TESIS DE MAESTRÍA EN MATEMÁTICA

Model Selection Techniques & Sparse Markov Chains

Nicolás Fraiman

Orientador: Gonzalo Perera

26 de Junio, 2008
Montevideo
Uruguay

PEDECIBA
Facultad de Ciencias
Universidad de la República

Resumen

Este trabajo trata sobre problemas de selección de modelo. El capítulo 0 plantea un estudio general de estos problemas estadísticos. Dados un proceso estocástico y una familia de clases de modelos, con cada clase determinada por un parámetro de estructura y cada modelo dentro de una clase descrito por un vector de parámetros en un espacio cuya dimensión depende de la estructura. Supongamos que dada una realización del proceso podemos estimar el vector de parámetros si la estructura es conocida. La tarea es estimar esta última.

Trabajamos usando el concepto de criterio de información, el parámetro de estructura es estimado mediante minimizar un valor asignado a cada clase de modelos. Los criterios más utilizados son el *Criterio de Información Bayesiano* (BIC) y el principio del *mínimo largo de descripción* (MDL). El BIC consiste de dos términos: menos el logaritmo de la máxima verosimilitud, esto mide la bondad de ajuste; y la mitad del número de parámetros libres por el logaritmo del tamaño muestral, esto penaliza modelos muy complejos.

En el capítulo 2, incluimos algunos resultados recientes en estimación de cadenas de Markov de alcance variable (VLMC), los cuales nos ayudarán a entender más en profundidad el problema planteado. Basados en Csiszár y Talata (2006) extendemos el concepto de árbol de contextos para procesos ergódicos arbitrarios y demostramos que los principios BIC y MDL dan estimadores fuertemente consistentes del árbol de contextos.

En el capítulo 3 presentamos una nueva e ingeniosa representación de los modelos Markovianos: los modelos de árbol de contexto disperso (STMs), una generalización de las cadenas de alcance variable, donde permitimos juntar conjuntos más generales de estados con distribuciones similares, y preservamos la útil estructura combinatoria de los árboles de contextos. El tema principal del trabajo es estudiar un método para estimar la estructura en esta clase de modelos parsimoniosos. Mostraremos resultados de consistencia para estimadores basados en el principio MDL, el objetivo es encontrar el menor árbol que determina las probabilidades de transición.

Finalmente, en el capítulo 4 describimos brevemente algunas aplicaciones en Biología y Teoría de la Información. Ilustramos como estas técnicas pueden ser utilizadas para clasificar familias de proteinas. Además mostramos como se pueden utilizar para comprimir imágenes bitonales, dando lugar a un método de compresión sin pérdida que mejora la performance de los métodos basados en árboles de contexto, y de varios algoritmos populares de compresión.

Palabras claves: Modelos de árboles dispersos, Cadenas de alcance variable, Mínimo largo de descripción, Árbol de contextos, Consistencia.

Abstract

The dissertation deals with model selection problems. Chapter 1 is a survey of these statistical problems. They can be formulated as follows. Let a stochastic process and a family of model classes be given, each class determined by a structure parameter. Each model in a class is described by a parameter vector from a subset of an Euclidean space whose dimension depends on the structure parameter. Suppose that based on a realization of the process, the statistical sample, we can estimate the parameter vector provided the structure parameter is known. The task is estimation of the latter.

We treat the model selection problem using the concept of information criterion, the structure parameter is estimated by minimizing a real number assigned to each model class. The mostly used information criteria are the *Bayesian Information Criterion* (BIC) and the *Minimum Description Length* (MDL). The BIC consists of two terms: the first one is the negative logarithm of the maximum likelihood, this measures the goodness of fit; the second term is half the number of free parameters in the model class times the logarithm of the sample size, this penalizes too complex models.

In Chapter 2, we include some recent results in estimation of variable length Markov chains (VLMC), that help to get a deeper understanding of the problem. Based on Csiszár and Talata (2006) the concept of context tree is extended to arbitrary stationary ergodic processes and the BIC and MDL principles are shown to be strongly consistent estimators of the context tree.

In Chapter 3 we present a new ingenious parametrization of Markov Models: the *Sparse Tree Models* (STMs), a generalization of tree models where more general sets of states with similar conditional distributions are allowed to merge, while preserving the attractive combinatorial properties of the tree structure. The main topic of this work is the study of a method to capture the "essential structure" within this class of parsimonious Markov models. Consistency results are shown for MDL estimation, where the objective is to find the smallest sparse tree needed to determine the transition probabilities.

Finally, in Chapter 4 we briefly present some applications in Biology and Information Theory. In the first case we illustrate how this techniques can be used to classify families of proteins. On the second case we show how STMs can be used to compress bi-level images, developing a method of Universal Source Coding which significantly outperforms conventional tree models, and some popular standard existing algorithms.

Keywords: Sparse Tree Models, Minimum Description Length Principle, Variable length Markov chains, Estimation of context trees, Strong consistency.

Agradecimientos

A Gonzalo por su incondicionalidad.
A Beatriz porque sin su apoyo este trabajo no hubiera sido posible.
A mis viejos por estar siempre.
A Amaia, Virgin, Ceci, Janine y JP por borrar nubes y dibujar risas y locuras.
A Yaiza por hacer que todo tenga sentido.

Table of Contents

CHAPTER 0

Introducción

0.1 El problema de selección de modelos

Sea $\{X_t, t \in T\}$ un proceso estocástico dado, donde cada X_t es una variable aleatoria con valores $a \in A$, y T es un conjunto de índices. A la distribución conjunta de las variables aleatorias X_t, $t \in T$ nos vamos a referir como la distribución del proceso y la denotaremos por Q. Un *modelo* del proceso determina una distribución, o una colección de distribuciones, hipotéticas sobre el proceso. Típicamente, un modelo está determinado por un parámetro de estructura k con valores en algún conjunto $\mathcal{K}$, y por un vector de parámetros $\theta_k \in \Theta_k \subset \mathbb{R}^{d_k}$; a un modelo tal lo denotaremos por $\mathcal{M}_{\theta_k}$. Dados los posibles modelos de un proceso, estos se pueden ordenar en clases de modelos de acuerdo a su parámetro de estructura: $\mathcal{M}_k = \{M_{\theta_k}, \theta_k \in \Theta_k \subset \mathbb{R}^{d_k}\}$. La inferencia estadística de un proceso está basada en una realización $\{x_t, t \in T\}$ del proceso observado en un rango $R_n \subset T$, donde R_n se va extendiendo con n. Por lo tanto, la n-ésima muestra es $x(n) = \{x_t, t \in R_n\}$. Algunos ejemplos típicos de procesos y sus modelos son descritos a continuación.

En el caso de *estimación de densidades*, $T = \mathbb{N}$ y las variables aleatorias X_t, $t \in \mathbb{N}$ son independientes e idénticamente distribuidas (i.i.d.) con densidad f_{θ_k}. La n-ésima muestra es $\{x_i, i = 1, \ldots, n\}$.

El problema de *ajuste polinomial* involucra un $T \subset \mathbb{R}$, que es un subconjunto numerable, $A = \mathbb{R}$, y el modelo

$$X_t = \theta_k[0] + \theta_k[1]t + \theta_k[2]t^2 + \cdots + \theta_k[k-1]t^{k-1} + Z_t,$$

donde Z_t, $t \in T$ son variables aleatorias independientes con distribución normal, de media cero, y varianza común desconocida, y $\theta_k[i]$ es el i-ésimo componente del vector de parámetros k-dimensional θ_k. En este caso, el parámetro de estructura $k \in \mathbb{N}$ es el grado del polinomio $\theta_k[0] + \theta_k[1]t + \theta_k[2]t^2 + \cdots + \theta_k[k-1]t^{k-1}$ más 1, y la n-ésima muestra es $\{x_t, t \in \{t_1, \ldots, t_n\} \subset T\}$.

Un proceso con $T = \mathbb{N}$, $A = \mathbb{R}$ se dice que es un *proceso autoregresivo (AR)* de orden k si

$$X_t = \sum_{i=1}^{k} a_i X_{t-i} + Z_t,$$

donde Z_t, $t \in \mathbb{N}$ son variables aleatorias independientes con distribución normal, media cero, varianza común desconocida, y $a_i \in \mathbb{R}$, $i = 1, \ldots, k$ forman el vector de parámetros θ_k. Aquí el parámetro de estructura $k \in \mathbb{N}$ es el número de coeficientes a_i, y la n-ésima muestra es $\{x_i, i = 1, \ldots, n\}$.

El *proceso autoregresivo con media móvil (ARMA)* es similar al proceso AR. En este caso tenemos que

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + Z_t + \sum_{j=1}^{q} b_j Z_{t-j}.$$

El vector de parámetros es $\theta_k = \{a_1, \ldots, a_p, b_1, \ldots, b_q\} \in \mathbb{R}^{p+q}$, y el parámetro de estructura k tiene dos componentes: $k = (p, q) \in \mathbb{N}^2$.

Un proceso con $T = \mathbb{N}$, $|A| < \infty$ es una *Cadena de Markov* de orden k si

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{i=k+1}^{n} Q(x_i | x_{i-k}^{i-1}), \quad n \geq k,\ x_1^n \in A^n, \tag{0.1}$$

con probabilidades de transición $Q(\cdot|\cdot)$ apropiadas. El símbolo x_i^j denota la secuencia $x_i, x_{i+1}, \ldots, x_j$. Como para cada $a_1^k \in A^k$ el vector $\{Q(a|a_1^k), a \in A\}$ es una distribución de probabilidad en A, el vector de parámetros $\theta_k \in \mathbb{R}^{d_k}$ consiste de $d_k = (|A| - 1)|A|^k$ probabilidades de transición $Q(a|a_1^k)$, $a \in A^\star$, $a_1^k \in A^k$, donde $|A^\star| = |A| - 1$. En este caso, el parámetro de estructura $k \in \mathbb{N}$ es el largo de la secuencia de la cual dependen las probabilidades de transición (en su segundo argumento). La muestra n-ésima es $\{x_i, i = 1, \ldots, n\}$.

Los procesos AR y ARMA, y las cadenas de Markov son ejemplos para los cuales la clase de modelos no determina una única distribución hipotética del proceso. En particular, para los procesos AR o las cadenas de Markov de orden k el modelo únicamente determina una distribución condicional para $X_{k+1}, X_{k+2}, \ldots$ dados $X_1, \ldots, X_k$.

El conjunto $\mathcal{K}$ de parámetros de estructura factibles k es un conjunto ordenado o parcialmente ordenado con respecto a la inclusión de las clases de modelos $\mathcal{M}_k$.

Cuando el modelo M_{θ_k} con parámetro de estructura k se corresponde con la verdadera distribución Q del proceso, un modelo más complejo con un parámetro de estructura k' más grande (en el sentido explicado arriba) puede también corresponder a la distribución Q con un vector de parámetros $\theta_{k'}$ apropiado. Por ejemplo, cualquier proceso AR o cadena de Markov de orden k es también de orden k', para cada $k' > k$. Vamos a referirnos como *verdadero modelo* M_{θ_0} al modelo minimal dentro de aquellos que se corresponden a la distribución Q del proceso (estos son comúnmente llamados modelos fieles), es decir, aquél modelo para el cual no existe otro modelo con la misma propiedad pero que tenga un parámetro de estructura más pequeño. El parámetro de estructura de tal modelo lo denotaremos por k_0.

El *problema de selección de modelo* consiste en estimar el verdadero parámetro de estructura k_0 basado en la observación estadística $x(n)$ del proceso.

El término *subestimación* refiere al caso cuando el parámetro de estructura k elegido es más chico que el verdadero parámetro k_0. En tal caso $\theta_0 \notin \Theta_k$, y por lo tanto el verdadero modelo no puede ser estimado con precisión; la estimación del vector de parámetros va a involucrar sesgo.

El término *sobreestimación* refiere al caso cuando el parámetro de estructura k elegido es más grande que el verdadero parámetro k_0. En este caso $M_{\theta_0} \in \mathcal{M}_{k_0} \subset \mathcal{M}_k$, y entonces $M_{\theta_0} = M_{\theta_k}$ para algún $\theta_k \in \Theta_k$, pero θ_k tiene más componentes que θ_0, esto implica que es más difícil estimar el verdadero escenario; la estimación del vector de parámetros va a tener una varianza más grande.

Esta tesis trata el problema de selección de modelo usando el concepto de criterio de información. Un *criterio de información* (IC) basado en la muestra $x(n)$ asigna un valor real a cada clase de modelos $\mathrm{IC} : \mathcal{K} \times \{x(n)\} \to \mathbb{R}$, y el estimador de k_0 es el parámetro de estructura con el menor valor del criterio:

$$\hat{k}(x(n)) = \arg\min_{k \in \mathcal{K}} \mathrm{IC}_k(x(n)).$$

Las siguientes secciones dan una descripción general sobre distintos criterios de información.

0.2 Reseña histórica

El problema de selección de modelo puede ser entendido como un problema de múltiples test de hipótesis, y el *test de cociente de verosimilitud* de Neyman y Pearson (1928) puede ser utilizado. Anderson estudio este procedimiento para el problema de ajuste polinomial (Anderson, 1962) y para procesos AR (Anderson, 1963). Estos procedimientos son secuencias de tests tomando los ordenes hipotéticos sucesivamente, comenzando por el más grande. La principal desventaja de estos procedimientos es la elección subjetiva de los niveles de significación de los tests para todos los ordenes hipotéticos del modelo.

Mallows (1964, 1973) introdujo, para seleccionar las verdaderas variables de modelos lineales, un método similar a un criterio de información. Consideremos el *modelo lineal*

$$X_t = \sum_{i=1}^{K} a_i u_{it} + a_0 + Z_t, \quad t \in \mathbb{Z},$$

donde a_i, $i = 1, \ldots, K$ son los parámetros del modelo, u_{it}, $i = 1, \ldots, K$ son variables (no aleatorias) independientes cuyos valores son dados en $t = 1, \ldots, n$, y las Z_t son variables aleatorias independientes con media nula y varianza desconocida σ^2. Dada la muestra $x(n) = \{x_t, t = 1, \ldots, n\}$, el problema consiste en estimar el conjunto $\{u_{i1}, \ldots, u_{ik}\}$ de variables del cual X_t efectivamente depende, es decir, $a_{i_\ell} \neq 0$ para $i_\ell \in \{i_1, \ldots, i_k\}$ y $a_{i_\ell} = 0$ sino.

Mallows le asigno a cada conjunto de índices $P = \{i_1, \ldots, i_k\}$ el valor

$$C_P = \frac{1}{\hat{\sigma}^2} \operatorname{RSS}_P - n + |P|,$$

donde RSS_P es la suma de los cuadrados de los residuos con respecto a P:

$$\mathrm{RSS}_P = \min_{a_{i_\ell}, i_\ell \in P} \sum_{t=1}^{n} \left(x_t - \sum_{a_{i_\ell}, i_\ell \in P} a_{i_\ell} u_{i_\ell t} - a_0 \right)^2,$$

además $\hat{\sigma}^2$ es un estimador adecuado de σ^2, por ejemplo, $\hat{\sigma}^2 = \mathrm{RSS}_{1,\ldots,k} / (n - k)$. El estimador del conjunto de índices P es aquél cuyo valor de C_P sea mínimo. Se puede probar que el valor esperado de C_P es igual a $|P|$ cuando P es el verdadero conjunto de índices, y que es mayor en cualquier otro caso.

Para procesos estacionarios, Davisson (1965) analizó el error cuadrático medio de predicción en el modelo AR de orden k, cuando los coeficientes del modelo están determinados por las n observaciones pasadas $x_1, \ldots, x_n$ y el modelo es utilizado para predecir la próxima observación. A saber, para el predictor $\hat{X}_n(k) = \sum_{i=1}^{k} \hat{a}_i X_{n-i}$ con coeficientes que minimizan el error cuadrático medio, o sea,

$$\{\hat{a}_1, \ldots, \hat{a}_k\} = \arg \min_{a_1, \ldots, a_k} \sum_{t=0}^{n} \left(x_t - \sum_{i=1}^{k} a_i x_{t-i} \right)^2,$$

el obtuvo

$$\mathbf{E} \left(X_n - \hat{X}_n(k) \right)^2 = \sigma^2(k) \left(1 + \frac{k}{n} \right) + o(1/n),$$

donde $\sigma^2(k)$ es el error cuadrático medio asintótico. Más aún, propuso utilizar el término principal de esta expresión para estimar el verdadero orden, mediante minimizarlo sobre los órdenes candidatos. Es claro que esto requiere la estimación de $\sigma^2(k)$.

Akaike (1970) encontró el mismo resultado, y pudo solucionar el problema de estimar $\sigma^2(k)$ con un método de estimación espectral adecuado. Definió un criterio llamado *error de predicción final* como

$$\mathrm{FPE}_k(x_1^n) = \frac{n+k}{n-k}\left(\hat{C}_0 - \hat{a}_1\hat{C}_1 - \cdots - \hat{a}_k\hat{C}_k\right),$$

donde $\hat{C}_i = (1/n)\sum_{t=1}^{n-1} x_{t+i}x_t$, $i = 0, \ldots, k$ son los coeficientes de correlación empíricos, y $\hat{a}_i$, $i = 1, \ldots, k$ son los coeficientes del modelo que minimizan el error cuadrático medio de predicción, como arriba. Éstos últimos valores pueden ser calculados a partir de los $\hat{C}_i$, resolviendo la ecuación de Yule-Walker. El estimador del orden es

$$\hat{k}(x_1^n) = \arg\min_{0\leq k\leq K} \mathrm{FPE}_k(x_1^n).$$

El único elemento subjetivo en este procedimiento es la determinación de la cota superior K de los ordenes candidatos. Akaike también demostró que este estimador sobreestima el verdadero orden asintóticamente con probabilidad positiva, es decir,

$$\liminf_{n\to\infty} Q\left(\hat{k}(x_1^n) > k_0\right) > 0.$$

Akaike (1972) introdujo un concepto general para resolver el problema de selección de modelo. Asumamos que cada modelo M_{θ_k} especifica una única distribución P_{θ_k} del proceso, y denotemos por $P_{\theta_k}^{(n)}$ a la distribución marginal de la muestra $x(n)$ (esta marginal respecto a la distribución del proceso, no es otra cosa que la distribución conjunta de la muestra). La *divergencia de Kullback-Leibler* entre $P_{\theta_k}^{(n)}$ y $P_{\theta_0}^{(n)}$ está dada por

$$\mathbf{D}(P_{\theta_0}^{(n)} \,\|\, P_{\theta_k}^{(n)}) = \int f_{\theta_0}^{(n)}(x(n)) \log \frac{f_{\theta_0}^{(n)}(x(n))}{f_{\theta_k}^{(n)}(x(n))} \lambda(dx(n)),$$

donde $f_{\theta_k}^{(n)}$ denota la densidad de $P_{\theta_k}^{(n)}$ con respecto a una medida dominante λ (típicamente, λ es la medida de Lebesgue o, en el caso discreto, la medida de conteo). Akaike apunto a minimizar esta cantidad para estimar el verdadero vector de parámetros θ_0 y el verdadero parámetro de estructura k_0. Encontró que este minimizador puede ser aproximado tomando los estimadores de máxima verosimilitud $\hat{\theta}_k = \arg\max_{\theta_k\in\Theta_k} f_{\theta_k}^{(n)}(x(n))$ en cada clase de modelos candidata, y luego seleccionando la clase de modelos cuyo parámetro de estructura minimice el valor

$$\mathrm{AIC}_k(x(n)) = -\log f^{(n)}_{\theta_k}(x(n)) + \dim \Theta_k.$$

Cuando los modelos no determinan unívocamente la distribución del proceso, se puede definir el AIC de manera similar, con funciones $f^{(n)}_{\theta_k}$ definidas apropiadamente. Por ejemplo, en el caso de los procesos AR de orden k podemos prefijar que $X_1, \ldots, X_k$ valgan 0, o fijar su distribución como la correspondiente marginal de la distribución estacionaria del proceso. Esto especifica una única distribución conjunta correspondiente al modelo, y podemos tomar su densidad como $f^{(n)}_{\theta_k}$. Notemos que una restricción adecuada del espacio de parámetros Θ_k puede garantizar la existencia de la distribución estacionaria. En el caso de las cadenas de Markov de orden k, podemos proceder de forma similar, o podemos definir $f^{(n)}_{\theta_k}$ como el el lado derecho de la igualdad (0.1) eliminando el factor $Q(X_1^k = x_1^k)$.

Este procedimiento de selección de modelo tiene una interpretación clara. El primero término del criterio de información es menos el logaritmo de la máxima verosimilitud. Mide la bondad de ajuste de la muestra a la clase de modelos $\mathcal{M}_k$. Este término disminuye cuando la complejidad del modelo aumenta. El segundo término del criterio de información, llamado el *término de penalización*, es el número de parámetros libres del modelo. Esto penaliza modelos demasiado complejos: aumenta con la complejidad del modelo. Por lo tanto, el modelo seleccionado tiene un buen compromiso entre una buena descripción de los datos y la complejidad del modelo (para no sobreajustar los datos).

Para los modelos AR, el AIC es asintóticamente (cuando el tamaño de la muestra tiene a infinito) idéntico al criterio FPE (Akaike, 1972, 1974). Por lo tanto, el estimador AIC también sobrestima el verdadero parámetro de estructura asintóticamente con probabilidad positiva. Shibata (1976) derivo la distribución asintótica exacta del orden seleccionado por estos estimadores.

El principio clásico de *validación cruzada* puede ser adaptado al problema de selección de modelo (ver por ejemplo, Stone 1974). El principio general requiere dividir el conjunto muestral en dos partes, y realizar la estimación del modelo con una de ellas solamente. Utilizando el otro subconjunto , el modelo candidato puede ser validado correctamente, esto es, la estimación y la validación van a ser independientes. Una formulación de este principio para el problema de ajuste polinomial es la siguiente. Dividimos la n-ésima muestra $x(n) = \{x_t, t = t_1, \ldots, t_n\}$ en subconjuntos dejando el p-ésimo elemento de lado: $x(n) = x_{\setminus p} \cup \{x_{t_p}\}$, donde $x_{\setminus p} = x(n) \setminus \{x_{t_p}\}$. Estimamos los coeficientes del polinomio de grado $k-1$ basados en la submuestra $x_{\backslash p}$:

$$\hat{\theta}_k^{(p)} = \arg\min_{\theta_k \in \Theta_k} \sum_{i \in \{1,\ldots,n\} \backslash \{p\}} (x_t - (\theta_k[0] + \theta_k[1]t_i + \cdots + \theta_k[k-1]t_i^{k-1}))^2,$$

y validamos basados en el p-ésimo elemento de la muestra x_{t_p}:

$$e_k(p) = x_{t_p} - (\theta_k[0] + \theta_k[1]t_p + \theta_k[2]t_p^2 + \cdots + \theta_k[k-1]t_p^{k-1}).$$

Calculamos este error de predicción para cada p, y minimizamos

$$e_k^2 = \sum_{p=1}^{n} e_k(p)^2$$

sobre los hipotéticos k para obtener el grado estimado del polinomio. Stone (1977) probó que el criterio de validación cruzada es asintóticamente equivalente al AIC.

0.3 Selección de modelo consistente

En este trabajo, la calidad de los métodos de selección de modelo va a ser considerada solamente desde el punto de vista asintótico; en la literatura este aspecto es el central.

Un estimador $\hat{k}(x(n))$ del parámetro de estructura k basado en la muestra $x(n)$ se dice que es *consistente* si la probabilidad de que el estimador sea igual al verdadero parámetro estructural k_0 tiende a 1 cuando el tamaño de la muestra n tiende a infinito:

$$Q\left(\hat{k}(x(n)) = k_0\right) \to 1 \qquad \text{si } n \to \infty.$$

Un estimador $\hat{k}(x(n))$ se dice que es *fuertemente consistente* si es igual al parámetro de estructura k_0 eventualmente casi seguro cuando el tamaño de la muestra n tiende a infinito:

$$\hat{k}(x(n)) = k_0, \qquad \text{eventualmente casi seguro para } n \to \infty.$$

Aquí y en el resto de la tesis, “eventualmente casi seguro” quiere decir que con probabilidad 1 existe un n_0 (dependiente de la realización $\{x_t, t \in T\}$) tal que la afirmación vale para todo $n \geq n_0$.

Para el caso de estimación de densidades, cuando las funciones de densidad factibles pertenecen a *familias exponenciales*, Schwarz (1978) derivó un criterio de información de la aproximación asintótica al estimador Bayesiano por Máximo A-Posteriori (MAP). Supongamos que la clase de modelos $\mathcal{M}_k$ consiste de funciones de densidad de la forma

$$f_{\theta_k}(x_i) = \exp(\pi\theta_k y_k(x_i) - b_k(\theta_k)), \quad \theta_k \in \Theta_k,$$

donde $\pi\cdot\cdot$ denota el producto interno en el espacio euclidiano $d_k = \dim \Theta_k$ dimensional, $y_k : \mathbb{R} \to \mathbb{R}^{d_k}$ son funciones dadas, y

$$b_k(\theta_k) = \log \int \exp(\pi\theta_k y_k(x_i))dx_i.$$

En este caso k varía sobre un conjunto finito $\mathcal{K}$. Una distribución a priori sobre el vector de parámetros puede ser escrita de la forma $\mu = \sum_{k\in\mathcal{K}} \alpha_k \mu_k$, donde α_k es la probabilidad a priori de que el modelo con parámetro estructural k sea el verdadero modelo, y μ_k es la distribución condicional a priori de θ_k bajo la condición de que el parámetro de estructura verdadero sea k; μ_k está concentrada en Θ_k. Schwarz demostró que, bajo ciertas hipótesis de regularidad, el estimador MAP del vector de parámetros θ_k a partir de una muestra i.i.d. x_1^n asintóticamente no depende de μ, y es equivalente al estimador de máxima verosimilitud $\hat{\theta}_k = \arg\max_{\theta_k\in\Theta_k} f_{\theta_k}^{(n)}(x(n))$ en la clase de modelos $\mathcal{M}_k$ cuyo parámetro de estructura k minimiza el valor

$$\mathrm{BIC}_k(x(n)) = -\log f_{\theta_k}^{(n)}(x(n)) + \frac{\dim \Theta_k}{2} \log n$$

sobre el conjunto $\mathcal{K}$. Este valor es llamado el *Criterio de Información Bayesiano (BIC)*.

