. \tag{21}$$

The objective derivatives are

$$\frac{\partial J}{\partial a} = \sum_{i,j}\left(\frac{\frac{\partial f_i(0,l_j)}{\partial a}-\frac{\partial f_i(l_j+1,l_j)}{\partial a}}{f_i(0,l_j)-f_i(l_j+1,l_j)} - \frac{\frac{\partial f_i(x_{ij},l_j)}{\partial a}-\frac{\partial f_i(x_{ij}+1,l_j)}{\partial a}}{f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)}\right), \tag{22}$$

$$\frac{\partial J}{\partial b_i} = \sum_{j}\left(\frac{\frac{\partial f_i(0,l_j)}{\partial b_i}-\frac{\partial f_i(l_j+1,l_j)}{\partial b_i}}{f_i(0,l_j)-f_i(l_j+1,l_j)} - \frac{\frac{\partial f_i(x_{ij},l_j)}{\partial b_i}-\frac{\partial f_i(x_{ij}+1,l_j)}{\partial b_i}}{f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)}\right). \tag{23}$$

Next, we calculate second derivatives of f_i:

$$\frac{\partial^2 f_i}{\partial a \partial a} = f_i(x,l)\log^2\left(\frac{x}{l}+b_i\right), \tag{24}$$

$$\frac{\partial^2 f_i}{\partial b_i \partial b_i} = f_i(x,l)\frac{a(a-1)}{\left(\frac{x}{l}+b_i\right)^2}. \tag{25}$$

Finally, we calculate the second derivatives of the objective,

$$\begin{aligned}\frac{\partial^2 J}{\partial a \partial a} = \sum_{i,j} & \frac{\frac{\partial^2 f_i(0,l_j)}{\partial a \partial a}-\frac{\partial^2 f_i(l_j+1,l_j)}{\partial a \partial a}}{f_i(0,l_j)-f_i(l_j+1,l_j)} - \frac{\left(\frac{\partial f_i(0,l_j)}{\partial a}-\frac{\partial f_i(l_j+1,l_j)}{\partial a}\right)^2}{\left(f_i(0,l_j)-f_i(l_j+1,l_j)\right)^2} - \\ & \frac{\frac{\partial^2 f_i(x_{ij},l_j)}{\partial a \partial a}-\frac{\partial^2 f_i(x_{ij}+1,l_j)}{\partial a \partial a}}{f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)} + \frac{\left(\frac{\partial f_i(x_{ij},l_j)}{\partial a}-\frac{\partial f_i(x_{ij}+1,l_j)}{\partial a}\right)^2}{\left(f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)\right)^2},\end{aligned} \tag{26}$$

$$\begin{aligned}\frac{\partial^2 J}{\partial b_i \partial b_i} = \sum_{j} & \frac{\frac{\partial^2 f_i(0,l_j)}{\partial b_i \partial b_i}-\frac{\partial^2 f_i(l_j+1,l_j)}{\partial b_i \partial b_i}}{f_i(0,l_j)-f_i(l_j+1,l_j)} - \frac{\left(\frac{\partial f_i(0,l_j)}{\partial b_i}-\frac{\partial f_i(l_j+1,l_j)}{\partial b_i}\right)^2}{\left(f_i(0,l_j)-f_i(l_j+1,l_j)\right)^2} - \\ & \frac{\frac{\partial^2 f_i(x_{ij},l_j)}{\partial b_i \partial b_i}-\frac{\partial^2 f_i(x_{ij}+1,l_j)}{\partial b_i \partial b_i}}{f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)} + \frac{\left(\frac{\partial f_i(x_{ij},l_j)}{\partial b_i}-\frac{\partial f_i(x_{ij}+1,l_j)}{\partial b_i}\right)^2}{\left(f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)\right)^2}.\end{aligned} \tag{27}$$

3.2 Exponent Per Word

Define $f_i(x,l) = \left(\frac{x}{l}+b\right)^{a_i}$. Then the data negative log-likelihood for a set of documents is

$$J = \sum_{i,j}\Big(\log[f_i(0,l_j)-f_i(l_j+1,l_j)] - \log[f_i(x_{ij},l_j)-f_i(x_{ij}+1,l_j)]\Big), \tag{28}$$

where l_j denotes the length of document j and x_{ij} denotes the frequency of word i in document j. First we calculate derivatives of f_i with respect to the parameters:

$$\frac{\partial f_i(x,l)}{\partial a_i} = f_i(x,l)\log\left(\frac{x}{l}+b\right), \tag{29}$$

$$\frac{\partial f_i(x,l)}{\partial b} = f_i(x,l)\frac{a_i}{\frac{x}{l}+b}. \tag{30}$$

