RC24969 (W1004-013) April 5, 2010
Other

IBM Research Report

Asymptotic Distribution for Sequence Alignment Scores

Daniel E. Platt
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

Asymptotic distribution for sequence alignment scores.

Daniel E. Platt

IBM T. J. Watson Research Center

Yorktown Hgts, NY

Abstract – Alignment score distributions have become very important for computing p-values of alignments built on database searches based on adjustable measures of similarity. Such scores provide a level of flexibility allowing for different measures of similarity, as well as for seeking candidates for RNA binding sites. However, computation of the probability distribution over its full range has proven complex, with a great deal of interest focusing on the scaling dependence of probabilities on alignment sequence lengths. This paper presents a distribution function derived from first principles that describes the sequence-length dependence and extreme-value behavior of an arbitrary matrix of alignment scores given their alignment species frequencies. The derivation is based on the asymptotic Sterling's formula applied in the continuum limit. The results are compared to simulation results for alignment sequence lengths ranging from 10 to 10,000, with observed binned frequencies down to the order of 10^{-7}, and seen to produce good approximations for probabilities of practical sequence alignment lengths. This distribution function provides a practical and fairly easy to compute baseline against which the behavior of real and simulated data may be measured.

Background

Sequence alignment algorithms have made it possible to identify and measure meaningful comparisons between protein or nucleotide sequences, as well as enabling wide-ranging database searching[1] which has ultimately driven the usefulness and demand for the large databases, and the laboratory technology to produce the data. Among these alignment techniques are scoring-matrix based algorithms, such as BLAST.[2] Such scoring-matrix based algorithms permit the interchange of scoring functions to handle different types of data, such as genetic vs. protein, as well as different features of scoring matrices, such as BLOSUM[3] or PAM[4] matrices, or of combinations of such methods.[5] Further, alignments may easily be formed to identify RNA binding candidates, for example in application to miRNA candidate searches, by appropriately adjusting the scoring matrices to represent binding energies.[6]

Assignment of probabilities to scores representing alignment candidates has been developed[7-9] showing that a Gumbel extreme value distribution is expected to describe this type of distribution. However, computational methods to estimate the Gumbel prefactor and scale factor have driven significant efforts to find faster or easier methods of computation for the better part of two decades.[10-13] Even then, there is some evidence and argument that the Gumbel distribution may not be entirely the best description of the probability distribution governing the probabilities of the highest scores.[14-16] One feature of these scoring approaches is the use of "gap penalties" to penalize the proliferation of gaps that searching algorithms would otherwise generate. However, assignment of probabilities to scores carrying these gap penalties highlights the fact that these probabilities measure the chances that the search algorithm would find such alignments among random sequences by chance.

An interesting feature of these approaches is the dependence, or independence, of the scoring distribution on the alignment length L.[16] One interesting path followed involves exploration of the alignment as a percolation transition.[17, 18] Heretofore, analytical expressions that capture the dependence of the entire distribution on alignment sequence length, even in restricted conditions or limiting cases, has not been obtained from first principles.

This paper presents a closed-form expression for the continuum asymptotic approximation to the distribution of scores at larger sequence lengths without gap penalties or indels, with the concomitant issues of excluding ambiguous combinations (multiple ways to represent the same alignment). Results are compared to Monte-Carlo simulations of randomly generated sequence alignments over a range of length scales, and with a Markov Chain Monte Carlo (MCMC) simulation of multinomially weighted alignment scores in the rare-event regime. In the first case, the test made no assumption about the multinomial character of alignment probabilities, while the second tested the ability of the asymptotic approximation to represent a distribution of scores generated by an MCMC simulation of multinomially weighted scores. Connections are developed relating these results to Carlin-Altschul distribution functions. Sequence-length invariant scaling forms are derived. Extreme-value distribution functions are derived.

Methods

Consider alignments between a query string and entries in a database. The entries in the database can be considered as all of the possible starting locations in each of the strings in that database. Characters are limited to an alphabet A (e.g., for nucleotides $A = \{"A","C","G","T"\}$). A sequence of length L is an ordered L-tuple of members of the

alphabet, so is a member of of an L-fold Cartesian product of the alphabet $A \otimes A \otimes A \cdots \otimes A$. So an alignment is composed of L-tuples of pairs of characters from the Alphabet $(A \otimes A) \otimes (A \otimes A) \otimes \cdots \otimes (A \otimes A)$. An example alignment may be

```
…GCAATAAACTGAAAATGTTTAGACGGGCTCACATCACCCCATAGACAAAT…
…GCAATACACTGAAAATGTTTAGACGAGCTCATATCACCTCATAAACAAAT…
```

In considering random sequence alignments, each of these alignment cases may be constructed by selecting alphabet members with typical probabilities p_j according to an index of alignment types k (e.g. for genetic sequence alignments, $k \in \{(AA),(CT),\ldots\}$). The score for any specific alignment for sites l is $s = \sum_l s_{k_l}$. This ultimately depends only on the number of each species of pairs, and is not sensitive to order. In the simple case of no indels, the probability of finding an alignment $k = (i,j)$ of a pair of characters $i, j \in A$ by chance is $p_{(i,j)} = f_i g_j$ where the frequencies f_i and g_j are the relative frequencies of the genetic or peptide species i and j in the alphabet.

The number of times each of these alignment pairs will have been constructed out of some L-tuple are n_k with the total number of alignments constrained to sum to the length of the alignment $L = \sum_k n_k$. Associated with each alignment type j is a score s_j. The total score is then a random variable $S = \sum_j n_j s_j$. Therefore, the scores depend only on the counts n_k, and do not depend on the order they were drawn. If there are no correlations among the sequence members, drawing alignments by chance are independent, and the probability of seeing any combination of n_k is the same as if L pairs were drawn with no sensitivity to sequence. This process can therefore be described by a multinomial distribution. The probability of observing any specific configuration of alignments is described by

$$p\left(\{n_j\}\right) = \frac{L!}{\prod_j n_j!} \prod_j p_j^{n_j}.$$

The problem to be solved is then to compute the p.d.f. for S. The probability density function for S may be computed

$$f(s)ds = E\left(I(s \le S \le s + ds)\right) = \sum_{\{n_j\}} I(s \le S \le s + ds) \frac{L!}{\prod_j n_j!} \prod_j p_j^{n_j}$$

given the indicator function $I(\cdot)$.

Before evaluating this function, the behavior under convolution may be derived. Consider l_j such that $\sum_j l_j = L_1$, and m_j such that $\sum_j m_j = L_2$ each distributed with the same corresponding p_j's Then the generating function for these variables is $E\left(e^{\sum_j l_j t_j}\right) = \left(\sum_j p_j e^{t_j}\right)^{L_1}$ and likewise for m_j. The generating function for $n_j = l_j + m_j$ is $E\left(e^{\sum_j (l_j + m_j) t_j}\right) = E\left(e^{\sum_j l_j t_j}\right) E\left(e^{\sum_j m_j t_j}\right) = \left(\sum_j p_j e^{t_j}\right)^{L_1 + L_2}$, so the convolution of the distributions for two lengths L_1 and L_2 is just the distribution corresponding to $L_1 + L_2$. Defining $S_1 = \sum_j s_j l_j$ and $S_2 = \sum_j s_j m_j$, the variable $S = S_1 + S_2$ will also satisfy the same convolution through its linear dependence on the $n_j = l_j + m_j$. Algorithms based on accumulating matching sequences may therefore be expected to be describable as renewal processes.

Evaluation of the distribution function

Sterling's approximation may be applied to the factorials in the multinomial distribution, and the discrete sum may be approximated by integrals in the continuum limit (larger L) to yield

$$f(s)ds = \int d\{n_j\} \delta\left(s - \sum_j n_j s_j\right) ds \, \delta\left(L - \sum_j n_j\right) \frac{\sqrt{2\pi L} L^L e^{-L}}{\prod_j \sqrt{2\pi n_j}} \exp\left(\sum_j \left(n_j \ln p_j - n_j \ln n_j + n_j\right)\right),$$

where the two delta functions represent the indicator function and impose the $L = \sum_j n_j$.

The term in the exponential may be expanded about the extremum subject to the constraints imposed by the δ functions. The arguments involving identification of the extremum are very similar to those describing the grand canonical ensemble in elementary statistical mechanics. [19] First, consider the expansion $n_j = N_j + \delta n_j$. Then

$$\sum_j \left(n_j \ln p_j - n_j \ln n_j + n_j\right) = \sum_j \left(N_j \ln p_j - N_j \ln N_j + N_j\right) + \sum_j \left(\ln p_j - \ln N_j\right) \delta n_j - \sum_j \frac{\delta n_j^2}{2N_j} + \sum_j \frac{\delta n_j^3}{2N_j^2} + \ldots$$

Imposing $s = \sum_j N_j s_j$, $L = \sum_j N_j$, $\sum_j \delta n_j s_j = 0$, and $\sum_j \delta n_j = 0$, if $\ln p_j - \ln N_j = \alpha + \beta s_j$,

as long as the δn_j's satisfy their constraints, then the linear combinations involving the α's and β's will be automatically satisfied, and those constraints can be guaranteed by finding the N_j's, α's and β's that match the constraints. In this case, $N_j = p_j e^{-\alpha-\beta s_j}$.

Then $L = \sum_j N_j = \sum_j p_j e^{-\alpha-\beta s_j} = e^{-\alpha}\sum_j p_j e^{-\beta s_j}$, $e^{-\alpha} = \dfrac{L}{\sum_j p_j e^{-\beta s_j}}$, $N_j = \dfrac{L p_j e^{-\beta s_j}}{\sum_k p_k e^{-\beta s_k}}$, and

$s = \sum_j N_j s_j = \dfrac{L\sum_j p_j s_j e^{-\beta s_j}}{\sum_j p_j e^{-\beta s_j}}$. Define $Z(\beta) = \sum_j p_j e^{-\beta s_j}$. Then $s = -L\dfrac{d\ln Z(\beta)}{d\beta}$.