La consistencia del estimador BIC en la situación descrita arriba fue probada por Haughton (1988). Es importante notar que para el problema de ajuste polinomial, Akaike (1977) introdujo el mismo criterio de información con la misma notación del BIC, de una manera heurística.

Para el modelo AR de orden k el criterio de información Bayesiano tiene la siguiente forma:

$$\mathrm{BIC}_k(x_1^n) = -\log f_{\theta_k}^{(n)}(x_1^n) + \frac{k}{2} \log n.$$

Hannan and Quinn (1979) probaron que el estimador BIC del orden de un proceso AR es fuertemente consistente. Para el modelo ARMA de orden (p, q) el BIC tiene una forma similar, pero k debe ser reemplazado por $p+q$. Hannan (1980) demostró que también es este caso, el estimador BIC es fuertemente consistente.

Para las cadenas de Markov de orden k, tenemos que

$$\mathrm{BIC}_k(x_1^n) = -\log P_{\hat{\theta}_k}^{(n)}(x_1^n) + \frac{(|A|-1)|A|^k}{2} \log n.$$

Finesso (1992) probó que en este caso también es un estimador fuertemente consistente del orden de la cadena.

Hay que enfatizar que todos los resultados citados arriba incluyen en sus hipótesis el hecho de que la cantidad de clases de modelos posibles es finita. Esto quiere decir que hay una cota superior conocida K para el verdadero orden k_0 o (p_0, q_0), y que la minimización del valor del BIC es realizada para los ordenes candidatos que cumplen $k \leq K$ o $p \leq K[1]$, $q \leq K[2]$.

0.4 Enfoque basado en Teoría de la Información

Rissanen (1978, 1983b, 1989) sugirió un enfoque basado en Teoría de la información para el problema de selección de modelo. De acuerdo al *Principio del mínimo largo de descripción* (MDL), el mejor modelo del proceso basado en los datos observados es aquel que da la menor descripción de los datos observados, tomando en cuenta que el modelo también debe ser descrito.

Para cada clase de modelos $\mathcal{M}_k$ definamos un código binario (de longitud variable) unívocamente decodificable $C_k^{(n)} : x(n) \to b(x(n))$ que mapea una muestra $x(n)$ a una secuencia binaria b cuya longitud puede variar con $x(n)$. La función largo del código $L_k^{(n)}(x(n))$ es simplemente el largo de la secuencia binaria $C_k^{(n)}(x(n))$. Más aún, sea $C : k \to b(k)$ un código para las clases de modelos $\mathcal{M}_k$ que mapea el parámetro de estructura k en una secuencia binaria b. Su función de largo de código será denotada por $L(k)$. Entonces, usando la clase de modelos $\mathcal{M}_k$, la muestra $x(n)$ puede ser codificada por $C_k^{(n)}(x(n))$ y un preámbulo $C(k)$ identificando $\mathcal{M}_k$. El criterio MDL es el largo total de esta descripción:

$$\mathrm{MDL}_k(x(n)) = L_k^{(n)}(x(n)) + L(k).$$

El estimador MDL selecciona la clase de modelos que provea la descripción más corta de la muestra:

$$\hat{k}(x_1^n) = \arg\min_{k \in \mathcal{K}} \mathrm{MDL}_k(x(n)).$$

Asumamos por simplicidad que A es finito y que cada modelo M_{θ_k} determina unívocamente una distribución hipotética del proceso; como antes, la distribución marginal para la muestra $x(n)$ es denotada por $P_{\theta_k}^{(n)}$. Un código binario unívocamente decodificable $C_k^{(n)}$ puede ser representado por una *distribución de codificación* $P_k^{(n)}$. Para ver esto, notamos el hecho bien conocido de que $L_k^{(n)}$ es una función de largo de código de algún código unívocamente decodificable $C_k^{(n)}$ si y sólo si satisface la desigualdad de Kraft $\sum_{x(n)} 2^{-L_k^{(n)}(x(n))} \leq 1$. Podemos asumir de hecho que se da la igualdad, pues de otra forma el código podría ser mejorado acortando algunas palabras. Claramente, para cualquier código $C_k^{(n)}$ con largo de código $L_k^{(n)}$ que satisface la desigualdad de Kraft con igualdad, podemos escribir $P_k^{(n)}(x(n)) = 2^{-L_k^{(n)}(x(n))}$. Por otro lado, para una distribución de probabilidad $P_k^{(n)}$ podemos construir un código unívocamente decodificable $C_k^{(n)}$ cuyo largo de código $L_k^{(n)}(x(n)) = \left\lceil -\log P_k^{(n)}(x(n)) \right\rceil$, llamado un código de Shannon. El código determinado por la distribución de codificación $P_k^{(n)}$ será referido como un $P_k^{(n)}$-código. Debe ser elegida de forma óptima en algún sentido bajo la hipótesis de que el verdadero modelo M_{θ_0} está en la clase de modelos $\mathcal{M}_k$. Notemos, que en general $P_k^{(n)}$ va a ser diferente de todas las $P_{\theta_k}^{(n)}$ en la clase de modelos $\mathcal{M}_k$.

La redundancia de un $P_k^{(n)}$-código relativa a la distribución verdadera $P_{\theta_0}^{(n)}$ es

$$R_{\theta_0}^{(n)}(x(n)) = -\log P_k^{(n)}(x(n)) + \log P_{\theta_0}^{(n)}(x(n)).$$

Esto es la diferencia de los largos de código por usar la distribución $P_k^{(n)}$ en lugar de la verdadera $P_{\theta_0}^{(n)}$. Como esta redundancia es una función de la muestra, para evaluar la calidad de $P_k^{(n)}$ es usual considerar o el máximo de $R_{\theta_0}^{(n)}(x(n))$ para todas las posibles muestras $x(n)$, o su valor esperado con respecto a la distribución verdadera $P_{\theta_0}^{(n)}$. Inclusive como la distribución verdadera $P_{\theta_0}^{(n)}$ es desconocida, como un criterio de optimalidad para $P_k^{(n)}$ bajo la hipótesis de que $M_{\theta_0} \in \mathcal{M}_k$ es usual considerar la redundancia media o máxima en el peor caso entre todas las distribuciones posibles $P_{\theta_k}^{(n)}$, $\theta_k \in \Theta_k$, en el rol de $P_{\theta_0}^{(n)}$.

Para la clase de modelos $\mathcal{M}_k$, la redundancia máxima en el peor caso de un $P_k^{(n)}$-código es

$$R^{(n)*} = \sup_{\theta_k \in \Theta_k} \max_{x(n)} R_{\theta_k}^{(n)}(x(n)).$$

Es fácil probar que la distribución de codificación $P_k^{(n)}$ que minimiza esta cantidad es la distribución de *máxima verosimilitud normalizada (NML)* definida como

$$\mathrm{NML}_k^{(n)}(x(n)) = P_{\hat{\theta}_k}^{(n)}(x(n)) \Bigg/ \sum_{x'(n)} P_{\hat{\theta}_k}^{(n)}(x'(n)) \ ,$$

donde $\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} P_{\theta_k}^{(n)}(x(n))$ es el estimador de máxima verosimilitud del vector de parámetros θ_k en la clase de modelos $\mathcal{M}_k$.

Usando esta distribución de codificación obtenemos el criterio MDL de esta forma

$$\mathrm{MDL}_k(x(n)) = -\log P_{\theta_k}^{(n)}(x(n)) + \log\left(\sum_{x'(n)} P_{\hat{\theta}_k}^{(n)}(x'(n))\right) + L(k).$$

Shtarkov (1977) demostró para el caso de cadenas de Markov que el término del medio es asintóticamente (cuando $n \to \infty$, con k fijo) igual a $(1/2)(\dim \Theta_k) \log n$. El mismo resultado es cierto también en otros casos, bajo ciertas hipótesis de regularidad, ver Rissanen (1996). Por lo tanto, cuando el número de clases de modelos posibles es finito, la versión NML del criterio MDL es asintóticamente equivalente al BIC.

Para la clase de modelos $\mathcal{M}_k$, la redundancia media en el peor caso de un $P_k^{(n)}$-código es

$$\begin{aligned}\overline{R}^{(n)} &= \sup_{\theta_k \in \Theta_k} \mathbf{E}_{\theta_k} \left(R_{\theta_k}^{(n)}(x(n)) \right) = \sup_{\theta_k \in \Theta_k} \sum_{x(n)} P_{\theta_k}^{(n)}(x(n)) \log \frac{P_{\theta_k}^{(n)}(x(n))}{P_k^{(n)}(x(n))} \\ &= \sup_{\theta_k \in \Theta_k} \mathbf{D}(P_{\theta_k}^{(n)} \,\|\, P_k^{(n)}),\end{aligned}$$

donde $\mathbf{E}_{\theta_k}(\cdot)$ denota el valor esperado con respecto a la distribución $P_{\theta_k}^{(n)}$, y $\mathbf{D}(\cdot \,\|\, \cdot)$ es la divergencia de información de Kullback-Leibler. La distribución de codificación que minimiza esta cantidad no puede ser dada de forma explícita como en el caso de arriba, pero una que es casi tan buena como aquella que minimiza muchas veces puede ser encontrada.

Concentrándonos en el caso de las cadenas de Markov, consideremos primero la clase de modelos $\mathcal{M}$ igual a $\mathcal{M}_k$ con $k = 0$, es decir la clase de procesos i.i.d. En este caso, una buena distribución de codificación es la distribución de *Krichevsky-Trofimov (KT)* (Krichevsky and Trofimov, 1981). Su redundancia media en el peor caso sobre $\mathcal{M}$ se aproxima al mínimo a menos de una constante que no depende de n. La distribución KT está definida como la mezcla de todas las distribuciones i.i.d. $P_\theta^{(n)}$, con respecto a la medida de Dirichlet μ de parámetros 1/2:

$$\mathrm{KT}_0(x_1^n) = \int P_\theta^{(n)} \mu(d\theta).$$

Mediante cálculos directos se obtiene la siguiente expresión explícita:

$$\mathrm{KT}_0(x_1^n) = \frac{\prod_{a:N_n(a)\geq 1}[(N_n(a) - \frac{1}{2})(N_n(a) - \frac{3}{2}) \cdots (\frac{1}{2})]}{(n - 1 + \frac{|A|}{2})(n - 2 + \frac{|A|}{2}) \cdots (\frac{|A|}{2})},$$

donde $N_n(a)$ denota el número de ocurrencias de $a \in A$ en la muestra x_1^n.

Para las cadenas de Markov de orden k tenemos un resultado similar. En particular, para la clase de modelos $\mathcal{M}_k$ una buena distribución de codificación es la distribución de Krichevsky-Trofimov de orden k, denotada por KT_k. Es una mezcla de todas las distribuciones de las forma

$$P_{\theta_k}^{(n)}(x_1^n) = \frac{1}{|A|^k} \prod_{i=k+1}^{n} P(x_i | x_{i-k}^{i-1}),$$

ver (0.1) con $Q(X_1^k = x_1^k) = |A|^{-k}$, donde el vector de parámetros θ_k especifica la matriz de probabilidades de transición $P(a|a_1^k)$, $a \in A$, $a_1^k \in A^k$. La medida de mezcla esta definida poniendo que las filas $\{P_{\theta_k}(a|a_1^k), a \in A\}$ de esta matriz sean independientes y con distribución de Dirichlet μ como arriba. También en este caso tenemos una forma explícita:

$$\mathrm{KT}_k(x_1^n) = \frac{1}{|A|^k} \prod_{a_1^k : N_n(a_1^k) \geq 1} \frac{\prod_{a_{k+1}:N_n(a_1^{k+1})\geq 1}[(N_n(a_1^{k+1}) - \frac{1}{2}) \cdots (\frac{1}{2})]}{(N_n(a_1^k) - 1 + \frac{|A|}{2})(N_n(a_1^k) - 2 + \frac{|A|}{2} \cdots \frac{|A|}{2})},$$

donde $N_n(a_1^k)$ denota el número de ocurrencias de $a_1^k \in A^k$ en la muestra x_1^n.

La distribución KT_k se puede calcular recursivamente en el tamaño de la muestra n de la forma

$$\mathrm{KT}_k(x_1^n) = \frac{N_{n-1}(x_{n-k}^n) + 1/2}{N_{n-1}(x_{n-k}^{n-1}) + |A|/2} \mathrm{KT}_k(x_1^{n-1}).$$

Las versiones basadas en las distribuciones NML y KT del criterio MDL son asintóticamente equivalentes, debido a que para los mínimos de la redundancia media y máxima en el peor caso tenemos

$$\frac{(|A|-1)|A|^k}{2} \log n - K_1 \leq \min_{P_k^{(n)}} \overline{R}^{(n)} \leq \min_{P_k^{(n)}} R^{(n)^*} \leq \frac{(|A|-1)|A|}{2} \log n - K_2, \tag{0.2}$$

donde K_1 y K_2 son constantes (que dependen de k).

El estimador MDL se puede pensar como un estimador MAP Bayesiano cuando, como arriba, la distribución de codificación $P_k^{(n)}$ es una mezcla de las distribuciones $P_{\theta_k}^{(n)}$, $\theta_k \in \Theta_k$, con una distribución de mezcla $\mu_k^{(n)}$ apropiadamente definida en Θ_k, es decir,

$$P_k^{(n)}(x(n)) = \int_{\Theta_k} P_{\theta_k}^{(n)}(x(n)) \mu_k^{(n)}(d\theta_k).$$

De hecho, representando el código $C(k)$ de la clase de modelos $\mathcal{M}_k$ con la distribución de codificación $P(k) = 2^{-L(k)}$, minimizar el largo de descripción

$$L_k^{(n)}(x(n)) + L(k) = -\log P_k^{(n)}(x(n)) - \log P(k)$$

es equivalente a maximizar $P(k)P_k^{(n)}(x(n))$. Esta última cantidad es proporcional a la probabilidad a posteriori del parámetro de estructura k, o sea, a la probabilidad condicional de k dada la muestra $x(n)$.

El principio MDL puede ser extendido al caso de un A general, por ejemplo $A = \mathbb{R}$, vía discretización, y esto conduce a resultados similares a los de arriba, ver Rissanen (1989), en particular, para los procesos AR Hamerly y Davis (1989), y para los procesos ARMA Gerencsér (1987).

0.5 Resultados recientes

Para varios procesos se ha probado que los estimadores BIC y MDL de los parámetros de estructura son fuertemente consistentes. Esto quiere decir que el argumento que minimiza el criterio BIC o MDL sobre los parámetros de estructura posibles es igual al verdadero parámetro de estructura, eventualmente casi seguro cuando el tamaño de la muestra tiene a infinito. La mayoría de las pruebas de consistencia en la literatura incluyen la hipótesis de que el número de parámetros de estructura posibles sea finito, es decir, que existe una cota superior conocida para el parámetro de estructura. Esta hipótesis es de naturaleza técnica y simplifica las pruebas. Sin embargo, esta hipótesis no es deseable, porque en la práctica no es usual tener información de antemano sobre el parámetro de estructura, además a medida que tenemos cantidades cada vez más grandes de datos, se deberían tener en cuenta más clases de modelos posibles, de mayor complejidad como candidatos. Por lo tanto, es una aspiración razonable la de eliminar esta hipótesis de una cota conocida del parámetro de estructura. Csiszár y Shields (2000) probaron que el estimador BIC del orden de una cadena de Markov es fuertemente consistente incluso sin la hipótesis de una cota constante sobre el orden de la cadena y basados en la n-ésima muestra x_1^n todos los ordenes posibles $0 \leq k < n$ son considerados como candidatos. Al mismo tiempo, Csiszár y Shields (2000) mostraron que el mismo resultado no era cierto para el estimador MDL. Consideremos un proceso i.i.d. con distribución uniforme en A. Este proceso es una cadena de Markov de orden 0. Para el criterio MDL

$$\mathrm{MDL}_k(x_1^n) = -\log P_k^{(n)}(x_1^n) + L(k),$$

donde la distribución de codificación $P_k^{(n)}$ es $\mathrm{NML}_k^{(n)}$ o KT_k, y el largo de código $L(k)$ para el orden de la cadena k satisface $L(k) = o(k)$, se tiene que

$$\hat{k}(x_1^n) = \arg \min_{0 \leq k \leq \alpha \log n} \mathrm{MDL}_k(x(n)) \to \infty \quad \text{si } n \to \infty,$$

donde $\alpha = 4/\log|A|$. Este contraejemplo muestra que el estimador MDL no es consistente cuando la cota superior del orden es eliminada.

Csiszár (2002) demostró la consistencia fuerte del estimador MDL para el orden de las cadenas de Markov cuando al conjunto de ordenes posibles considerado se le permite extenderse cuando el tamaño de la muestra n aumenta, la cota para los ordenes tenidos en cuenta es $o(\log n)$ en el caso de la distribución de KT, y $\alpha \log n$ con $\alpha < 1/\log|A|$ en el caso de la distribución NML. Observemos que estos estimadores MDL no necesitan una cota previa en el verdadero orden de la cadena. El resultado de consistencia fue probado para el criterio MDL sin el término $L(k)$, lo cual es un resultado aún más fuerte.

0.6 Estimación de árboles de contexto

El modelo llamado *fuente árbol* o *cadena de Markov de alcance variable* (VLMC) es un refinamiento del modelo de cadena de Markov. Dada una cadena de Markov de orden k, para una secuencia $a_1^k \in A^k$ las probabilidades de transición $Q(a|a_1^k)$, $a \in A$ pueden no depender de la secuencia entera $a_1, \ldots, a_k$, sino solamente de una subsecuencia $a_l, \ldots, a_k$; esto admite una parametrización más parsimoniosa.

Consideremos un proceso con $T = \mathbb{Z}$ y $|A| < \infty$. Por simplicidad, asumamos que todas las distribuciones marginales finito dimensionales de la distribución Q del proceso son estrictamente positivas. La secuencia $s = a_{-\ell}^{-1} \in A^\ell$ es un *contexto* para el proceso Q si

$$Q(X_i = a|X_{-\infty}^{i-1} = x_{-\infty}^{i-1}) = Q(a|s) \quad \text{para todo } i \in \mathbb{Z}, a \in A,$$

con probabilidades de transición apropiadas $Q(\cdot|s)$, donde $x_{i-\ell}^{i-1} = a_{-\ell}^{-1}$, y ninguna subsecuencia $a_{-\ell'}^{-1}$, $\ell' < \ell$ tiene esta propiedad. El conjunto de todos los contextos es llamado el *árbol de contextos*, y será denotado por $\mathcal{T}$. Asumamos que para cada secuencia $x_{i-\ell}^{i-1}$ existe un contexto s de largo finito ℓ, y que el supremo de estos largos es un número finito k. Un proceso con un tal árbol de contextos $\mathcal{T}$ es una cadena de Markov de orden k, pero una colección de $(|A|-1)|\mathcal{T}|$ probabilidades de transición son suficientes para describirlo, en lugar de las $(|A|-1)|A|^k$ requeridas para una cadena de Markov general de orden k.

El término árbol de contextos refiere a su visualización. Los contextos s, escritos de adelante hacia atrás, pueden ser entendidos como hojas de un árbol, donde el camino desde la raíz a la hoja está determinado por la secuencia s. Este árbol de contextos es completo, es decir, cada nodo excepto las hojas tiene exactamente $|A|$ hijos.

Para un proceso con árbol de contextos $\mathcal{T}$, la probabilidad de la realización x_1^n puede ser escrita como

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{s \in \mathcal{T}, a \in A} Q(a|s)^{N_n(s,a)},$$

donde $N_n(s,a)$ denota el número de ocurrencias de $a \in A$ en la secuencia x_{k+1}^n precedido del contexto $s \in \mathcal{T}$. Dada la muestra x_1^n, el máximo del segundo factor es alcanzado por $Q(a|s) = \frac{N_n(s,a)}{N_n(s)}$, $a \in A$, $s \in \mathcal{T}$, donde $N_n(s) = \sum_{a \in A} N_n(s,a)$. Es decir, los estimadores de máxima verosimilitud de las probabilidades de transición son las probabilidades de transición empíricas, como para las cadenas de Markov.

Para la clase de modelos descrita arriba, el árbol de contextos $\mathcal{T}$ juega el papel del parámetro de estructura. Por analogía al caso de las cadenas de Markov, el criterio de información Bayesiano para la clase de modelos determinada por $\mathcal{T}$ es

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) = -\sum_{s\in\mathcal{T},a\in A} N_n(s,a)\log\frac{N_n(s,a)}{N_n(s)} + \frac{(|A|-1)|\mathcal{T}|}{2}\log n.$$

El principio MDL también se puede formular para modelos de fuente árbol. Se pueden definir como antes las distribuciones NML y KT para la clase de modelos con árbol de contextos $\mathcal{T}$, aquí nos concentramos en la distribución KT. Esta tiene la siguiente forma explícita:

$$KT_{\mathcal{T}}(x_1^n) = \frac{1}{|A|^k}\prod_{s:N_n(s)\geq 1}\frac{\prod_{a:N_n(s,a)\geq 1}[(N_n(s,a)-\frac{1}{2})(N_n(s,a)-\frac{3}{2})\dots(\frac{1}{2})]}{(N_n(s)-1+\frac{|A|}{2})(N_n(s)-2+\frac{|A|}{2})\dots(\frac{|A|}{2})} \tag{0.3}$$

Podemos calcularla recursivamente en n así

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = \frac{N_{n-1}(s,x_n)+1/2}{N_{n-1}(s)+|A|/2}\,\mathrm{KT}_{\mathcal{T}}(x_1^{n-1}),$$

donde s es el contexto para el pasado $x_{-\infty}^{n-1}$. De forma similar a la clase de cadenas de Markov de orden k, la distribución $\mathrm{KT}_{\mathcal{T}}$ minimiza la redundancia media en el peor caso para la clase de modelos con árbol de contexto $\mathcal{T}$, a menos de una constante. Además, las cotas (0.2) también valen si uno reemplaza $|A|^k$ por $|\mathcal{T}|$. Es importante notar también que existe un código simple C (ver Willems, Shtarkov, y Tjalkens 1995) para describir un árbol de contextos $\mathcal{T}$ con largo de código

$$L(\mathcal{T}) = \frac{|A||\mathcal{T}|-1}{|A|-1}. \tag{0.4}$$

Para el problema de selección de modelo en este entorno, en lugar de utilizar criterios de información como arriba, se utilizaban en la mayoría de los casos variantes del algoritmo "context" de Rissanen (1983a). La razón para esto es la complejidad computacional: la cantidad de árboles de contexto posibles de altura como máximo D es enorme si D es grande, y no es posible calcular un criterio de información para todos los árboles de contexto posibles y elegir el candidato de valor mínimo. Este problema fue parcialmente resuelto por Willems, Shtarkov, y Tjalkens (2000). Ellos probaron que el estimador MDL con la distribución de codificación KT (0.3) y el largo de código para el árbol dado por (0.4) es fuertemente consistente, cuando se conoce una cota a priori en la altura D de los árboles de contexto posibles. Al mismo tiempo presentaron un algoritmo, llamado *Context Tree Maximizing (CTM)*, para calcular el estimador sin computar y comparar los valores para todos los árboles candidatos. Para la n-ésima muestra x_1^n, el número de operaciones elementales requeridas por el algoritmo es proporcional a nD, es decir, el algoritmo es de orden lineal.

0.7 Modelos de árboles dispersos

Para una cadena de Markov de orden k, la probabilidad de una muestra x_t en una secuencia $x_1^n = x_1, x_2, \ldots, x_n$ definida sobre un alfabeto finito A está dada por una distribución discreta condicional $Q(x_t|x_{t-k}^{t-1})$, que depende del valor de la k-úpla inmediatamente anterior a x_t. Por lo tanto, el proceso puede ser parametrizado por un modelo de Markov de orden k con $|A|^k(|A|-1)$ parámetros libres.[1] En un *modelo árbol* (Rissanen, 1983a; Buhlmann and Wyner, 1998) del mismo proceso, al alcance de la memoria se le permite variar para diferentes posiciones de los datos. Estos modelos son parametrizaciones más eficientes de los procesos, ya que la cantidad exponencial de parámetros estadísticos en un modelo de Markov muchas veces puede ser reducida drásticamente en un modelo árbol bien ajustado. Además de capturar redundancias típicas en datos reales de una manera económica, una gran ventaja de los modelos árbol es que la información estadística necesaria para optimizar el modelo puede ser guardada en un árbol de contextos, que va aumentando a medida que la secuencia es observada, guardando esencialmente todas las ocurrencias de cada letra en cada contexto, esta estructura combinatoria recursiva es la clave para obtener eficiencia algorítmica.

