

Uniform Accuracy of the Maximum Likelihood Estimates for Probabilistic Models of Biological Sequences

Svetlana Ekisheva · Mark Borodovsky

Received: 4 September 2008 / Revised: 31 January 2009 /
Accepted: 11 February 2009
© Springer Science + Business Media, LLC 2009

Abstract Probabilistic models for biological sequences (DNA and proteins) have many useful applications in bioinformatics. Normally, the values of parameters of these models have to be estimated from empirical data. However, even for the most common estimates, the maximum likelihood (ML) estimates, properties have not been completely explored. Here we assess the uniform accuracy of the ML estimates for models of several types: the independence model, the Markov chain and the hidden Markov model (HMM). Particularly, we derive rates of decay of the maximum estimation error by employing the measure concentration as well as the Gaussian approximation, and compare these rates.

Keywords Maximum likelihood estimate · Asymptotic properties of estimates · Hidden Markov model · Concentration of measure

AMS 2000 Subject Classifications 62M05 · 60J10 · 62F10 · 11L07

1 Introduction

A biological sequence studied in bioinformatics is a sequence of symbols from a finite alphabet \mathcal{A} , with $|\mathcal{A}| = 4$ for a DNA sequence and $|\mathcal{A}| = 20$ for a sequence of amino acids. If a finite empirical sequence $x = (x_1, \dots, x_N)$ is considered as a realization of a random vector $X = (X_1, \dots, X_N)$, then the distribution of X is called the probabilistic (stochastic) model of the sequence x . Since in practice the probabilistic model of a

S. Ekisheva
Department of Mathematics, Syktyvkar State University, Oktjabrskii pr., 55,
Syktyvkar, 167000, Russia

M. Borodovsky (✉)
Wallace H. Coulter Department of Biomedical Engineering and Computational Science and
Engineering Division, Georgia Institute of Technology, Atlanta, GA 30332-0535, USA
e-mail: borodovsky@gatech.edu

biological sequence is *a priori* unknown, a standard model selection test is typically carried out together with the parameter estimation. Stochastic models for biological sequences have been studied in many publications, particularly in Gatlin (1972), Almagor (1983), Borodovsky et al. (1986a, b), Churchill (1989), Tavaré and Song (1989), Karlin and Macken (1991), Karlin et al. (1992), Durbin et al. (1998).

The simplest model, called an independence model in bioinformatics literature, assumes that the components of random vector X are i.i.d. random variables. If p_α designates the probability of the occurrence of symbol α , $\alpha \in \mathcal{A}$, at any sequence position, then the maximum likelihood (ML) estimate of the parameter p_α is given by the ratio $\hat{p}_\alpha = N(\alpha)/N$, where $N(\alpha)$ is the number of symbols α observed in the random vector X .

A more general model is a homogeneous ergodic Markov chain of m -order with transition probabilities $p_{\alpha_1, \dots, \alpha_m, \alpha_{m+1}} = P(x_i = \alpha_{m+1} | x_{i-1} = \alpha_m, \dots, x_{i-m} = \alpha_1)$, $\alpha_k \in \mathcal{A}$, $i = m + 1, \dots, N$, and stationary distribution π . The ML estimate of the transition probability $p_{\alpha, \beta}$, $\alpha, \beta \in \mathcal{A}$, of the first-order Markov chain is given by the ratio $\hat{p}_{\alpha, \beta} = N(\alpha\beta)/N(\alpha\bullet)$. Here $N(\alpha\beta)$ is the number of occurrences of the pair (α, β) in the sequence X and $N(\alpha\bullet) = \sum_{\gamma \in \mathcal{A}} N(\alpha\gamma)$. Similarly, for the m -order Markov chain, the ML estimate of the transition probability $p_{\alpha_1, \dots, \alpha_m, \alpha_{m+1}}$ is

$$\hat{p}_{\alpha_1, \dots, \alpha_m, \alpha_{m+1}} = \frac{N(\alpha_1 \dots \alpha_m \alpha_{m+1})}{N(\alpha_1 \dots \alpha_m \bullet)},$$

where $N(\alpha_1 \dots \alpha_k)$ designates the number of occurrences of k -letter word $(\alpha_1, \dots, \alpha_k)$ in the sequence X and $N(\alpha_1 \dots \alpha_m \bullet) = \sum_{\gamma \in \mathcal{A}} N(\alpha_1 \dots \alpha_m \gamma)$. Further in the text we assume that $\mathcal{A} = \{\alpha_1, \dots, \alpha_k\}$, $p_i = p_{\alpha_i}$, $i = 1, \dots, k$, for the independence model, while $\pi(\alpha_i) = \pi_i$ and $p_{\alpha_i, \alpha_j} = p_{ij}$, $i, j = 1, \dots, k$ for the Markov chain model.

Our goal is to determine, as the length N of the sequence increases, the rate of convergence to zero (in probability) of the maximum error in estimation, Δ , where $\Delta = \max_{i=1, \dots, k} |\hat{p}_i - p_i|$ for the independence model, and $\Delta = \max_{i=1, \dots, k} |\hat{\pi}_i - \pi_i|$ or $\Delta = \max_{i, j=1, \dots, k} |\hat{p}_{ij} - p_{ij}|$ for the Markov chain. Furthermore, we apply these results to assess the rate of convergence of the parameter estimation errors for a hidden Markov model (HMM). We explore the asymptotic behavior of $\tilde{P} = P(\Delta \leq \epsilon)$ as N grows by employing both the normal approximation of the ML estimates and the method based on the measure concentration. This theoretical study has been motivated by bioinformatic applications. Many bioinformatics algorithms (e.g. Lawrence et al. 1993; Borodovsky and McIninch 1993; Burge and Karlin 1997) require computations of the probabilities of sequence fragments or the logarithm of these probabilities under a given type of probabilistic model, and these computations use the ML estimates of unknown model parameters. Knowledge of the bounds on the maximum error in estimation Δ allows us to assess error in the computations for such algorithms.

The bounds on the estimation error can be expressed in terms of the uniform confidence intervals for the model parameters. If the model in question is a Markov chain or an HMM, then the confidence intervals can be combined with the perturbation bounds derived in Mitrophanov (2005) and in Mitrophanov et al. (2005) for further investigation of the model properties.

In the earlier paper by Ekisheva and Borodovsky (2006), the authors derived lower bounds on \tilde{P} for the same set of probabilistic models by employing the property of the asymptotic normality of the vector of the ML estimates. In the present

work, the measure concentration approach allows us to show that the convergence of the probability \tilde{P} to one takes place with the faster rate than was proved in Ekisheva and Borodovsky (2006).

2 Independence Model

Throughout this section we assume that the empirical sequence $x = (x_1, \dots, x_N)$ is a realization of the random vector $X = (X_1, \dots, X_N)$ with i.i.d. components X_j , $j = 1, \dots, N$, such that $P(X_j = \alpha_i) = p_i$, $j = 1, \dots, N$, $i = 1, \dots, k$, and $\sum_{i=1}^k p_i = 1$.