Next, $\sum_j \left(N_j \ln p_j - N_j \ln N_j + N_j\right) = L - L\ln L + L\dfrac{d\left(\beta^{-1}\ln Z\right)}{d(1/\beta)}$. If $\delta n_j^2 / N_j = O(1)$, the cubic term is contributing $N_j^{3/2}/N_j^2 = O\left(L^{-1/2}\right)$, and when the δn_j are large enough that the cubic term is of order 1, then the quadratic term is $N_j^{4/3}/N_j = O\left(L^{1/3}\right)$. Contributions to the integral outside of this region are negligible. The sequence in the exponent may be replaced by the truncated series terminating with the quadratic term. The integral then reduces to

$$f(s)ds = \frac{\sqrt{2\pi L}\exp\left(L\dfrac{d\left(\beta^{-1}\ln Z\right)}{d(1/\beta)}\right)}{\prod_j \sqrt{2\pi N_j}} ds \int d\{\delta n_j\}\delta\left(-\sum_j \delta n_j s_j\right)\delta\left(-\sum_j \delta n_j\right)\exp\left(-\sum_j \frac{\delta n_j^2}{2N_j}\right)$$

$$= \frac{1}{(2\pi)^2}\frac{\sqrt{2\pi L}\exp\left(L\dfrac{d\left(\beta^{-1}\ln Z\right)}{d(1/\beta)}\right)}{\prod_j \sqrt{2\pi N_j}} ds$$

$$\cdot \int d\omega_1 \int d\omega_2 \int d\{\delta n_j\}\exp\left(-\sum_j\left(\frac{\delta n_j^2}{2N_j} + i\omega_1\delta n_j + i\omega_2 s_j \delta n_j\right)\right)$$

using Fourier integral representations of the δ-function. Then the integral reduces to

$f(s)ds = \sqrt{\dfrac{L}{2\pi\sigma_s^2}}\exp\left(L\dfrac{d\left(\beta^{-1}\ln Z\right)}{d(1/\beta)}\right)ds$, where $\sigma_s^2 = \dfrac{L^2}{Z(\beta)}\sum_j s_j^2 p_j e^{-\beta s_j} - s^2 = L^2\dfrac{d^2\ln Z(\beta)}{d\beta^2}$.

To summarize:

$$Z(\beta) = \sum_j p_j e^{-\beta s_j}$$

$$N_j = \frac{L p_j e^{-\beta s_j}}{Z(\beta)}$$

$$s = -L \frac{d \ln Z(\beta)}{d\beta}$$

$$\sigma_s^2 = L^2 \frac{d^2 \ln Z}{d\beta^2}$$

$$f(s)ds = \sqrt{\frac{L}{2\pi\sigma_s^2}} \exp\left(L \frac{d\left(\beta^{-1} \ln Z\right)}{d(1/\beta)} \right) ds = \sqrt{\frac{L}{2\pi\sigma_s^2}} Z^L e^{\beta s} ds$$

Comparison with Monte-Carlo simulations

A distribution for a simple model with four characters, frequencies $p_j = (0.22, 0.22, 0.28, 0.28)$ and scores $s_{ij} = \begin{pmatrix} 0.8 & -1.0 & -1.0 & -1.5 \\ -1.0 & 1.0 & -1.0 & -1.5 \\ -1.0 & -1.0 & 1.2 & -1.5 \\ -1.5 & -1.5 & -1.5 & 0.5 \end{pmatrix}$ was computed from the formula and by simulation. Scores were sampled 5,000,000 times, for random strings of lengths 10, 21, 100, 1000, and 10,000 using simple direct Monte-Carlo sampling. The random number generator employed was the Tausworthe generator[20], employing the improved seeding algorithm[21] as implemented in the Gnu Scientific Library (GSL)[22] (http://www.gnu.org/software/gsl/). The theoretical distribution function was computed at the midpoint of each bin, with ds representing the bin-width in comparing $f(s)ds$ with the sampled frequencies. The expected variance due to sampling from the multinomial distribution is expected to be $Nq_l(1-q_l) \approx Nq_l$ given probability q_l that the randomly generated sample would land in bin l. Therefore the expected standard error would roughly scale as the square root of the counts in each bin. The sample error bars are too narrow at 5,000,000 samples to plot. Figure 1 shows normal and logarithmic scales for $f(s)ds$ for each of L = 10, 21, 100, 1000, and 10,000. Cutoff for the theoretical curve occurred where computation was outside of the range of exponents for double precision floating point representation. Cutoffs on the log-scales occurred where no samples were generated out of the 5,000,000 samples computed.

Rare event probabilities were estimated for L=100 by performing a Metropolis-Hastings MCMC simulation with importance sampling. The importance weighting function employed was

a simple exponential in the score s. Given an array of occupancies for alignment species, a transition of one item from n_k to $n_{k'}$ derived from the multinomial distribution was $\frac{n_k p_{k'} e^{\lambda \Delta s}}{(n_{k'}+1)p_k}$, where λ is the importance sampling weighting parameter, and the values of the n's are those prior to the actual movement of the count from n_k to $n_{k'}$. 5×10^8 samples were collected. Values of λ ranged from 0.1287, 0.1545, 0.1802, 0.1931, and 0.2060. This produced results for probabilities within the bins spanning a range roughly of 10^{-5} to 10^{-13}. The selected initial values of $N_{ij} = \begin{pmatrix} 9 & 4 & 5 & 4 \\ 4 & 10 & 5 & 4 \\ 5 & 5 & 17 & 5 \\ 4 & 4 & 5 & 12 \end{pmatrix}$ were derived from an initial state corresponding to $\beta = -0.1287$. Figure 2 shows the comparison of the asymptotic approximation to MCMC with importance sampling where $L = 100$.

Figure 3 shows the value of β describing the distribution for L=100. The primary range representing most of the realized alignment scores shows smooth and relatively slow variation compared to the endpoints representing the extreme maximum and minimum alignment scores that can be realized.

Results

The plots show that the distribution function correctly captures the scaling of the simulations over 3 orders of magnitude, from L = 10 to 10,000. While the number of Monte-Carlo samples collected reduces the error expected in the frequencies due to stochastic variations, the bin sizes also provides some deviation between the continuum estimate at midpoint and the total collected within the bin where the distribution function varies rapidly within the bin. The samples collected probe the distribution over orders of magnitude below 10^{-6}, testing that the accuracy of the expression to ranges expected for searching databases containing of the order of 10^6 entries. Further, the shape of the distribution shifts over changes in L, which the multinomial asymptotic approximation captures.

The distribution shows some deviations at small L = 10 in part due to the discrete character of the score sums. It is expected that this minimum number would be larger with a more complicated alphabet (e.g. peptides rather than nucleotides). The approximation is essentially based on Sterling's approximation, and replacing discrete sums with integrals over the

continuum. The most important contributors to the distribution will tend to be those with the larger counts, which is where Sterling's approximation applies most accurately. So the distribution approximation describes the data surprisingly well down to scales as small as L = 10. The approximation is much better at L = 21, typical of miRNA alignments.

The function $\beta = \beta(s/L)$ is essentially independent of L. The plot for L = 100 for the model computed here is shown in Figure 3. This shows that β varies smoothly and with relatively low variation over much of the range where scores are likely to be observed, with significant swings in value near the end-points of the distribution where almost all the alignments approach extreme score values. This result suggests that simple expansions about a score of interest may be applied to explore the distribution of results in that immediate vicinity of that score, which would be appropriate for the best score results in a database search for which extreme-value distribution behavior is of interest.

Discussion

The results show that the asymptotic distribution presented above adequately describes the form of the distribution over a wide range of length scale, and down to the rare-event regime. As such, analytical behavior of the distribution may be expected to be informative in application to understanding the Karlin-Altschul results[2, 7-10], as well as conditions where approximate agreement may be expected.

Distribution in intermediate ranges

At more intermediate ranges, where probabilities would be more typical of random sequences, but for larger scores where $\beta < 0$, the behavior of the exponent satisfies $\frac{d}{ds}\left(L\ln Z(\beta) + \beta s\right) = \beta$, and $\frac{d^2}{ds^2}\left(L\ln Z(\beta) + \beta s\right) = \frac{d\beta}{ds} = -\frac{L}{\sigma_s^2}$, so that

$$L\ln Z(\beta) + \beta s = L\ln Z(\beta_0) + \beta s_0 + \beta_0 \cdot (s - s_0) - \frac{L}{2\sigma_s^2}(s - s_0)^2 + \ldots \text{ If } \beta \cdot (s - s_0) \approx 1,$$

then $\frac{L}{2\sigma_s^2}(s - s_0)^2 \approx \frac{L}{2\beta^2\sigma_s^2} = O\left(\frac{1}{L}\right)$. Also, the variation of $\sigma_s^2(\beta)$ in the denominator will depend on $\frac{d\sigma_s^2(\beta)}{ds}\frac{1}{\beta} = O(L)$ compared to $\sigma_s^2(\beta) = O(L^2)$. Then

$$P(s \le S \le s+ds) = \sqrt{\frac{L}{2\pi\sigma_s^2(\beta_0)}} e^{L\ln Z_0 + \beta_0 s_0 + \beta_0(s-s_0)} ds,$$

so that

$$P(S \ge s_0) = \int_{s_0}^{\infty} \sqrt{\frac{L}{2\pi\sigma_s^2}} e^{L\ln Z + \beta s} ds \approx -\frac{1}{\beta_0} \sqrt{\frac{L}{2\pi\sigma_s^2(\beta_0)}} e^{L\ln Z(\beta_0) + \beta_0 s_0},$$

recalling that $\beta_0 < 0$.