El ahorro en el número de parámetros estadísticos obtenido por los modelos árbol puede verse como el resultado de agrupar estados equivalentes, es decir, k-úplas que inducen la misma probabilidad de transición en el modelo de Markov completo. Por lo tanto, un estado de largo $k - \ell$ en un modelo árbol para una cadena de orden k corresponde a agrupar $|A|^{k-\ell}$ estados equivalentes en el modelo de Markov completo. Esta observación, de hecho, caracteriza la estructura de los conjuntos de estados que pueden agruparse en un modelo árbol: cada conjunto de este tipo debe estar formado de todas las extensiones de una cadena de largo $k-\ell$, con $0 \leq \ell \leq k$. Cuando trabajamos con datos reales, otros conjuntos de estados equivalentes pueden aparecer, y es natural preguntarse si es posible optimizar modelos donde se permitan agrupar conjuntos de estados más generales. En la práctica, las distribuciones que se agrupan son empíricas y no necesariamente idénticas, y una búsqueda de la partición óptima del espacio de estados es irrealizable. Este problema es conocido como el problema de *cuatización de contextos*, y ha sido estudiado con varias técnicas, aunque la mayoría de las soluciones han sido ad-hoc y de distintos grados de complejidad en función de k, que en general es pequeño (ver por ejemplo, Forchhammer et al. 2004, para un marco actual).

[1] Distinguimos entre el proceso (que es siempre una cadena de Markov) y sus parametrizaciones. Por lo tanto, un modelo de Markov de orden k es la parametrización más natural de un proceso de memoria finita de orden k, pero nuestro objetivo principal es estudiar otras parametrizaciones más eficientes.

En otras palabras las cadenas de Markov de alcance variable reducen la cantidad de parámetros agrupando conjuntos completos de estados equivalentes que tienen un sufijo común. Estos estados tienen (o se estima que tienen) iguales o muy parecidas distribuciones condicionales. A pesar de todo, las VLMC tienen la restricción que los contextos deben estar dados por una secuencia de símbolos contiguos, y que dos estados cualesquiera agrupados deben tener un sufijo común con la misma distribución para poder formar un árbol completo.

Por ejemplo si tenemos que $Q(a|x_{n-\ell}^{n})$ solo depende de x_n y $x_{n-\ell}$ el modelo árbol de este proceso estaría representado por un árbol completo de profundidad ℓ. Este modelo tendría $|A|^{\ell}$ hojas pero sólo habrían $|A| \times |A|$ parámetros diferentes entre los asociados a cada hoja. Nos gustaría representar este modelo de una forma más parsimoniosa, de ser posible con sólo $|A| \times |A|$ estados distintos. Estamos interesados en esquemas que agrupen conjuntos de estados más generales, que no necesariamente cumplen la condición de formar un subárbol completo.

En muchas aplicaciones tiene sentido buscar las dependencias estadísticas en las observaciones contiguas, esto hace que los modelos de Markov sean muy populares. Sin embargo, el costo de describir el modelo crece exponencialmente con la memoria del proceso. Una manera de solucionar este problema es la idea de modelos árbol que permiten que el alcance de la memoria dependa de los valores de las observaciones contiguas. A pesar de esto, el ejemplo de arriba muestra que la condición de contigüidad puede ser muy restrictiva. Nuestro enfoque permite que hayan posiciones no contiguas en el contexto para poder capturar dependencias entre variables distantes sin aumentar el tamaño del modelo. Proponemos optimizar el conjunto de posiciones condicionantes para los datos dados, asumiendo un alcance de dependencia máximo k y algunas restricciones.

En esta tesis vamos a proponer un nuevo tipo de modelos Markovianos con dependencias dispersas. Estudiamos los *modelos de árboles dispersos* (STMs), una generalización de los modelos árbol donde conjuntos de estados más generales con distribuciones condicionales similares pueden ser agrupados, mientras se preservan las propiedades combinatorias atractivas de la estructura de árbol. En algunos casos, esta estructura nos permitirá encontrar de forma eficiente el mejor modelo para una secuencia dada, de una clase amplia de STMs. En un STM, las variables van a estar condicionadas por cadenas finitas de símbolos no necesariamente contiguos, y el contexto va a determinar no solo cuan lejos llega la memoria, sino también cuales son las posiciones de variables anteriores en las que condicionamos. Modelos con contextos no contiguos han sido estudiados en Eskin et al. (2000); Bourguignon and Robelin (2004); Zhao et al. (2004); Leonardi (2006), la mayoría de estos artículos son aplicaciones de modelos similares y de su estimación en Biología, particularmente en modelización de ADN y proteínas.

Elegimos un símbolo fijo $\phi \notin A$, y denotamos por A_ϕ al *alfabeto expandido* $A_\phi = A \cup \{\phi\}$. Nos referiremos a las secuencias sobre A_ϕ como *patrones.* Vamos a interpretar el símbolo ϕ como un "comodín", dado un patrón $w_1^m \in A_\phi^m$, las secuencias en el conjunto

$$\mathcal{C}(w_1^m) = \{\, u_1^m \in A^m \,|\, u_i = w_i \;\text{ cuando } w_i \neq \phi \,\}$$

se dice que son *consistentes* o *conformes* con w_1^m.

Decimos que un patrón v es un *ϕ-sufijo* de w, lo que denotamos por $v \prec_\phi w$, cuando $l(v) \leq l(w)$ y $v_{n-i} = w_{m-i}$ para todo $i \leq l(v)$ tal que $v_{n-i} \neq \phi$. Un conjunto $\mathcal{T}$ de patrones forma un *árbol de contexto disperso* si ningún $w \in \mathcal{T}$ es un ϕ-sufijo de otro $v \in \mathcal{T}$ ($\mathcal{T}$ es un conjunto libre de ϕ-sufijos). Más aún, un árbol disperso se dice que es *completo* cuando cada secuencia x_1^n suficientemente larga tiene un ϕ-sufijo en el árbol.

Consideremos un proceso de Markov estacionario $\{X_n\}_{n\in\mathbb{N}}$ tomando valores en A, dado un árbol de contextos disperso $\mathcal{T}$ decimos que el proceso es *$\mathcal{T}$-adaptado* si $\forall x_1^n$ tal que $w \prec_\phi x_1^n$ (como $\mathcal{T}$ es un conjunto libre de ϕ-sufijos, w es único) tenemos que

$$\mathbf{P}(X_{n+1} = a | X_1^n = x_1^n) = Q(a|w), \qquad \forall a \in A.$$

Definimos la distribución KT de x_1^n correspondiente a $\mathcal{T}$ como

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^k} \prod_{w\in\mathcal{T}} \frac{\prod_{a\in A}(N_n(w,a)-\frac{1}{2})(N_n(w,a)-\frac{3}{2})\ldots\frac{1}{2}}{(N_n(w)-1+\frac{|A|}{2})(N_n(w)-2+\frac{|A|}{2})\ldots\frac{|A|}{2}}$$

donde $N_n(w,a)$ denota el número de ocurrencias del patrón w seguido del símbolo a como en el caso de las VLMC.

Mostraremos que el estimador MDL basado en la distribución KT es fuertemente consistente para la estimación de la clase de árboles de contextos dispersos. Además, proponemos una manera eficiente de calcular el estimador. Algunos de los resultados están basados en una parte de mi tesis de Maestría en Informática "Universal Coding via Sparse Tree Models" (Fraiman, 2008).

CHAPTER 1

Introduction

1.1 The model selection problem

Let a stochastic process $\{X_t, t \in T\}$ be given, where each X_t is a random variable with values $a \in A$, and T is an index set. The joint distribution of the random variables X_t, $t \in T$ will be referred to as the distribution of the process and will be denoted by Q. A *model* of the process determines a hypothetical distribution of the process or a collection of hypothetical distributions. Typically, a model is determined by a structure parameter k with values in some set $\mathcal{K}$, and by a parameter vector $\theta_k \in \Theta_k \subset \mathbb{R}^{d_k}$; this model is denoted by $\mathcal{M}_{\theta_k}$. Given the feasible models of the process, they can be arranged into model classes according to the structure parameter: $\mathcal{M}_k = \{M_{\theta_k}, \theta_k \in \Theta_k \subset \mathbb{R}^{d_k}\}$. Statistical inference about the process is drawn based on a realization $\{x_t, t \in T\}$ of the process observed in the range $R_n \subset T$, where R_n extends with n. Thus the n'th sample is $x(n) = \{x_t, t \in R_n\}$. Some typical examples for processes and their models are listed below.

In the case of *density function estimation*, $T = \mathbb{N}$ and the random variables X_t, $t \in \mathbb{N}$ are independent and identically distributed (i.i.d.) with density function f_{θ_k}. The n'th sample is $\{x_i, i = 1, \ldots, n\}$.

The *polynomial fitting* involves $T \subset \mathbb{R}$, where T is a countable set, $A = \mathbb{R}$, and the model

$$X_t = \theta_k[0] + \theta_k[1]t + \theta_k[2]t^2 + \cdots + \theta_k[k-1]t^{k-1} + Z_t,$$

where Z_t, $t \in T$ are independent random variables with normal distribution, zero mean, unknown common variance, and $\theta_k[i]$ is the i'th component of the k-dimensional parameter vector θ_k. Here the structure parameter $k \in \mathbb{N}$ is the degree of the polynomial $\theta_k[0] + \theta_k[1]t + \theta_k[2]t^2 + \cdots + \theta_k[k-1]t^{k-1}$ plus 1, and the n'th sample is $\{x_t, t \in \{t_1, \ldots, t_n\} \subset T\}$.

The process with $T = \mathbb{N}$, $A = \mathbb{R}$ is an *autoregressive (AR) process* of order k if

$$X_t = \sum_{i=1}^{k} a_i X_{t-i} + Z_t,$$

where Z_t, $t \in \mathbb{N}$ are independent random variables with normal distribution, zero mean, unknown common variance, and $a_i \in \mathbb{R}$, $i = 1, \ldots, k$ form the parameter vector θ_k. Here the structure parameter $k \in \mathbb{N}$ is the number of coefficients a_i, and the n'th sample is $\{x_i, i = 1, \ldots, n\}$.

The *autoregressive moving average (ARMA) process* is similar to the AR process. In this case we have

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + Z_t + \sum_{j=1}^{q} b_j Z_{t-j}.$$

The parameter vector is $\theta_k = \{a_1, \ldots, a_p, b_1, \ldots, b_q\} \in \mathbb{R}^{p+q}$, and the structure parameter k has two components: $k = (p, q) \in \mathbb{N}^2$.

The process with $T = \mathbb{N}$, $|A| < \infty$ is a *Markov chain* of order k if

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{i=k+1}^{n} Q(x_i | x_{i-k}^{i-1}), \quad n \geq k,\ x_1^n \in A^n, \tag{1.1}$$

with suitable transition probabilities $Q(\cdot|\cdot)$. Here x_i^j denotes the sequence $x_i, x_{i+1}, \ldots, x_j$. Since for each $a_1^k \in A^k$ the vector $\{Q(a|a_1^k), a \in A\}$ gives a probability distribution on A, the parameter vector $\theta_k \in \mathbb{R}^{d_k}$ consists of $d_k = (|A| - 1)|A|^k$ transition probabilities $Q(a|a_1^k)$, $a \in A^\star$, $a_1^k \in A^k$, where $|A^\star| = |A| - 1$. Here the structure parameter $k \in \mathbb{N}$ is the length of the sequence that the transitional probabilities depend on in their second argument. The n'th sample is $\{x_i, i = 1, \ldots, n\}$.

The AR and ARMA processes, and Markov chains are examples for the case when the model does not determine a unique hypothetical distribution of the process. In particular, for AR processes or Markov chains of order k the model determines only a hypothetical conditional distribution for $X_{k+1}, X_{k+2}, \ldots$ given $X_1, \ldots, X_k$.

The set $\mathcal{K}$ of feasible structure parameters k is an ordered or partially ordered set with respect to the inclusion of the model classes $\mathcal{M}_k$. When the model M_{θ_k} with structure parameter k corresponds to the true distribution

Q of the process, a more complex model with (in the above sense) larger structure parameter k' may also correspond to the distribution Q with a suitable parameter vector $\theta_{k'}$. For example, any AR process or Markov chain of order k is also of order k', for each $k' > k$. We mean by the *true model* M_{θ_0} the minimal model among those that correspond to the true distribution Q, that is, for which there exists no other model with the same property that has a smaller structure parameter in the above sense. The structure parameter of this true model will be denoted by k_0.

The *model selection problem* consists in estimating the true structure parameter k_0 based on the statistical observation $x(n)$ of the process.

The term *underestimation* refers to the case when a smaller structure parameter k is selected than the true one k_0. In such a case $\theta_0 \notin \Theta_k$, hence the true model cannot be estimated accurately; the estimation of the parameter vector will involve bias.

The term *overestimation* refers to the case when a greater structure parameter k is selected than the true one k_0. In this case $M_{\theta_0} \in \mathcal{M}_{k_0} \subset \mathcal{M}_k$, thus $M_{\theta_0} = M_{\theta_k}$ for some $\theta_k \in \Theta_k$, but θ_k has more components than θ_0, hence it is more difficult to estimate the true setting; the estimation of the parameter vector will have larger variance.

The dissertation treats the model selection problem using the concept of information criterion. An *information criterion* (IC) based on the sample $x(n)$ assigns a real value to each model class: $\mathrm{IC} : \mathcal{K} \times \{x(n)\} \to \mathbb{R}$, and the estimator of k_0 equals the structure parameter with the minimum value of the criterion:

$$\hat{k}(x(n)) = \arg\min_{k\in\mathcal{K}} \mathrm{IC}_k(x(n)).$$

The next sections give an overview about information criteria.

1.2 Historical review

The model selection problem can be regarded as multiple hypothesis testing, and the *likelihood ratio test* procedure of Neyman and Pearson (1928) can be used. Anderson worked out this procedure for polynomial fitting (Anderson, 1962) and for AR processes (Anderson, 1963). These procedures are sequences of tests taking the hypothetical orders successively, starting at the highest one. The main disadvantage of these procedures is the subjective choice of the significance levels of the tests for all hypothetical model orders.

Mallows (1964, 1973) introduced, for selecting the true variables of linear models, a method similar to the information criteria. Consider the *linear model*

$$X_t = \sum_{i=1}^{K} a_i u_{it} + a_0 + Z_t, \quad t \in \mathbb{Z},$$

where a_i, $i = 1, \dots, K$ are the parameters of the model, u_{it}, $i = 1, \dots, K$ are (non-random) independent variables whose values are given at $t = 1, \dots, n$, and Z_t's are independent random variables with zero mean and unknown common variance σ^2. Given the sample $x(n) = \{x_t, t = 1, \dots, n\}$, the problem is to estimate the set $\{u_{i1}, \dots, u_{ik}\}$ of variables that X_t effectively depends on, that is, $a_{i_\ell} = 0$ for $i_\ell \in \{i_1, \dots, i_k\}$ and $a_{i_\ell} = 0$ otherwise.

Mallows assigned to each hypothetical index set $P = \{i_1, \dots, i_k\}$ the value

$$C_P = \frac{1}{\hat{\sigma}^2} \operatorname{RSS}_P - n + |P|,$$

where RSS_P is the residual sum of squares according to P:

$$\operatorname{RSS}_P = \min_{a_{i_\ell}, i_\ell \in P} \sum_{t=1}^{n} \left(x_t - \sum_{a_{i_\ell}, i_\ell \in P} a_{i_\ell} u_{i_\ell t} - a_0 \right)^2,$$

moreover $\hat{\sigma}^2$ is a suitable estimate of σ^2, e.g., $\hat{\sigma}^2 = \operatorname{RSS}_{1,\dots,k} / (n - k)$. The estimator is the index set P with minimum C_P. It can be shown that the expected value of C_P is equal to $|P|$ when P is the true index set, and it is greater otherwise.

For stationary processes, Davisson (1965) analyzed the mean square prediction error of the AR model of order k, when the coefficients of the model are determined based on the past n observations $x_1, \dots, x_n$ and this model is applied to predict the next observation. Namely, for the predictor $\hat{X}_n(k) = \sum_{i=1}^{k} \hat{a}_i X_{n-i}$ with coefficients which minimize the mean square prediction error, that is,

$$\{\hat{a}_1, \dots, \hat{a}_k\} = \arg \min_{a_1, \dots, a_k} \sum_{t=0}^{n} \left(x_t - \sum_{i=1}^{k} a_i x_{t-i} \right)^2,$$

he obtained

$$\mathbf{E} \left(X_n - \hat{X}_n(k) \right)^2 = \sigma^2(k) \left(1 + \frac{k}{n} \right) + o(1/n),$$

where $\sigma^2(k)$ is the asymptotic mean square error. Moreover, he suggested using the main term of the above expression to estimate the true order, via minimizing it over the candidate orders. Of course, this requires the estimation of $\sigma^2(k)$.

Akaike (1970) arrived at the same result, and he overcame the problem of estimating $\sigma^2(k)$ by a suitable spectral estimation method. He defined a criterion called *final prediction error* as

$$\mathrm{FPE}_k(x_1^n) = \frac{n+k}{n-k}\left(\hat{C}_0 - \hat{a}_1\hat{C}_1 - \cdots - \hat{a}_k\hat{C}_k\right),$$

where $\hat{C}_i = (1/n)\sum_{t=1}^{n-1} x_{t+i}x_t$, $i = 0, \ldots, k$ are the empirical correlation coefficients, and $\hat{a}_i$, $i = 1, \ldots, k$ are the model coefficients which minimize the least square prediction error, as above. The latter values can be calculated from the $\hat{C}_i$'s, solving the Yule-Walker equation. The order estimator is

$$\hat{k}(x_1^n) = \arg\min_{0\le k\le K} \mathrm{FPE}_k(x_1^n).$$

The only subjective element in this procedure is the determination of the upper bound K of candidate orders. Akaike also showed that this estimator overestimates the true order asymptotically with positive probability, that is,

$$\liminf_{n\to\infty} Q\left(\hat{k}(x_1^n) > k_0\right) > 0.$$

Akaike (1972) introduced a general concept for solving the model selection problem. Assume that each model M_{θ_k} specifies a unique distribution P_{θ_k} of the process, and let $P_{\theta_k}^{(n)}$ denote its marginal equal to the distribution of the sample $x(n)$. The *Kullback-Leibler information divergence* between $P_{\theta_k}^{(n)}$ and $P_{\theta_0}^{(n)}$ is

$$\mathbf{D}(P_{\theta_0}^{(n)} \,\|\, P_{\theta_k}^{(n)}) = \int f_{\theta_0}^{(n)}(x(n)) \log \frac{f_{\theta_0}^{(n)}(x(n))}{f_{\theta_k}^{(n)}(x(n))} \lambda(dx(n)),$$

where $f_{\theta_k}^{(n)}$ denotes the density of $P_{\theta_k}^{(n)}$ with respect to a dominating measure λ (typically, λ is either the Lebesgue measure or, in the discrete case, the counting measure). Logarithms are to the base e. Akaike aimed at minimization of this quantity for estimating the true parameter vector θ_0 and the true structure parameter k_0. He found that this minimizer can be approximated by taking the maximum likelihood estimator $\hat{\theta}_k = \arg\max_{\theta_k\in\Theta_k} f_{\theta_k}^{(n)}(x(n))$ in each candidate model class, and then selecting the model class whose structure parameter minimizes the value

$$\mathrm{AIC}_k(x(n)) = -\log f_{\theta_k}^{(n)}(x(n)) + \dim\Theta_k.$$

When the models do not determine uniquely the distribution of the process, we can define the AIC similarly, with suitably defined $f_{\theta_k}^{(n)}$. For example, in the case of AR process of order k we can prescribe $X_1, \dots, X_k$ to be 0, or to have the marginal distribution of the stationary distribution of the process. This specifies a unique joint distribution corresponding to the model, and we can take its density as $f_{\theta_k}^{(n)}$. Note that suitable restriction on the parameter set Θ_k can guarantee the existence of the stationary distribution. In the case of Markov chains of order k, we can either proceed similarly, or we can define $f_{\theta_k}^{(n)}$ as the right hand side of (1.1) dropping the factor $Q(X_1^k = x_1^k)$.

This model selection procedure has a clear interpretation. The first term of the information criterion is the negative logarithm of the maximum likelihood. It measures the goodness of fit of the sample to the model class $\mathcal{M}_k$. This term decreases when the complexity of the model increases. The second term of the information criterion, called the *penalty term*, is the number of free parameters of the model. This penalizes too complex models: it increases with the model complexity. Thus, the selected model has a good trade-off between good description of the data and the model complexity.

For AR models, AIC is asymptotically (i.e., as the sample size tends to infinity) identical to the FPE criterion (Akaike, 1972, 1974). Therefore, the AIC estimator also overestimates the true structure parameter asymptotically with positive probability. Shibata (1976) derived the exact asymptotic distribution of the order selected by these estimators.

The classical *cross-validation* principle can be adopted to the model selection problem (e.g., Stone 1974). The general principle requires dividing the sample set into two subsets, and performing the model estimation based on one subset only. Using the other subset, the candidate model can be validated correctly, that is, the estimation and the validation will be independent. A formulation of this principle for the polynomial fitting problem is the following. Divide the n'th sample $x(n) = \{x_t, t = t_1, \dots, t_n\}$ into subsets via leaving out the p'th element: $x(n) = x_{\setminus p} \cup \{x_{t_p}\}$, where $x_{\setminus p} = x(n) \setminus \{x_{t_p}\}$. Estimate the coefficients of the polynomial of degree $k-1$ based on the sample set $x_{\setminus p}$:

$$\hat{\theta}_k^{(p)} = \arg\min_{\theta_k \in \Theta_k} \sum_{i \in \{1,\dots,n\} \setminus \{p\}} (x_t - (\theta_k[0] + \theta_k[1]t_i + \dots + \theta_k[k-1]t_i^{k-1}))^2,$$

and validate it based on the p'th sample element x_{t_p}:

$$e_k(p) = x_{t_p} - (\theta_k[0] + \theta_k[1]t_p + \theta_k[2]t_p^2 + \dots + \theta_k[k-1]t_p^{k-1}).$$

Calculate this prediction error for all p, and minimize

$$e_k^2 = \sum_{p=1}^{n} e_k(p)^2$$

over the hypothetical k's to obtain the estimated degree of the polynomial. Stone (1977) showed that the cross-validation criterion is asymptotically equivalent to the AIC.

1.3 Consistent model selection

In this work, the goodness of model selection will be considered only from the asymptotical point of view; in the literature, this aspect is in the focus.

An estimator $\hat{k}(x(n))$ of the structure parameter k based on the sample $x(n)$ is said to be *consistent* if the probability that the estimator equals the true structure parameter k_0 approaches 1 when the sample size n tends to infinity:

$$Q\left(\hat{k}(x(n)) = k_0\right) \to 1 \qquad \text{if } n \to \infty.$$

The estimator $\hat{k}(x(n))$ is said to be *strongly consistent* if it equals the true structure parameter k_0 eventually almost surely as the sample size n tends to infinity:

$$\hat{k}(x(n)) = k_0, \qquad \text{eventually almost surely as } n \to \infty.$$

Here and in the sequel, "eventually almost surely" means that with probability 1 there exists a threshold n_0 (depending on the realization $\{x_t, t \in T\}$) such that the claim holds for all $n \geq n_0$.

For the case of density estimation, when the feasible density functions belong to *exponential families*, Schwarz (1978) derived an information criterion from the asymptotic approximation of the Bayesian Maximum A-posteriori Probability (MAP) estimator. Suppose the model class $\mathcal{M}_k$ consists of density functions

$$f_{\theta_k}(x_i) = \exp(\pi\theta_k y_k(x_i) - b_k(\theta_k)), \quad \theta_k \in \Theta_k,$$

where $\pi\cdot\cdot$ denotes the inner product in the $d_k = \dim\Theta_k$ dimensional Euclidean space, $y_k : \mathbb{R} \to \mathbb{R}^{d_k}$ are given functions, and

$$b_k(\theta_k) = \log \int \exp(\pi\theta_k y_k(x_i))dx_i.$$

Here k ranges over a finite set $\mathcal{K}$. A prior distribution of the parameter vector can be written in the form $\mu = \sum_{k\in\mathcal{K}} \alpha_k\mu_k$, where α_k is the a priori probability that a model with structure parameter k is the true one, and μ_k is the conditional a priori distribution of θ_k under the condition that the true structure parameter is k; μ_k is concentrated on Θ_k. Schwarz showed that, under regularity conditions, the MAP estimator of the parameter vector θ_k from an i.i.d. sample x_1^n asymptotically does not depend on μ, and is equivalent to the maximum likelihood estimator $\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} f_{\theta_k}^{(n)}(x(n))$ in the model class $\mathcal{M}_k$ whose structure parameter k minimizes the value

$$\mathrm{BIC}_k(x(n)) = -\log f_{\theta_k}^{(n)}(x(n)) + \frac{\dim \Theta_k}{2} \log n$$

over the set $\mathcal{K}$. This value is called *Bayesian Information Criterion (BIC).*

The consistency of the BIC estimator in the above situation has been proved by Haughton (1988). Note that for the polynomial fitting problem, Akaike (1977) introduced the same information criterion with the same notation BIC, in a heuristic way.

For the AR model of order k the Bayesian Information Criterion has the following form:

$$\mathrm{BIC}_k(x_1^n) = -\log f_{\theta_k}^{(n)}(x_1^n) + \frac{k}{2} \log n.$$

Hannan and Quinn (1979) proved that the BIC estimator of the order of AR processes is strongly consistent. For the ARMA model of order (p, q) the BIC has the similar form, but k is replaced by $p + q$. Hannan (1980) proved that also in this case, the BIC estimator is strongly consistent.

For the Markov chain of order k, we have

$$\mathrm{BIC}_k(x_1^n) = -\log P_{\hat{\theta}_k}^{(n)}(x_1^n) + \frac{(|A| - 1)|A|^k}{2} \log n.$$

Finesso (1992) proved that this gives a strongly consistent order estimator.