The ML estimate \hat{p}_i of the parameter p_i , $i = 1, \dots, k$, possesses the following well-known properties (Cox and Hinkley 1974): it is unbiased, strongly consistent, and asymptotically normally distributed (i.e. $\frac{(\hat{p}_i - p_i)\sqrt{N}}{\sqrt{p_i(1-p_i)}} \rightarrow^d N(0, 1)$ if $p_i \in (0, 1)$). The strong consistency of \hat{p}_i implies that the error in estimation $\delta_i = |\hat{p}_i - p_i|$, $i = 1, \dots, k$, converges to zero almost surely as the sample size N increases to infinity, and the same holds true for the maximum error in estimation over all parameters, $\Delta = \max_{i=1, \dots, k} |\hat{p}_i - p_i|$. Therefore, the convergence $\Delta \xrightarrow{P} 0$ also takes place. The exponential lower bound on the rate of the convergence is given by the following theorem.

Theorem 2.1 *The inequality*

$$P(\Delta \leq \varepsilon) \geq 1 - 2k \exp(-2N\varepsilon^2) = B_1 \tag{2.1}$$

holds true for any positive ε .

Proof To get an explicit result on the uniform closeness of \hat{p}_i to p_i , $i = 1, \dots, k$, we start with the absolute error of a single ML estimate, $\delta_i = |\hat{p}_i - p_i|$. For a given i we consider random variables $v_1(\alpha_i), \dots, v_N(\alpha_i)$, where $v_l(\alpha_i)$ stands for the indicator of occurrence of symbol α_i at site l of the sequence X . The random variable $S_N = \hat{p}_i N = \sum_{l=1}^N v_l(\alpha_i)$ has the binomial distribution with parameters N and p_i , and, therefore, we can apply to S_N the concentration result by Chernoff (1952): for any $\varepsilon > 0$,

$$\begin{aligned} P(\delta_i \geq \varepsilon) &= P\left(\left|\frac{\sum_{l=1}^N v_l(\alpha_i)}{N} - p_i\right| \geq \varepsilon\right) = P\left(\left|\frac{S_N}{N} - p_i\right| \geq \varepsilon\right) \\ &= P(|S_N - p_i N| \geq \varepsilon N) \leq 2 \exp(-2N\varepsilon^2). \end{aligned} \tag{2.2}$$

Now, for the maximum estimation error Δ the inequality (2.2) implies that

$$P(\Delta \leq \varepsilon) \geq 1 - \sum_{i=1}^k P(\delta_i \geq \varepsilon) \geq B_1. \tag{2.3}$$

This completes the proof. □

The inequality (2.1) also yields confidence intervals $(\hat{p}_i - \varepsilon, \hat{p}_i + \varepsilon)$ for p_i , $i = 1, \dots, k$, at the confidence level at least B_1 , or, equivalently, k -dimensional confidence region for the k -dimensional vector of parameters $\mathbf{P} = (p_1, \dots, p_k)$.

Alternatively, the error in estimation δ_i can be studied by applying the property of asymptotic normality of the ML estimate. Let p_i be positive, then for $S_N = \sum_{i=1}^N v_l(\alpha_i)$, the Central Limit Theorem implies that, for any $x \in \mathbf{R}$,

$$P_N(x) = P\left(\frac{S_N - \mathbf{E}S_N}{\sqrt{\mathbf{Var}S_N}} \leq x\right) = P\left(\frac{S_N - p_i N}{\sqrt{N p_i(1 - p_i)}} \leq x\right) \rightarrow \Phi(x) \tag{2.4}$$

as $N \rightarrow \infty$. Here $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt$ is the cumulative distribution function of the standard normal distribution. Let $\Delta_N(x)$ be the absolute error of the normal approximation (2.4), i.e. $\Delta_N(x) = |P_N(x) - \Phi(x)|$. Then

$$\begin{aligned} P(\delta_i \leq \varepsilon) &= P(|\hat{p}_i - p_i| \leq \varepsilon) = P\left(\frac{|S_N - p_i N|}{\sqrt{N p_i(1 - p_i)}} \leq \frac{\varepsilon \sqrt{N}}{\sqrt{p_i(1 - p_i)}}\right) \\ &\geq \Phi(y) - \Phi(-y) - \Delta_N(y) - \Delta_N(-y), \end{aligned} \tag{2.5}$$

where $y = \frac{\varepsilon \sqrt{N}}{\sqrt{p_i(1 - p_i)}}$. Next, we use the result for non-uniform bounds $\Delta_N(y)$ obtained by Nagaev (1965) and the inequality

$$\Phi(y) \geq 1 - \frac{1}{\sqrt{2\pi} y} \exp(-y^2/2), \tag{2.6}$$

for $y > 0$, to derive that

$$P(\delta_i \leq \varepsilon) \geq 1 - \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\sqrt{p_i(1 - p_i)}}{\varepsilon \sqrt{N}} \exp\left(-\frac{\varepsilon^2 N}{2 p_i(1 - p_i)}\right) - \frac{C p_i(1 - p_i)(1 - 2 p_i + 2 p_i^2)}{\varepsilon^3 N^2} \tag{2.7}$$

with a positive absolute constant C . From inequalities (2.7) and $p_i(1 - p_i)(1 - 2 p_i + 2 p_i^2) \leq 1/8, i = 1, \dots, k$, it follows, similarly to Eq. 2.3, that, for any $\varepsilon > 0$,

$$P(\Delta \leq \varepsilon) \geq 1 - \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{k}{\varepsilon \sqrt{N}} \exp(-2 N \varepsilon^2) - \frac{Ck}{8 N^2 \varepsilon^3} = B_2. \tag{2.8}$$

There are other non-uniform estimates of the accuracy of the normal approximation applicable to this case, e.g. in Feller (1945) and in Osipov (1967) (the former paper deals specifically with the binomial distribution). However, for $y = O(\sqrt{N})$ the substitution of any of these bounds $\Delta_N(y)$ to Eq. 2.5 does not improve the polynomial lower bound on the rate of convergence obtained in Eq. 2.8.

