DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 904

August 1993

# A NOTE ON BOOTSTRAP LARGE DEVIATIONS AND DISCRETE PARAMETER SPACES

by

**Michael A. Newton**

# A note on bootstrap large deviations and discrete parameter spaces

Michael A. Newton

August 1993

**Abstract**

A bootstrap large deviation result for the mean is shown to be a consequence of a classical large deviation result due to Sievers and later improved by Plachky and Steinebach. The result implies a level of asymptotic correctness for nonparametric bootstrapping of the maximum likelihood estimator in models having a discrete parameter space. The application of bootstrap methods in molecular evolution gains theoretical support from this result.

## 1 Bootstrap deviations

Let $X_1, X_2, \ldots, X_n$ be a sample of independent and identically $P-$distributed real-valued random variables. Let $P_n = (1/n)\sum_i \delta_{X_i}$ be the empirical measure determined by the sample, i.e. the measure putting mass $1/n$ at each sample point. A nonparametric bootstrap sample $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n}$ is a set of conditionally independent and identically $P_n-$distributed random variables, given the original sample (Efron, 1979). The empirical measure of the bootstrap sample deviates from $P_n$ in a manner similar to how $P_n$ deviates from the unknown $P$. Laws of large numbers and central limit theory for these bootstrap deviations are known (Bickel and Freedman, 1981, Athreya, 1983, Arcones and Giné, 1989, Csörgő and Mason, 1989, Csörgő, 1990).

Suppose that $P$ has a finite moment generating function (MGF)

$$\psi(t) = \int_{-\infty}^{\infty} e^{tx}\, dP(x)$$

for $t \in (-b, b)$, and $b > 0$. Further, and without loss of generality, suppose that $X_1$ has mean 0, and to avoid trivialities suppose that $P$ is not degenerate at 0. It follows that $\psi(t)$ is differentiable to all orders and strictly convex in $(-b, b)$, and that $X_1$ has moments of all orders. (See Billingsley, 1986, pg 285, for example.) As well as converging strongly to 0, and having a central limit theorem, the sample mean $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ has a large deviation rate. That is, for certain $a > 0$,

$$p_n = P(\bar{X}_n > a)$$

satisfies

$$\begin{aligned} p_n &> 0 \qquad \forall n, \\ p_n &\to 0, \end{aligned}$$

and

$$p_n^{1/n} \to \rho(a) = \inf_{t \geq 0} e^{-at}\psi(t) \in (0,1) \tag{1}$$

as $n \to \infty$. Allowable values of $a$ live in the set

$$A = \{a > 0 : a = \psi^{(1)}(t)/\psi(t) \quad t \in (-b, b)\}$$

where $\psi^{(1)}$ is the derivative of $\psi$. Having $a \in A$ ensures that $p_n > 0$. (To see this, note that $a$ is the mean of the conjugate distribution having density $e^{tx}/\psi(t)$ with respect to $P$.)

Exponential decay to zero (1) of the large deviation probabilities has been known for some time. It follows from Mills' ratio if $X_1$ is normal. Important modern extensions are due to Cramér (1938), Chernoff (1952), and Bahadur and Rao (1960). See Book (1985) for a historical account. Our main result is that the same large deviation rate is attained by the nonparametric bootstrap.

**Theorem 1** *Under the above assumptions, the bootstrap sample mean*

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_{n,i}$$

*satisfies, for $a \in A$, and as $n \to \infty$,*

$$q_n^{1/n} \to \rho(a) \qquad a.s.[P]$$

*where*

$$q_n = P(\bar{Y}_n > a | X_1, \ldots, X_n).$$

Hall (1990) proves accuracy of the bootstrap for smaller deviations; where $a$ is replaced by a sequence $a_n$ converging to 0 at a certain rate. It is perhaps surprising that the bootstrap picks up extreme tail probabilities, given that it does not put mass beyond the data. On the other hand, the bootstrap consistently estimates the moment generating function, which relates directly to these tail probabilities. It shall be immediate from the proof that the same result holds if the bootstrap sample size is $m_n$, as long as $m_n \to \infty$ as $n \to \infty$.