Invariance of the distribution under displacements of the scoring function

Define $s_j' = s_j - \bar{s}$ and $Z'(\beta) = \sum_j p_j e^{-\beta s_j'} = e^{\beta\bar{s}} \sum_j p_j e^{-\beta s_j} = e^{\beta\bar{s}} Z(\beta)$ with the same β.

Then $s' = -L\frac{d\ln Z'}{d\beta} = -L\bar{s} + s$. Under this transformation, $L\ln Z' + \beta s' = L\ln Z + \beta s$, and $\sigma_s'^2 = L^2 \frac{d^2 \ln Z'}{d\beta^2} = L^2 \frac{d^2 \ln Z}{d\beta^2} = \sigma_s^2$. It is always possible to choose an $\bar{s}$ such that $\ln Z' = \beta\bar{s} + \ln Z = 0$, which occurs at $\bar{s} = -\frac{\ln Z}{\beta}$. In this case, it is clear that $L\ln Z + \beta s = L\ln Z' + \beta s' = \beta s' < 0$ as long as $s' > 0$. Essentially, the structure of the spectrum of scores is arbitrary. Further, it is always possible to transform the scores so that $Z(\beta) = 1$ at the region of scores being explored. Yet, the distribution function and β parameter are invariant under these transformations.

Length dependence

The distribution function $f(s)ds = \sqrt{\frac{L}{2\pi\sigma_s^2}} Z^L e^{\beta s} ds$ suggests that the distribution becomes insensitive to variations in L when $Z(\beta) = 1$. This occurs at some $\beta = \beta_0$, at which point the distribution appears to satisfy an exponential distribution $f(s)ds = \sqrt{\frac{L}{2\pi\sigma_s^2}} e^{\beta_0 s} ds$.

There are a number of points to note, most of which have been indicated before. The alignment scores s_j must satisfy some specific conditions in order to guarantee that a root for $Z(\beta) = 1$ exists. Further, that root is not invariant under transformations since a transformation of

alignment scores $s_j' = s_j - \bar{s}$ produces a new $Z'(\beta) = e^{\beta\bar{s}} Z(\beta)$ so that the root $Z(\beta_0) = 1$ corresponds to $Z'(\beta_0) = e^{\beta_0 \bar{s}}$. The greatest difficulty is that the exponential expansion approximates the distribution only for a width much smaller than $\sigma_s(\beta_0) L^{-1/2} \propto L^{1/2}$ (the effective size of the region where the quadratic contribution is less than 1), and the score in this region is $s = -L \dfrac{d \ln Z(\beta)}{d\beta} \propto L$. The region where the distribution is relatively insensitive to L is itself L dependent, though there is some tolerance within that region. Further, any particular sample of alignments may themselves not be contained in a region characterized by this particular β_0.

More generally, the distribution function depends on L in several positions. However, the simplest parts involve an L in the exponent. Expansion about the extremum, which occurs at $\beta = 0$ suggests a first order approximation to the width proportional to $L^{-1/2}$. However, this must be propagated back through the $\dfrac{d(\beta^{-1} \ln Z)}{d(1/\beta)}$ in the exponent to determine exactly how the width in $\hat{s}$ scales. The scaling of the fall-off in the tails is proportional to L^{-1}.

An alternative probe to the randomness of relatively improbable scores still far from the endpoints is to note that

$$P\left(\hat{S}_L \geq \hat{s} \middle| \hat{S}_L \geq \hat{s}_0\right) = e^{\beta(\hat{s}_0)(s - s_0)} = e^{L\beta(s_0)(\hat{s} - \hat{s}_0)},$$

for s_0 near the set of best scores retrieved, which gives a view of what would be expected by random scores larger than a threshold. In this case, it may be possible to determine if the distribution of scores reflects random substitutions within that region of scores. Then it may be expected that

$$\left[P\left(\hat{S}_L \geq \hat{s} \middle| \hat{S}_L \geq \hat{s}_0\right)\right]^{1/L} = e^{\beta(\hat{s}_0)(\hat{s} - \hat{s}_0)}$$

will be an L independent measure of how scores from multiple alignment lengths may behave.

Extreme value distribution in intermediate ranges

Given the above results considering fixed L, defining $S_N = \max\{S_j\}$ for N samples of iid S_j's, it follows that $P(S_N \leq s_0) = \left[P(S \leq s_0)\right]^N \approx \exp\left(-N \dfrac{-1}{\beta_0} \sqrt{\dfrac{L}{2\pi\sigma_s^2(s_0)}} e^{L \ln Z_0 + \beta_0 s_0}\right)$.

Given results from a query of length m and a database with n loci in it, it may be expected that $N \approx mn$. If the quadratic contribution to the expansion of $L\ln Z + \beta s$ satisfies

$$\frac{L}{2\sigma_s^2}(s-s_0)^2 \Big/ \left[\beta_0(s-s_0)\right] = \frac{L(s-s_0)}{2\beta_0\sigma_s^2} \approx \frac{L\ln N}{2\beta_0^2\sigma_s^2} << 1,$$

then $P(S_N \le s_0) = \exp\left(-\frac{-1}{\beta_0}\sqrt{\frac{L}{2\pi\sigma_s^2\left(s_0+\frac{\ln N}{\beta_0}\right)}}e^{L\ln Z_0+\beta_0\left(s_0+\frac{\ln N}{\beta_0}\right)}\right)$, which corresponds to a Karlin-Altschul -- like result, though with an L dependence, with parameter $K^* \approx \frac{-1}{\beta_0}\sqrt{\frac{L}{2\pi\sigma_s^2}}e^{L\ln Z_0}$. Note this does not apply to scores with gap penalties.

Specifically, the largest of N retrieved scores is distributed as

$$P\left(S_N \le s \middle| S_j \ge s_0\right) = \left[P\left(S \le s \middle| S \ge s_0\right)\right]^N \approx \exp\left[-Ne^{\beta_0\cdot(s-s_0)}\right] = \exp\left[-e^{\beta_0\cdot\left(s+\frac{\ln N}{\beta_0}-s_0\right)}\right].$$

Conclusions

The distribution function presented here shows good agreement with simulation over twelve orders of magnitude in frequency, and over the three orders of magnitude tested for aligned sequence lengths derived from first principles. The formulation also provides a measure of the number of alignment species N_j contributing most to the distribution as a function of score. It provides a practical way to compute probabilities for scores representing gapless alignments as in scoring sequences retrieved from a database. The distribution provides a reasonable approximation down to relatively small sequence lengths (10 bases in the example computed here) at which point the discrete character of the scores starts to become visible, which is smaller than many practical lengths of interest (e.g. miRNAs). Lastly, it provides an analytical expression for a distribution against which behavior of real or simulated data may be compared. This type of tool may find an expanding utility in exploring the diversity and variability of the human genome as more samples are analyzed and more sequences throughout the genome become available.

1. Mitrophanov AY, Borodovsky M: **Statistical significance in biological sequence analysis**. *Brief Bioinform* 2006, **7**(1):2-24.

2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
3. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**(22):10915-10919.
4. Dayhoff MO, Schwartz R, Orcutt BC: **A model of Evolutionary Change in Proteins**. In: *Atlas of protein sequence and structure, supplemental 3rd edition.* Edited by Dayhoff MO, vol. 5, supplement 3: National Biomedical Research Foundation; 1978: 1059-1063.
5. Agrawal A, Huang X: **Pairwise statistical significance of local sequence alignment using multiple parameter sets and empirical justification of parameter set change penalty**. *BMC Bioinformatics* 2009, **10 Suppl 3**:S1.
6. Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, Storz G: **Target prediction for small, noncoding RNAs in bacteria**. *Nucleic Acids Res* 2006, **34**(9):2791-2802.
7. Karlin S, Dembo A: **Limit distributions of maximal segmental score among Markov-dependent partial sums**. *Advances in Applied Probability* 1992, **24**(1):113-140.
8. Dembo A, Karlin S, Zeitouni O: **Critical phenomena for sequence matching with scoring**. *The Annals of Probability* 1994, **22**(4):1993-2021.
9. Dembo A, Karlin S, Zeitouni O: **Limit distribution of maximal non-aligned two-sequence segmental score**. *The Annals of Probability* 1994, **22**(4):2022-2039.
10. Altschul SF, Bundschuh R, Olsen R, Hwa T: **The estimation of statistical parameters for local alignment score distributions**. *Nucleic Acids Res* 2001, **29**(2):351-361.
11. Poleksic A, Danzer JF, Hambly K, Debe DA: **Convergent Island Statistics: a fast method for determining local alignment score significance**. *Bioinformatics* 2005, **21**(12):2827-2831.
12. Sheetlin S, Park Y, Spouge JL: **The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment**. *Nucleic Acids Res* 2005, **33**(15):4987-4994.
13. Park Y, Sheetlin S, Spouge JL: **Estimating the Gumbel scale parameter for local alignment of random sequences by importance sampling with stopping times**. *The Annals Of Statistics* 2009, **37**(68):3697-3714.
14. Wolfsheimer S, Burghardt B, Hartmann AK: **Local sequence alignments statistics: deviations from Gumbel statistics in the rare-event tail**. *Algorithms Mol Biol* 2007, **2**:9.
15. Pang H, Tang J, Chen SS, Tao S: **Statistical distributions of optimal global alignment scores of random protein sequences**. *BMC Bioinformatics* 2005, **6**:257.
16. Eddy SR: **A probabilistic model of local sequence alignment that simplifies statistical significance estimation**. *PLoS Comput Biol* 2008, **4**(5):e1000069.
17. Sardiu ME, Alves G, Yu YK: **Score statistics of global sequence alignment from the energy distribution of a modified directed polymer and**