Finally, from the asymptotic normality of the random vector Y with components $Y_i = (\hat{p}_i - p_i) \sqrt{N}, i = 1, \dots, k$, (Cox and Hinkley 1974) and the inequality of Li and Shao (2002), the following result was proved in Ekisheva and Borodovsky (2006):

$$\begin{aligned} P(\Delta \leq \varepsilon) &\geq (2\Phi(2\varepsilon \sqrt{N}) - 1)^k - 2^{k-2} k(k - 1) \exp(-2\varepsilon^2 N) \\ &\geq 1 - [k/2]! \frac{k}{2\sqrt{2\pi} N \varepsilon} \exp(-2\varepsilon^2 N) - 2^{k-2} k(k - 1) \exp(-2\varepsilon^2 N) \\ &= B_3, \end{aligned} \tag{2.9}$$

where $[z]$ denotes the biggest integer that does not exceed z . This inequality, however, does not take into account the error of the multidimensional normal approximation; moreover, even without the error, $B_1 > B_3$. Comparison of the lower

bounds B_1 and B_2 shows that while B_1 tends to one faster than B_2 as N grows to infinity, for small N (i.e. for some real biological sequences) B_2 might be greater than B_1 . However, since the exact value of the constant C is unknown, for any sequence length N the inequality (2.1) is more useful for practical purposes of empirical sequence analysis (as B_1 has an explicit form) and it is more interesting as a theoretical result providing an exponential bound on the convergence rate rather than a polynomial one in Eq. 2.8.

An example of important application of Theorem 2.1 is an evaluation of the error in estimation of the probability of local appearance of specific DNA and protein sequences fragments in the functionally inhomogeneous polymers. (The logarithms of the ratios of sequence fragment probabilities computed with two alternative models make log-odds scores, the important quantities used in many algorithms of biological sequence analysis; further discussions of the log-odds scores applications are available in Section 2.2 of Durbin et al. 1998, and in Sections 4.2 and 5.2.1 of Borodovsky and Ekisheva 2006). The bounds of the error can be estimated from inequality (2.1) for the maximum estimation error Δ .

Let us assume that k positive unknown parameters of the independence model are estimated from the database of size N . Then an error in estimation of the log-probability S of the sequence fragment $x = (x_1, \dots, x_L)$ is

$$|\hat{S} - S| = |\log \hat{P}(x_1, \dots, x_L) - \log P(x_1, \dots, x_L)| = \left| \sum_{j=1}^L \log \frac{\hat{p}_{x_j}}{p_{x_j}} \right|. \tag{2.10}$$

If l_i is the number of symbols of type i in the fragment x , $i = 1, \dots, k$, then Eq. 2.10 continues as

$$|\hat{S} - S| = \left| \sum_{i=1}^k l_i \log \frac{\hat{p}_i}{p_i} \right| = \left| \sum_{i=1}^k l_i \log \left(1 + \frac{\hat{p}_i - p_i}{p_i} \right) \right|. \tag{2.11}$$

Theorem 2.1, equality (2.11) and standard calculations lead to the following result: for any positive ε such that $\varepsilon < \frac{3L}{4}$, we have

$$\begin{aligned} P(|\hat{S} - S| \leq \varepsilon) &\geq P\left(\Delta \leq \frac{P_*}{2}(\sqrt{1 + 4\varepsilon/L} - 1)\right) \\ &\geq 1 - 2k \exp\left(-\frac{Np_*^2}{2}(\sqrt{1 + 4\varepsilon/L} - 1)^2\right). \end{aligned} \tag{2.12}$$

Here $p_* = \min_i p_i$. Note that the right-hand side of the inequality (2.12) converges to 1 as N grows only if the length of the fragment $L = o(\sqrt{N})$. In this case Eq. 2.12 provides uniform weak convergence rate over all segments of the length L .

3 Markov Chain Model

3.1 ML Estimates of Stationary Probabilities

We consider the first-order homogeneous ergodic Markov chain $X = (X_1, \dots, X_N)$ with transition probabilities p_{ij} , $i, j = 1, \dots, k$. (A Markov chain is said to be ergodic if it is irreducible and aperiodic). Then there exists a unique stationary distribution

$\pi = (\pi_1, \dots, \pi_k), \pi_i > 0, i = 1, \dots, k$, and the ML estimates of the stationary probabilities are given by ratios $\hat{\pi}_i = \frac{N(\alpha_i)}{N}, i = 1, \dots, k$. From the theory of Markov processes it is known that estimates $\hat{\pi}_i$ are consistent, and the random vector with components $(\hat{\pi}_i - \pi_i)\sqrt{N}, i = 1, \dots, k$, is asymptotically a centered Gaussian vector (Lemma 3.2 and Theorem 3.3, Billingsley 1961). It is easy to show that $\hat{\pi}_i$ are unbiased if the initial distribution coincides with the stationary one. The consistency of $\hat{\pi}_i$ implies that the error in estimation $\delta_i = |\hat{\pi}_i - \pi_i|, i = 1, \dots, k$, converges to zero in probability and the same holds true for the maximum error in estimation over all parameters, $\Delta = \max_{i=1, \dots, k} |\hat{\pi}_i - \pi_i|$.

To study the behavior of $\tilde{P} = P(\Delta \leq \varepsilon), \varepsilon > 0$, we will use the concentration methods and some known results for some special types of Markov chains.

Definition 3.1 A homogeneous Markov chain Y is called uniformly ergodic if

$$\sup_{\alpha} \|P(Y_{n+1}|Y_1 = \alpha) - \pi\| \rightarrow 0 \tag{3.1}$$

as $n \rightarrow \infty$. The supremum is taken over all states α of the Markov chain, $\|v\|$ designates the total variation of the measure v , and π stands for the stationary distribution of Y .

Obviously, a homogeneous ergodic Markov chain with a finite state space is uniformly ergodic. In addition, for a uniformly ergodic Markov chain the convergence in Eq. 3.1 takes place at a uniform geometric rate (Theorem 16.0.2 in Meyn and Tweedie 1993). Therefore, for the homogeneous ergodic Markov chain X there exist $\rho, 0 \leq \rho < 1$, and $R, 0 \leq R < +\infty$, such that

$$\sup_i \|P(X_{m+n}|X_m = i) - \pi\| = \max_{i=1, \dots, k} \sum_{j=1}^k |p_{ij}^{(n)} - \pi_j| \leq R\rho^n, \tag{3.2}$$

where $p_{ij}^{(n)}$ is the probability of transition in n steps from state i to state j .

Theorem 3.1 For a homogeneous ergodic Markov chain X and a positive ε , the following inequality holds

$$P(\Delta \leq \varepsilon) \geq 1 - 2k \exp\left(-\frac{\varepsilon^2 N}{2(1 + \frac{2R\rho}{1-\rho} + a^*)^2}\right) = D_1. \tag{3.3}$$

Here a^* can be chosen to be equal to either $\pi^* = \max_i \pi_i, p^* = \max_{i,j} p_{ij}$, or 1 , while R, ρ are defined by Eq. 3.2.