Before presenting a proof, we apply this result to an inference problem.

## 2 Likelihoods and discrete parameter spaces

Traditionally, large deviation theorems have been used in statistics to compare hypothesis tests in terms of asymptotic efficiency. Large deviation probabilities also arise when studying the maximum likelihood estimator in a discrete parameter space, and we focus on this problem.

Consider a parametric model $\mathcal{P}$ for the distribution $P$ of the data, which is indexed by points $\theta$ in a parameter space $\Theta$. For notation, suppose that $P = P_{\theta_0}$ is the actual measure generating the data. Assume that the parameter is identifiable. That is, if $\theta_1 \neq \theta_2$ are both in $\Theta$, then the distributions $P_{\theta_1}$ and $P_{\theta_2}$ are distinct. Further, assume that each distribution $P_\theta \in \mathcal{P}$ has a density $f_\theta$ with respect to a common measure on the line.

Based on a random sample $X_1, X_2, \ldots, X_n$ from $P_{\theta_0}$, the loglikelihood of $\theta$ is

$$L_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i) .$$

It is convenient to work with transformed variables

$$Z_i = \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}$$

defined for some particular alternative $\theta \neq \theta_0$. The chance that the likelihood is lower at the truth than at $\theta$ is

$$\begin{aligned} p_n &= P\left(L_n(\theta_0) < L_n(\theta)\right) \\ &= P\left(\frac{1}{n}\sum_{i=1}^{n} Z_i > 0\right). \end{aligned}$$

By Jensen's inequality and identifiability, the expectation of $Z_i$ is strictly negative (and possibly $-\infty$). We thus observe the well-known consequence of the weak law of large numbers that $p_n \to 0$ as $n \to \infty$. In other words, the likelihood tends to be higher the truth than at any other point. (See Lehmann, 1983, pg 409.) By the large deviation theory outlined in Section 1, the rate at which $p_n$ goes to zero is

$$p_n^{1/n} \to \inf_{t \geq 0} \psi_Z(t) \in (0, 1) \tag{2}$$

where $\psi_Z(t)$ is the MGF of the $Z_i$, as long as this MGF exists in a neighborhood of 0. For instance, if $X_i$ have a normal distribution with mean $\theta_0 = 0$, then

$$p_n^{1/n} \to e^{-\theta^2/8}$$

as $n \to \infty$. This agrees with our intuition that the limit should be decreasing in $|\theta|$.

An immediate consequence of Theorem 1, in the context of likelihoods, is that the likelihood based on a nonparametric bootstrap sample will also tend to be higher at $\theta_0$ than at any other point. The loglikelihood $L_n^*$ from a bootstrap sample $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n}$ satisfies

$$L^*(\theta)_n - L_n^*(\theta_0) = \sum_{i=1}^{n} \log \frac{f_\theta(Y_{n,i})}{f_{\theta_0}(Y_{n,i})}$$

Thus, with notation as above:

**Corollary 1** *As $n \to \infty$, and for $\theta \neq \theta_0$,*

$$(P(L^*(\theta) > L^*(\theta_0)|X_1, X_2, \ldots, X_n))^{1/n} \to \inf_{t \geq 0} \psi_Z(t) \qquad a.s.[P_{\theta_0}]$$

*as long as the MGF $\psi_Z(t)$ is finite in a neighborhood of the origin.*

In regular parameter spaces, the sampling distribution of the maximum likelihood estimator (MLE) is approximated, to first order, using a central limit theorem. In some applications, however, the parameter space is discrete, and so it makes no sense to consider $1/\sqrt{n}$-neighborhoods of $\theta_0$. Large deviation probabilities, on the other hand, can give information about this sampling distribution. For example, it is natural to ask about the chance that the MLE, denoted $\hat{\theta}_n$, equals any particular value in the parameter space.