It should be emphasized that all consistency results above include the assumption that the number of candidate model classes is finite. This means that there is a known upper bound K on the true order k_0 or (p_0, q_0), and the minimization of the value BIC is for the candidate orders $k \le K$ or $p \le K[1]$, $q \le K[2]$.

1.4 Information theoretical approach

Rissanen (1978, 1983b, 1989) suggested an information theoretical approach to the model selection problem. According to the *Minimum Description Length* (MDL) principle, the best model of the process based on the observed data is the one that gives the shortest description of the observed data, taking into account that the model itself must also be described.

Let each model class $\mathcal{M}_k$ be assigned a uniquely decodable, variable-length binary code $C_k^{(n)} : x(n) \to b(x(n))$ which maps a sample $x(n)$ to a binary sequence b whose length can vary with $x(n)$. The codelength function $L_k^{(n)}(x(n))$ is the length of the binary sequence $C_k^{(n)}(x(n))$. Moreover, let $C : k \to b(k)$ be a code of the model classes $\mathcal{M}_k$ which maps a structure parameter k to a binary sequence b. Its codelength function will be denoted by $L(k)$. Thus, using a model class $\mathcal{M}_k$, the sample $x(n)$ can be encoded by $C_k^{(n)}(x(n))$ and a preamble $C(k)$ identifying $\mathcal{M}_k$. The MDL criterion is the total length of this description:

$$\mathrm{MDL}_k(x(n)) = L_k^{(n)}(x(n)) + L(k).$$

The MDL estimator selects the model class which provides the shortest description of the sample:

$$\hat{k}(x_1^n) = \arg\min_{k \in \mathcal{K}} \mathrm{MDL}_k(x(n)).$$

Assume for simplicity that A is finite and also that each model M_{θ_k} uniquely determines a hypothetical distribution of the process; as before, its marginal for the sample $x(n)$ is denoted by $P_{\theta_k}^{(n)}$. A uniquely decodable, variable-length binary code $C_k^{(n)}$ can be represented by a *coding distribution* $P_k^{(n)}$. To see this, note the well-known fact that $L_k^{(n)}$ is the codelength function of some uniquely decodable code $C_k^{(n)}$ if and only if it satisfies the Kraft inequality $\sum_{x(n)} 2^{-L_k^{(n)}(x(n))} \leq 1$. We may assume that here the equality holds, for otherwise the code could be improved by shortening some codewords. Clearly, for any code $C_k^{(n)}$ with codelength $L_k^{(n)}$ which satisfies the Kraft inequality with the equality, we can write $P_k^{(n)}(x(n)) = 2^{-L_k^{(n)}(x(n))}$. On the other hand, for any probability distribution $P_k^{(n)}$ we can construct a uniquely decodable code $C_k^{(n)}$ with codelength $L_k^{(n)}(x(n)) = \left\lceil -\log P_k^{(n)}(x(n)) \right\rceil$, called a Shannon code. The code determined by the coding distribution $P_k^{(n)}$ will be referred to as $P_k^{(n)}$-code. It should be chosen to be optimal in some sense under the assumption that the true model M_{θ_0} is in the model class $\mathcal{M}_k$. Note, however, that $P_k^{(n)}$ will typically differ from each $P_{\theta_k}^{(n)}$ in the model class $\mathcal{M}_k$.

The redundancy of a $P_k^{(n)}$-code relative to the true distribution $P_{\theta_0}^{(n)}$ is

$$R_{\theta_0}^{(n)}(x(n)) = -\log P_k^{(n)}(x(n)) + \log P_{\theta_0}^{(n)}(x(n)).$$

This is the difference of the codelength due to using the coding distribution $P_k^{(n)}$ instead of the true $P_{\theta_0}^{(n)}$. Since the redundancy is a function of the sample, to evaluate the goodness of $P_k^{(n)}$ one usually considers either the maximum of $R_{\theta_0}^{(n)}(x(n))$ for all possible $x(n)$, or its expectation with respect to $P_{\theta_0}^{(n)}$. Moreover, since the true distribution $P_{\theta_0}^{(n)}$ is unknown, as an optimality criterion for $P_k^{(n)}$ under the assumption $M_{\theta_0} \in \mathcal{M}_k$ it is usual to consider worst case maximum or expected redundancy for all feasible distributions $P_{\theta_k}^{(n)}$, $\theta_k \in \Theta_k$, in the role of $P_{\theta_0}^{(n)}$.

For the model class $\mathcal{M}_k$, the worst case maximum redundancy of a $P_k^{(n)}$-code is

$$R^{(n)*} = \sup_{\theta_k \in \Theta_k} \max_{x(n)} R_{\theta_k}^{(n)}(x(n)).$$

It is easy to show that the coding distribution $P_k^{(n)}$ minimizing this quantity is the *Normalized Maximum Likelihood (NML)* distribution defined as

$$\mathrm{NML}_k^{(n)}(x(n)) = P_{\hat{\theta}_k}^{(n)}(x(n)) \Bigg/ \sum_{x'(n)} P_{\hat{\theta}_k}^{(n)}(x'(n)) \ ,$$

where $\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} P_{\theta_k}^{(n)}(x(n))$ is the maximum likelihood estimator of the parameter vector θ_k in the model class $\mathcal{M}_k$.

Using this coding distribution we get the MDL criterion

$$\mathrm{MDL}_k(x(n)) = -\log P_{\theta_k}^{(n)}(x(n)) + \log \left(\sum_{x'(n)} P_{\hat{\theta}_k}^{(n)}(x'(n)) \right) + L(k).$$

Shtarkov (1977) showed that for the case of Markov chains the middle term is asymptotically (as $n \to \infty$, with k fixed) equal to $(1/2)(\dim \Theta_k) \log n$. The same holds also in other cases, under suitable regularity conditions, see Rissanen (1996). Hence, when the number of the candidate model classes is finite, the NML version of the MDL criterion is asymptotically equivalent to BIC.

For the model class $\mathcal{M}_k$, the worst case expected redundancy of a $P_k^{(n)}$-code is

$$\begin{aligned}\overline{R}^{(n)} &= \sup_{\theta_k \in \Theta_k} \mathbf{E}_{\theta_k} \left(R_{\theta_k}^{(n)}(x(n)) \right) = \sup_{\theta_k \in \Theta_k} \sum_{x(n)} P_{\theta_k}^{(n)}(x(n)) \log \frac{P_{\theta_k}^{(n)}(x(n))}{P_k^{(n)}(x(n))} \\ &= \sup_{\theta_k \in \Theta_k} \mathbf{D}(P_{\theta_k}^{(n)} \,\|\, P_k^{(n)}),\end{aligned}$$

where $\mathbf{E}_{\theta_k}(\cdot)$ denotes the expected value with respect to the distribution $P_{\theta_k}^{(n)}$, and $\mathbf{D}(\cdot \,\|\, \cdot)$ is the Kullback-Leibler information divergence. The coding distribution minimizing this quantity can not be given explicitly as above, but one is almost as good as the exact minimizer can often be.

Concentrating on the Markov chain case, consider first the model class $\mathcal{M}$ equal to $\mathcal{M}_k$ with $k = 0$, the class of i.i.d. processes. In this case, a good coding distribution is the *Krichevsky-Trofimov (KT)* distribution (Krichevsky and

Trofimov, 1981). Its worst case expected redundancy over $\mathcal{M}$ approaches the minimum up to a constant not depending on n. The KT distribution is defined as the mixture of all i.i.d. distributions $P_\theta^{(n)}$, with respect to the Dirichlet distribution μ of parameters 1/2:

$$\mathrm{KT}_0(x_1^n) = \int P_\theta^{(n)} \mu(d\theta).$$

Direct calculation gives the following explicit expression:

$$\mathrm{KT}_0(x_1^n) = \frac{\prod_{a:N_n(a)\geq 1}[(N_n(a)-\frac{1}{2})(N_n(a)-\frac{3}{2})\cdots(\frac{1}{2})]}{(n-1+\frac{|A|}{2})(n-2+\frac{|A|}{2})\cdots(\frac{|A|}{2})},$$

where $N_n(a)$ denotes the number of occurrences of $a \in A$ in the sample x_1^n.

For the Markov chains of order k we have a similar result. For the model class $\mathcal{M}_k$ a good coding distribution is the Krichevsky-Trofimov distribution of order k, denoted by KT_k. It is a mixture of all distributions of form

$$P_{\theta_k}^{(n)}(x_1^n) = \frac{1}{|A|^k} \prod_{i=k+1}^{n} P(x_i|x_{i-k}^{i-1}),$$

see (1.1) with $Q(X_1^k = x_1^k) = |A|^{-k}$, where the parameter vector θ_k specifies the matrix of transition probabilities $P(a|a_1^k)$, $a \in A$, $a_1^k \in A^k$. The mixing distribution is defined by letting the rows $\{P_{\theta_k}(a|a_1^k), a \in A\}$ of this matrix independent and having the Dirichlet distribution μ as above. We also have an explicit form:

$$\mathrm{KT}_k(x_1^n) = \frac{1}{|A|^k} \prod_{a_1^k:N_n(a_1^k)\geq 1} \frac{\prod_{a_{k+1}:N_n(a_1^{k+1})\geq 1}[(N_n(a_1^{k+1})-\frac{1}{2})\cdots(\frac{1}{2})]}{(N_n(a_1^k)-1+\frac{|A|}{2})(N_n(a_1^k)-2+\frac{|A|}{2}\cdots\frac{|A|}{2})},$$

where $N_n(a_1^k)$ denotes the number of occurrences of $a_1^k \in A^k$ in the sample x_1^n.

The distribution KT_k can be calculated recursively in the sample size n as

$$\mathrm{KT}_k(x_1^n) = \frac{N_{n-1}(x_{n-k}^n)+1/2}{N_{n-1}(x_{n-k}^{n-1})+|A|/2}\,\mathrm{KT}_k(x_1^{n-1}).$$

The NML and KT versions of the MDL criterion are asymptotically equivalent, because for the minimizers of the worst case maximum and expected redundancy we have

$$\frac{(|A|-1)|A|^k}{2}\log n - K_1 \leq \min_{P_k^{(n)}} \overline{R}^{(n)} \leq \min_{P_k^{(n)}} R^{(n)^*} \leq \frac{(|A|-1)|A|}{2}\log n - K_2, \tag{1.2}$$

where K_1 and K_2 are constants (depending on k).

The MDL estimator can be regarded as a Bayesian MAP estimator when, as above, the coding distribution $P_k^{(n)}$ is a mixture of the distributions $P_{\theta_k}^{(n)}$, $\theta_k \in \Theta_k$, with a suitable mixing distribution $\mu_k^{(n)}$ defined on Θ_k, that is,

$$P_k^{(n)}(x(n)) = \int_{\Theta_k} P_{\theta_k}^{(n)}(x(n))\mu_k^{(n)}(d\theta_k).$$

Indeed, representing the code $C(k)$ of the model class $\mathcal{M}_k$ with the coding distribution $P(k) = 2^{-L(k)}$, minimization of the description length

$$L_k^{(n)}(x(n)) + L(k) = -\log P_k^{(n)}(x(n)) - \log P(k)$$

is equivalent to maximization of $P(k)P_k^{(n)}(x(n))$. The latter quantity is proportional to the posterior probability of the structure parameter k, that is, to the conditional probability of k given the sample $x(n)$.

The MDL principle can be extended to the case of general A, say $A = \mathbb{R}$, via discretization, and this leads to similar results as above, see Rissanen (1989), in particular, for AR processes Hamerly and Davis (1989), and for ARMA processes Gerencsér (1987).

1.5 Recent results

For various processes it has been proved that BIC and MDL estimators of the structure parameter are strongly consistent. This means that the minimizer of BIC or MDL criterion over the candidate structure parameters is equal to the true structure parameter, eventually almost surely as the sample size tends to infinity. Most consistency proofs in the literature include the assumption that the number of candidate structure parameters is finite, that is, there is a known prior bound on the structure parameter. This assumption is of technical nature and it simplifies the proof. However, it is undesirable, because in practice usually there is no prior information on the structure parameter, moreover when we have increasing amount of data, we would require to take into account more and more complex hypothetical model classes as candidate ones. Therefore, it is a reasonable aim to drop the assumption of prior bound on the structure parameter. Csiszár and Shields (2000) proved that the BIC estimator of the order of Markov chains is strongly consistent even if the assumption of the prior constant bound on the order is dropped and based on the n'th sample x_1^n all possible orders $0 \le k < n$ are considered as candidate orders. At the same time, Csiszár and Shields (2000) pointed out that the same result cannot hold for the MDL estimator. Consider the i.i.d. process with uniform distribution. This process is a Markov chain of order 0. For the MDL criterion

$$\mathrm{MDL}_k(x_1^n) = -\log P_k^{(n)}(x_1^n) + L(k),$$

where the coding distribution $P_k^{(n)}$ is either $\mathrm{NML}_k^{(n)}$ or KT_k, and the codelength $L(k)$ of the order k satisfies $L(k) = o(k)$, we have

$$\hat{k}(x_1^n) = \arg \min_{0 \le k \le \alpha \log n} \mathrm{MDL}_k(x(n)) \to \infty \quad \text{as } n \to \infty,$$

where $\alpha = 4/\log|A|$. This counterexample shows that the MDL estimator fails to be consistent when the prior bound on the order is totally dropped.

Csiszár (2002) proved strong consistency of the MDL estimator of the order of Markov chains when the set of candidate orders is allowed to extend as the sample size n increases, namely, the bound on the orders taken into account is $o(\log n)$ in the KT case, and $\alpha \log n$ with $\alpha < 1/\log|A|$ in the NML case. Let us observe that these MDL estimators need no prior bound on the true order. The consistency was proved for the MDL criterion without the term $L(k)$, which is a stronger result.

1.6 Context tree estimation

The model called *tree model* or *variable length Markov chain* (VLMC) is a refinement of the Markov chain model. Given a Markov chain of order k, for a sequence $a_1^k \in A^k$ the transition probabilities $Q(a|a_1^k)$, $a \in A$ may depend not on the whole sequence $a_1, \dots, a_k$, but only on a subsequence $a_l, \dots, a_k$; this admits a more parsimonious parameterization.

Consider a process with $T = \mathbb{Z}$ and $|A| < \infty$. For simplicity, assume that all finite dimensional marginals of the distribution Q of the process are strictly positive. The string $s = a_{-\ell}^{-1} \in A^\ell$ is a *context* for the process Q if

$$Q(X_i = a | X_{-\infty}^{i-1} = x_{-\infty}^{i-1}) = Q(a|s) \quad \text{for all } i \in \mathbb{Z}, a \in A,$$

with suitable transition probabilities $Q(\cdot|s)$, whenever $x_{i-\ell}^{i-1} = a_{-\ell}^{-1}$, and no substring $a_{-\ell'}^{-1}$, $\ell' < \ell$ has this property. The set of all contexts is called *context tree*, it will be denoted by $\mathcal{T}$. Assume that for every past sequence $x_{i-\ell}^{i-1}$ there exists a context s of finite length ℓ, and the supremum of these lengths is a finite number k. A process with such context tree $\mathcal{T}$ is a Markov chain of order k, but a collection of $(|A|-1)|\mathcal{T}|$ transition probabilities suffices to describe it, instead of $(|A|-1)|A|^k$ ones required for a general Markov chain of order k.

The term context tree refers to its visualization. The contexts s, written backwards, can be regarded as leaves of a tree, where the path from the root to a leaf is determined by the string s. This context tree is complete, that is, each node except the leaves has exactly $|A|$ children.

For a process with context tree $\mathcal{T}$, the probability of a realization x_1^n can be written as

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{s \in \mathcal{T}, a \in A} Q(a|s)^{N_n(s,a)},$$

where $N_n(s,a)$ denotes the number of that occurrences of $a \in A$ in the sequence x_{k+1}^n when the context $s \in \mathcal{T}$ precedes a. Given the sample x_1^n, the maximum of the second factor above is attained for $Q(a|s) = \frac{N_n(s,a)}{N_n(s)}$, $a \in A$, $s \in \mathcal{T}$, where $N_n(s) = \sum_{a \in A} N_n(s,a)$. That is, the maximum likelihood estimates of the transition probabilities are the empirical probabilities, as for Markov chains.

For the class of tree models as above, the context tree $\mathcal{T}$ plays the role of the structure parameter. To the analogy of Markov chains, the Bayesian information criterion for the model class determined by $\mathcal{T}$ is

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) = - \sum_{s \in \mathcal{T}, a \in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)} + \frac{(|A|-1)|\mathcal{T}|}{2} \log n.$$

The MDL principle can also be formulated for tree models. One can define as before the NML and KT distributions for the class of tree models with context tree $\mathcal{T}$, here we concentrate on the KT distribution. This has the following explicit form:

$$KT_{\mathcal{T}}(x_1^n) = \frac{1}{|A|^k} \prod_{s: N_n(s) \geq 1} \frac{\prod_{a: N_n(s,a) \geq 1} [(N_n(s,a) - \frac{1}{2})(N_n(s,a) - \frac{3}{2}) \dots (\frac{1}{2})]}{(N_n(s) - 1 + \frac{|A|}{2})(N_n(s) - 2 + \frac{|A|}{2}) \dots (\frac{|A|}{2})} \tag{1.3}$$

We can calculate it recursively in n as

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = \frac{N_{n-1}(s, x_n) + 1/2}{N_{n-1}(s) + |A|/2} \mathrm{KT}_{\mathcal{T}}(x_1^{n-1}),$$

where s is the context for the past $x_{-\infty}^{n-1}$. Similarly as for the class of k'th order Markov chains, the coding distribution $\mathrm{KT}_{\mathcal{T}}$ minimizes the worst case expected redundancy for the class of tree models with context tree $\mathcal{T}$, up to an additive constant. Moreover, the bounds (1.2) also hold if $|A|^k$ is replaced by $|\mathcal{T}|$. Note also that there is a simply code C (see, e.g., Willems, Shtarkov, and Tjalkens 1995) which describes a context tree $\mathcal{T}$ with codelength

$$L(\mathcal{T}) = \frac{|A||\mathcal{T}| - 1}{|A| - 1}. \tag{1.4}$$

For model selection in the tree model setting, instead of using the information criteria above, mostly variants of Rissanen's "context" algorithm (1983a) have been used. The reason is computational complexity: as the number of possible context trees of depth at most D is very large if D is large, it is not feasible to calculate an information criterion for all candidate context trees and choose the context tree with minimal value. This problem has been partially overcome by Willems, Shtarkov, and Tjalkens (2000). They proved that the MDL estimator with the KT coding distribution (1.3) and the context tree codelength (1.4) is strongly consistent, when there is a prior bound D on the depth of candidate context trees. At the same time they presented an algorithm, called *Context Tree Maximizing* (CTM) method, to calculate the estimator without actually computing and comparing the KT values for all candidate context trees. For the n'th sample x_1^n, the number of elementary computations required by the algorithm is proportional to nD, that is, the algorithm is of linear time.

1.7 Sparse tree models

In a kth order Markov chain, the probability of a sample x_t in a sequence $x_1^n = x_1, x_2, \ldots, x_n$ defined over a finite alphabet A is given by a discrete conditional distribution $Q(x_t|x_{t-k}^{t-1})$, conditioned on the value of the consecutive K-tuple immediately preceding x_t. Thus, the process can be parametrized by a kth order Markov model with $|A|^k(|A|-1)$ free parameters.[1] In a tree model (Rissanen, 1983a; Buhlmann and Wyner, 1998) of the same process, the memory length is allowed to vary from location to location in the data. These models are a more efficient parametrizations of the process, as the exponential number of statistical parameters in the Markov model can often be dramatically reduced in a well tuned tree model. Besides economically capturing redundancies typical of real life data, a major advantage of tree models is that the statistical information needed to optimize the model can be stored in a context tree, which grows as the sequence is observed, recording essentially all the occurrences of each letter in every context, this recursive combinatorial structure is the key for algorithmic efficiency.

The savings in the number of statistical parameters realized by tree models can be seen as the result of lumping together equivalent states, i.e., k-tuples that induce the same conditional distribution in the full Markov model. Thus, a conditioning state (leaf) at depth $k-\ell$ in a tree model for a kth order process corresponds to the merging of $|A|^{k-\ell}$ equivalent states in the full Markov model.

[1] We distinguish between the process (which is always a Markov chain) and its parametrizations. Thus, a kth order Markov model is the most natural parametrization, but our main goal is to study other, more efficient, parametrizations.

This observation, in fact, characterizes the special structure of the sets of full-length states that can be lumped together in a tree model: each such set must consist of all the extensions of a given string of length $k - \ell$, $0 \leq \ell \leq k$. With real data, other sets of equivalent states might arise, and it is natural to ask whether it is possible to optimize models where more general sets states are allowed to merge. In practice, the distributions that are merged are empirical and not necessarily identical, and a search for the best state space partition is unfeasible. The problem is also known as the *context quantization* problem, and it has been studied in various settings, with the proposed solutions being generally ad-hoc, and of varying degrees of complexity as a function of k, which is usually kept relatively small (see, e.g., Forchhammer et al. (2004), for a recent setting).

In other words variable length Markov chains reduce the number of parameters by merging complete sets of equivalent states descending from a common suffix. These states have (or are estimated to have) equal or "close enough" conditional distributions. Still, VLMC have the restriction that the contexts must be a sequence of contiguous symbols and that any two states merged must have a common suffix with the same distribution in order to form a complete tree.

For example if we have that $Q(a|x_{n-\ell}^{n})$ only depends on x_n and $x_{n-\ell}$ the tree model would be represented by a complete tree of depth ℓ. This would have $|A|^{\ell}$ leaves but there would be only $|A| \times |A|$ different parameters among the ones associated with each leaf. We would like to represent this model in a more parsimonious way, if possible with only $|A| \times |A|$ different states. We are interested in schemes that merge more general state sets, not necessarily obeying the "complete tree" restriction.

In many applications it makes sense to look for statistical dependencies in a close contiguous neighborhood, that makes Markov models very popular. Nevertheless, the model description cost grows exponentially with the memory of the model. One way to overcome this problem is the idea of tree models which allow a variable memory whose length depends on the values of the closest contiguous samples. However, the example above shows the condition of contiguity can be very restrictive. Our approach is to allow non-contiguous positions in the context in order to capture dependencies between distant samples without increasing the model cost. We propose to optimize the set of conditioning positions for some given data, assuming a maximum memory length k and some restrictions.

In this dissertation we are going to propose a new type of Markov models with sparse dependencies. We study *sparse tree models* (STMs), a generalization of tree models where more general sets of states with similar conditional distributions are allowed to merge, while preserving the attractive combinatorial properties of the tree structure. In some cases, this structure will allow us to find, efficiently, the best model for a given input sequence, from a broad subclass of STMs (in a sense to be made precise in the sequel). In an STM, samples will be conditioned on finite strings of *not necessarily contiguous* past symbols, and the context will determine not only how far into the past the conditioning samples are, but also what their (possibly non-contiguous) location are. Non-contiguous conditioning contexts have also been studied in Eskin, Grundy, and Singer (2000); Bourguignon and Robelin (2004); Zhao, Huang, and Speed (2004); Leonardi (2006), most of them are applications of similar models and their estimation mainly in Biology, particularly in DNA and protein modeling.

We choose a fixed symbol $\phi \notin A$, and we denote by A_ϕ the *expanded alphabet* $A_\phi = A \cup \{\phi\}$. Strings over A_ϕ will be referred to as *patterns*. We will interpret ϕ as a wildcard, given a pattern $w_1^m \in A_\phi^m$, sequences in the set

$$\mathcal{C}(w_1^m) = \{\, u_1^m \in A^m \,|\, u_i = w_i \ \text{ whenever } w_i \neq \phi \,\}$$

are said to be *consistent* or *conformal* with w_1^m.

We say a pattern v is a *ϕ-suffix* of w, denoted by $v \prec_\phi w$, when $l(v) \leq l(w)$ and $v_{n-i} = w_{m-i}$ for all $i \leq l(v)$ such that $v_{n-i} \neq \phi$. A set $\mathcal{T}$ of patterns is called a *sparse context tree* if no $w \in \mathcal{T}$ is a ϕ-suffix of any other $v \in \mathcal{T}$ ($\mathcal{T}$ is a ϕ-suffix free set). Moreover, a sparse tree is said to be *complete* when every sequence x_1^n large enough has a ϕ-suffix in the tree.

Consider a stationary Markov process $\{X_n\}_{n\in\mathbb{N}}$ taking values on A, given a sparse context tree $\mathcal{T}$ we say that the process is *$\mathcal{T}$-adapted* if $\forall x_1^n$ such that $w \prec_\phi x_1^n$ (since $\mathcal{T}$ is a ϕ-suffix free set, w is unique) we have that

$$\mathbf{P}(X_{n+1} = a | X_1^n = x_1^n) = Q(a|w), \qquad \forall a \in A.$$

We define the KT distribution of x_1^n corresponding to $\mathcal{T}$ as

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^k} \prod_{w\in\mathcal{T}} \frac{\prod_{a\in A}(N_n(w,a) - \frac{1}{2})(N_n(w,a) - \frac{3}{2})\dots\frac{1}{2}}{(N_n(w) - 1 + \frac{|A|}{2})(N_n(w) - 2 + \frac{|A|}{2})\dots\frac{|A|}{2}}$$

where $N_n(w,a)$ denotes the number of occurrences of the pattern w followed by the symbol a as in the VLMC case.

We show that the MDL estimator based on the KT distribution is strongly consistent for the estimation of sparse context model class. Furthermore, we propose an efficient way to calculate the estimator. Some of the results are based on part of my Masters thesis in Computer Science "Universal Coding via Sparse Tree Models" (Fraiman, 2008).