Proof We will use a concentration-type technique, the method of bounded martingale differences (McDiarmid, 1998). We start with an absolute error of a single ML estimate, $\delta_i = |\hat{\pi}_i - \pi_i|$. For a given $i, i = 1, \dots, k$, we define a random variable

$$f = f_i = N(\alpha_i) - \pi_i N = \sum_{l=1}^N v_l(\alpha_i) - \pi_i N,$$

where $v_l(\alpha_i)$ denotes the indicator of occurrence of symbol α_i at site l of the sequence X . Next, we define Doob’s martingale sequence f^1, \dots, f^N as follows. Let (Ω, \mathcal{F}, P)

be a probability space where X_1, \dots, X_N are defined and let $\mathcal{F}_m, m = 1, \dots, N$, be a σ -field generated by X_1, \dots, X_m , i.e. $\mathcal{F}_m = \sigma(X_1, \dots, X_m)$. We define $\mathcal{F}_0 = \{\emptyset, \mathcal{F}\}$ and $f^m = \mathbf{E}(f|\mathcal{F}_m), m = 0, \dots, N$. Then $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}_N \subset \mathcal{F}$ and f^0, \dots, f^N is Doob's martingale sequence. We have to show that the martingale differences

$$\begin{aligned}
 f^m - f^{m-1} &= \mathbf{E}(f|\mathcal{F}_m) - \mathbf{E}(f|\mathcal{F}_{m-1}) \\
 &= v_m(\alpha_i) + \sum_{l=m+1}^N \sum_{j_m, j_{m+1}, \dots, j_{l-1}} v_m(\alpha_{j_m}) P_{j_m j_{m+1}} P_{j_{m+1} j_{m+2}} \times \dots \times P_{j_{l-1} i} \\
 &\quad - \sum_{l=m}^N \sum_{j_{m-1}, j_m, \dots, j_{l-1}} v_{m-1}(\alpha_{j_{m-1}}) P_{j_{m-1} j_m} P_{j_m j_{m+1}} \times \dots \times P_{j_{l-1} i} \tag{3.4}
 \end{aligned}$$

are uniformly bounded. Indeed, by plugging in Eq. 3.4 possible values of the indicators $v_m(\alpha_i), v_{m-1}(\alpha_j), v_m(\alpha_j), j = 1, \dots, k$, we obtain from Eq. 3.2 that

$$\begin{aligned}
 |f^m - f^{m-1}| &\leq 1 + \max_{q, j=1, \dots, k} \sum_{l=1}^{N-m} |p_{qi}^{(l)} - p_{ji}^{(l)}| + \max_{j=1, \dots, k} p_{ji}^{(N-m+1)} \\
 &\leq 1 + 2R \sum_{l=1}^{N-m} \rho^l + \pi_i + R\rho^{N-m+1} \\
 &\leq 1 + 2R \sum_{l=1}^{N-m+1} \rho^l + \pi_i < 1 + \frac{2R\rho}{1-\rho} + \pi_i
 \end{aligned}$$

for $m = 1, \dots, N$. Next, we apply inequality in Remark 1 by Azuma (1967) to the martingale differences and, due to $f = f^N = \mathbf{E}(f|\mathcal{F}_N)$ and $\mathbf{E} f = f^0 = \mathbf{E}(f|\mathcal{F}_0) = 0$, obtain

$$\begin{aligned}
 P(\delta_i \geq \varepsilon) &= P(|N(\alpha_i) - \pi_i N| \geq \varepsilon N) = P(|f| \geq \varepsilon N) \\
 &= P\left(\left|\sum_{m=1}^N (f^m - f^{m-1})\right| \geq \varepsilon N\right) \leq 2 \exp(-\varepsilon^2 N \beta_i), \tag{3.5}
 \end{aligned}$$

where $\beta_i = 1/(2(1 + \frac{2R\rho}{1-\rho} + \pi_i)^2)$. At last step, application of Eq. 3.5 for $i = 1, \dots, k$ gives inequality (3.3). □

Remark 3.1 Theorems 3.1 remains true for a finite state Markov chain with one aperiodic ergodic class (thus, the unique stationary distribution will contain zero components corresponding to non-essential states) since such a chain is uniformly ergodic.

There exists an extensive literature on bounding geometric and subgeometric rate of convergence for Markov chains and Markov processes (see, e.g. Tuominen and Tweedie 1994; Roberts and Tweedie 1999; Fort and Roberts 2005, and Chapters 4-5 in Montenegro and Tetali 2006 and references in them).

To formulate the next theorem, we need the following definition.

Definition 3.2 A homogeneous Markov chain Y is called contracting if

$$\frac{1}{2} \sup_{\alpha, \beta} \|P(Y_2|Y_1 = \alpha) - P(Y_2|Y_1 = \beta)\| = \lambda < 1. \tag{3.6}$$

Here the supremum is taken over all pairs of states α, β of the Markov chain Y .

For the Markov chain X the condition (3.6) becomes

$$\frac{1}{2} \max_{i, j=1, \dots, k} \sum_{l=1}^k |p_{il} - p_{jl}| = \lambda < 1. \tag{3.7}$$

A proof of the following theorem uses the measure concentration result for Markov chains proved by Samson (2000); a similar statement was proved in Glynn and Ormoneit (2002).

Theorem 3.2 *If a homogeneous ergodic Markov chain X is also contracting with a contracting constant λ defined by Eq. 3.7, then for any $\varepsilon > 0$ we have*

$$P(\Delta \leq \varepsilon) \geq 1 - 2k \exp(-\varepsilon^2 N(1 - \sqrt{\lambda})^2/2) = D_2. \tag{3.8}$$

Proof First, we fix $i, i = 1, \dots, k$, and define a Markov chain \tilde{X} as follows: $\tilde{X}_i = 0$ if $X_i = \alpha_i$, and $\tilde{X}_i = a_j, j \neq i$, if $X_i = \alpha_j$, where $a_j, j \neq i$, are distinct numbers from $(0, 1]$. Then the Markov chain \tilde{X} is contracting (with the same λ) since \tilde{X} has the same matrix of transition probabilities as X . Now, we define a function $h : [0, 1]^N \rightarrow \mathbf{R}$ by $h(x_1, \dots, x_N) = N(0)/\sqrt{N}$, where $N(0)$ is the number of zeros among (x_1, \dots, x_N) . It is easy to check that h is a convex Lipschitz function on $[0, 1]^N$ with Lipschitz constant 1. According to Corollary 4 Samson (2000), for any positive t

$$P(|h(\tilde{X}) - \mathbf{E}h(\tilde{X})| \geq t) \leq 2 \exp(-t^2(1 - \sqrt{\lambda})^2/2). \tag{3.9}$$

Then, for any positive ε , the inequality

$$\begin{aligned} P(\delta_i \geq \varepsilon) &= P\left(\left|\frac{N(\alpha_i)}{N} - \pi_i\right| \geq \varepsilon\right) = P(|h(\tilde{X}) - \mathbf{E}h(\tilde{X})| \geq \varepsilon\sqrt{N}) \\ &\leq 2 \exp(-\varepsilon^2 N(1 - \sqrt{\lambda})^2/2) \end{aligned} \tag{3.10}$$

holds true. Finally, Eq. 3.10 implies the statement of the theorem. □

Now, let X be a homogeneous ergodic Markov chain with a uniform geometric rate of convergence with parameters R and ρ defined by Eq. 3.2. Let m be a minimum positive integer such that

$$\max_{i=1, \dots, k} \sum_{j=1}^k |p_{ij}^{(m)} - \pi_j| \leq R\rho^m < 1. \tag{3.11}$$

Then, Eqs. 3.11 and 3.7 together imply that an m -skeleton of Markov chain X (a Markov chain Y with transition probabilities $p_{ij}^{(m)}$) is contracting (with contracting constant $\lambda, \lambda \leq R\rho^m < 1$) even if X is not (i.e. $m \geq 2$).