Consider a finite or countably infinite parameter space

$$\Theta = \{\theta_0, \theta_1, \theta_2, \ldots\}$$

for the model above. Upon sampling $n$ times from $P_{\theta_0}$, the chance that the MLE equals a particular wrong value $\theta_j \neq \theta_0$ is

$$\begin{aligned} p_n &= P(\hat{\theta}_n = \theta_j) \\ &= P\left(\cap_{k \neq j}[L(\theta_j) > L(\theta_k)]\right) \\ &\leq P(L(\theta_j) > L(\theta_0)) \,. \end{aligned}$$

Since $p_n$ involves the joint distribution of $L(\theta_k)$ for all $k$, the one-dimensional large deviation result from Section 1 is not applicable directly to this probability. However, from the upper bound on $p_n$ and (2), we have

$$\frac{p_n}{\rho^n} \leq 1 + o(1) \qquad \text{as } n \to \infty \tag{3}$$

where $\rho \in (0,1)$ is the infimum for $t \geq 0$ of $\psi_Z(t)$, the MGF of $\log(f_{\theta_j}(X_1)/f_{\theta_0}(X_1))$. This gives us an approximation to the chance that the MLE equals any particular value $\theta_j$. Of course this approximation, and indeed the actual chance, depend on the unknown $\theta_0$. By applying

the nonparametric bootstrap, and Corollary 1, we see that the conditional probability $q_n$ that the bootstrap MLE $\hat{\theta}_n^*$ equals $\theta_j$ is within the same bound (3) as $p_n$. Thus the nonparametric bootstrap approximates exponentially small probabilities in the sampling distribution of the MLE in discrete parameter spaces. While these small probabilities depend on the true $\theta_0$, the bootstrap approximation does not, and can usually be computed by simulation.

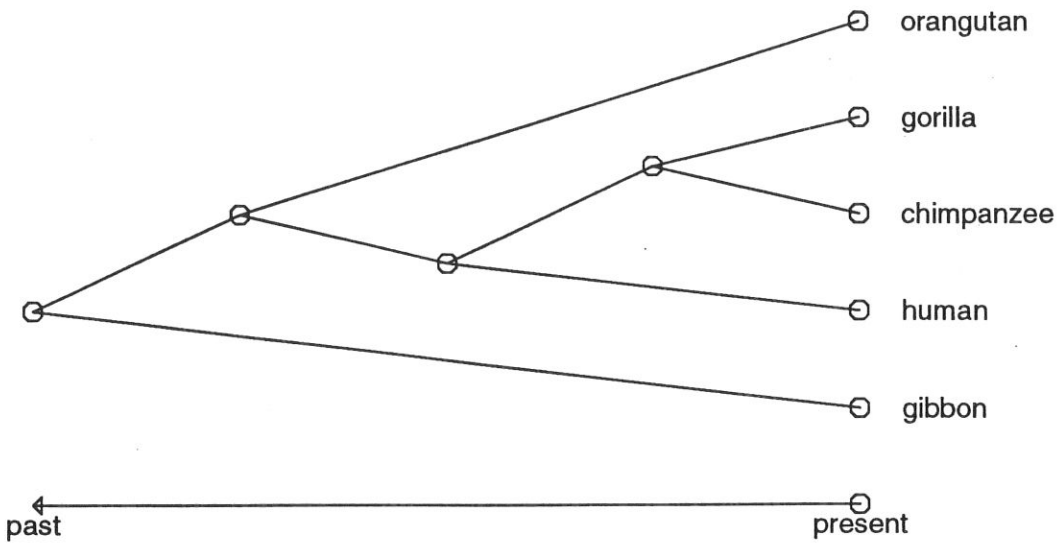
## 3 An application

An important statistical problem is how to infer past evolution using data from living species. Under the theory of common descent, ancestors of any set of $k$ species belonged, at some time in the past, to a single species. The phylogeny is the set of relationships between the $k$ species from the time they were one until the present. In recent years, vast amounts of molecular data (e.g. DNA) have become available to address this problem. Through parametric statistical modeling, Felsenstein (1981, 1983, 1992a) has advocated the use of maximum likelihood to infer phylogenies using molecular data. Further, Felsenstein (1985) applies nonparametric bootstrapping to assess the uncertainty in phylogenetic reconstructions. Others have studied bootstrapping in this context: Zharkikh and Li (1992) and Hedges (1992).