CHAPTER 2

Context tree estimation for unbounded memory processes

2.1 Introduction

In this chapter, *process* always means a stationary ergodic stochastic process with finite alphabet. Processes are often described by the collection of the conditional probabilities of the possible symbols given the infinite pasts. When these probabilities depend on at most k previous symbols, the process is a Markov chain of order k.

The number of parameters of a general Markov chain grows exponentially with the order. A more efficient description is possible if the strings determining the conditional probabilities -referred to as *contexts*- are of variable length, sometimes substantially shorter than the order k. Models of this kind and the term context tree date back to Rissanen (1983a). These models are also called finite memory sources or tree sources (Weinberger, Lempel, and Ziv, 1992), (Weinberger, Rissanen, and Feder, 1995), (Willems, Shtarkov, and Tjalkens, 1995) or variable length Markov chains (Buhlmann and Wyner, 1998). We note that the terms context and context tree appear in the literature in various senses. Here, the context tree of a finite memory process means, in effect, the minimal tree admitting a tree source representation of the process; the exact definition will be given in the next Section.

As indicated above, the context tree model is typically used to more efficiently describe certain Markov chains (of finite order k) and, accordingly, the context tree has finite depth k. In this work, we drop the finite depth requirement, admitting also non-Markov processes. The term “infinite-depth context tree" appears in (Willems, 1998) in a different sense, as a tree assigned to an observed sequence, with an "indeterminate symbol" ε such that infinitely many ε's may precede a finite number of symbols of the true alphabet. A concept of generalized context tree, (see Martín et al., 2004, and references therein), admits edges labeled by strings rather than single symbols. That concept is not used here, but similarly to (Martín, Seroussi, and Weinberger, 2004) we drop the completeness requirement, often made in the literature, that each non-leaf node of the context tree has as many children as the alphabet size. If some strings have zero probability for the given process, these can not be contexts, and then the context tree need not be complete.

We address the problem of statistical estimation of the context tree in the indicated generality, based on an observed finite realization of the process, of length $n \to \infty$. This task, for finite depth context trees, has been considered, among others, in the references above. Variants of Rissanen's "context" algorithm (1983a) are popular. In particular, Buhlmann and Wyner (1998) proved the consistency of such an algorithm not assuming a known prior bound on the depth of the context tree, but using a bound allowed to grow with n. They asserted that standard statistical methods as the Bayesian information criterion (BIC) of (Schwarz, 1978) and the minimum description length (MDL) principle of Rissanen (1989), (Baron, Rissanen, and Yu, 1998), were inappropriate for context tree estimation, due to computational infeasibility of comparing a very large number of hypothetical models. Still, Willems, Shtarkov, and Tjalkens (2000) showed that time-consuming comparisons can be avoided by clever use of tree techniques. Their context tree maximizing (CTM) algorithm computes in linear time the context tree estimator obtained by the version of MDL that uses the Krichevsky-Trofimov (KT) code length (Krichevsky and Trofimov, 1981), and this estimator is consistent, as they proved assuming a known upper bound on the depth of the context tree. Similar results were obtained also by Nohre (1994). Recent results on consistent context tree estimation in linear time, assuming finite depth but no known upper bound on it, appear in (Baron and Bresler, 2004) and Martín et al. (2004). These references use tools as the Burrows-Wheeler transform or generalized context trees.

Recently Csiszár and Talata (2006) showed the consistency of a context tree estimation method via BIC. While BIC is commonly regarded as an approximate version of MDL, this is justified only when a finite number of model classes is considered, see (Csiszár and Shields, 2000). We note that much of the literature of context tree models is motivated by universal source coding. In particular, CTM is a modification of the celebrated Context Tree Weighting data compression algorithm of Willems, Shtarkov, and Tjalkens (1995).

In this chapter, we prove that both MDL with KT code length and BIC provide strongly consistent estimators of the context tree if the set of candidate context trees is suitably chosen; finiteness or completeness of the true context tree is not required. Moreover, these estimators can be implemented in linear time. The set of candidate context trees is specified by a bound on the length of the hypothetical contexts, allowed to grow as $o(\log n)$, and in one case by an additional condition on their occurrences in the observed sample. Strong consistency means in the finite depth case that the estimated context tree is equal to the true one, eventually almost surely as $n \to \infty$, while otherwise, that the estimated context tree truncated at any fixed level is equal to the true one truncated at the same level, eventually almost surely as $n \to \infty$. This chapter is strongly based on Csiszár and Talata (2006), Csiszár (2002), and Csiszár and Shields (2000).

For order estimation of Markov chains, it is well known that BIC and MDL, both with the KT and the normalized maximum likelihood (NML) code length, are strongly consistent when the number of candidate model classes is finite, that is, when there is a known upper bound on the order (Finesso, 1992). The consistency of the BIC order estimator without such prior bound has been proved by Csiszár and Shields (2000). That paper also contains a counterexample to the consistency of the KT and NML versions of MDL without any bound on the order, or with a bound depending on the sample size n, equal to a sufficiently large constant times $\log n$. The consistency of the latter order estimators with bound $o(\log n)$, respectively, $O(\log n)$, on the order was proved by Csiszár (2002).

Linear time implementation of the context tree estimators is achieved via the CTM method (Willems, Shtarkov, and Tjalkens, 2000). This has been developed for the KT version of MDL, Csiszár and Talata (2006) proved that also the BIC estimator admits a CTM-like implementation. The same does not seem to hold for the NML version of MDL, this is why the latter is not considered in this work.

By the consistency result, if the context tree of a process has finite depth, it can be exactly recovered, with probability 1, when the sample size is large enough; the sample size actually needed remains, however, unknown. A heuristic rule might be to stop when the estimated context tree "stabilizes", that is, it remains unchanged when the sample size n runs over a large interval. The last result in this chapter shows that (slightly modified versions of) the estimators can be calculated on-line in such a way that $o(n \log n)$ time suffices to calculate them for all sample sizes $i \leq n$. This implies that the above stopping rule can be implemented with only a small increment in the order of required computations.

The structure of the chapter is the following. In Section 2.2, we introduce the notation and definitions, on Sections 2.3 and 2.4 and formulate and prove the results for the BIC estimator and the KT estimator about strong consistency. In Section 2.5 we prove the under and overestimation lemmas used in the proof of the consistency theorems. Sections 2.6 and 2.7 contain auxiliary results on typicality and martingale techniques. Finally, in Section 2.8, we introduce the algorithms for calculating the estimators, and establish their computational complexity both for off-line and on-line calculations.

2.2 Notation and Statement of the Main Results

For a finite set A we denote its cardinality by $|A|$. A string $s = a_m a_{m+1} \dots a_n$ (with $a_i \in A$, $m \le i \le n$) is denoted also by an a_m^n; its length is $l(s) = n-m+1$. The empty string is denoted by λ, its length is $l(\lambda) = 0$. The concatenation of the strings u and v is denoted by uv. We say that a string v is a suffix of a string s, denoted by $s \succeq v$, when there exists a string u such that $s = uv$. For a proper suffix, that is, when $s \neq v$, we write $s \succ v$. A suffix of a semi-infinite sequence $a_{-\infty}^{-1} = \dots a_{-k} \dots a_{-1}$ is defined similarly. Note that in the literature $\succ$ more often denotes the prefix relation.

A set $\mathcal{T}$ of strings, and perhaps also of semi-infinite sequences, is called a tree if no $s_1 \in \mathcal{T}$ is a suffix of any other $s_2 \in \mathcal{T}$.

Each string $s = a_1^k \in \mathcal{T}$ is visualized as a path from a leaf to the root (drawn with the root at the top), consisting of k edges labeled by the symbols $a_1 \dots a_k$. A semi-infinite sequence $a_{-\infty}^{-1} \in \mathcal{T}$ is visualized as an infinite path to the root, see Figure 2.2. The strings $s \in \mathcal{T}$ are identified also with the leaves of the tree $\mathcal{T}$, leaf s is the leaf connected with the root by the path visualizing s as above. Similarly, the nodes of the tree $\mathcal{T}$ are identified with the finite suffixes of all (finite or infinite) $s \in \mathcal{T}$, the root being identified with the empty string λ. The children of a node s are those strings as, $a \in A$, that are themselves nodes, that is, suffixes of some $s' \in \mathcal{T}$.

The tree $\mathcal{T}$ is complete if each node except the leaves has exactly $|A|$ children. A weaker property we shall need is irreducibility, which means that no $s \in \mathcal{T}$ can be replaced by a proper suffix without violating the tree property. The family of irreducible trees will be denoted by $\mathcal{I}$.

Denote $d(\mathcal{T})$ the depth of the tree $\mathcal{T}$: $d(\mathcal{T}) = \max\{l(s), s \in \mathcal{T}\}$. Let $\mathcal{T}|_K$ denote the tree $\mathcal{T}$ truncated at level K:

$$\mathcal{T}|_K = \{s' : s' \in \mathcal{T} \text{ with } l(s') \le K \text{ or } l(s') = K \text{ and } s' \preccurlyeq s \in \mathcal{T}\}. \tag{2.1}$$

Let us be given a stationary ergodic stochastic process $\{X_i, -\infty < i < \infty\}$ with finite alphabet A. Write

$$Q(a_m^n) = \mathbf{P}(X_m^n = a_m^n)$$

and, if $s \in A^k$ has $Q(s) > 0$, write

$$Q(a|s) = \mathbf{P}(X_0 = a|X_{-k}^{-1} = s).$$

A process as above will be referred to as process Q.

Definition 2.1. A string $s \in A^k$ is a *context* for a process Q if $Q(s) > 0$ and

$$\mathbf{P}(X_0 = a|X_{-\infty}^{-1} = x_{-\infty}^{-1}) = Q(a|s), \qquad \text{for all } a \in A$$

whenever s is a suffix of the semi-infinite sequence $x_{-\infty}^{-1}$, and no proper suffix of s has this property. An infinite context is a semi-infinite sequence $x_{-\infty}^{-1}$ whose suffixes x_{-k}^{-1}, $k = 1, 2, \ldots$ are of positive probability but none of them is a context.

Clearly, the set of all contexts is a tree. It will be called the context tree of the process Q, denoted by $\mathcal{T}_0$.

Remark. The context tree $\mathcal{T}_0$ has to be complete if $Q(s) > 0$ for all strings s. In general, for each node s of $\mathcal{T}_0$ which is not a leaf, exactly those as, $a \in A$, are the children of s for which $Q(as) > 0$. Moreover, Definition 2.1 implies that the context tree is always irreducible, $\mathcal{T}_0 \in \mathcal{I}$.

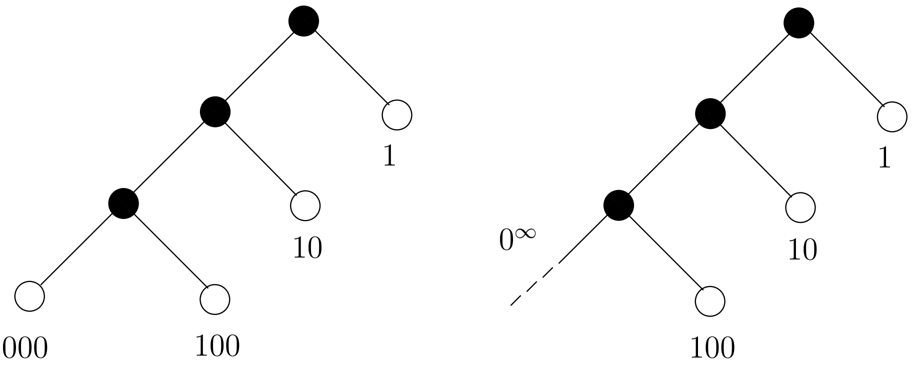
When the context tree has depth $d(\mathcal{T}_0) = k_0 < \infty$, the process Q is a Markov chain of order k_0. In this case the context tree leads to a parsimonious description of the process, because a collection of $(|A| - 1)|\mathcal{T}_0|$ transition probabilities suffices to describe the process, instead of $(|A| - 1)|A|^{k_0}$ ones. Note that the context tree of an i.i.d. process consists of the root λ only, thus, $|\mathcal{T}_0| = 1$.

Example (Renewal Process)**.** Let $A = \{0, 1\}$ and suppose that the distances between the occurrences of 1's are i.i.d. Denote p_j the probability that this distance is j, that is, $p_j = Q(10^{j-1}1)/Q(1)$, and let $q_k = \sum_{j=k}^{\infty} p_j$, $k \leq 1$. Then for $k \leq 1$ we have

$$Q(1|10^{k-1}) = p_k/q_k \overset{\text{def}}{=} Q_k$$

(undefined if $q_k = 0$). Setting $Q_0 = Q(1)$, denote by k_0 the smallest integer such that Q_k is constant or undefined for $k \geq k_0$, or $k = \infty$ if no such integer exists. Then the contexts are the strings 10^{i-1}, $i \leq k_0$, and the string 0^{k_0} (if $k_0 < \infty$) or the semi-infinite sequence 0^{∞} (if $k_0 = 1$), see Figure 2.2.

In this work, we are concerned with the statistical estimation of the context tree $\mathcal{T}_0$ from the sample x_1^n, a realization of X_1^n. We demand strongly consistent estimation. We mean by this in the case $d(\mathcal{T}_0) < \infty$ that the estimated context



(a) $k_0 = 3$

(b) $k_0 = \infty$

Figure 2.1: Context tree of a renewal process.

tree equals $\mathcal{T}_0$ eventually almost surely as $n \to \infty$, while otherwise that the estimated context tree truncated at any fixed level K equals $\mathcal{T}_0|_K$ eventually almost surely as $n \to \infty$, see (2.1). Here and in the sequel, "eventually almost surely" means that with probability 1 there exists a threshold n_0 (depending on the infinite realization x_1^∞) such that the claim holds for all $n \geq n_0$.

Let $N_n(s, a)$ denote the number of occurrences of the string $s \in A^{l(s)}$ followed by the letter $a \in A$ in the sample x_1^n, where s is supposed to be of length at most $D(n)$, specified later, and-for technical reasons-only the letters in positions $i > D(n)$ are considered

$$N_n(s, a) = \#\{i : D(n) < i \leq n, x_{i-l(s)}^{i-1} = s, x_i = a\}.$$

The number of such occurrences of s is denoted by $N_n(s)$

$$N_n(s) = \#\{i : D(n) < i \leq n, x_{i-l(s)}^{i-1} = s\}.$$

Given a sample x_1^n, a feasible tree is any tree $\mathcal{T}$ of depth $d(\mathcal{T}) \leq D(n)$ such that $N_n(s) \geq 1$ for all $s \in \mathcal{T}$, and each string s' with $N_n(s') \geq 1$ is either a suffix of some $s \in \mathcal{T}$ or has a suffix $s \in \mathcal{T}$. A feasible tree $\mathcal{T}$ is called r-frequent if $N_n(s) \geq r$ for all $s \in \mathcal{T}$. The family of all feasible, respectively, r-frequent trees is denoted by $\mathcal{F}_1(x_1^n, D(n))$, respectively, $\mathcal{F}_r(x_1^n, D(n))$. Clearly,

$$\sum_{a \in A} N_n(s, a) = N_n(s), \quad \text{and} \quad \sum_{s \in \mathcal{T}} N_n(s) = n - D(n)$$

for any feasible tree $\mathcal{T}$. Regarding such a tree $\mathcal{T}$ as the context tree of a hypothetical process Q', the probability of the sample x_1^n can be written as

$$Q'(x_1^n) = Q'(x_1^{D(n)}) \prod_{s \in \mathcal{T}, a \in A} Q'(a|s)^{N_n(s,a)}.$$

With some abuse of terminology, for a hypothetical context tree $\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n))$ we define the maximum likelihood $P_{\mathrm{ML},\mathcal{T}}(x_1^n)$ as the maximum in $Q'(a|s)$ of the second factor above, that is,

$$P_{\mathrm{ML},\mathcal{T}}(x_1^n) = \prod_{s\in\mathcal{T}, N_n(s)\geq 1} \prod_{a\in A} \left(\frac{N_n(s,a)}{N_n(s)} \right)^{N_n(s,a)} \tag{2.2}$$

2.3 Consistency of the BIC criterion

We investigate two information criteria to estimate $\mathcal{T}_0$, both motivated by the MDL principle. An information criterion assigns a score to each hypothetical model (here, context tree) based on the sample, and the estimator will be that model whose score is minimal.

Definition 2.2. Given a sample x_1^n, the BIC for a feasible tree $\mathcal{T}$ is

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) = \mathrm{ML}_{\mathcal{T}}(x_1^n) + \frac{(|A|-1)|\mathcal{T}|}{2} \log n,$$

where $\mathrm{ML}_{\mathcal{T}}(x_1^n) = -\log P_{\mathrm{ML},\mathcal{T}}(x_1^n)$.

Remark. Characteristic for BIC is the "penalty term" half the number of free parameters times $\log n$. Here, a process Q with context tree $\mathcal{T}$ is described by the conditional probabilities $Q(a|s)$, $a \in A$, $s \in \mathcal{T}$, and $(|A|-1)|\mathcal{T}|$ of these are free parameters when the tree $\mathcal{T}$ is complete. For a process with an incomplete context tree, the probabilities of certain strings must be 0, hence, the number of free parameters is typically smaller than $(|A|-1)|\mathcal{T}|$ when $\mathcal{T}$ is not complete. Thus, Definition 2.2 involves a slight abuse of terminology. We note that replacing $(|A|-1)/2$ in Definition 2.2 by any $c > 0$ would not affect the results below and their proofs. In the literature, context trees are often required to be complete. This can be achieved by adding dummy edges if necessary, but this increases the penalty term in Definition 2.2, and the analog of Theorem 2.1 below appears a weaker result for completed context trees.

It is known (Csiszár and Shields, 2000) that for estimating the order of Markov chain, the BIC estimator is consistent without any restriction on the hypothetical orders. The following theorem does need a bound on the depth of the hypothetical context trees. Still, as this bound grows with the sample size n, no a priori bound on the size of the unknown $\mathcal{T}_0$ is required; in fact, even $d(\mathcal{T}_0) = \infty$ is allowed. Note also that the presence of this bound decreases computational complexity.

Remark. In Theorems 2.1 and 2.2 later, the indicated minimum is certainly attained, as the number of feasible trees is finite, but the minimizer is not necessarily unique; in that case, either minimizer can be taken as arg min.

Theorem 2.1. *In the case $d(\mathcal{T}_0) < \infty$, the* BIC *estimator*

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}} \mathrm{BIC}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$, satisfies

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \mathcal{T}_0$$

eventually almost surely as $n \to \infty$.

In the general case, this estimator satisfies for any constant K

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n)|_K = \mathcal{T}_0|_K$$

eventually almost surely as $n \to \infty$.

Proof. It suffices to prove the second assertion of the theorem. Fix an arbitrary constant K. It suffices to show that if $\mathcal{T}|K \neq \mathcal{T}_0|K$ for some $\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$ then there exists a modification $\mathcal{T}'$ of $\mathcal{T}$ also satisfying $\mathcal{T}' \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$ such that

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) > \mathrm{BIC}_{\mathcal{T}}(x_1^n) \tag{2.3}$$

simultaneously for all considered trees T, eventually almost surely as $n \to \infty$.

According to (2.2), the maximum likelihood factorizes as

$$P_{\mathrm{ML},\mathcal{T}}(x_1^n) = \prod_{s \in \mathcal{T}} \tilde{P}_{\mathrm{ML},s}(x_1^n)$$

where

$$\tilde{P}_{\mathrm{ML},s}(x_1^n) = \begin{cases} \prod_{a \in A} \left(\frac{N_n(s,a)}{N_n(s)} \right)^{N_n(s,a)} & \text{if } N_n(s) \geq 1, \\ 1 & \text{if } N_n(s) = 0. \end{cases} \tag{2.4}$$

Using this and the definition of BIC, see Definition 2.2, (2.3) is equivalent to

$$\sum_{s \in \mathcal{T}} \log P_{\mathrm{ML},s}(x_1^n) - \sum_{s' \in \mathcal{T}'} \log P_{\mathrm{ML},s'}(x_1^n) < \frac{(|A|-1)}{2} (|\mathcal{T}| - |\mathcal{T}'|) \log n. \tag{2.5}$$

Since $\mathcal{T}$ is a feasible tree by assumption, so is also $\mathcal{T}|_K$ defined by (2.1). For n sufficiently large, so that $N_n(s) \geq 1$ for all s with $l(s) \leq K$, $Q(s) > 0$, it follows by the Remark just after Definition 2.1 that $\mathcal{T}_0|_K$ is feasible, as well.

Hence, the indirect assumption $\mathcal{T}|_K \neq \mathcal{T}_0|_K$ implies that there exist strings $\tilde{s} \in \mathcal{T}|_K$ and $\tilde{s}_0 \in \mathcal{T}_0|_K$ such that either $\tilde{s} \prec \tilde{s}_0$ (underestimation) or $\tilde{s} \succ \tilde{s}_0$ (overestimation). Equivalently, there exist $s \in \mathcal{T}$ and $s_0 \in \mathcal{T}_0$ such that either **(a)** $l(s) < K$, $s \prec s_0$ or **(b)** $l(s') < K$, $s_0 \prec s$.

We claim that a modification $\mathcal{T}'$ of $\mathcal{T}$ with the required properties is

$$\mathcal{T}' = (\mathcal{T} \backslash \{s\}) \cup \tilde{\mathcal{T}} \tag{2.6}$$

in case **(a)**, with $\tilde{T}$ as in Lemma 2.4 below, and

$$\mathcal{T}' = (\mathcal{T} \backslash \tilde{\mathcal{T}}) \cup \{w\} \tag{2.7}$$

in case **(b)**, with $\tilde{\mathcal{T}}$ and w as in Lemma 2.5 below. The properties of $\tilde{\mathcal{T}}$ in Lemmas 2.4 and 2.5 immediately imply that the condition $\mathcal{T}' \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}$ is satisfied in both cases **(a)** and **(b)** (in case **(a)**, $\mathcal{T}' \in \mathcal{F}_1(x_1^n, D(n))$ holds by the ergodic theorem, eventually almost surely as $n \to \infty$). Thus, it remains to check (2.5) for this choice of $\mathcal{T}'$.

In case **(a)**, for $\mathcal{T}'$ given by (2.6), we have $|\mathcal{T}| - |\mathcal{T}'| = 1 - |\tilde{\mathcal{T}}|$, and the left-hand side of (2.5) is equal to that of (2.13) below. By Lemma 2.4, the latter is less than $-cn$, eventually almost surely as $n \to \infty$, and thus (2.5) certainly holds. Regarding simultaneity for all considered trees $\mathcal{T}$, note that $\tilde{\mathcal{T}}$ corresponding to a particular $\mathcal{T}$ may be chosen depending on s only, and the number of strings s with $l(s) \leq K$ is finite.

In case **(b)**, for $\mathcal{T}'$ given by 2.7, we have $|\mathcal{T}| - |\mathcal{T}'| = |\tilde{\mathcal{T}}| - 1$, and the left-hand side of (2.5) is equal to that of (2.14) below. Hence, by Lemma 2.5, (2.5) is satisfied also in this case, eventually almost surely as $n \to \infty$ for all considered $\mathcal{T}$. □

2.4 Consistency of the MDL principle

The other information criterion we consider is the MDL based on the the Krichevsky-Trofimov code length (Krichevsky and Trofimov, 1981), (Willems, Shtarkov, and Tjalkens, 1995).

Definition 2.3. Given a sample x_1^n, the KT criterion for a feasible tree $\mathcal{T}$ is

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = -\log P_{\mathrm{KT},\mathcal{T}}(x_1^n);$$

where

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^{D(n)}} \prod_{s \in \mathcal{T}} \frac{\prod_{a: N_n(s,a) \geq 1} [(N_n(s,a) - \frac{1}{2})(N_n(s,a) - \frac{3}{2}) \cdots (\frac{1}{2})]}{(N_n(s) - 1 + \frac{|A|}{2})(N_n(s) - 2 + \frac{|A|}{2}) \cdots (\frac{|A|}{2})}$$

is the KT-probability of x_1^n corresponding to $\mathcal{T}$.

Remark. The coding distribution $P_{\mathrm{KT},\mathcal{T}}$ is nearly optimal for the class of processes with context tree $\mathcal{T}$, in the sense that the code lengths $\lceil \log P_{\mathrm{KT},\mathcal{T}}(x_1^n)\rceil$ minimize the worst case average redundancy for this class, up to an additive constant.

For estimating the order of Markov chains, the consistency of the KT estimator has been proved when the hypothetical orders are $o(\log n)$ (Csiszár, 2002), while without any bound on the order, or with a bound equal to a sufficiently large constant times $\log n$, a counterexample to its consistency is known (Csiszár and Shields, 2000).

Remark. Strictly speaking, the MDL principle would require to minimize the "code length" $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ incremented by an additional term, the "code length of $\mathcal{T}$" (called the cost of $\mathcal{T}$ in (Willems, Shtarkov, and Tjalkens, 1995)). This additional term is omitted, since this does not affect the consistency result.

Theorem 2.2. *In the case $d(\mathcal{T}_0) < \infty$, the* KT *estimator*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \arg\min_{\mathcal{T}\in\mathcal{F}_1(x_1^n,D(n))\cap\mathcal{I}} \mathrm{KT}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$, satisfies

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \mathcal{T}_0$$

eventually almost surely as $n \to \infty$.

In the general case, the KT *estimator*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \arg\min_{\mathcal{T}\in\mathcal{F}_{n^\beta}(x_1^n,D(n))\cap\mathcal{I}} \mathrm{KT}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$ and arbitrary $0 < \beta < 1$, satisfies for any constant K

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n)|_K = \mathcal{T}_0|_K$$

eventually almost surely as $n \to \infty$.