Since Y has the same stationary distribution $\pi = (\pi_1, \dots, \pi_k)$ as the Markov chain X , and satisfies all conditions of Theorem 3.2, the following assertion holds true.

Theorem 3.3 *For a homogeneous ergodic Markov chain X with R , ρ and m defined by Eqs. 3.2 and 3.11, and $\varepsilon > 0$,*

$$P(\Delta' \leq \varepsilon) \geq 1 - 2k \exp\left(-\frac{\varepsilon^2 N(1 - \sqrt{R\rho^m})^2}{2m}\right) = D_3. \tag{3.12}$$

Here $\Delta' = \max_{i=1, \dots, k} |\hat{\pi}'_i - \pi_i|$ with $\hat{\pi}'_i$, $i = 1, \dots, k$, defined as a relative frequency of symbol α_i in the Markov chain Y , $Y = (X_1, X_{m+1}, X_{2m+1}, \dots, X_{\lfloor N/m \rfloor \times m+1})$.

Theorems 3.1, 3.2 and 3.3 provide exponential lower bounds D_1 , D_2 and D_3 on the rate of decay of the maximum error of the ML estimation of stationary probabilities π_i , $i = 1, \dots, k$, of a Markov chain X under different sets of conditions (in Theorem 3.3 the ML estimation is defined for m -skeleton of X). The inequalities (3.3), (3.8) and (3.12) also yield the k -dimensional confidence region for the stationary distribution (π_1, \dots, π_k) at the confidence level at least D_1 , D_2 , and D_3 , respectively.

Since for a homogeneous ergodic Markov chain X both inequalities (3.3) and (3.12) are valid, it is natural to find out which one provides the fastest rate of convergence. Obviously, an answer will depend on the values of constants m , R and ρ . Comparison of D_1 and D_3 shows that $D_1 > D_3$ if $\frac{\sqrt{m}}{1 - \sqrt{R\rho^m}} > 1 + a^* + \frac{2R\rho}{1 - \rho}$, while $D_3 > D_1$ if $\frac{\sqrt{m}}{1 - \sqrt{R\rho^m}} < 1 + a^* + \frac{2R\rho}{1 - \rho}$.

If, additionally, X is contracting with a contraction constant λ , then inequalities (3.3) and (3.8) hold true. Theorem 3.1 provides faster uniform rate of convergence (i.e. $D_1 > D_2$) if $\sqrt{\lambda} > \frac{(1-\rho)a^* + 2R\rho}{(1-\rho)(1+a^*) + 2R\rho}$; otherwise, $D_2 > D_1$, and Theorem 3.2 guarantees the faster uniform rate of convergence of ML estimates to the stationary probabilities.

Remark 3.2 Both ergodicity and the contraction property do not seem to be an excessively restrictive condition on a Markov chain X being the probabilistic model of an empirical biological sequence, such as DNA sequence. A DNA sequence $x = (x_1, \dots, x_N)$ normally possesses the property that $N(\alpha\beta) > 0$ for all $\alpha, \beta \in \mathcal{A}$. For the Markov chain X (with realization x) this property implies that all transition probabilities p_{ij} are positive. Therefore, X is ergodic as well as contracting.

Remark 3.3 Note that we do not obtain any lower bound for $P(|\hat{\pi}_i - \pi_i| \geq \varepsilon)$ from the asymptotic normality of π_i , $i = 1, \dots, k$ (such as inequality (2.7) for the independence model) since, to best of our knowledge, the existing results on the accuracy of the one- or multi-dimensional normal approximation for the Markov chains (e.g. Dembo and Zeitouni 1998; Saulis and Statulevičius 1991, 2000, and references thereof) do not cover our case. For example, the result on large deviations, Theorem 4 in Gudynas (2000), allows to estimate the error of the normal approximation for

$$P\left(\frac{N(\alpha_i)}{\sqrt{N}\sigma} \leq x\right) = P\left(\hat{\pi}_i - \pi_i \leq \frac{\sigma x}{\sqrt{N}}\right),$$

$$P\left(\frac{N(\alpha_i)}{\sqrt{N}\sigma} \leq -x\right) = P\left(\hat{\pi}_i - \pi_i \leq -\frac{\sigma x}{\sqrt{N}}\right),$$

where $\sigma^2 = \pi_i(1 - \pi_i) + \frac{\pi_i p_{ii}}{1 - p_{ii}} < \infty$ and positive $x = o(\sqrt{N})$. We, however, are interested in the values $x = \frac{\varepsilon\sqrt{N}}{\sigma} = O(\sqrt{N})$, that are not covered by this theorem.

3.2 ML Estimates of Transition Probabilities

The ML estimates $\hat{p}_{ij} = \frac{N(\alpha_i\alpha_j)}{N(\alpha_i\bullet)}$, $i, j = 1, \dots, k$, of the transition probabilities possess many useful properties. Billingsley (1961) proved that these estimates are consistent and that the k^2 -dimensional random vector $\eta = (\eta_{11}, \dots, \eta_{1k}, \dots, \eta_{k1}, \dots, \eta_{kk})$ with components

$$\eta_{ij} = \frac{N(\alpha_i\alpha_j) - N(\alpha_i)p_{ij}}{\sqrt{N(\alpha_i)}} = (\hat{p}_{ij} - p_{ij})\sqrt{N(\alpha_i)},$$

is an asymptotically centered normal vector with covariance matrix $R = (r_{ij,sl}) = (\mathbf{E}\eta_{ij}\eta_{sl})$, $i, j, s, l = 1, \dots, k$: $r_{ij,sl} = 0$ for $i \neq s, j, l = 1, \dots, k$; $r_{i,j} = p_{ij} - p_{ij}^2$ for $i, j = 1, \dots, k$; $r_{i,j,l} = -p_{ij}p_{il}$ for $i, j, l = 1, \dots, k, j \neq l$. The consistency of \hat{p}_{ij} implies that the error in estimation $\delta_{ij} = |\hat{p}_{ij} - p_{ij}|$, $i, j = 1, \dots, k$, converges to zero in probability and the same statement holds true for the maximum error in estimation over all transition probabilities, $\Delta = \max_{i,j=1,\dots,k} |\hat{p}_{ij} - p_{ij}|$. Below we obtain lower bounds on the rate of decay (in probability) of the maximum error Δ .