Inference for phylogenies is a nonstandard statistical problem because the parameter space is a set of possible relationships rather than a flat Euclidean space. Figure 1 shows a possible phylogeny relating five primate species. A phylogeny is composed of a tree topology $\theta$ and a set of branch lengths $\eta$. The set $\Theta$ of all possible topologies is finite, with cardinality depending on precisely how you define a point $\theta$. One way to build $\Theta$ is to perform all $\prod_{j=0}^{k-1}(k-j)\,C\,2$ of the following $k-1$ step constructions: In step one, join 2 of the $k$ species; in step two, join two of the remaining $k-1$, and so on. With 5 species, there are 180 distinct tree topologies.

A maximum likelihood reconstruction produces an estimate $\hat{\theta}$ of the topology along with an estimate of the $\hat{\eta}$ of the branch lengths. Felsenstein's (1985) bootstrap method simulates an estimate of the sampling distribution of $\hat{\theta}$. From corollary 1, we see that this method quite accurately approximates the true sampling distribution. Thus, we have demonstrated a theoretical underpinning of the bootstrap in this nonstandard problem. This goes beyond the heuristic central limit theory argument presented in Felsenstein (1985). Felsenstein (1992b) uses the bootstrap in a different context to approximate an integral. Our result says nothing about the theoretical justification of that procedure.

Figure 1: A phylogeny estimated from mitochondrial DNA data (Felsenstein, 1992a) for five primate species. The time scale is not estimated, and may not be linear.




## 4 Proof of the Theorem 1

The proof is a straightforward application of a large deviation theorem due to Sievers (1969), and improved by Plachky (1971) and then Plachky and Steinebach (1975). In proving a converse, Lynch (1978) also states the general result:

**Theorem 2** *Let $S_1, S_2, \ldots, S_n$ be a sequence of random variables with moment generating functions $\psi_1(t), \psi_2(t), \ldots, \psi_n(t)$ which are finite for $t \in [0, d)$, $d > 0$. Suppose that for all $t \in (c, d)$ where $0 \leq c < d$, we have pointwise convergence of $(1/n)\log\psi_n(t)$ to a limit $\phi(t) = \log\psi(t)$ as $n \to \infty$. Let*

$$A = \{a = \phi^{(1)}(t) : \phi^{(1)} \text{ exists, is right-continuous, and strictly monotonic for } t \in (c, d)\}.$$

*Then, for any sequence $a_n$ converging to $a \in A$,*

$$P(S_n > na_n)^{1/n} \to \inf_{t \geq 0} e^{-at}\psi(t)$$

*as $n \to \infty$.*

Note that $\{S_n\}$ need not be sample sums, as in Section 1, but can be arbitrary variables, subject to the constraints of the theorem.

Associate the bootstrap mean $\bar{Y}_n$ with the random variable $S_n/n$ of the theorem. By conditional independence, the MGF of $S_n = n\bar{Y}_n$, given the data, is

$$\psi_n(t) = \left(\int_{-\infty}^{\infty} e^{ty}\, dP_n(y)\right)^n$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} e^{tX_i}\right)^n$$

where again $X_1, X_2, \ldots, X_n$ form the sample of data. For every fixed $t$, by the strong law of large numbers,

$$\frac{1}{n}\log\psi_n(t) \to \log\psi(t) \qquad a.s.[P]$$

where $\psi(t)$ is the MGF of $X_1$. Recall that $\psi$ is differentiable to all orders and strictly convex in $(-b, b)$. Since a countable set of null sets is again a null set, $(1/n)\log\psi_n(t)$ converges to $\log\psi(t)$ for all $t$ in a countable set $B \subset (-b, b)$, for all but a null set $N$ of data sequences. Choosing $B$ to be dense in $(-b, b)$, and using convexity of $\psi(t)$, it follows that except for data sequences in $N$, $(1/n)\log\psi_n(t)$ converges pointwise to $\log\psi(t)$ for all $t \in (-b, b)$. (Use Theorem 10.8, pg 70, Rockafellar, 1970.)