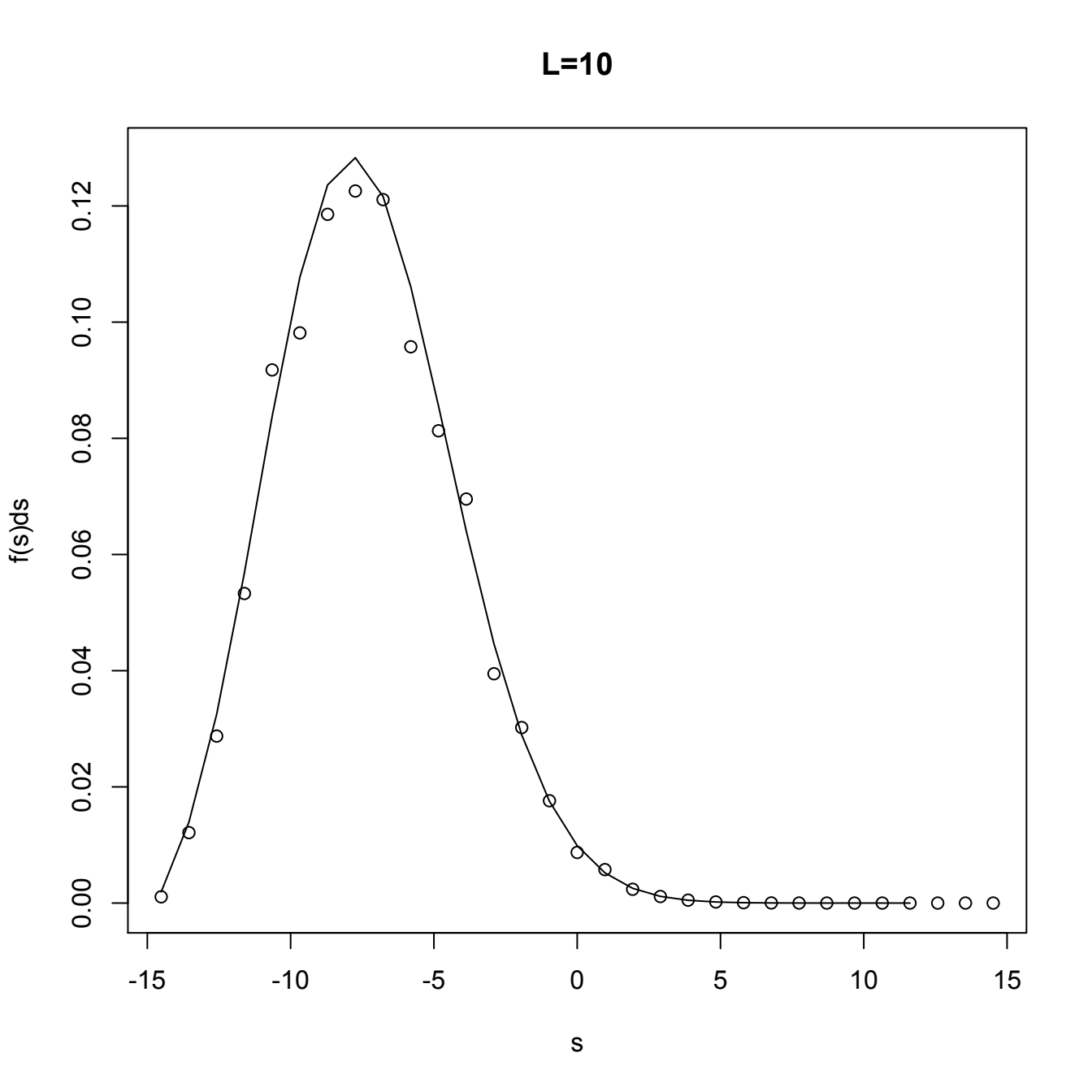
directed percolation problem. *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **72**(6 Pt 1):061917.
18. Wolfsheimer S, Melchert O, Hartmann AK: **Finite-temperature local protein sequence alignment: Percolation and free-energy distribution**. *Physical Review E* 2009, **80**(Copyright (C) 2010 The American Physical Society):061913.
19. Sears FW, Salinger GL: **Thermodynamics, Kinetic Theory, and Statistical Mechanics**, 3rd Edition edn: Addison Wesley; 1975.
20. L'Ecuyer P: **Maximally Equidistributed Combined Tausworthe Generators**. *Mathematics of Computation* 1996, **65**(213):203-213.
21. L'Ecuyer P: **Tables of Maximally Equidistributed Combined LFSR Generators**. *Mathematics of Computation* 1999, **68**(225):261-269.
22. Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F: **GNU Scientific Library Reference Manual**; 2007.