Proof. If $d(\mathcal{T}_0) < \infty$, the assumptions $\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n))$, $D(n) = o(\log n)$ imply that $\mathcal{T} \in \mathcal{F}_{n^\beta}(x_1^n, D(n))$ eventually almost surely as $n \to \infty$, by Corollary 2.8 in next Section. Hence, it suffices to prove the second assertion of the theorem.

The proof is similar to that of Theorem 2.1. It has to be checked that if $\mathcal{T}|_K \neq \mathcal{T}_0|_K$ for some $\mathcal{T} \in \mathcal{F}_{n^\beta}(x_1^n, D(n)) \cap \mathcal{I}$ with $d(\mathcal{T}) \leq D(n)$, then the modification $\mathcal{T}'$ defined by (2.6) or (2.7) satisfies $\mathcal{T}' \in \mathcal{F}_{n^\beta}(x_1^n, D(n)) \cap \mathcal{I}$ and

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) > \mathrm{KT}_{\mathcal{T}'}(x_1^n), \tag{2.8}$$

simultaneously for all considered trees $\mathcal{T}$, eventually almost surely as $n \to \infty$. Let $\tilde{P}_{\mathrm{KT},s}(x_1^n)$ denote

$$\frac{\prod_{a:N_n(s,a)\geq 1}[(N_n(s,a)-\frac{1}{2})(N_n(s,a)-\frac{3}{2})\cdots(\frac{1}{2})]}{(N_n(s)-1+\frac{|A|}{2})(N_n(s)-2+\frac{|A|}{2})\cdots(\frac{|A|}{2})}$$

if $N_n(s) \geq 1$, and 1 if $N_n(s) = 0$. Then the KT probability $P_{\mathrm{KT},\mathcal{T}}(x_1^n)$ in Definition 2.3 factorizes as

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^{D(n)}} \prod_{s\in\mathcal{T}} \tilde{P}_{\mathrm{KT},s}(x_1^n). \tag{2.9}$$

It follows that (2.8) is equivalent to

$$\sum_{s\in\mathcal{T}} \log \tilde{P}_{\mathrm{KT},s}(x_1^n) - \sum_{s'\in\mathcal{T}'} \log \tilde{P}_{\mathrm{KT},s'}(x_1^n) < 0.$$

Substituting $\mathcal{T}'$ given by (2.6) or (2.7), this reduces to

$$\log \tilde{P}_{\mathrm{KT},s}(x_1^n) - \sum_{u\in\tilde{\mathcal{T}}} \log \tilde{P}_{\mathrm{KT},u}(x_1^n) < 0 \tag{2.10}$$

in case **(a)**, where $\tilde{\mathcal{T}}$ is as in Lemma 2.4, respectively, to

$$\sum_{u\in\tilde{\mathcal{T}}} \log \tilde{P}_{\mathrm{KT},u}(x_1^n) - \log \tilde{P}_{\mathrm{ML},w}(x_1^n) < 0 \tag{2.11}$$

in case **(b)**, where $\tilde{\mathcal{T}}$ and w are as in Lemma 2.5.

To deduce (2.10) and (2.11) from Lemmas 2.4 and 2.5 (in the required eventually almost sure sense), we use the bound from lemma 3.6 adapted for the context tree case

$$\left| \log \tilde{P}_{\mathrm{KT},u}(x_1^n) - \sum_{a\in A} N_n(u,a) \log \frac{N_n(u,a)}{N_n(u)} + \frac{|A|-1}{2} \log N_n(u) \right| < C$$

for any string u with $N_n(u) \geq 1$, where C is a constant depending only on the alphabet size $|A|$. Using the notation from (2.2) the last bound can be equivalently written as

$$\left| \log \tilde{P}_{\mathrm{KT},u}(x_1^n) - \log \tilde{P}_{\mathrm{ML},u}(x_1^n) + \frac{|A|-1}{2} \log N_n(u) \right| < C \tag{2.12}$$

The claim (2.10) immediately follows from (2.13) by (2.12) and the trivial bounds $0 \leq \log N_n(u) \leq \log n$. Also, (2.12) gives for the left-hand side of (2.11), the upper bound

$$\sum_{u\in\tilde{\mathcal{T}}}\left(\log\tilde{P}_{\mathrm{ML},u}(x_1^n) - \frac{|A|-1}{2}\log N_n(u) + C\right) \\ - \left(\log\tilde{P}_{\mathrm{ML},w}(x_1^n) - \frac{|A|-1}{2}\log N_n(w) - C\right).$$

For $\tilde{\mathcal{T}}$ in Lemma 2.5, the assumption $\mathcal{T} \in \mathcal{F}_{n^\beta}(x_1^n, D(n))$ implies $N_n(u) \geq n^\beta$ for each $u \in \tilde{\mathcal{T}}$, and since the sum of $N_n(u)$ for $u \in \tilde{\mathcal{T}}$ is equal to $N_n(w)$, we have $N_n(u) \geq N_n(w)/|\tilde{\mathcal{T}}|$ for at least one $u \in \tilde{\mathcal{T}}$. Using these facts in the last bound, and denoting the left-hand side of (2.14) in Lemma 2.5 by Δ, it follows that the left-hand side of (2.11) is bounded above by

$$\Delta - (|\tilde{\mathcal{T}}|-1)\frac{|A|-1}{2}\alpha\log n - \frac{|A|-1}{2}\log\frac{N_n(w)}{|\tilde{\mathcal{T}}|} \\ + \frac{|A|-1}{2}\log N_n(w) + (|\tilde{\mathcal{T}}+1)C.$$

By Lemma 2.5, here $\Delta < \delta|\tilde{\mathcal{T}}|\log n$ eventually almost surely as $n \to \infty$, for arbitrary $\delta > 0$, simultaneously for all considered $\mathcal{T}$ and s, and thus, the claim (2.11) follows. □

Corollary 2.3. *The vector of the empirical conditional probabilities*

$$\widehat{Q}_{\widehat{\mathcal{T}}}(a|s) = \frac{N_n(s,a)}{N_n(s)}, \quad a \in A, s \in \widehat{\mathcal{T}}$$

converges to that of the true conditional probabilities $Q(a|s)$, $a \in A$, $s \in \mathcal{T}_0$ almost surely as $n \to \infty$, where $\widehat{\mathcal{T}}$ is either the BIC *estimator or the* KT *estimator.*

Proof. Immediate from Theorems 2.1, 2.2 and the ergodic theorem. □

2.5 Underestimation and overestimation lemmas

Lemma 2.4. *For any proper suffix s of some $s_0 \in \mathcal{T}_0$, there exists an irreducible tree $\tilde{\mathcal{T}}$ with $d(\tilde{\mathcal{T}}) < \infty$ such that $u \succ s$ and $Q(u) > 0$ for each $u \in \tilde{\mathcal{T}}$, each $v \succeq s$ with $Q(v) > 0$ has a suffix in $\tilde{\mathcal{T}}$, and*

$$\log\tilde{P}_{\mathrm{ML},s}(x_1^n) - \sum_{u\in\tilde{\mathcal{T}}}\log\tilde{P}_{\mathrm{ML},u}(x_1^n) < -cn \qquad (2.13)$$

Proof. Given $s \prec s_0 \in \mathcal{T}_0$, denote by $s_0^{(\ell)}$ the ℓ-length suffix of s_0. Let

$$\mathcal{T} = \{s_0^{(L+1)}\} \cup \{as_0^{(\ell)} : l(s) \leq \ell \leq L, a \in A, as_0^{(\ell)} \neq s_0^{(\ell+1)}, Q(as_0^{(\ell)}) > 0\}.$$

We show that if $L = l(s_0) - 1$ when $l(s_0) < \infty$, or L is sufficiently large with the property $Q(s_0^{(L+1)}) < Q(s_0^{(L)})$ when $l(s_0) = \infty$, this tree $\tilde{\mathcal{T}}$ satisfies the assertions of the lemma.

Now, using (2.4), the inequality (2.13) can be written as

$$\sum_{u \in \tilde{\mathcal{T}}, a \in A} N_n(u,a) \log \frac{N_n(u,a)}{N_n(u)} - \sum_{a \in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)} > cn.$$

Due to the ergodic theorem, $N_n(v,a)/n \to Q(va)$ for any string v, almost surely as $n \to \infty$. Hence, it is enough to show that

$$\sum_{u \in \tilde{\mathcal{T}}, a \in A} Q(ua) \log \frac{Q(ua)}{Q(u)} - \sum_{a \in A} Q(sa) \log \frac{Q(sa)}{Q(s)} > 0.$$

Jensen's inequality implies

$$Q(s) \sum_{u \in \tilde{\mathcal{T}}} \frac{Q(u)}{Q(s)} \left(\frac{Q(ua)}{Q(u)} \log \frac{Q(ua)}{Q(u)} \right) \geq Q(sa) \log \frac{Q(sa)}{Q(s)}, \quad a \in A$$

where the strict inequality holds for some $a \in A$, unless $Q(a|s) = Q(a|u)$ for each $a \in A$ and $u \in \mathcal{T}$, in particular, for $u = s_0^{(L+1)}$. In the case $l(s_0) < \infty$, we have $s_0^{(L+1)} = s_0$, hence, the last contingency is ruled out by $s \prec s_0 \in \mathcal{T}_0$ and the definition of context tree $\mathcal{T}_0$. In the case $l(s_0) = \infty$, if $Q(a|s)$ were equal to $Q(a|s_0^{(L+1)})$ for each $a \in A$ and all L satisfying $Q(s_0^{(L+1)}) < Q(s_0^{(L)})$, letting $L \to \infty$ would give $Q(a|s) = Q(a|s_0)$, again contradicting $s \prec s_0 \in \mathcal{T}_0$.

The irreducibility of $\tilde{\mathcal{T}}$ is obvious when $l(s_0) = \infty$, and in the case $l(s_0) < \infty$ it only requires checking that for $L = l(s_0) - 1$ there exists $a \in A$ with $as_0^{(L)} \neq s_0$, $Q(as_0^{(L)}) > 0$; this follows from $s_0 \in \mathcal{T}_0$ by Definition 2.1.

Moreover, we have $Q(u) > 0$ for each $u \in \tilde{\mathcal{T}}$, and each $v \succeq s$ with $Q(v) > 0$ has a suffix in $\tilde{\mathcal{T}}$ by construction. □

Lemma 2.5. *For any irreducible tree $\mathcal{T}$ with $d(\mathcal{T}) \leq D(n)$, $D(n) = o(\log n)$, and $s \in \mathcal{T}$ that has a proper suffix $s_0 \in \mathcal{T}_0$ with $l(s_0) \leq K$, there exists w satisfying $s \succ w \succ s_0$ such that, for $\tilde{\mathcal{T}} = \{u \in \mathcal{T} : u \succ w\}$ and arbitrary $\varepsilon > 0$*

$$\sum_{u \in \tilde{\mathcal{T}}} \log \tilde{P}_{\mathrm{ML},u}(x_1^n) - \log \tilde{P}_{\mathrm{ML},w}(x_1^n) < \varepsilon |\tilde{\mathcal{T}}| \log n \tag{2.14}$$

holds simultaneously for all $\mathcal{T}$ and s as above, eventually almost surely as $n \to \infty$. Moreover, here $w = a_{-k}a_{-k+1} \dots a_{-1}$ can be chosen such that $a_{-k+1} \dots a_{-1}$ is a proper suffix of some $u \in \mathcal{T} \backslash \tilde{\mathcal{T}}$.

Proof. Let $w = a_{-k}a_{-k+1}\dots a_{-1}$ be the longest suffix of s with $k < l(s)$ for which there exists a string in $\mathcal{T}$ not equal to w but having the suffix $a_{-k+1}\dots a_{-1}$ Then $\mathcal{T}_0 \in \mathcal{I}$ implies that $w \succeq s_0$, and hence, $a_{-k+1}\dots a_{-1} \prec u$ for some $u \in \mathcal{T}\backslash\tilde{\mathcal{T}}$.

Since

$$\prod_{a\in A}\left(\frac{N_n(w,a)}{N_n(w)}\right)^{N_n(w,a)} \geq \prod_{a\in A} Q(a|w)^{N_n(w,a)}$$

the left-hand side of the claimed inequality can be bounded above by

$$\begin{aligned}
&\sum_{u\in\tilde{\mathcal{T}},a\in A} N_n(u,a)\log\frac{N_n(u,a)}{N_n(u)} - \sum_{a\in A} N_n(w,a)\log Q(a|w)\\
&\qquad\overset{(*)}{=} \sum_{u\in\tilde{\mathcal{T}},a\in A} N_n(u,a)\log\frac{N_n(u,a)}{N_n(u)} - \sum_{u\in\tilde{\mathcal{T}},a\in A} N_n(u,a)\log Q(a|u)\\
&\qquad= \sum_{u\in\tilde{\mathcal{T}}} N_n(u)\sum_{a\in A}\frac{N_n(u,a)}{N_n(u)}\log\frac{N_n(u,a)/N_n(u)}{Q(a|u)}\\
&\qquad= \sum_{u\in\tilde{\mathcal{T}}} N_n(u)\,\mathbf{D}\left(\frac{N_n(u,\cdot)}{N_n(u)}\,\middle\|\, Q(\cdot|u)\right)
\end{aligned}$$

Here $(*)$ follows as $u \succ w \succ s_0 \in \mathcal{T}_0$ implies $Q(a|u) = Q(a|w) = Q(a|s_0)$ by Definition 2.1. Using Corollary 2.12 and Lemma 2.6, this can be further bounded above, eventually almost surely for all considered $\mathcal{T}$ and s, by

$$\begin{aligned}
&\sum_{u\in\tilde{\mathcal{T}}} N_n(u)\frac{1}{q_{\min}}\sum_{a\in A}\left(\frac{N_n(u,a)}{N_n(u)} - Q(a|u)\right)^2\\
&\qquad\qquad < \sum_{u\in\tilde{\mathcal{T}}} N_n(u)\frac{1}{q_{\min}}|A|\frac{\delta\log n}{N_n(u)} \leq \frac{\delta|A|}{q_{\min}}|\tilde{\mathcal{T}}|\log n
\end{aligned}$$

where $q_{\min}$ is the minimum of the nonzero conditional probabilities $Q(a|s_0)$, $a \in A$, $s_0 \in \mathcal{T}_0$, $l(s_0) \leq K$, and $\delta > 0$ is arbitrary small. □

Lemma 2.6. *For probability distributions P and Q on A*

$$\mathbf{D}(P\,\|\,Q) \leq \sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}.$$

Proof. We use the bound $\log x \leq x - 1$

$$\begin{aligned}\mathbf{D}(P\,\|\,Q) &= \sum_{a\in A} P(a)\log\frac{P(a)}{Q(a)}\\ &\leq \sum_{a\in A} P(a)\left(\frac{P(a)}{Q(a)}-1\right)\\ &\leq \sum_{a\in A}\frac{P(a)^2-P(a)Q(a)}{Q(a)}\\ &\leq \sum_{a\in A}\frac{P(a)^2-P(a)Q(a)}{Q(a)}+\sum_{a\in A}Q(a)-\sum_{a\in A}P(a)\\ &\leq \sum_{a\in A}\frac{P(a)^2-P(a)Q(a)}{Q(a)}+\sum_{a\in A}\frac{Q(a)^2-P(a)Q(a)}{Q(a)}\\ &\leq \sum_{a\in A}\frac{(P(a)-Q(a))^2}{Q(a)}.\end{aligned}$$ □

2.6 Typicality and conditional typicality

Let $X_1, X_2, \ldots,$ be a stochastic process with finite alphabet A, satisfying the mixing condition that for each k, m and each $s \in A^{l(s)}$, $s' \in A^m$, with $\mathbf{P}(X_1^m = s') > 0$,

$$|\mathbf{P}(X_n^{n+l(s)-1} = s | X_1^m = s') - Q(s)| \leq \psi(\ell) Q(s), \quad n > m + \ell, \tag{2.15}$$

where $\psi(\ell) \downarrow 0$ as $\ell \to \infty$. Note that the process need not be stationary, but (2.15) obviously implies that $\lim_{n\to\infty} \mathbf{P}(X_n^{n-l(s)-1} = s) = Q(s)$. Irreducible, aperiodic Markov chains (of any order) satisfy the condition (2.15), with $\psi(\ell) = \delta^\ell$, for suitable $0 < \delta < 1$.

Theorem 2.7.

(a) *Assume the mixing condition* (2.15). *To any $\eta > 0$ there exists $C > 0$ such that eventually almost surely as $n \to \infty$,*

$$\left|\frac{N_n(s)/n - l(s)}{Q(s)} - 1\right| \leq \eta,$$

for all s for which $nQ(s) \geq C \log^2 n$.

(b) *Assume* (2.15) *holds with* $\psi(\ell) = \delta^\ell$, $0 < \delta < 1$. *There exists a constant* C *such that eventually almost surely as* $n \to \infty$,

$$\left| \frac{N_n(s)/n - l(s)}{Q(s)} - 1 \right| \leq \sqrt{\frac{C \log^2 n}{nQ(s)}}$$

for all s *for which* $nQ(s) \geq C \log^2 n$.

(c) *The conclusion of* ***(b)*** *holds for an irreducible Markov chain of any order and* Q *equal to its stationary distribution.*

Proof. We suppose that Q is non-degenerate, i.e., that $Q_{\max} = \max_{a \in A} Q(a) < 1$; in the (uninteresting) degenerate case the proof is much simpler.

Fix first $\ell > 0$ and s with $l(s) \leq \ell$ and $Q(s) > 0$ and write

$$Z_j^{(i)} = \begin{cases} 1 & \text{if } X_{i+j\ell+1}^{i+j\ell+l(s)} = s \\ 0 & \text{otherwise} \end{cases}, 0 \leq i < \ell,\ j \geq 0$$

The mixing condition (2.15) implies that for each i, the random variables $Z_j^{(i)}$ satisfy the hypotesis of Lemma 2.10 with $p_* = (1 - \psi(\ell - l(s))Q(s)$, $p^* = (1 + \psi(\ell - l(s))Q(s)$. We assume $\ell - l(s)$ is sufficiently large such that $\psi(\ell - l(s)) < \min\{1, \frac{1-Q_{\max}}{Q_{\max}}\}$. Since clearly $Q(s) \leq Q_{\max}$, this assumption guarantees that $0 < p_* \leq p^* < 1$.

Suppose for convenience that $n - l(s) = m\ell$ (the obvious modifications needed when $n - l(s)$ is not divisible by ℓ are omitted). Then

$$N_n(s) = \sum_{i=0}^{\ell-1} \sum_{j=0}^{m-1} Z_j^{(i)}$$

and Lemma 2.10 imply, for $\psi = \psi(\ell - l(s)) < \eta < \min\{1, \frac{1-Q_{\max}}{Q_{\max}}\}$, that

$$\begin{aligned} \mathbf{P}\Big(N_n(s) > (1+\eta)(n - l(s))Q(s)\Big) &\leq \sum_{i=0}^{\ell-1} \mathbf{P}\left(\sum_{j=0}^{\ell-1} Z_j^{(i)} > (1+\eta)mQ(s)\right) \\ &\leq \ell \exp\Big(-m\mathbf{D}((1+\eta)Q(s) \,\|\, (1+\psi)Q(s))\Big) \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}\Big(N_n(s) < (1-\eta)(n - l(s))Q(s)\Big) &\leq \sum_{i=0}^{\ell-1} \mathbf{P}\left(\sum_{j=0}^{\ell-1} Z_j^{(i)} > (1-\eta)mQ(s)\right) \\ &\leq \ell \exp\Big(-m\mathbf{D}((1-\eta)Q(s) \,\|\, (1-\psi)Q(s))\Big). \end{aligned}$$

It follows from these two bounds and Lemma 2.9, that for $\eta \in (\psi, \min\{1, \frac{1-Q_{\max}}{Q_{\max}}\})$, $\psi = \psi(\ell - l(s))$ that

$$\mathbf{P}\left(\left|\frac{N_n(s)}{n-l(s)} - Q(s)\right| > \eta Q(s)\right) \le 2\ell \exp\left(-\frac{m(\eta-\psi)^2 Q(s)}{4}\right).$$

If in addition, $\psi = \psi(\ell - l(s)) < \eta/2$, this bound and $m = \frac{n-l(s)}{\ell} > \frac{n}{\ell} - 1$ give that

$$\mathbf{P}\left(\left|\frac{N_n(s)}{n-l(s)} - Q(s)\right| > \eta Q(s)\right) \le 2\ell \exp\left(-\frac{nQ(s)\eta^2}{16\ell} + \frac{1}{16}\right). \tag{2.16}$$

The mixing condition (2.15) implies that $\max_s Q(s)$ decreases exponentially as $l(s) \to \infty$. In particular, there exists α^* such that $nQ(s) < 1$ for $l(s) > \alpha^* \log n$, hence in the probability bound below, attention may be restricted to $k \le \alpha^* \log n$.

It follows from (2.16) with $\ell = 2\alpha^* \log n$ that for $\eta < \min\{1, \frac{1-Q_{\max}}{Q_{\max}}\}$ fixed and n so large that $\psi(\alpha^* \log n) < \eta/2$ (making sure that $\psi = \psi(\ell - l(s)) < \eta/2$ always holds when $l(s) \le \alpha^* \log n$) we have

$$\begin{aligned}
&\mathbf{P}\left(\left|\frac{N_n(s)/n - l(s)}{Q(s)} - 1\right| > \eta \text{ for some } k \text{ and } s \text{ with } nQ(s) > C\log^2 n\right) \\
&\le \sum_{l(s)=1}^{\alpha^* \log n} \sum_{nQ(s) > C\log^2 n} \mathbf{P}\left(\left|\frac{N_n(s)}{n-l(s)} - Q(s)\right| > \eta Q(s)\right) \\
&\le (\alpha^* \log n)|A|^{\alpha^* \log n} 4\alpha^* \log n \cdot \exp\left(-\frac{C\log^2 n \cdot \eta^2}{32\alpha^* \log n} + \frac{1}{16}\right).
\end{aligned}$$

Clearly this last bound is summable in n if C is sufficiently large, for example, if $C\eta^2/32\alpha^* > \alpha^* \log|A| + 1$. An application of Borel-Cantelli completes the proof of part **(a)**.

To prove part **(b)**, we use (2.16) with $\eta = \sqrt{\frac{C\log^2 n}{nQ(s)}}$, $\ell = (\alpha^* + \gamma)\log n$, choosing γ so large that $\psi(\gamma \log n) < 1/\sqrt{n}$ (which is possible by the hypothesis $\psi(\ell) = \delta^\ell$); with this choice, the requirement $\psi(\ell - l(s)) < \eta/2$, needed for

(2.16), is satisfied for all $k \le \alpha^* \log n$. Thus we obtain

$$\mathbf{P}\left(\left|\frac{N_n(s)/n - l(s)}{Q(s)} - 1\right| > \sqrt{\frac{C\log^2 n}{nQ(s)}} \text{ for some } s \text{ with } nQ(s) > C\log^2 n\right)$$
$$\le \sum_{l(s)=1}^{\alpha^*\log n} \sum_{nQ(s)>C\log^2 n} \mathbf{P}\left(\left|\frac{N_n(s)}{n-l(s)} - Q(s)\right| > \sqrt{\frac{C\log^2 n}{nQ(s)}}Q(s)\right)$$
$$\le (\alpha^* \log n)|A|^{\alpha^* \log n} 2(\alpha^*+\gamma)\log n \cdot \exp\left(-\frac{C\log^2 n}{16(\alpha^*+\gamma)\log n} + \frac{1}{16}\right),$$

which is clearly summable in n if C is sufficiently large. This completes the proof of part **(b)**.

To prove part **(c)**, note that part **(b)** immediately applies to aperiodic (irreducible) Markov chains. For a periodic chain, say of period p, the mixing condition (2.15), with $\psi(\ell) = \delta^\ell$, is satisfied only when $n \equiv i \mod p$ for some $1 \le i \le p$, the role of $Q(s)$ in (2.16) is then played by $Q_i(s) = \lim_{m\to\infty} \mathbf{P}(X_{mp+i}^{mp+i+l(s)} = s)$. The stationary distribution is now given by $Q(s) = \frac{1}{p}\sum_{i=1}^{p} Q_i(s)$. The proof is completed by applying part **(b)** to

$$N_{n,i}(s) = \#\{j : X_j^{j+l(s)} = s,\ 0 \le j \le n - l(s),\ j \equiv i \mod p\}$$

and $Q_i(s)$ in the role of $N_n(s)$ and $Q(s)$, respectively. □

Corollary 2.8. *Given a process Q with context tree of finite depth, for any $0 < \beta < 1$ there exists $\kappa > 0$ such that, eventually almost surely as $n \to \infty$*

$$N_n(s) \ge n^\beta$$

simultaneously for all strings s with $Q(s) > 0$, $l(s) \le \kappa \log n$.

Proof. Since the process is a Markov chain there exists $\xi > 0$ such that $Q(s) > \xi^{l(s)}$, whenever $Q(s) > 0$, and hence $Q(s) > n^{-\alpha}$ if $l(s) \le \frac{-\alpha}{\log \xi}\log n$. Moreover if $n \ge n_0$ we have that $n^{-\alpha} \ge \frac{C\log^2 n}{n}$ where n_0 only depends on C and α.