For correspondence, the set $(c, d)$ in Sievers theorem is $(0, b)$ for our result. The set $A$ in Sievers theorem is precisely the same as the set $A$ in the statement of Theorem 1. Strict convexity of $\psi(t)$ ensures strict monotonicity of $\phi^{(1)}(t)$. Suppose the constants $a_n$ all equal $a$.

With this correspondence, $q_n$, the conditional probability that the bootstrap mean exceeds $a$, is in fact $P(S_n > na)$, where the probability is conditioned on a particular data sequence not in $N$. The corollary follows immediately, noting that $E(Z_i) < 0$.

## Acknowledgements

This paper is a partial response to questions raised by Joseph Felsenstein during several discussions on bootstrapping and evolutionary genetics. The author is indebted to David Mason for suggesting Sievers' theorem as a method of proof. Conversations with Charles Geyer and Peter Guttorp are gratefully acknowledged.

## References

Arcones, M. A. and Giné, E. (1989). The bootstrap of the mean with arbitrary bootstrap sample size, *Annals of the Institute of Henri Poincaré* **25**: 457–481.

Athreya, K. B. (1983). Strong law for the bootstrap, *Statist. Probab. Letters* **1**: 147–150.

Bahadur, R. R. and Rao, R. R. (1960). On deviations of the sample mean, *Ann. Math. Statist* **31**: 1015–1027.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Annals of Statistics* **9**: 1196–1217.

Billingsley, P. (1986). *Probability and Measure*, John Wiley & Sons, New York.

Book, S. A. (1985). Large deviations and applications, *in* S. Kotz and N. L. Johnson (eds), *Encyclopedia of statistical sciences*, Vol. 6, John Wiley & Sons.

Chernoff, H. (1952). A measure of asymptotic efficiency of tests based on sums of observations, *Ann. Math. Statist.* **23**: 493–507.

Cramér, H. (1938). *Actualités Scientifiques et Industrielles* (736): 5–23.

Csörgő, S. (1990). On the law of large numbers for the bootstrap mean, *Technical Report 2029*, Institute of Statistics, Consolidated University of North Carolina.

Csörgő, S. and Mason, D. M. (1989). Bootstrapping empirical functions, *Annals of Statistics* **17**: 1447–1471.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics* **7**: 1–26.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution* **17**: 368–376.

Felsenstein, J. (1983). Statistical inference of phylogenies, *J. Roy. Statist. Soc. A* **146**: 246–272.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap, *Evolution.*

Felsenstein, J. (1992b). Estimating effective population size from samples of sequences: a bootstrap monte carlo integration method, *Genetical Research* **60**: 209–220.

Felsenstein, J. (1992a). Phylogenies from restriction sites: A maximum likelihood approach, *Evolution* **46**: 159–173.

Hall, P. (1990). On the relative performance of bootstrap and Edgeworth approximations of a distribution function, *Journal of Multivariate Analysis* **35**: 108–129.

Hedges, S. B. (1992). The number of replications needed for accurate estimation of the bootstrap p-value in phylogenetic studies, *Mol. Biol. Evol.* **9**: 366–369.

Lehmann, E. (1983). *Theory of Point Estimation*, John Wiley & Sons, New York.

Lynch, J. (1978). A curious converse to Siever's *(sic)* theorem, *Annals of Probability* **6**: 169–173.

Plachky, D. (1971). On a theorem of G. L. Sievers, *Ann. Math. Statist.* **42**: 1442–1443.

Plachky, D. and Steinebach, J. (1975). A theorem about probabilities of large deviations with an application to queuing theory, *Periodica Mathematica Hungarica* **6**: 343–345.

Rockafellar, R. T. (1970). *Convex Analysis*, Princeton University Press, Princeton.

Sievers, G. L. (1969). On the probability of large deviations and exact slopes, *Ann. Math. Statist* **40**: 1908–1921.

Zharkikh, A. and Li, W.-H. (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. four taxa with a molecular clock, *Molecular Biology and Evolution* **9**: 1119–1147.