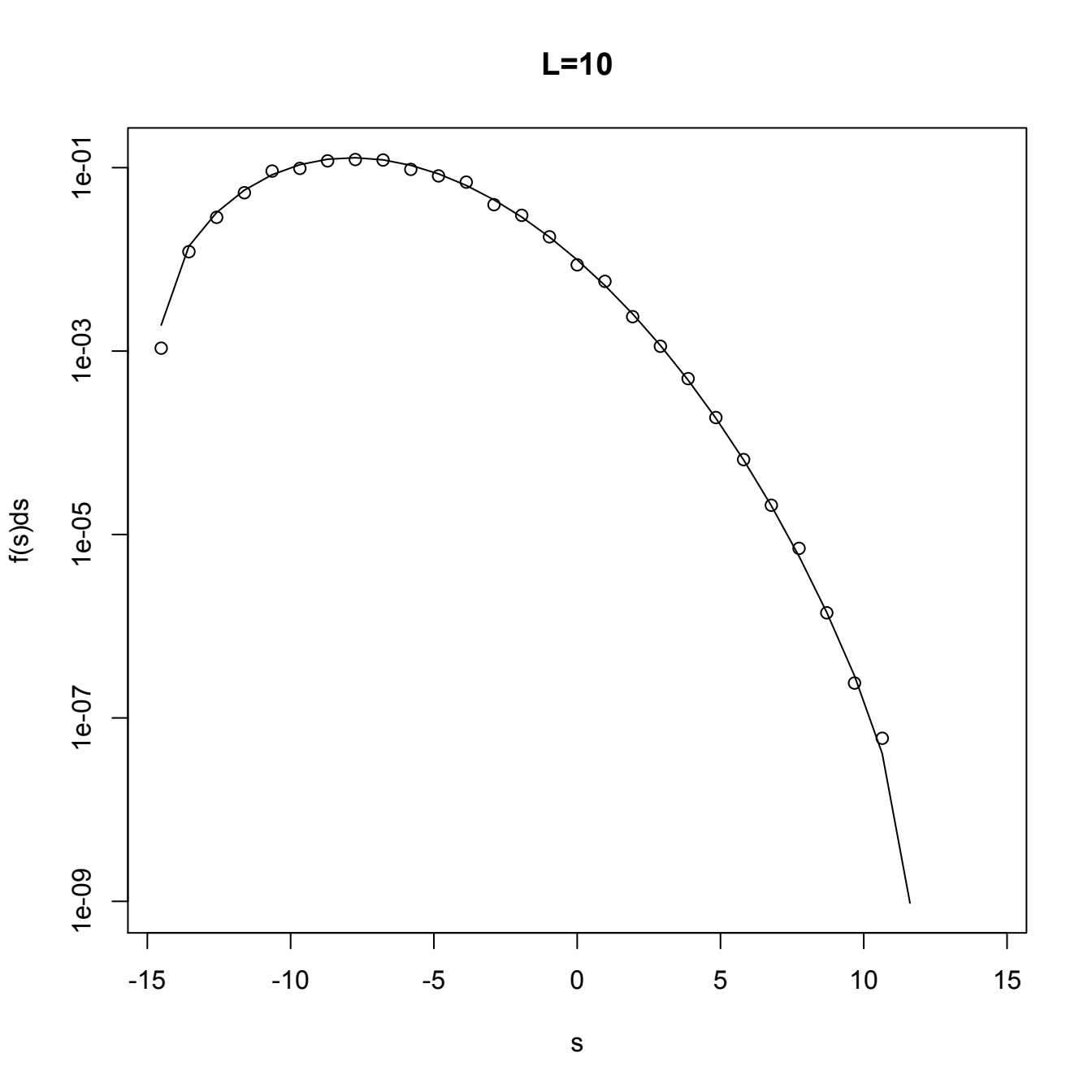
Figure Captions

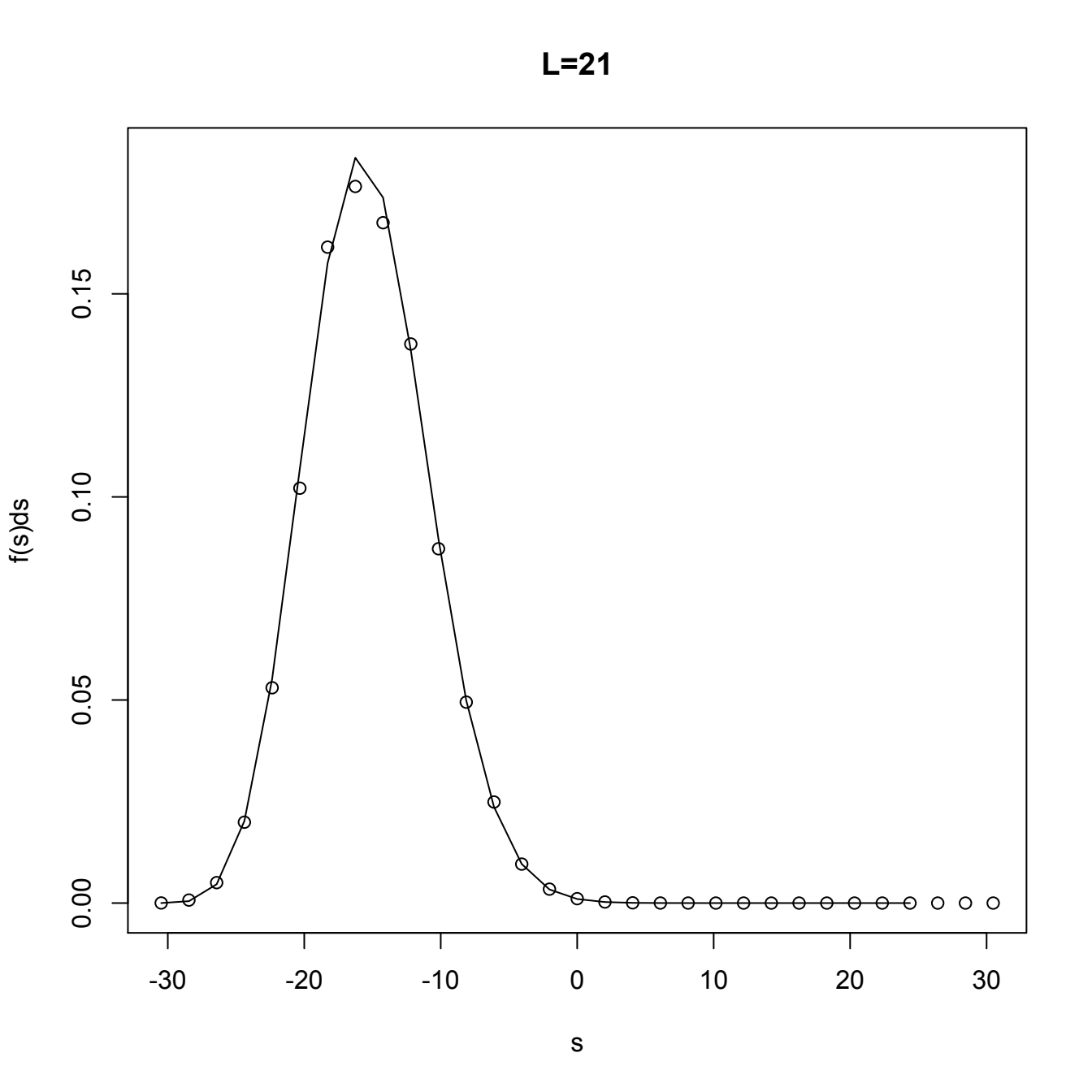
Figure 1. Distributions generated by simple Monte-Carlo sampling of alignments against the asymptotic approximation in both normal and log scales, where alignments are of length 10 (a) and (b), 21, (c) and (d), 100 (e) and (f), 1000 (g) and (h), and 10,000 (i), and (j).

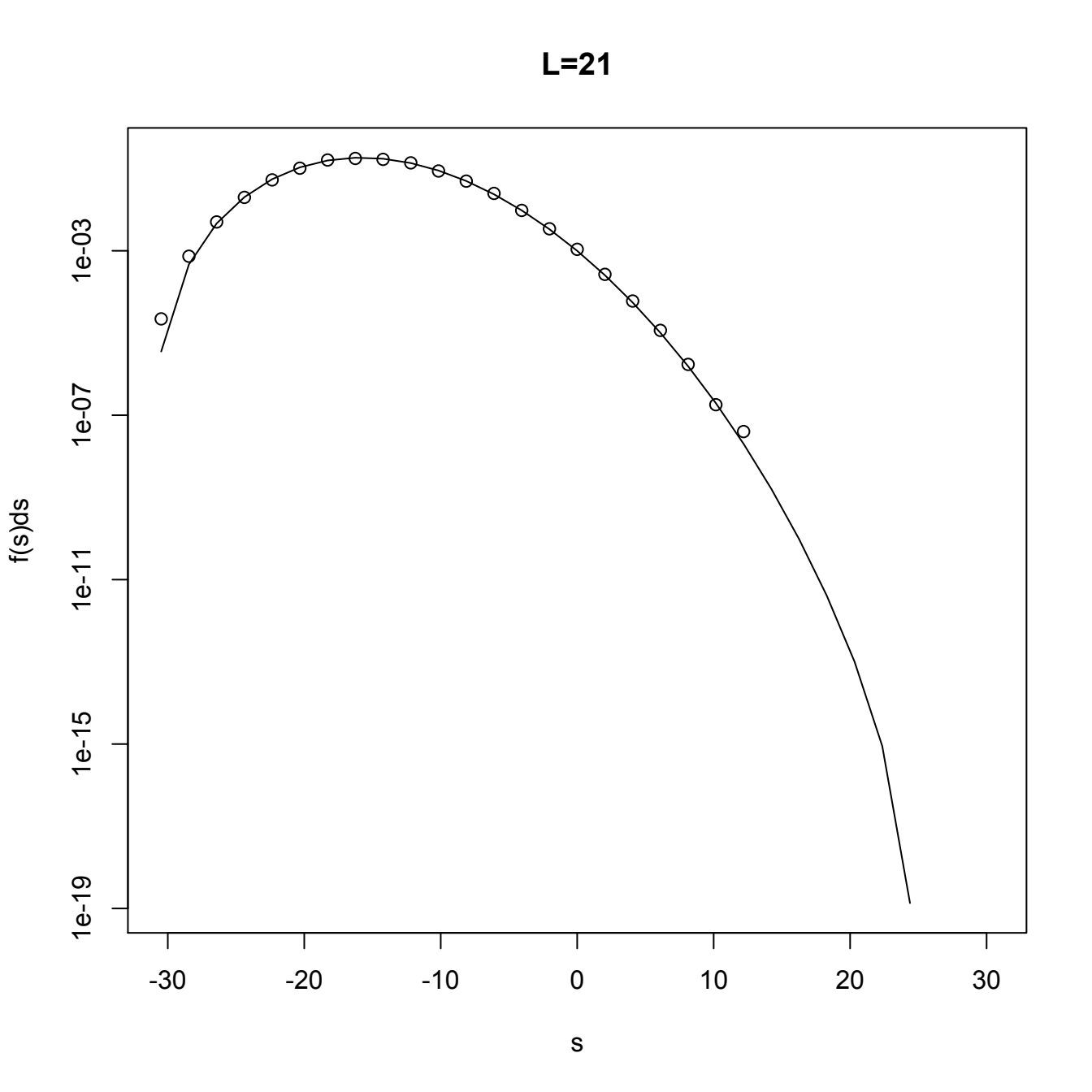
Figure 2. Distribution for $L = 100$ in a rare-event tail using MCMC with importance sampling against the asymptotic approximation in log-scale.

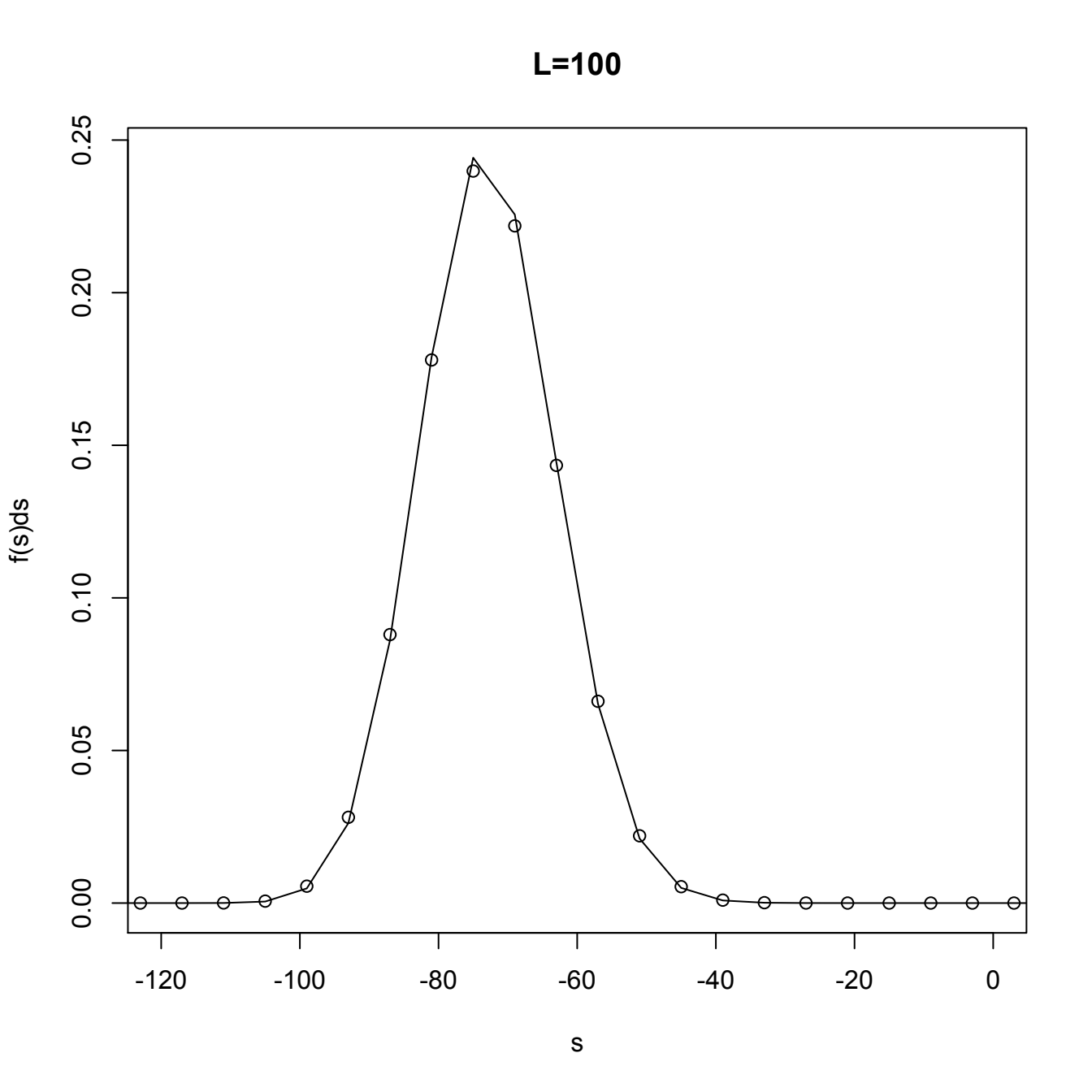
Figure 3. Plot of $\beta = \beta(\hat{s})$ for $L = 100$.