Theorem 3.4 *For a homogeneous ergodic Markov chain X and positive ε the following inequality holds true*

$$P(\Delta \leq \varepsilon) \geq 1 - 3k^2 \exp\left(-\frac{(\varepsilon\pi_*)^2 N}{2(1 + \varepsilon(1 + \frac{2R\rho}{1-\rho} + a^*))^2}\right) = E_1. \tag{3.13}$$

where $\pi_* = \min_i \pi_i$. The term a^* can be chosen to be equal to either $\pi^* = \max_i \pi_i$, or $p^* = \max_{i,j} p_{ij}$, or 1. The constants R, ρ determine a geometric rate of convergence to the stationary distribution π and are defined as in Eq. 3.2.

Proof To study the asymptotic behavior of Δ , we start with the estimation error δ_{ij} of the individual transition probability p_{ij} and once again turn to the method of bounded martingale differences from McDiarmid (1998). Let us select a pair of indices $i, j = 1, \dots, k$, and consider the family of increasing σ -fields $\mathcal{F}_m, m = 0, \dots, N$, the same as in proof of Theorem 3.1. Next, we define Doob’s martingale sequence $g^m = \mathbf{E}(g|\mathcal{F}_m), m = 0, \dots, N$, associated with the random variable

$$g = g_{ij} = N(\alpha_i\alpha_j) - N(\alpha_i)p_{ij} = \sum_{l=1}^{N-1} (v_l(\alpha_i\alpha_j) - v_l(\alpha_i)p_{ij}),$$

where $v_l(\alpha_i\alpha_j)$ stands for the indicator of occurrence of the two-letter word $\alpha_i\alpha_j$ starting at site l of the sequence X . Then the martingale differences

$$g^m - g^{m-1} = \mathbf{E}(g|\mathcal{F}_m) - \mathbf{E}(g|\mathcal{F}_{m-1}) = v_{m-1}(\alpha_i\alpha_j) - p_{ij}v_{m-1}(\alpha_i)$$

are uniformly bounded: $-p_{ij} \leq g^m - g^{m-1} \leq 1 - p_{ij}$, $m = 1, \dots, N$. To make use of these martingale differences, we will need the one-sided version of Eq. 3.5: for $\delta > 0$,

$$P(N(\alpha_i) \leq (\pi_i - \delta)N) \leq \exp\left(-\frac{\delta^2 N}{2(1 + \frac{2R\rho}{1-\rho} + \pi_i)^2}\right). \tag{3.14}$$

Next, after using the equalities $g = g^N$, $\mathbf{E}g = g^0 = \pi_i p_{ij}(N - 1) - \pi_i p_{ij}(N - 1) = 0$, and Eq. 3.14, we apply Theorem 3.12 McDiarmid (1998) to the martingale differences and derive

$$\begin{aligned} P(|\hat{p}_{ij} - p_{ij}| \geq \varepsilon) &= P(|N(\alpha_i \alpha_j) - \pi_i N(\alpha_i)| \geq \varepsilon N(\alpha_i)) = P(|g| \geq \varepsilon N(\alpha_i)) \\ &= P(|g| \geq \varepsilon N(\alpha_i), N(\alpha_i) \leq (\pi_i - \delta)N) \\ &\quad + P(|g| \geq \varepsilon N(\alpha_i), N(\alpha_i) > (\pi_i - \delta)N) \\ &\leq P(N(\alpha_i) \leq (\pi_i - \delta)N) + P(|g| \geq \varepsilon(\pi_i - \delta)N) \\ &\leq \exp\left(-\frac{\delta^2 N}{2(1 + \frac{2R\rho}{1-\rho} + \pi_i)^2}\right) + 2 \exp(-\varepsilon^2 N(\pi_i - \delta)^2/2). \end{aligned} \tag{3.15}$$

If we choose

$$\delta = \frac{\varepsilon \pi_i (1 + \frac{2R\rho}{1-\rho} + \pi_i)}{1 + \varepsilon (1 + \frac{2R\rho}{1-\rho} + \pi_i)},$$

then the inequality

$$P(\delta_{ij} \geq \varepsilon) = P(|\hat{p}_{ij} - p_{ij}| \geq \varepsilon) \leq 3 \exp\left(-\frac{(\varepsilon \pi_i)^2 N}{2(1 + \varepsilon (1 + \frac{2R\rho}{1-\rho} + \pi_i))^2}\right) \tag{3.16}$$

holds for any positive ε . Inequality (3.16) verifies the consistency of the ML estimate \hat{p}_{ij} of the transition probability p_{ij} , $i, j = 1, \dots, k$ and shows that the lower bound on the convergence rate of the estimation error δ_{ij} depends on the parameters of the Markov chain.

Finally, inequality (3.16) yields result (3.13). The proof is now complete. □

Theorem 3.5 *For a homogeneous ergodic Markov chain X which is also contracting with the contraction constant λ defined by Eq. 3.7, the decay rate of the maximum estimation error $\Delta = \max_{i,j=1,\dots,k} |\hat{p}_{ij} - p_{ij}|$ is given by*

$$P(\Delta \leq \varepsilon) \geq 1 - 3k^2 \exp\left(-\frac{(\varepsilon \pi_*)^2 N (1 - \sqrt{\lambda})^2}{2(1 - \sqrt{\lambda} + \varepsilon)^2}\right) = E_2, \tag{3.17}$$

where $\pi_* = \min_i \pi_i$.

Proof The proof is similar to the proof of Theorem 3.4. By applying the one-sided version of Eq. 3.10,

$$P(N(\alpha_i) \leq (\pi_i - \delta)N) \leq \exp\left(-\delta^2 N (1 - \sqrt{\lambda})^2 / 2\right), \tag{3.18}$$

repeating the same arguments as in Eq. 3.15, and selecting $\delta = \frac{\varepsilon\pi_i}{1-\sqrt{\lambda}+\varepsilon} > 0$, we arrive at the inequality

$$P(\delta_{ij} \geq \varepsilon) = P(|\hat{p}_{ij} - p_{ij}| \geq \varepsilon) \leq 3 \exp\left(-\frac{\varepsilon^2\pi_i^2 N(1-\sqrt{\lambda})^2}{2(1-\sqrt{\lambda}+\varepsilon)^2}\right). \tag{3.19}$$

Together Eq. 3.19 and $\pi_i \geq \pi_*$ imply Eq. 3.17. □

In Theorems 3.4 and 3.5 we derived lower bounds E_1 and E_2 on the rate of decay of the maximum error of the ML estimation of transition probabilities which hold true for a Markov chain under different sets of conditions. The inequalities (3.13) and (3.17) yield confidence intervals $(\hat{p}_{ij} - \varepsilon, \hat{p}_{ij} + \varepsilon)$ for parameters $p_{ij}, i, j = 1, \dots, k$, or, equivalently, the k^2 -dimensional confidence interval for k^2 -dimensional parameter (p_{11}, \dots, p_{kk}) at the confidence level at least E_1 and E_2 , respectively.