Theorem 2.7 implies that for all s with $nQ(s) \ge C\log^2 n$ eventually almost surely as $n \to \infty$

$$\left|\frac{N_n(s)/n - l(s)}{Q(s)} - 1\right| \le \frac{1}{2},$$

multiplying by $Q(s)$ we have

$$-\frac{1}{2}Q(s) \le \frac{N_n(s)}{n - l(s)} - Q(s) \le \frac{1}{2}Q(s). \tag{2.17}$$

Then for all strings s with $Q(s) > 0$ and $l(s) \leq \kappa \log n$ where $\kappa = \frac{-\alpha}{\log \xi}$ we have that $Q(s) \geq n^{-\alpha}$ and if n is big enough also equation (2.17) holds.

Since both inequalities hold for n sufficiently large, combining them we can write

$$\frac{N_n(s)}{n - l(s)} \geq \frac{1}{2} Q(s) \geq \frac{1}{2} n^{-\alpha},$$

choosing $(1 - \alpha) > \beta$ we obtain

$$N_n(s) \geq \frac{1}{2} n^{-\alpha}(n - l(s)) \geq n^{\beta}.$$

and the proof is complete. □

Lemma 2.9. *For arbitrary $s \in (0,1)$ and $\eta, \psi \in (-1, \min\{1, \frac{1-s}{s}\})$ we have*

$$\mathbf{D}((1+\eta)s \,\|\, (1+\psi)s) \geq \frac{s}{4}(\eta - \psi)^2.$$

Proof. $f(\eta) = \mathbf{D}((1+\eta)s \,\|\, (1+\psi)s) - \frac{s}{4}(\eta - \psi)^2$ is a convex function of $\eta \in (-1, 1)$, since $f''(\eta) \geq 0$. This and $f(\psi) = f'(\psi) = 0$ imply that $f(\eta) \geq 0$. □

Remark. Here, for p and q in $(0,1)$, $\mathbf{D}(p \,\|\, q)$ denotes the information divergence of $(p, 1-p)$ from $(q, 1-q)$, i.e.

$$\mathbf{D}(p \,\|\, q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.$$

Lemma 2.10. *Let $Z_1, Z_2, \ldots,$ be $\{0,1\}$-valued random variables such that*

$$0 < p_* \leq \mathbf{P}(Z_j = 1 | Z_1^{j-1}) \leq p^* < 1, \quad j \geq 1, \tag{2.18}$$

with probability 1. Then

$$\mathbf{P}\left(\sum_{j=1}^{m} Z_j > \gamma m\right) \leq \exp(-m\mathbf{D}(\gamma \,\|\, p^*)), \quad p^* < \gamma < 1,$$

$$\mathbf{P}\left(\sum_{j=1}^{m} Z_j < \gamma m\right) \leq \exp(-m\mathbf{D}(\gamma \,\|\, p_*)), \quad 0 < \gamma < p_*.$$

Proof. Chernoff bounding gives

$$\mathbf{P}\left(\sum_{j=1}^{m} Z_j > \gamma m\right) \leq \exp(-\lambda\gamma m)\mathbf{E}\left(\exp\left(\lambda \sum_{j=1}^{m} Z_j\right)\right), \quad \lambda > 0,$$

$$\mathbf{P}\left(\sum_{j=1}^{m} Z_j < \gamma m\right) \leq \exp(-\lambda\gamma m)\mathbf{E}\left(\exp\left(\lambda \sum_{j=1}^{m} Z_j\right)\right), \quad \lambda < 0.$$

Here

$$\mathbf{E}\left(\exp\left(\lambda\sum_{j=1}^{m} Z_j\right)\right) = \mathbf{E}\left(\exp\left(\lambda\sum_{j=1}^{m-1} Z_j\right)\mathbf{E}(\exp(\lambda Z_m)|Z_1^{m-1})\right)$$

where

$$\mathbf{E}(\exp(\lambda Z_m)|Z_1^{m-1}) = 1 + (e^\lambda - 1)\mathbf{P}(Z_m = 1|Z_1^{m-1})$$

is upper bounded by $1 + (e^\lambda - 1)p^*$ if $\lambda > 0$ and by $1 + (e^\lambda - 1)p_*$ if $\lambda < 0$, on account of the assumption (2.18). Hence

$$\mathbf{E}\left(\exp\left(\lambda\sum_{j=1}^{m} Z_j\right)\right) \leq \mathbf{E}\left(\exp\left(\lambda\sum_{j=1}^{m-1} Z_j\right)\right) \cdot \begin{cases} 1 + (e^\lambda - 1)p^* & \text{if } \lambda < 0 \\ 1 + (e^\lambda - 1)p_* & \text{if } \lambda > 0 \end{cases}$$

A repetition of the argument leads to

$$\mathbf{E}\left(\exp\left(\lambda\sum_{j=1}^{m} Z_j\right)\right) \leq \begin{cases} \left(1 + (e^\lambda - 1)p^*\right)^m & \text{if } \lambda < 0 \\ \left(1 + (e^\lambda - 1)p_*\right)^m & \text{if } \lambda > 0 \end{cases}$$

Thus the Chernoff bounds give

$$\mathbf{P}\left(\sum_{j=1}^{m} Z_j > \gamma m\right) \leq \exp\Big(-m(\lambda\gamma - \log(1 + (e^\lambda - 1)p^*))\Big), \quad \lambda > 0,$$
$$\mathbf{P}\left(\sum_{j=1}^{m} Z_j < \gamma m\right) \leq \exp\Big(-m(\lambda\gamma - \log(1 + (e^\lambda - 1)p_*))\Big), \quad \lambda < 0.$$

The assertions of the lemma follow with the respective choices $\lambda = \log\frac{\gamma(1-p^*)}{p^*(1-\gamma)}$ (which is positive if $\gamma > p^*$), and $\lambda = \log\frac{\gamma(1-p_*)}{p_*(1-\gamma)}$ (which is negative if $\gamma < p_*$). □

Theorem 2.11. *To arbitrary $\xi > \log|A|/2$ there exist $\eta > 0$ and $c > 0$ such that, eventually almost surely as $n \to \infty$*

$$\left|\frac{N_n(s,a)}{N_n(s)} - Q(a|s)\right| < \sqrt{\frac{\max\{\xi l(s), \eta \log\log N_n(s)\}}{N_n(s)}}$$

simultaneously for all strings s with $N_n(s) \geq cl(s)$ which have a suffix s_0 in the context tree $\mathcal{T}_0$ of Q.

Proof. For fixed ξ, η and c, let $B_n(s,a)$ denote the event

$$N_n(s) > cl(s), \quad \left| \frac{N_n(s,a)}{N_n(s)} - Q(a|s) \right| < \sqrt{\frac{\max\{\xi l(s), \eta \log\log N_n(s)\}}{N_n(s)}} \tag{2.19}$$

We have to show that for $\xi > \log|A|/2$ and η and c sufficiently large, only a finite number of the events

$$B_n = \bigcup_{s \succeq s_0 \in \mathcal{T}_0} B_n(s,a)$$

can occur, with probability 1. The proof will involve several constants, arbitrary first, and suitably specified later. For orientation, we note that in addition to ξ, η, c above our main constants will be θ, ε and an explicit function of θ and ε, denoted by μ. The later specification of θ and ε will depend on ξ only, that c of will depend on ξ and ε, and of η on ξ and μ. Fixing $\theta > 1$, write

$$\begin{aligned} C_m(s,a) &= \bigcup_{n=l(s)}^{\infty} \Big(B_n(s,a) \cap \{\theta^m < N_n(s) \le \theta^{m+1}\} \Big), \qquad (2.20) \\ C_m &= \bigcup_{s \succeq s_0 \in \mathcal{T}_0} C_m(s,a). \end{aligned}$$

Claim: Let B^* and C^* denote the event that infinitely many of the events B_n, respectively, C_m occur. Then

$$\mathbf{P}(B^* \backslash C^*) = 0.$$

Consider an infinite sample path for which B^* occurs, i.e., there exist infinitely many n, $\{s_n\}$ with the property $s_n \succeq s_{n-1}$, and a_n, such that $B_n(s_n, a_n)$ occurs for this sample path. If $l(s_n)$ here is unbounded then C^* certainly occurs, because the first condition in (2.19) implies

$$B_n(s_n, a_n) \subset \bigcup_{m=m_k}^{\infty} C_m(s_n, a_n), \quad m_k = \left\lfloor \frac{\log cl(s)}{\log\theta} \right\rfloor. \tag{2.21}$$

If $l(s)$ is bounded, there must be an s such that $B_n(s,a)$ occurs for infinitely many n. This again implies that C^* occurs, provided that $N_n(s) \to \infty$. The latter holds with probability 1 for every s with $Q(s) > 0$, since the Markov chain is assumed to be irreducible. This completes the proof of the claim.

By the claim it is enough to show that $\mathbf{P}(C^*) = 0$, or, by Borel-Cantelli, that

$$\sum_m \mathbf{P}(C_m) < \infty,$$

for a suitable choice of $\theta > 1$ in the definition of C_m, see (2.20). The proof of this makes use of a large deviations bound valid for any martingale $\{Z_n\}$ with $Z_0 = 0$ for which the differences $Z_i - Z_{i-1}$ are almost surely bounded above by 1. See Lemmas 2.13 and 2.14, we have that

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty}\{\exp(\lambda Z_n - \phi_c(\lambda)A_n) > \exp(\alpha\lambda)\}\right) \leq \exp(-\alpha\lambda)$$

if λ and $\alpha \in \mathbb{R}^+$, as $\exp(\lambda Z_0 - \phi_c(\lambda)A_0) = 1$. Taking logarithms, choosing $c = 1$ and using the fact that $\lambda^{-1}\phi_c(\lambda) = (e^\lambda - 1 - \lambda)/\lambda$ in this case the inequality can also be written

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty}\left\{Z_n > \alpha + \frac{e^\lambda - 1 - \lambda}{\lambda}A_n\right\}\right) < \exp(-\alpha\lambda)$$

Here $\{A_n\}$ denotes the increasing process associated with the submartingale $\{Z_n^2\}$ by the Doob decomposition and $\alpha > 0$, $\lambda > 0$ are arbitrary positive numbers. Note that if $|Z_i - Z_{i-1}| \leq 1$, almost surely, then the same bound holds also for the martingale $\{-Z_n\}$, and therefore

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty}\left\{|Z_n| > \alpha + \frac{e^\lambda - 1 - \lambda}{\lambda}A_n\right\}\right) < 2\exp(-\alpha\lambda). \tag{2.22}$$

The martingale in this setting is defined for fixed $s \succeq s_0 \in \mathcal{T}_0$ by

$$Z_n = \begin{cases} N_n(s,a) - Q(a|s)N_n(s) & n \geq l(s) \\ 0 & 0 \leq n < l(s) \end{cases} \tag{2.23}$$

It is easy to see that $|Z_i - Z_{i-1}| \leq 1$ and a direct calculation gives

$$\begin{aligned} A_n &= \sum_{i=k}^{\infty} \mathbf{E}((Z_i - Z_{i-1})^2 | X_1^{i-1}) \\ &= N_n(s)Q(a|s)(1 - Q(a|s)) \leq \frac{N_n(s)}{4}. \end{aligned}$$

Fix $\theta > 1$ and $\varepsilon > 0$. For any $m \geq 0$, and $0 < \lambda < \lambda_0(\varepsilon)$ where $\lambda_0(\varepsilon)$ is the solution of the equation $(e^\lambda - 1 - \lambda)\lambda^{-1} = (1/2 + \varepsilon)\lambda$, we can be sure that

$$\begin{aligned} &\left\{\theta^m < N_n(s) \leq \theta^{m+1}\right\} \cap \left\{|Z_n| > \alpha + \left(\frac{1}{2} + \varepsilon\right)\lambda\frac{\theta^{m+1}}{4}\right\} \\ &\qquad \subset \left\{|Z_n| > \alpha + \frac{e^\lambda - 1 - \lambda}{\lambda}A_n\right\}. \end{aligned} \tag{2.24}$$

We will use this with α and λ depending on m as

$$\alpha = \frac{\theta^{m/2}\sqrt{\max\{\xi l(s), \eta \log(m \log \theta)\}}}{2(1+\varepsilon)}, \quad \lambda = \frac{4\sqrt{\max\{\xi l(s), \eta \log(m \log \theta)\}}}{(1+\varepsilon)\theta^{1+m/2}}.$$

To ensure $\lambda < \lambda_0(\varepsilon)$ needed for (2.24), we assume that m satisfies

$$\theta^{m+1} > \tau(\varepsilon)\xi l(s), \quad \tau(\varepsilon) = \left(\frac{4}{\lambda_0(\varepsilon)}\right)^2$$

and also exceeds a threshold m_0 depending on ε, θ, and η.

With α and λ as above, we have

$$\begin{aligned}\alpha + \left(\frac{1}{2} + \varepsilon\right)\lambda\frac{\theta^{m+1}}{4} &= \theta^{m/2}\sqrt{\max\{\xi l(s), \eta \log(m \log \theta)\}}\\ &\leq \sqrt{N_n(s)\max\{\xi l(s), \eta \log\log N_n(s)\}}\end{aligned}$$

when $\theta^m < N_n(s)$, and

$$\alpha\lambda = \mu \max\{\xi l(s), \eta \log(m \log \theta)\}, \quad \text{where } \mu = \frac{2}{(1+\varepsilon)^2\theta}.$$

Hence, by the inclusion (2.24) and the martingale bound (2.22), the event

$$\bigcup_{n=l(s)}^{\infty} \left(\left\{\theta^m < N_n(s) \leq \theta^{m+1}\right\} \cap \left\{|Z_n| > \sqrt{N_n(s)\max\{\xi l(s), \eta \log\log N_n(s)\}}\right\}\right)$$

has probability less than $2\exp(-\mu \max\{\xi l(s), \eta \log(m \log \theta)\})$, if $m \geq m_0$ and $\theta^{m+1} > \tau(\varepsilon)\xi l(s)$.

Recalling the definition (2.19) of the event $B_n(s,a)$, and that (2.23) imply

$$\frac{Z_n}{N_n(s)} = \frac{N_n(s,a)}{N_n(s)} - Q(a|s),$$

it follows that $C_m(s,a)$ defined in (2.20) equals the intersection of the last union with the event $\{N_n(s) > cl(s)\}$. When that intersection is non-empty then $\theta^{m+1} > \tau(\varepsilon)\xi l(s)$ automatically holds, providing $c \geq \tau(\varepsilon)\xi$. Thus we have shown that if c in (2.19) is sufficiently large, for each $s \succeq s_0 \in \mathcal{T}_0$ and each $m \geq m_0$

$$\begin{aligned}\mathbf{P}(C_m(s,a)) &\leq 2\exp(-\mu \max\{\xi l(s), \eta \log(m \log \theta)\})\\ &\leq \begin{cases} 2\exp(-\mu\xi l(s)) & \text{if } l(s) \geq \frac{\eta}{\xi}\log(m \log \theta)\\ 2(m \log \theta)^{-\mu\eta} & \text{if } l(s) < \frac{\eta}{\xi}\log(m \log \theta).\end{cases}\end{aligned} \tag{2.25}$$

Assume now that the so far arbitrary $\theta > 1$ and $\varepsilon > 0$ are selected such that $\mu\xi = 2\xi/(1+\varepsilon)^2\theta > \log|A|$ (possible by the assumption $\xi > \log|A|/2$). Then if follows from (2.25) that

$$\begin{aligned}
\mathbf{P}(C_m) &= \sum_{s\succeq s_0\in\mathcal{T}_0} \mathbf{P}(C_m(s,a)) \\
&\le 2\left[\sum_{l(s)\ge\frac{\eta}{\xi}\log(m\log\theta)} |A|^{l(s)}\exp(-\mu\xi l(s)) + \left(\sum_{l(s)<\frac{\eta}{\xi}\log(m\log\theta)} |A|^{l(s)}\right)(m\log\theta)^{-\mu\eta}\right] \\
&\le 2\left[\sum_{l(s)=\lceil\frac{\eta}{\xi}\log(m\log\theta)\rceil}^{\infty} \exp\Big(-(\mu\xi-\log|A|)l(s)\Big) + |A|^{\frac{\eta}{\xi}\log(m\log\theta)+1}(m\log\theta)^{-\mu\eta}\right] \\
&\le 2\left[\frac{\exp\Big(-(\mu\xi-\log|A|)\frac{\eta}{\xi}\log(m\log\theta)\Big)}{1-\exp\Big(-(\mu\xi-\log|A|)\Big)} + |A|(m\log\theta)^{\frac{\eta}{\xi}\log|A|-\mu\eta}\right] \\
&\le K(m\log\theta)^{-(\mu-\frac{\log|A|}{\xi})\eta}.
\end{aligned}$$

This is summable in m if $\eta > \xi/(\mu\xi - \log|A|)$ and this completes the proof. □

Corollary 2.12. *Given a process Q, to any $\delta > 0$ there exists $\kappa > 0$ such that, eventually almost surely as $n \to \infty$*

$$\left|\frac{N_n(s,a)}{N_n(s)} - Q(a|s)\right| < \sqrt{\frac{\delta\log n}{N_n(s)}}$$

simultaneously for all strings s with $l(s) \le \kappa\log n$ and $N_n(s) \ge 1$ which have a suffix in the context tree of Q.

Proof. By Theorem 2.11, for $\xi > (\log|A|)/2$ there exist $\eta > 0$ and $c > 0$ such that, eventually almost surely,

$$\left|\frac{N_n(s,a)}{N_n(s)} - Q(a|s)\right| < \sqrt{\frac{\max\{\xi l(s), \eta\log\log N_n(s)\}}{N_n(s)}} \tag{2.26}$$

simultaneously for all strings s with $N_n(s) \ge cl(s)$ which have a suffix in the context tree of Q.

Then the choice $\kappa = \delta/\max\{\xi, c/4\}$ is suitable. Indeed, if $N_n(s) \ge cl(s)$, the bound (2.26) holds and gives the assertion, because

$$\begin{aligned}
\eta\log\log N_n(s) \le \eta\log\log n &\le \delta\log n \quad \text{for } n \ge n_0 \\
\xi l(s) \le \xi\kappa\log n &\le \delta\log n
\end{aligned}$$

While in the opposite case $N_n(s) < cl(s) \leq c\kappa \log n$ we have

$$\sqrt{(\delta \log n)/N_n(s)} \geq \sqrt{\delta/(c\kappa)} \geq 2$$

and the assertion holds trivially. □

2.7 Auxiliary results on Martingales

All martingales considered are defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ endowed with an increasing filtration $\{\mathcal{B}_n\}_{n\in\mathbb{N}}$ with respect to which the variables are measurable, we considered $\mathcal{B}_n = \sigma(X_1, \dots, X_n)$ on the theorems above.

Lemma 2.13. *For every real number $c > 0$, let $\phi_c : \mathbb{R}^+ \to \mathbb{R}^+$ be the function defined by $\phi_c(\lambda) = c^{-2}(\exp(\lambda c) - 1 - \lambda c)$; it satisfies $\phi_c(\lambda) = \frac{1}{2}\lambda^2 + o(\lambda^2)$ at the neighborhood of the origin. For every square-integrable martingale $\{Z_n\}_{n\in\mathbb{N}}$ such that* $\sup(Z_{n+1} - Z_n) \leq c$ *a.s., the sequence*

$$\exp(\lambda Z_n - \phi_c(\lambda) A_n), \quad n \in \mathbb{N}$$

of random variables is then a positive supermartingale for every $\lambda \in \mathbb{R}^+$.

Proof. We will be using the following elementary inequality, valid for every real number $y \in (\infty, c]$ and every real $c > 0$:

$$\exp(\lambda y) \leq 1 + \lambda y + \phi_c(\lambda) y^2.$$

This inequality is easily proved from the series expansion of the exponential function when $y \in [0, c]$,

$$\begin{aligned}\exp(\lambda y) = 1 + \lambda y + \sum_{n\geq 2} \frac{\lambda^n y^n}{n!} &\leq 1 + \lambda y + y^2 \sum_{n\geq 2} \frac{\lambda^n c^{n-2}}{n!} \\ &\leq 1 + \lambda y + y^2 \phi_c(y).\end{aligned}$$

To prove the above inequality when $y \leq 0$ we integrate the inequality $\exp(\lambda y) \leq 1$ twice over the interval $[y, 0]$; we find that

$$\exp(\lambda y) - (1 + \lambda y) \leq \frac{1}{2}\lambda^2 y^2, \quad y \leq 0,$$

but by definition $\frac{1}{2}\lambda^2 \leq \phi_c(\lambda)$.

The inequality that we have just proved implies that under the assumptions of the lemma

$$\begin{aligned}\mathbf{E}(\exp(\lambda(Z_{n+1} - Z_n))|\mathcal{B}_n) &\leq \mathbf{E}(1 + \lambda(Z_{n+1} - Z_n) + \phi_c(\lambda)(Z_{n+1} - Z_n)^2|\mathcal{B}_n) \\ &\leq 1 + \phi_c(\lambda)(A_{n+1} - A_n) \leq \exp(\phi_c(\lambda)(A_{n+1} - A_n))\end{aligned}$$

for every $n \in \mathbb{N}$. But this can also be written

$$\mathbf{E}(\exp(\lambda Z_{n+1} - \phi_c(\lambda) A_{n+1}) | \mathcal{B}_n) \leq \exp(\lambda Z_n - \phi_c(\lambda) A_n), \quad n \in \mathbb{N}$$

and the lemma is therefore proved. □

Lemma 2.14 (Maximal inequality). *For every positive supermartingale $\{Z_n\}_{n\in\mathbb{N}}$, the random variable $\sup_{n\in\mathbb{N}} Z_n$ is a.s. finite on the set $\{Z_0 < \infty\}$, and, more precisely, satisfies the following inequality*

$$\mathbf{P}(\sup_{n\in\mathbb{N}} Z_n \geq a) \leq \min\left\{\frac{Z_0}{a}, 1\right\}$$

for all constants $a > 0$.

Proof. Let us associate with the positive supermartingale $\{Z_n\}_{n\in\mathbb{N}}$ of the proposition the stopping time defined by

$$\tau_a = \begin{cases} \min\{n : Z_n > a\} & \text{if } \sup_{n\in\mathbb{N}} Z_n > a \\ \infty & \text{if } \sup_{n\in\mathbb{N}} Z_n \leq a \end{cases} \qquad a > 0.$$

Since $Z_{\tau_a} > a$ on $\{\tau_a < \infty\}$ and since the constant a can be considered a supermartingale, the formula

$$Y_n = \begin{cases} Z_n & \text{if } n < \tau_a \\ a & \text{if } n \geq \tau_a \end{cases} \qquad n \in \mathbb{N}$$

defines a new supermartingale by Lemma 2.15. Then we may write $Y_0 \geq \mathbf{E}(Y_n|\mathcal{B}_0)$; since Y_0 takes the value Z_0 or a according as $Z_0 \leq a$ or $Z_0 > a$, and since $Y_n \geq a\mathbf{1}_{\{\tau \leq n\}}$, the preceding inequality implies that

$$a\mathbf{E}(\mathbf{1}_{\{\tau_a \leq n\}}|\mathcal{B}_0) \leq \min\{Z_0, a\}.$$

Letting $n \to \infty$ and dividing by a, we obtain

$$\mathbf{E}(\mathbf{1}_{\{\sup Z_n > a\}}|\mathcal{B}_0) \leq \min\left\{\frac{Z_0}{a}, 1\right\}$$

since $\{\tau_a < \infty\} = \{\sup_{n\in\mathbb{N}} Z_n > a\}$. This inequality implies that of the lemma; it suffices to replace a by $a(1-1/k)$ in the inequality above and then let $k \to \infty$ to obtain the same inequality with $\geq$ instead of $>$ on the left-hand side, and then use the fact that $\mathbf{E}(\mathbf{E}(X|\mathcal{B}_0)) = \mathbf{E}(X)$ the inequality from the lemma follows. Finally, let us integrate both sides of this inequality over the event $\{Z_0 < \infty\}$ which belongs to $\mathcal{B}_0$; we find that

$$\mathbf{P}(Z_0 < \infty, \sup_{n\in\mathbb{N}} Z_n > a) \le \int_{\{Z_0<\infty\}} \min\left\{\frac{Z_0}{a}, 1\right\} dP.$$

When $a \to \infty$, the right-hand side tends to zero by the dominated convergence theorem and we have proved that

$$\mathbf{P}(Z_0 < \infty, \sup_{n\in\mathbb{N}} Z_n = \infty) = 0,$$

which completes the proof. □

Lemma 2.15 (Switching principle)**.** *Given two positive supermartingales* $\{Z_n^{(i)}\}_{n\in\mathbb{N}}$*,* $i = 1, 2$ *and a stopping time* τ *such that* $Z_\tau^{(1)} \ge Z_\tau^{(2)}$ *on* $\{\tau < \infty\}$*, the formula*

$$Z_n(\omega) = \begin{cases} Z_n^{(1)}(\omega) & \text{if } n < \tau(\omega) \\ Z_n^{(2)}(\omega) & \text{if } n \ge \tau(\omega) \end{cases}, \quad n \in \mathbb{N}$$

defines a new positive supermartingale.