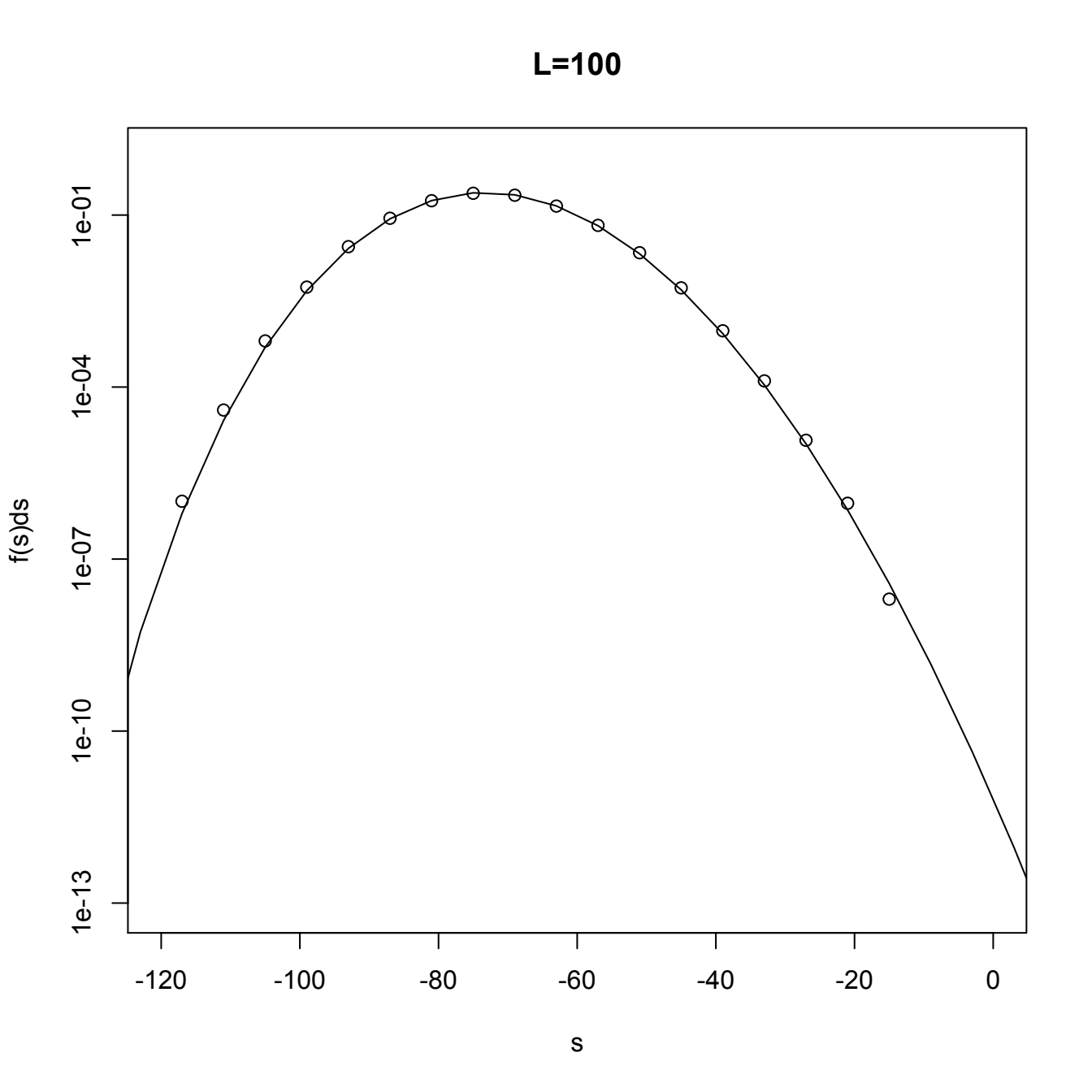

L=10
f(s)ds
0.00
0.02
0.04
0.06
0.08
0.10
0.12
-15
-10
-5
0
5
10
15
s

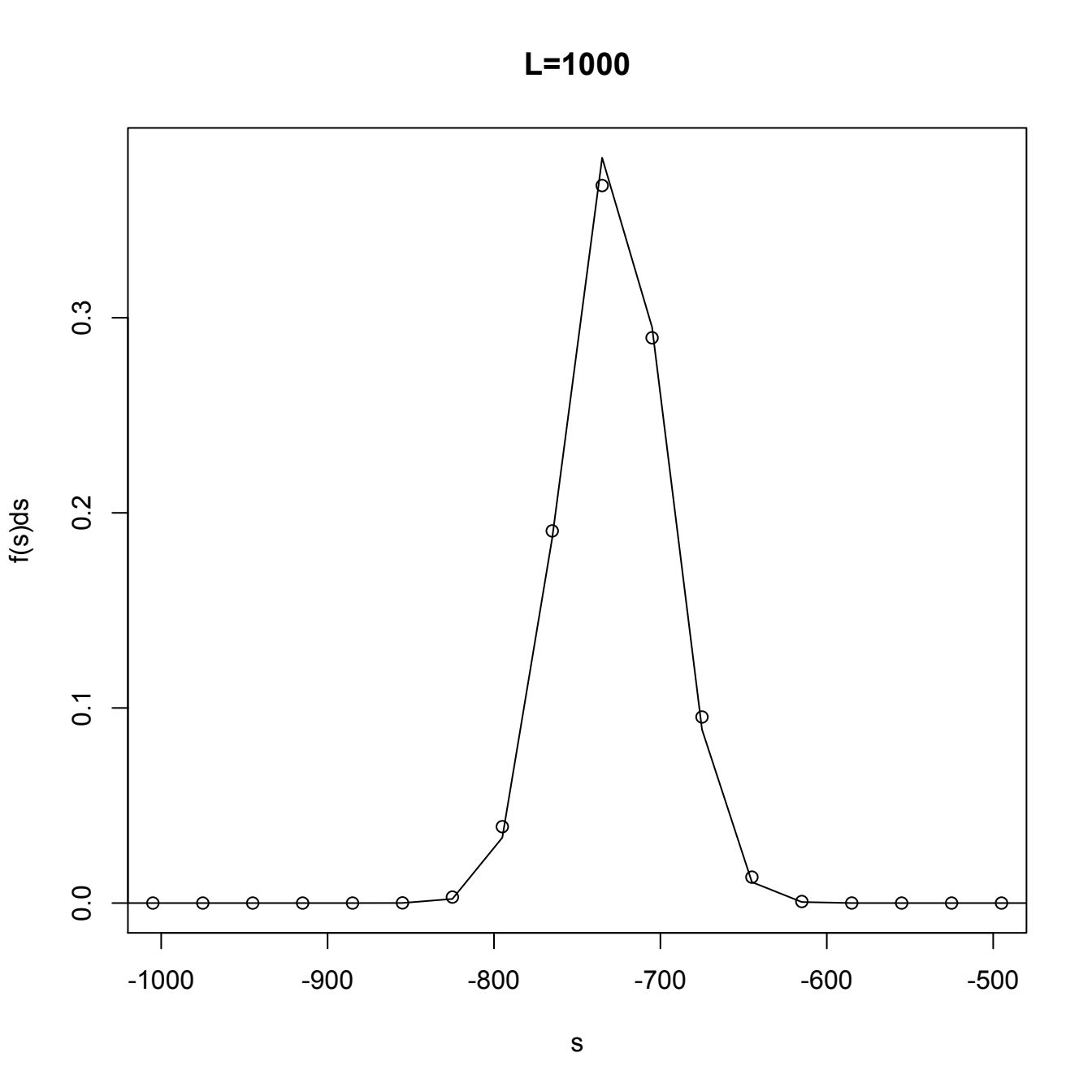

L=10
f(s)ds
1e-01
1e-03
1e-05
1e-07
1e-09
-15
-10
-5
0
5
10
15
s

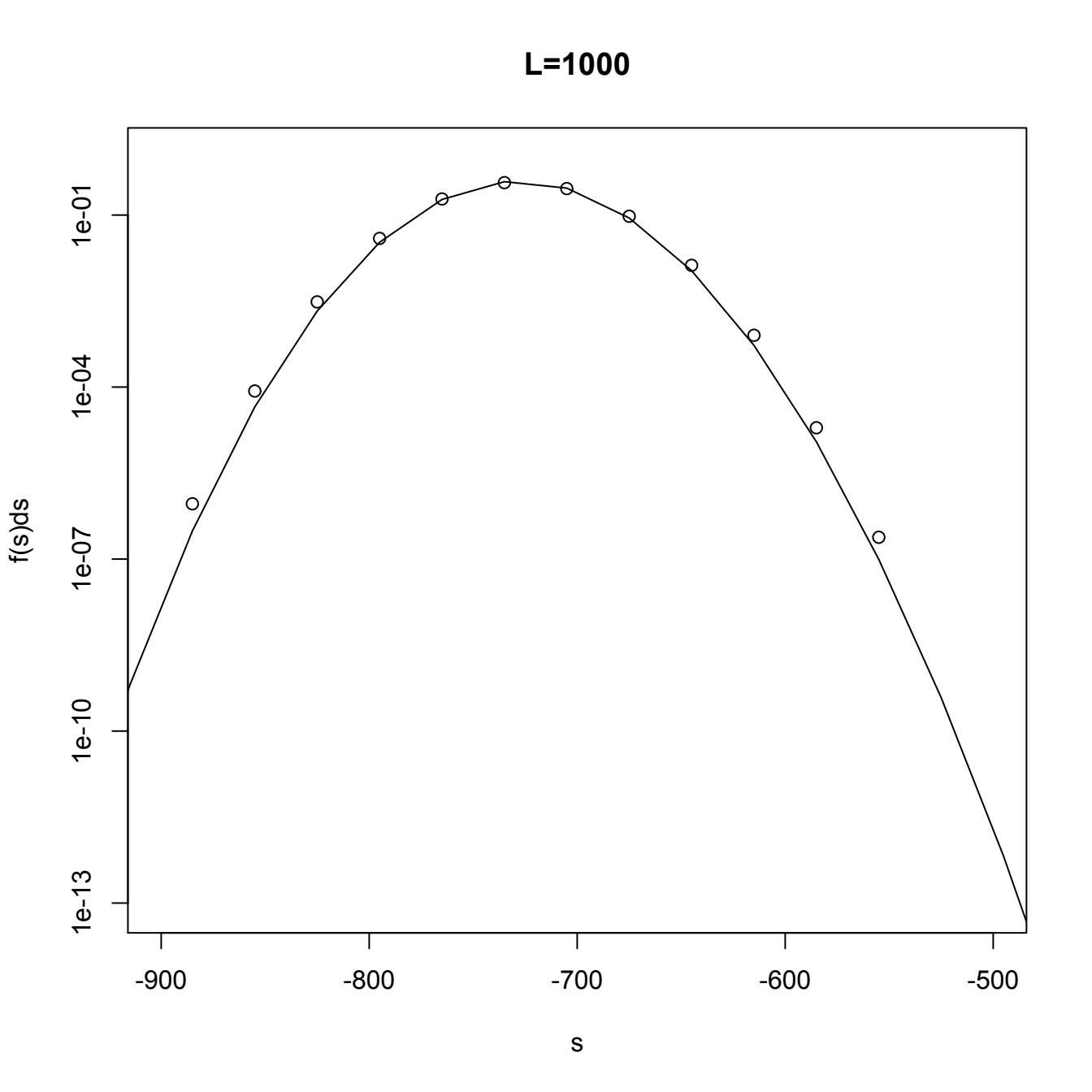

L=21
f(s)ds
0.00
0.05
0.10
0.15
-30
-20
-10
0
10
20
30
s

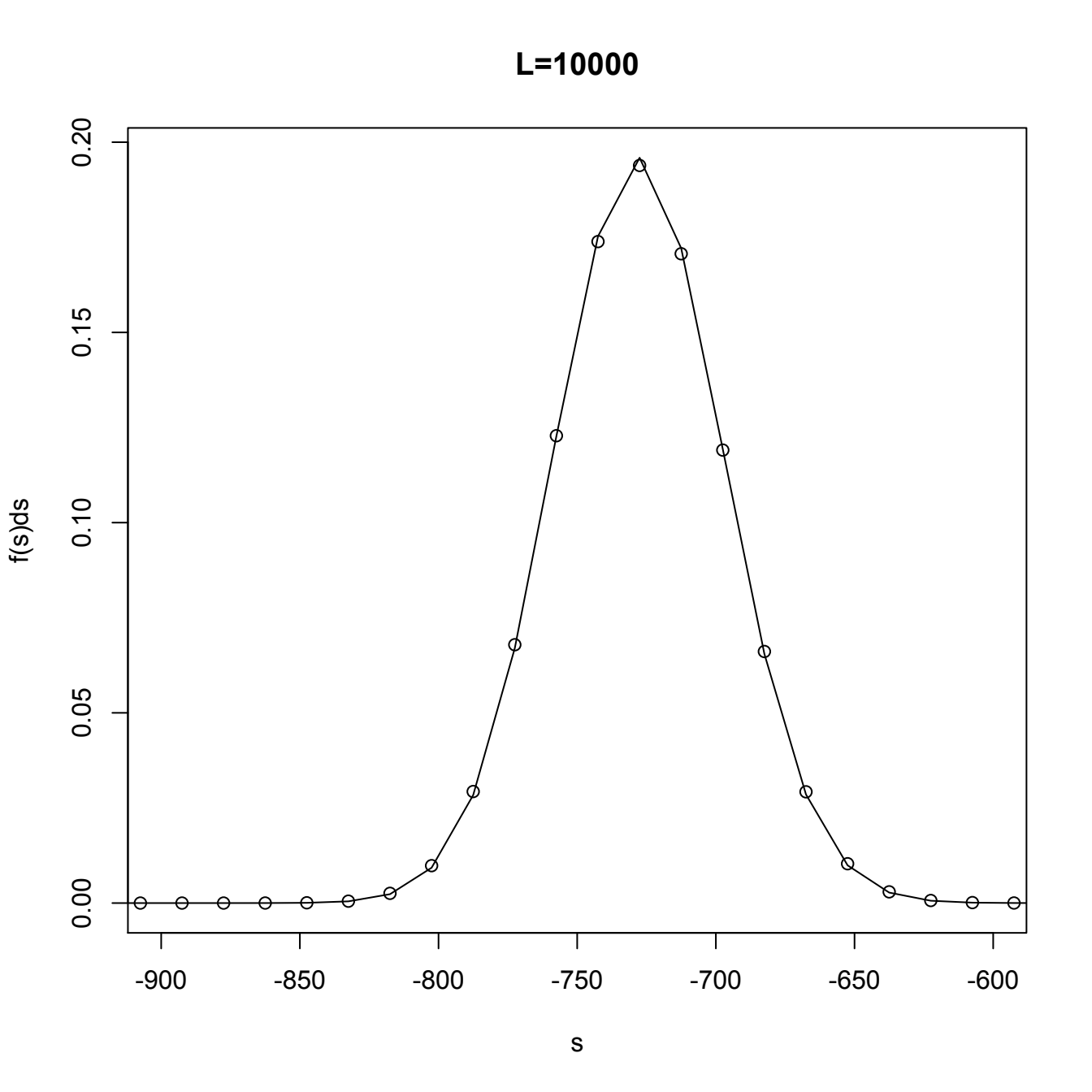

L=21
f(s)ds
1e-03
1e-07
1e-11
1e-15