Which bound, E_1 or E_2 , provides the fastest rate of convergence if a homogeneous Markov chain is both ergodic and contracting, depends on values of the contraction constant λ and constants R and ρ in inequality (3.2) that determines the uniform geometric rate of convergence of $p_{ij}^{(n)}$ to the stationary probabilities $\pi_j, i, j = 1, \dots, k$. Comparison of E_1 and E_2 shows that $E_1 > E_2$ if $\sqrt{\lambda} > \frac{(1-\rho)a^*+2R\rho}{(1-\rho)(1+a^*)+2R\rho}$, while $E_2 > E_1$ if $\sqrt{\lambda} < \frac{(1-\rho)a^*+2R\rho}{(1-\rho)(1+a^*)+2R\rho}$. Note that the relationships between the bounds E_1 and E_2 are similar to the ones obtained for the bounds D_1 and D_2 on the rate of convergence of the maximum error of the ML estimates of the stationary probabilities.

Remark 3.4 For a Markov chain with $p_{ij} > 0, i, j = 1, \dots, k$, Ekisheva and Borodovsky (2006) have obtained a lower bound in the form:

$$\begin{aligned} P(\Delta \leq \varepsilon) &\geq \left(2\Phi(2\varepsilon\sqrt{N\pi_*}) - 1\right)^{k^2} - 2^{k^2-2}k^2(k-1)\exp(-2\varepsilon^2 N\pi_*) \\ &\geq 1 - \left\{ [k^2/2]! \frac{k^2}{2\sqrt{2\pi\pi_*N}\varepsilon} + 2^{k^2-2}k^2(k-1) \right\} \exp(-2\varepsilon^2 N\pi_*) \tag{3.20} \\ &= E_3. \end{aligned}$$

The proof has used the asymptotic normality of vector η and the inequality proved by Li and Shao (2002). A Markov chain with positive transition probabilities is both ergodic and contracting; therefore, the lower bounds E_1 and E_2 remain valid. Comparison of bounds E_1, E_2 and E_3 reveals that E_3 converges to one faster than $E_i, i = 1, 2$, for any values of R, ρ, π_*, a^* , and λ . Inequality (3.20), however, does not take into account an error of the multidimensional normal approximation and, therefore, should be used with caution (see also Remark 3.4 on the accuracy of Gaussian approximation for parameters of the Markov chain).

Finally, note that for a Markov chain with $p_{ij} > 0, i, j = 1, \dots, k$, term $\pi_* = \min_i \pi_i$ in formulas for lower bounds E_1, E_2, E_3 may be replaced by $p_* = \min_{i,j} p_{ij}$.

4 Hidden Markov Models

The results obtained in the previous sections can also be used to determine bounds on the rate of convergence in probability for the ML estimates of the HMM

parameters. Such bounds complement the obtained earlier results on consistency and asymptotic normality of these estimates (Petrie 1969; Bickel et al. 1998). We assume that both a sequence of symbols and the corresponding sequence of hidden states are experimentally determined and available to use for parameter estimation. For instance, for the HMM based algorithm for protein secondary structure prediction, the training set consists of protein sequences with known secondary structures. In another example, parameters of the HMM describing gene organization in DNA can be estimated from the set of genomic sequences with known (annotated) genes (Durbin et al. 1998, p. 62).

We consider an HMM with k_1 hidden states $1, 2, \dots, k_1$, emitting k_2 distinct symbols x_1, \dots, x_{k_2} . The parameters of the HMM are transition probabilities $p_{ij}, i, j = 1, \dots, k_1$, and emission probabilities $e_i(x_j), i = 1, \dots, k_1, j = 1, \dots, k_2$. We assume that the homogeneous Markov chain X of hidden states is both ergodic and contracting. The ML estimates \hat{p}_{ij} of the transition probabilities are the same as for a (non-hidden) Markov chain, and the ML estimates $\hat{e}_i(x_j)$ of emission probabilities $e_i(x_j), i = 1, \dots, k_1, j = 1, \dots, k_2$, are given by the equation

$$\hat{e}_i(x_j) = \frac{E_i(x_j)}{\sum_{l=1}^{k_2} E_i(x_l)},$$

where $E_i(x_j)$ designates the number of times that symbol x_j was emitted from hidden state i in the training sequence (Durbin et al. 1998, p. 62). A subsequence S^i of the training sequence that includes only symbols emitted from a given hidden state i is generated by the independence model with parameters $e_i(x_j), j = 1, \dots, k_2$.

- (1) Assuming that sequence S^i has length N_i , from inequality (2.1) for any $i = 1, \dots, k_1$ and $\varepsilon > 0$ we have

$$P(\max_{1 \leq j \leq k_2} |\hat{e}_i(x_j) - e_i(x_j)| \leq \varepsilon) \geq 1 - 2k_2 \exp(-2N_i \varepsilon^2).$$

Since sequences $S^i, i = 1, \dots, k_1$, are independent, the following inequality for the ML estimates of emission probabilities holds:

$$P(\max_{1 \leq i \leq k_1} \max_{1 \leq j \leq k_2} |\hat{e}_i(x_j) - e_i(x_j)| \leq \varepsilon) \geq \prod_{i=1}^{k_1} (1 - 2k_2 \exp(-2N_i \varepsilon^2)) = G_1. \quad (4.1)$$

- (2) To obtain a similar result for the transition probabilities, we turn to the underlying sequence of hidden states whose probabilistic model is the Markov chain X . The lower bounds derived in Theorems 3.4 and 3.5 are valid for Markov chain X . Therefore, inequalities (3.13) and (3.17) imply that for the maximum estimation error over all transition probabilities, we have

$$\begin{aligned} & P\left(\max_{1 \leq i, j \leq k_1} |\hat{p}_{ij} - p_{ij}| \leq \varepsilon\right) \\ & \geq 1 - 3k_1^2 \min\left\{\exp\left(-\frac{(\varepsilon\pi_*)^2 N(1 - \sqrt{\lambda})^2}{2(1 - \sqrt{\lambda} + \varepsilon)^2}\right), \exp\left(-\frac{(\varepsilon\pi_*)^2 N}{2(1 + \varepsilon(1 + \frac{2R\rho}{1-\rho} + a^*))^2}\right)\right\} \\ & = G_2. \end{aligned} \quad (4.2)$$

Here $\pi_* = \min_i \pi_i$, the term a^* can be chosen to be equal to either $\pi^* = \max_i \pi_i$, $p^* = \max_{i,j} p_{ij}$, or 1; and N is the length of the training sequence (equal to the length of the sequence of underlying hidden states). The constants R , ρ determine a geometric rate of convergence of $p_{ij}^{(n)}$ to the stationary distribution π and are defined as in Eq. 3.2, while the contraction coefficient λ is defined by Eq. 3.7.