The objective derivatives are

$$\frac{\partial J}{\partial a_i} = \sum_j \left(\frac{\frac{\partial f_i(0,l_j)}{\partial a_i} - \frac{\partial f_i(l_j+1,l_j)}{\partial a_i}}{f_i(0,l_j) - f_i(l_j+1,l_j)} - \frac{\frac{\partial f_i(x_{ij},l_j)}{\partial a_i} - \frac{\partial f_i(x_{ij}+1,l_j)}{\partial a_i}}{f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)} \right), \tag{31}$$

$$\frac{\partial J}{\partial b} = \sum_{i,j} \left(\frac{\frac{\partial f_i(0,l_j)}{\partial b} - \frac{\partial f_i(l_j+1,l_j)}{\partial b}}{f_i(0,l_j) - f_i(l_j+1,l_j)} - \frac{\frac{\partial f_i(x_{ij},l_j)}{\partial b} - \frac{\partial f_i(x_{ij}+1,l_j)}{\partial b}}{f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)} \right). \tag{32}$$

Next, we calculate second derivatives of f_i:

$$\frac{\partial^2 f_i}{\partial a_i \partial a_i} = f_i(x,l)\log^2\left(\frac{x}{l}+b\right), \tag{33}$$

$$\frac{\partial^2 f_i}{\partial b \partial b} = f_i(x,l)\frac{a_i(a_i-1)}{\left(\frac{x}{l}+b\right)^2}. \tag{34}$$

Finally, we calculate the second derivatives of the objective,

$$\begin{aligned}\frac{\partial^2 J}{\partial a_i \partial a_i} = \sum_j & \frac{\frac{\partial^2 f_i(0,l_j)}{\partial a_i \partial a_i} - \frac{\partial^2 f_i(l_j+1,l_j)}{\partial a_i \partial a_i}}{f_i(0,l_j) - f_i(l_j+1,l_j)} - \frac{\left(\frac{\partial f_i(0,l_j)}{\partial a_i} - \frac{\partial f_i(l_j+1,l_j)}{\partial a_i}\right)^2}{\left(f_i(0,l_j) - f_i(l_j+1,l_j)\right)^2} - \\ & \frac{\frac{\partial^2 f_i(x_{ij},l_j)}{\partial a_i \partial a_i} - \frac{\partial^2 f_i(x_{ij}+1,l_j)}{\partial a_i \partial a_i}}{f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)} + \frac{\left(\frac{\partial f_i(x_{ij},l_j)}{\partial a_i} - \frac{\partial f_i(x_{ij}+1,l_j)}{\partial a_i}\right)^2}{\left(f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)\right)^2},\end{aligned} \tag{35}$$

$$\begin{aligned}\frac{\partial^2 J}{\partial b \partial b} = \sum_{i,j} & \frac{\frac{\partial^2 f_i(0,l_j)}{\partial b \partial b} - \frac{\partial^2 f_i(l_j+1,l_j)}{\partial b \partial b}}{f_i(0,l_j) - f_i(l_j+1,l_j)} - \frac{\left(\frac{\partial f_i(0,l_j)}{\partial b} - \frac{\partial f_i(l_j+1,l_j)}{\partial b}\right)^2}{\left(f_i(0,l_j) - f_i(l_j+1,l_j)\right)^2} - \\ & \frac{\frac{\partial^2 f_i(x_{ij},l_j)}{\partial b \partial b} - \frac{\partial^2 f_i(x_{ij}+1,l_j)}{\partial b \partial b}}{f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)} + \frac{\left(\frac{\partial f_i(x_{ij},l_j)}{\partial b} - \frac{\partial f_i(x_{ij}+1,l_j)}{\partial b}\right)^2}{\left(f_i(x_{ij},l_j) - f_i(x_{ij}+1,l_j)\right)^2}.\end{aligned} \tag{36}$$

References

[1] J. D. M. Rennie. A class of convex functions. http://people.csail.mit.edu/~jrennie/writing, May 2005.

[2] J. D. M. Rennie. Learning a log-log term frequency model. http://people.csail.mit.edu/∼jrennie/writing, May 2005.

[3] J. D. M. Rennie. A role-reversal in the log-log model. http://people.csail.mit.edu/∼jrennie/writing, May 2005.

[4] J. D. M. Rennie. Using a log-log distribution to model term frequency rates. http://people.csail.mit.edu/∼jrennie/writing, May 2005.