1e-19
-30
-20
-10
0
10
20
30
s

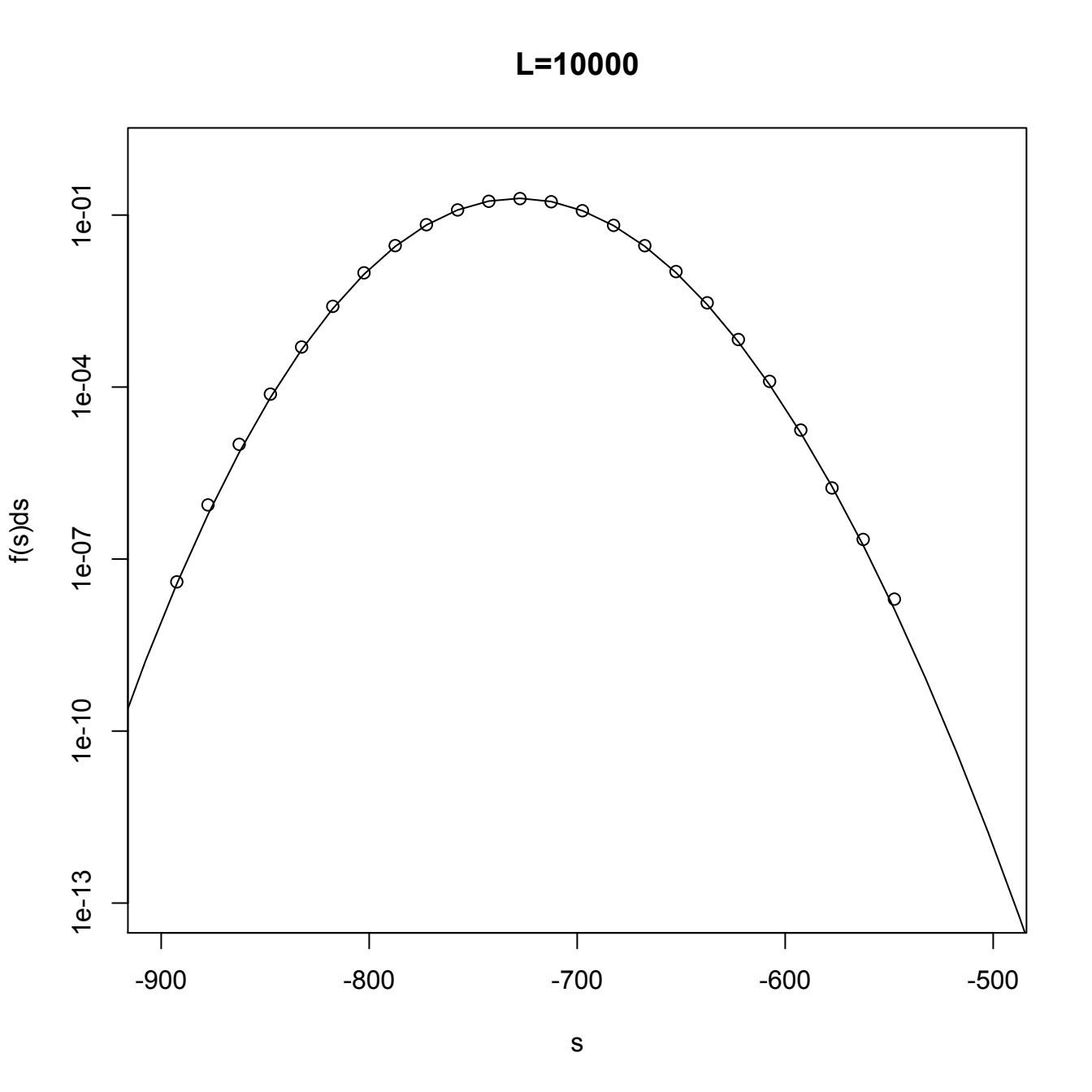

L=100
f(s)ds
0.00
0.05
0.10
0.15
0.20
0.25
-120
-100
-80
-60
-40
-20
0
s

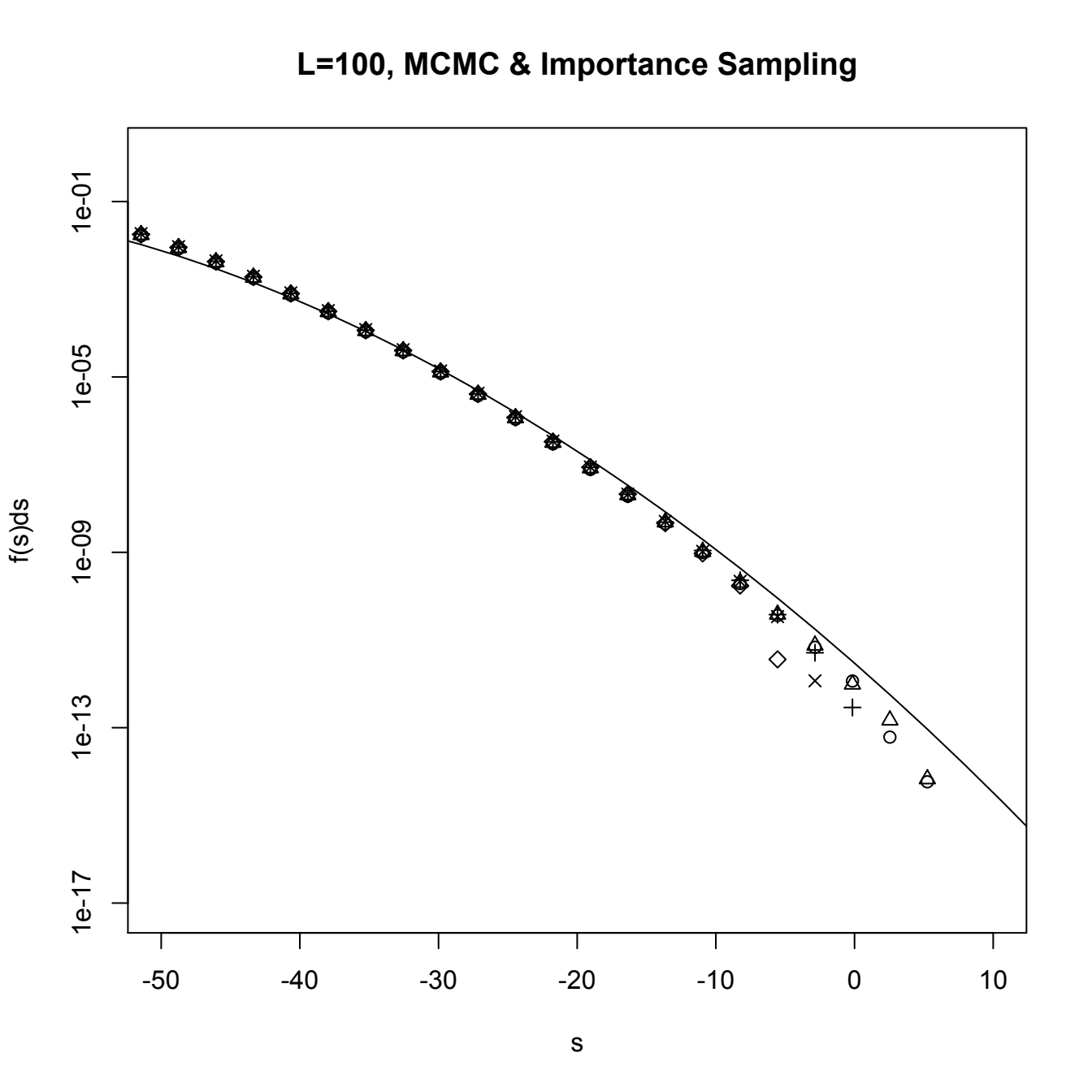

L=100
f(s)ds
1e-01
1e-04
1e-07
1e-10
1e-13
-120
-100
-80
-60
-40
-20
0
s

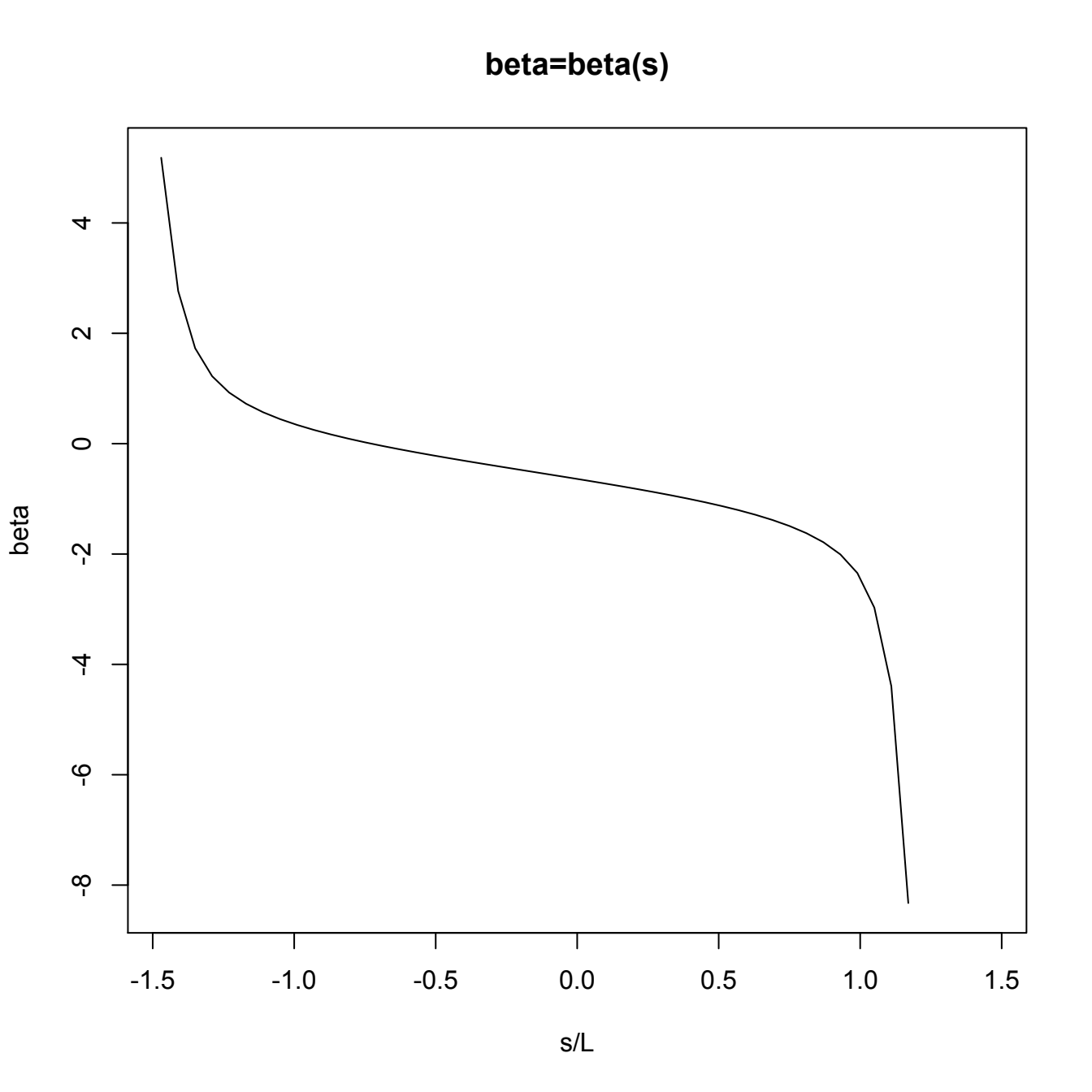

L=1000
f(s)ds
0.0
0.1
0.2
0.3
-1000
-900
-800
-700
-600
-500
s


L=1000
f(s)ds
1e-01
1e-04
1e-07
1e-10
1e-13
-900
-800
-700
-600
-500
s


L=10000
f(s)ds
0.00
0.05
0.10
0.15
0.20
-900
-850
-800
-750
-700
-650
-600
s


L=10000
f(s)ds
1e-01
1e-04
1e-07
1e-10
1e-13
-900
-800
-700
-600
-500
s


L=100, MCMC & Importance Sampling
f(s)ds
1e-01
1e-05
1e-09
1e-13
1e-17
-50
-40
-30
-20
-10
0
10
s


beta=beta(s)
beta
4
2
0
-2
-4
-6
-8
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
s/L