Note that the lower bound G_1 depends on both k_1 and k_2 , while G_2 depends on k_1 only, as the Markov chain of hidden states is completely independent of emissions and does not depend on the size k_2 of the alphabet of emitted symbols. Formulas (4.1) and (4.2) can yield confidence intervals for true values of parameters $e_l(x_l)$, $l = 1, \dots, k_2$, $i = 1, \dots, k_1$, and p_{ij} , $i, j = 1, \dots, k_1$, at the specified confidence level G_m , $m = 1, 2$, respectively.

5 Conclusion

We have derived lower bounds on the rate of convergence in probability of the maximum error Δ for ML estimates of parameters of the statistical models frequently used in applications, particularly in bioinformatics: the independence model, the Markov chain model, and the HMM. For these models, the inequalities (2.1), (2.9), (3.3), (3.8), (3.13), (3.17), (3.20)–(4.2) provide exponential lower bounds, while inequality (2.8) gives a polynomial bound. These inequalities also yield the uniform confidence intervals for unknown values of parameters at a specified confidence level. Lower bounds B_2 , B_3 and E_3 are derived from the normal approximation for the ML estimates, while the other lower bounds are found from the concentration properties of the corresponding probability measures. If more than one lower bound is obtained for the maximum estimation error for parameters of a particular model (e.g. $B_1 - B_3$ for the probabilities of symbols in the independence model), we compare the lower bounds in order to identify the best (highest) of them. Finding the bounds on the decay rate of the maximum error in estimation of parameters of several stochastic models constitutes a new theoretical result revealing yet another property of the maximum likelihood estimates.

A possible important application of Theorems 2.1, 3.1–3.5 is an evaluation of the error in estimation of the probability of biological sequence fragments or the logarithm of this probability used in many bioinformatic algorithms (e.g. described in Lawrence et al. 1993; Borodovsky and McIninch 1993; Burge and Karlin 1997; Durbin et al. 1998; Borodovsky and Ekisheva 2006). The bound of the error is obtained in Eq. 2.12 from Theorem 2.1 for the independence model. Similarly, such bounds can be derived for a Markov chain and an HMM.

Acknowledgements The authors thank Dr. Gennadii Chistyakov for helpful discussions and Dr. Alexander Mitrophanov for critical comments. We are grateful to an anonymous reviewer whose helpful suggestions led to an introduction of Theorem 3.3 as well as other improvements of the text.

This work was supported in part by the US National Institutes of Health grant awarded to MB.

References

- Almagor H (1983) A Markov analysis of DNA sequences. *J Theor Biol* 104:633–645
 Azuma K (1967) Weighted sums of certain dependent random variables. *Tôhoku Math J* 19:357–367

- Bickel PJ, Ritov Y, Rydén T (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann Stat* 26:1614–1635
- Billingsley P (1961) Statistical methods in Markov chains. *Ann Math Stat* 32:12–40
- Borodovsky M, Ekisheva S (2006) Problems and solutions in biological sequence analysis. Cambridge University Press, Cambridge
- Borodovsky MY, Sprizhitsky YA, Golovanov EI, Alexandrov AA (1986a) Statistical patterns in the primary structure of the functional regions of the *Escherichia coli* genome. I. Frequency characteristics. *Mol Biol* 20:826–833 (English translation)
- Borodovsky MY, Sprizhitsky YA, Golovanov EI, Alexandrov AA (1986b) Statistical patterns in the primary structure of the functional regions of the *Escherichia coli* genome. II. Nonuniform Markov models. *Mol Biol* 20:833–840 (English translation)
- Borodovsky M, McIninch J (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput Chem* 17:123–133
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268:78–94
- Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann Math Stat* 23:493–509
- Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* 51:79–94
- Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London
- Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd edn. Springer, New York
- Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge
- Ekisheva S, Borodovsky M (2006) Probabilistic models for biological sequences: selection and maximum likelihood estimation. *Int J Bioinformatics Res Appl* 2:305–324
- Feller W (1945) On the normal approximation to the binomial distribution. *Ann Math Stat* 16: 319–329
- Fort G, Roberts GO (2005) Subgeometric ergodicity of strong Markov processes. *Ann Appl Probab* 15:1565–1589
- Gatlin LL (1972) Information theory and the living system. Columbia University Press, New York
- Glynn PW, Ormoneit D (2002) Hoeffding's inequality for uniformly ergodic Markov chains. *Stat Probab Lett* 56:143–146
- Gudynas P (2000) Refinements of the central limit theorem for homogeneous Markov chains. In: Prokhorov YV, Statulevičius V (eds) Limit theorems of probability theory. Springer, Berlin, pp 167–183
- Karlin S, Burge C, Campbell AM (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res* 20:1363–1370
- Karlin S, Macken C (1991) Assessment of inhomogeneities in an *E. Coli* physical map. *Nucleic Acids Res* 19:4241–4246
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262: 208–214
- Li WV, Shao Q-M (2002) A normal comparison inequality and its applications. *Probab Theory Relat Fields* 122:494–508
- McDiarmid C (1998) Concentration. In: Probabilistic methods for algorithmic discrete mathematics. Algorithms in combinatorics, vol 16. Springer, Berlin, pp 195–248
- Meyn SP, Tweedie RL (1993) Markov chains and stochastic stability. Springer, London
- Mitrophanov AY (2005) Sensitivity and convergence of uniformly ergodic Markov chains. *J Appl Probab* 42:1003–1114
- Mitrophanov AY, Lomsadze A, Borodovsky M (2005) Sensitivity of hidden Markov models. *J Appl Probab* 42:632–642
- Montenegro R, Tetali P (2006) Mathematical aspects of mixing times in Markov chains. In: Sudan M (ed) Book in series *foundations and trends in theoretical computer science*, vol 1:3. NOW, Boston
- Nagaev SV (1965) Some limit theorems for large deviations. *Theor Probab Appl* 10:214–235
- Osipov LV (1967) Asymptotic expansion in the central limit theorem. *Vestn Leningr Univ Ser I* 19:45–62 (in Russian)
- Petrie T (1969) Probabilistic functions of finite state Markov chains. *Ann Math Stat* 40:97–115
- Roberts GO, Tweedie RL (1999) Bounds on regeneration times and convergence rates for Markov chains. *Stoch Process their Appl* 80:211–229

- Samson P-M (2000) Concentration of measure inequalities for Markov chains and ϕ -mixing processes. *Ann Probab* 28:416–461
- Saulis L, Statulevičius VA (1991) Limit theorems for large deviations. Kluwer Academic, Dordrecht
- Saulis L, Statulevičius VA (2000) Limit theorems on large deviations. In: Prokhorov YV, Statulevičius V (eds) *Limit theorems of probability theory*. Springer, Berlin, pp 185–266
- Tavaré S, Song B (1989) Codon preference and primary sequence structure in protein coding regions. *Bull Math Biol* 51:95–115
- Tuominen P, Tweedie RL (1994) Subgeometric rates of convergence of f -ergodic Markov chains. *Adv Appl Probab* 26:775–798