Proof. Indeed, the defining formula of the Z_n can also be written

$$Z_n = \mathbf{1}_{\{\tau>n\}} Z_n^{(1)} + \mathbf{1}_{\{\tau\le n\}} Z_n^{(2)}$$

and it is clear that Z_n is $\mathcal{B}_n$-measurable for all $n \in \mathbb{N}$. The supermartingale property of the $Z^{(i)}$ allows us to write

$$\begin{aligned} Z_n &= \mathbf{1}_{\{\tau>n\}} Z_n^{(1)} + \mathbf{1}_{\{\tau\le n\}} Z_n^{(2)} \\ &\ge \mathbf{1}_{\{\tau>n\}} \mathbf{E}(Z_{n+1}^{(1)} | \mathcal{B}_n) + \mathbf{1}_{\{\tau\le n\}} \mathbf{E}(Z_{n+1}^{(2)} | \mathcal{B}_n) \\ &= \mathbf{E}(\mathbf{1}_{\{\tau>n\}} Z_{n+1}^{(1)} + \mathbf{1}_{\{\tau\le n\}} Z_{n+1}^{(2)} | \mathcal{B}_n). \end{aligned}$$

But the assumption $Z_\tau^{(1)} \ge Z_\tau^{(2)}$ on $\{\tau < \infty\}$ implies that $Z_{n+1}^{(1)} \ge Z_{n+1}^{(2)}$ on $\{\tau = n+1\}$; it then follows that

$$\mathbf{1}_{\{\tau>n\}} Z_{n+1}^{(1)} + \mathbf{1}_{\{\tau\le n\}} Z_{n+1}^{(2)} \ge \mathbf{1}_{\{\tau>n+1\}} Z_{n+1}^{(1)} + \mathbf{1}_{\{\tau\le n+1\}} Z_{n+1}^{(2)} = Z_{n+1}.$$

This proves that $Z_n \ge \mathbf{E}(X_{n+1} | \mathcal{B}_n)$. □

2.8 Computation of the MDL and BIC estimators

In practice, it is infeasible to calculate estimators via computing the value of an information criterion for each model, since the number of the hypothetical context trees is very large. However, an algorithm in this Section admits finding the considered estimators with practical computational complexity.

We consider both off-line and on-line methods, in the latter case with a slight modification of the estimators. Note that on-line calculation of the estimator is useful when the sample size is not fixed but we keep sampling until the estimator becomes "stable" say it remains constant when the sample size is doubled.

As usual, see (Baron and Bresler, 2004) and (Martín, Seroussi, and Weinberger, 2004), we assume that the computations are done in registers of size $O(\log n)$.

The estimators $\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n)$ and $\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n)$ in Theorems 2.1 and 2.2, the latter for the case $d(\mathcal{T}_0) < \infty$, can be represented as

$$\widehat{\mathcal{T}}(x_1^n) = \arg \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n, D(n)) \cap \mathcal{I}} \prod_{s \in \mathcal{T}} \tilde{P}_s(x_1^n) \tag{2.27}$$

where $\tilde{P}_s(x_1^n) = \tilde{P}_{\mathrm{KT},s}(x_1^n)$ in the KT case, and $\tilde{P}_s(x_1^n) = n^{-\frac{|A|-1}{2}} \tilde{P}_{\mathrm{ML},s}(x_1^n)$ in the BIC case, see (2.9), Definition 2.3, (2.2), Definition 2.3. In what follows, $\tilde{P}_s(x_1^n)$ will denote either possibility.

These facts admit a joint treatment of the computations of the BIC and KT estimators, via an extension of the CTM algorithm of Willems, Shtarkov, and Tjalkens (2000) developed for the KT case in Csiszár and Talata (2006). This algorithm has the following construction.

Consider the full tree A^D, where $D = D(n) = o(\log n)$, and let $\mathcal{S}_D$ denote the set of its nodes, i.e., the set of all strings of length at most D. Based on the sample x_1^n we assign to each node a value and a binary indicator. This assignment is recursive, that is, the value and the indicator assigned to a node are calculated from the values assigned to the children of this node. The desired estimator will be the subtree determined by the indicators as specified below.

Definition 2.4. Given a sample x_1^n, to each string $s \in \mathcal{S}_D$ with $N_n(s) \geq 1$, $D = D(n)$ we assign recursively, starting from the leaves of the full tree A^D, the value

$$V_s^D(x_1^n) = \begin{cases} \max\left\{ \tilde{P}_s(x_1^n), \prod_{a \in A : N_n(as) \geq 1} V_a^D s(x_1^n) \right\} & \text{if } 0 \leq l(s) < D, \\ \tilde{P}_s(x_1^n) & \text{if } l(s) = D, \end{cases}$$

and the indicator

$$\chi_s^D(x_1^n) = \begin{cases} 1 & \text{if } 0 \le l(s) < D \text{ and } \prod_{a\in A: N_n(as)\ge 1} V_{as}^D(x_1^n) > \tilde{P}_s(x_1^n), \\ 0 & \text{if } 0 \le l(s) < D \text{ and } \prod_{a\in A: N_n(as)\ge 1} V_{as}^D(x_1^n) \le \tilde{P}_s(x_1^n), \\ 0 & \text{if } l(s) = D. \end{cases}$$

Using these indicators, we assign to each $s \in \mathcal{S}_D$, $D = D(n)$ a maximizing tree $\mathcal{T}_s^D(x_1^n)$ consisting of strings $u \succeq s$. The term "maximizing" is justified by Lemma 2.17 below.

Definition 2.5. Given $s \in \mathcal{S}_D$, let $\mathcal{T}_s^D(x_1^n)$ equal to

$$\{u \in \mathcal{S}_D : \chi_u^D(x_1^n) = 0, \chi_v^D(x_1^n) = 1 \text{ for all } s \preceq v \prec u\}$$

if $\chi_s^D(x_1^n) = 1$, and to $\{s\}$ if $\chi_s^D(x_1^n) = 0$.

The maximizing tree $\mathcal{T}_s^D(x_1^n)$ is irreducible unless it equals $\{s\}$. Indeed, if $N_n(s) = N_n(as)$ holds for a string $s \in \mathcal{S}_{D-1}$ and a letter a (and thus $N_n(bs) = 0$ for all $b \neq a$, $b \in A$) then $\chi_s^D(x_1^n) = 1$ implies $\chi_{as}^D(x_1^n) = 1$.

Proposition 2.16. *The context tree estimator $\widehat{\mathcal{T}}(x_1^n)$ in (2.27) equals the maximizing tree assigned to the root, that is,*

$$\widehat{\mathcal{T}}(x_1^n) = \mathcal{T}_\lambda^D(x_1^n).$$

Proof. The claimed equality follows from the next lemma by substituting $s = \lambda$, on account of (2.27) and the fact that $\mathcal{T}_\lambda^D(x_1^n)$ is irreducible. □

For any $s \in \mathcal{S}_D$ with $N_n(s) \ge 1$, define $\mathcal{F}_1(x_1^n|s)$ as the family of all trees $\mathcal{T}$ of depth $d(\mathcal{T}) \le D$, consisting of strings $u \succeq s$ with $N_n(u) \ge 1$, such that each $s' \succ s$ with $N_n(s') \ge 1$ is either a suffix of some $u \in \mathcal{T}$ or has a suffix in $\mathcal{T}$.

Lemma 2.17. *For any $s \in \mathcal{S}_D$ with $N_n(s) \ge 1$*

$$V_s^D(x_1^n) = \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n|s)} \prod_{u\in\mathcal{T}} \tilde{P}_u(x_1^n) = \prod_{u\in\mathcal{T}_s^D(x_1^n)} \tilde{P}_u(x_1^n).$$

Proof. By induction on the length of the string s. For $l(s) = D$ the statement is obvious. Supposing the assertion holds for all strings of length d; we have for any s with $l(s) = d - 1$

$$\begin{aligned} \prod_{a\in A:N_n(as)\ge 1} V_{as}^D(x_1^n) &= \prod_{a\in A:N_n(as)\ge 1} \left(\max_{\mathcal{T}_a\in\mathcal{F}_1(x_1^n|as)} \prod_{u\in\mathcal{T}_a} \tilde{P}_u(x_1^n) \right) \\ &= \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n|s): d(\mathcal{T})\ge 1} \prod_{u\in\mathcal{T}} \tilde{P}_u(x_1^n). \end{aligned}$$

Here the second equality holds since any family of trees $\mathcal{T}_a$, $a \in A$, $N_n(as) \geq 1$, satisfying the indicated constraints, uniquely corresponds to a tree $\mathcal{T} \in \mathcal{F}_1(x_1^n|s)$ with $d(\mathcal{T}) \geq 1$ via $\mathcal{T} = \cup_a \mathcal{T}_a$.

It follows by Definition 2.4 that

$$V_s^D(x_1^n) = \max\left\{ \tilde{P}_s(x_1^n), \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n|s): d(\mathcal{T}) \geq 1} \prod_{u \in \mathcal{T}} \tilde{P}_u(x_1^n) \right\} = \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n|s)} \prod_{u \in \mathcal{T}} \tilde{P}_u(x_1^n),$$

proving the first equality in the lemma. The second equality also follows from the last identity, by the induction hypothesis and Definitions 2.4 and 2.5. □

Remark. For the KT case, Lemma 2.17 above with the condition $\mathcal{T} \in \mathcal{F}_1(x_1^n|s)$ replaced by the condition that $\mathcal{T}$ is complete, is a result of Willems, Shtarkov, and Tjalkens (2000) (with the minor difference that the trees there also had "costs"), and the above proof is similar to theirs.

The KT estimator in Theorem 2.2 for the general case can still be represented as in (2.27), with $\tilde{P}_s(x_1^n) = \tilde{P}_{\mathrm{KT},s}(x_1^n)$, the only difference is that $\mathcal{F}_1(x_1^n, D(n))$ in (2.27) is replaced by $\mathcal{F}_r(x_1^n, D(n))$ with $r = n^\beta$. For this case, Definition 2.4 is modified by setting $V_s^D(x_1^n) = 0$ for all $s \in \mathcal{S}_D$ with $N_n(s) < r$. The definition remains unchanged for $s \in \mathcal{S}_D$ with $N_n(s) \geq r$, but of course the values $V_s^D(x_1^n)$ may change also for these strings s. In particular, if $N_n(s) \geq r$ but $1 \leq N_n(as) < r$ for some $a \in A$, the modified definition gives $V_s^D(x_1^n) = \tilde{P}_s(x_1^n)$ and $\chi_s^D(x_1^n) = 0$.

Adopting this modified Definition 2.4, it is easy to see that Proposition 2.16 still holds, that is, the maximizing tree of Definition 2.5 assigned to the root equals the KT estimator in Theorem 2.2 for the general case.

Next we show that the computation of the estimators in Theorems 2.1 and 2.2 via the above method has the asserted complexity in the off-line case.

Theorem 2.18. *The number of computations needed to determine the* BIC *estimator and the* KT *estimator in Theorems 2.1 and 2.2 for a given sample x_1^n is $O(n)$, and this can be achieved storing $O(n^\varepsilon)$ data, where $\varepsilon > 0$ is arbitrary.*

Proof. Since $D(n) = o(\log n)$, we may write $D(n) = \varepsilon_n \log n$, where $\varepsilon_n \to 0$. For each string $s \in \mathcal{S}_D$, $D = D(n) = \varepsilon_n \log n$, the counts $N_n(s,a)$, $a \in A$, as well as $\tilde{P}_s(x_1^n)$, $V_s^D(x_1^n)$, $\chi_s^D(x_1^n)$ are stored. The number of stored data is proportional to the cardinality of $\mathcal{S}_D$, which is

$$\sum_{j=0}^{D} |A|^j = \frac{|A|^{D+1} - 1}{|A| - 1} \leq 2|A|^D = O(n^\varepsilon). \tag{2.28}$$

To get the indicators $\chi_s^D(x_1^n)$, $s \in \mathcal{S}_D$ which give rise to the trees $\mathcal{T}_s^D(x_1^n)$ according to Definition 2.5, first we need the counts $N_n(s,a)$, $s \in \mathcal{S}_D$, $a \in A$.

The counts $N_n(s,a)$ for $s \in A^D$, $a \in A$ can be determined by successively processing the sample x_1^n from position $j = D(n)$ to $j = n$, and at instance j incrementing the count $N_n(x_{j-D(n)}^{j-1}, x_j)$ by 1 (the starting values of all counts being 0). This is $O(n)$ calculations. The other counts $N_n(s,a)$, $s \in \mathcal{S}_{D-1}$, $a \in A$ can be determined recursively, as $N_n(s,a) = \sum_{b \in A} N_n(bs,a)$. This is $|A||S_{D-1}| = o(n)$ calculations.

Then, from these counts the values $\tilde{P}_s(x_1^n)$ are determined by $O(n)$ multiplications. The calculation of the values $V_s^D(x_1^n)$ and $\chi_s^D(x_1^n)$ requires calculations proportional to the cardinality of $\mathcal{S}_D$, which is less than $2|A|^D = o(n)$. □

On-line algorithms are considered with the following minor modifications of the estimators, which obviously do not affect the consistency. In the BIC penalty term, $\log n$ is replaced by $\lfloor \log_{|A|} n \rfloor \log |A|$, and in the second kind of KT estimator in Theorem 2.2 $\mathcal{F}_{n^\beta}(x_1^n, D(n))$ is replaced by $\mathcal{F}_r(x_1^n, D(n))$ with $r = e^{\beta \lfloor \log_{|A|} n \rfloor}$. No modification is needed in the first kind of KT estimator whose consistency has been proved for the case $d(\mathcal{T}_0) < \infty$.

Consider next the on-line versions of the estimators, with the modifications described in the passage above. In the BIC case, the representation 2.27 holds with

$$\tilde{P}_s(x_1^n) = e^{-\frac{|A|-1}{2} \lfloor \log_{|A|} n \rfloor \log |A|} \tilde{P}_{\mathrm{ML},s}(x_1^n).$$

In the KT case, the same estimator is used as for the off-line computation, when $d(\mathcal{T}_0) < \infty$. The on-line version of the KT estimator for the general case is analogous to the off-line version, with $r = e^{\beta \lfloor \log_{|A|} n \rfloor}$ instead of $r = n^\beta$.

Finally, we show that these algorithms have the asserted computational complexity in the on-line case.

Theorem 2.19. *Suppose $D(n) = o(\log n)$ is a nondecreasing function of n. Adopting the above modifications, the number of computations needed to determine the* BIC *estimator in Theorem 2.1 or the* KT *estimator in Theorem 2.2, simultaneously for all subsamples x_1^i $i \le n$, is $o(n \log n)$, and this can be achieved storing $O(n^\varepsilon)$ data at any time, where $\varepsilon > 0$ is arbitrary.*

Proof. The calculations required by the algorithm in Definition 2.4 can be performed recursively in the sample size n.

Suppose at instant i, for each string $s \in \mathcal{S}_{D(i)}$, the counts $N_i(s,a)$, $a \in A$, as well as $\tilde{P}_s(x_1^i)$, $V_s^D(x_1^i)$, $\chi_s^D(x_1^i)$ are stored, where $D = D(i)$. The number of stored data is proportional to the cardinality of $\mathcal{S}_{D(i)}$, which is $O(i^\varepsilon)$, see (2.28).

Consider first those instances i when the sample size increases from $i-1$ to i but $\lfloor \log_{|A|}(i-1) \rfloor = \lfloor \log_{|A|} i \rfloor$, and the depth does not change, $D(i) = D(i-1)$. If $\tilde{P}_s(x_1^{i-1})$ at a node s is known, $\tilde{P}_s(x_1^i)$ can be calculated using, for the KT case, that

$$\tilde{P}_{\mathrm{KT},s}(x_1^i) = \frac{N_i(s,x_i)+1/2}{N_i(s)+|A|/2} \tilde{P}_{\mathrm{KT},s}(x_1^{i-1})$$

and in the BIC case that in the expression of $\tilde{P}_{\mathrm{ML},s}(x_1^{i-1})$ only the counts $N_i(s,x_i)$ and $N_i(s)$ were incremented to obtain $\tilde{P}_{\mathrm{ML},s}(x_1^i)$. From $\tilde{P}_s(x_1^i)$ the values $V_s^D(x_1^i)$ and $\chi_s^D(x_1^i)$ can be computed in constant number of steps. These values are different for x_1^{i-1} and x_1^i only when s is a suffix of x_1^{i-1}, hence updating is needed at $D(i)$ nodes only. Thus, the number of required computations is proportional to $D(i)$.

Consider those instances i when the sample size increases from $i-1$ to i such that $\lfloor \log_{|A|} i \rfloor = \lfloor \log_{|A|}(i-1) \rfloor + 1$ but the depth does not change. The additional task compared to the previous case is that recalculation of $V_s^D(x_1^i)$ and $\chi_s^D(x_1^i)$ is needed for all nodes $s \in \mathcal{S}_{D(i)}$, which requires calculations proportional to the cardinality of $\mathcal{S}_{D(i)}$.

Consider next those instances i when the depth increases, $D(i) = D(i-1)+1$. In this case we have three tasks. We have to update $\tilde{P}_s(x_1^{i-1})$ at those nodes s that already existed at instance $i-1$, namely, where $l(s) < D(i)$. In addition, we have to calculate them for the new terminal nodes s, $l(s) = D(i)$, and recalculate $V_s^D(x_1^i)$ and $\chi_s^D(x_1^i)$ at all nodes s of the new full tree. The former needs $O(i)$ calculations. Indeed, the counts $N_i(s,a)$, $l(s) = D(i)$, can be determined by successively processing the sample x_1^i from position $j = D(i)$ to $j = i$, and at instance j incrementing the count $N_i(x_{j-D(i)}^{j-1}, x_j)$ by 1, the starting values of all counts being 0; from these counts, the values $\tilde{P}_s(x_1^i)$ are determined by $O(i)$ multiplications. The recalculation of the values $V_s^D(x_1^i)$ and $\chi_s^D(x_1^i)$ requires calculations proportional to the cardinality of $\mathcal{S}_{D(i)}$.

Finally, the total number of computations performed on a sample x_1^n is bounded as follows. The number of computations needed for the updating at all instances $i \leq n$ is proportional to

$$\sum_{i=1}^{n} D(i) = \sum_{i=1}^{n} \lfloor \varepsilon_i \log i \rfloor = o(n \log n).$$

The number of computations to recalculate V_s^D, χ_s^D for all nodes in the full tree $A^{D(i)}$ at the instances when $\lfloor \log_{|A|} i \rfloor$ increases is of order

$$\sum_{D=0}^{\lfloor \log_{|A|} n \rfloor} 2|A|^D = O(|A|^{\log_{|A|} n}) = O(n).$$

The number of computations to calculate $\tilde{P}_s$ for the new terminal nodes at the instances when $D(i)$ increases is proportional to

$$\begin{aligned}\sum_{D=0}^{\lfloor\varepsilon_n \log n\rfloor} \min\{i : D \leq \varepsilon_i \log i\} &= \sum_{D=0}^{\lfloor\varepsilon_n \log n\rfloor} \min\{i : e^{D/\varepsilon_i} \leq i\} \\ &\leq \sum_{D=0}^{\lfloor\varepsilon_n \log n\rfloor} e^{D/\varepsilon_i} + 1 \\ &\leq O(e^{\frac{1}{\varepsilon_n}\varepsilon_n \log n}) + \varepsilon_n \log n = O(n).\end{aligned}$$

The number of computations to recalculate V_s^D, χ_s^D for all nodes in the full tree $A^{D(i)}$ at the instances when $D(i)$ increases is of order

$$\sum_{D=0}^{\lfloor\varepsilon_n \log n\rfloor} 2|A|^D = O(|A|^{\varepsilon_n \log n}) = o(n).$$

□

Remark. Of course, the $O(n^\varepsilon)$ storage does not include storage of the context tree estimators for all $i \leq n$; note that for the indicated purpose of deciding when to stop sampling, it suffices to keep track of the last instance when the estimator has changed.

CHAPTER 3

Estimation of Sparse tree models

3.1 Markov models and different parametrizations

For the purpose of fitting them to data, we must consider special types of processes, for otherwise it would take an infinite number of conditional probabilities to specify them. The most familiar and important subclasses of processes are those of finite memory.

Definition 3.1. A process $X_1, X_2, \ldots$ is a *Markov chain* (or has the finite-memory property), if there exists a nonnegative integer k such that, for all $n \geq k$, $a \in A$, and $x_1^n \in A^n$, $\mathbf{P}(X_{n+1} = a | X_1^n = x_1^n)$ satisfies

$$\mathbf{P}(X_{n+1} = a | X_1^n = x_1^n) = \mathbf{P}(X_{n+1} = a | X_{n-k+1}^n = x_{n-k+1}^n)$$

The minimum integer k for which the finite-memory property holds for $\{X_n\}$ is referred to as the *order* of the process.

If, in particular, the probabilities $\mathbf{P}(X_{n+1} = a | X_{n-k}^n = x_{n-k}^n)$ do not change as a function of n, the process is called *time-invariant*, or often by an abuse of language "stationary". In that case we denote $Q(a | x_{n-K}^n) = \mathbf{P}(X_{n+1} = a | X_{n-k}^n = x_{n-k}^n)$ and $Q(x_1^n) = \mathbf{P}(X_1^n = x_1^n)$ so we can then write the recursion

$$\begin{aligned} Q(x^{n+1}) &= Q(x_{n+1} | x_{n-k+1}^n) Q(x^n) \\ &= Q(x_{n+1} | x_{n-k+1}^n) Q(x_n | x_{n-k}^{n-1}) \ldots Q(x_{k+1} | x_1^k) Q(x_1^k) \end{aligned}$$

where the segments x_{n-k}^n are sometimes called states, the conditional probabilities $Q(a | x_{i-k}^i)$ are called *transition probabilities* and the value $Q(x_1^k)$ is known as the initial distribution.

In a finite-memory process, the conditional probability assigned to the next emitted symbol, given all the past, depends only on a finite number k of contiguous past observations. This class of processes can be parametrized with the usual Markov model of order k, but since for practical data the actual memory length often varies from location to location, such parametrizations can be very inefficient. The number of model parameters, which grows exponentially with k in a Markov model, can be dramatically reduced by lumping together equivalent states (i.e., k-vectors) that yield identical conditional distributions.

There are two characteristics of particular importance in Markov chains: the order and the number of free parameters needed to define the process. A Markov chain of order k with an alphabet size $|A|$ has $|A|^k$ states, each having $|A| - 1$ probability parameters. However, there are important applications where the number of states is very large but the number of parameters required to specify the process is small, which makes their estimation relatively easy.

A subclass of the Markov chains of order k, often used in statistical modeling, is specified by the assumption that the transition probabilities $Q(a|x_1^n)$ depend on x_1^n through a "context function" $\sigma(x_1^n)$ that takes values on a finite set say $\mathcal{S}$ which is called the state space. What we are doing is lumping together transition probabilities for different past strings

$$\mathbf{P}(X_{n+1} = a | X_1^n = x_1^n) = Q(a|\sigma(x_1^n)) \quad \forall a \in A.$$

3.2 Recursive Models

One way of generating a process is by use of a *recursive model*, see Rissanen and Langdon (1981). Specifically, given a set of states $\mathcal{S}$, consider a *context function* $\sigma : A^* \to \mathcal{S}$ and a set of conditional probability functions (CPF) $\{Q(\cdot|s)\}_{s\in\mathcal{S}}$. For an arbitrary sequence $x_1^n \in A^n$, let the state sequence be given by $s_i = \sigma(x^i)$, $0 \leq i \leq n$, and define the process by

$$\mathbf{P}(X_1^n = x_1^n) \overset{\text{def}}{=} Q(x_1^n) = \prod_{i=1}^{n} Q(x_i|s_{i-1}), \quad n \geq 1. \tag{3.1}$$

Clearly, this assignment defines a process. We say that the model, denoted $\langle\sigma, Q\rangle$, generates the process $\{X_n\}$. The process is then defined by the $|\mathcal{S}| \times |A|$ conditional probabilities and the context function σ.

For any state s, and x_1^n such that $\sigma(x_1^n) = s$, we say that x_1^n *selects* s, and that s *accepts* x_1^n, we also say that x_1^n occurs in *context* s. A state is called *permanent* if it accepts arbitrarily long sequences; otherwise, the state is called *transient*.

Clearly, the finite-memory property holds for k if $\{X_n\}$ can be generated with a recursive model such that, for all $n \geq k$ and $x_1^n \in A^n$, x_1^n selects the same state as x_{n-k+1}^n. In particular, if the context function is the restriction $\sigma(x_1^n) = x_{n-k+1}^n$ then we recover the usual Markov chains, in this case $\mathcal{S} = A^k$ but the advantage of this models is that state space has usually less than $|A|^k$ possible values.

3.3 Tree Models (VLMC)

For each particular Markov chain of order k, other recursive representations may involve less than $|A|^k$ (permanent) states. A *tree model* (see, e.g., Rissanen 1983a; Weinberger et al. 1995) is another type of recursive model which may involve less states than the "basic" Markov chain representation. In a tree model, the permanent states are not necessarily all of the same length k. Specifically, given a complete suffix free set over A^*, the state selected by x_1^n is given by the (unique) suffix of x_1^n in the set, if n is large enough for such a suffix to exist (permanent state), or by x_1^n otherwise (transient state). Thus, the set of permanent states is most naturally represented by the leaves of a complete $|A|$-ary tree. We will represent the states as strings, which we will call *contexts*, $\sigma(x^n) = x_{n-j} \ldots x_n$ where j depends on the actual symbols.

Roughly speaking, a tree model consists of *context tree* (which is a complete $|A|$-ary tree), and a set of conditional probability distributions over the alphabet, one associated with each leaf of the tree (the *states*). This models are also known as *Variable length Markov Chains* (VLMC) in the statistics literature, a term coined by Buhlmann and Wyner (1998).

Consider a $|A|$-ary complete tree $\mathcal{T}$ (each node either is a leaf or has exactly $|A|$ offspring), where the branches are labeled by the symbols in the alphabet. Each context defines a node in the tree reached by the path starting at the root with the branch x_n, followed by the branch x_{n-1}, and so on. The tree $\mathcal{T}$ permits finding a distinguished context for each symbol x_t in the string x_1^n. We will think of the tree $\mathcal{T}$ both as a combinatorial structure and as suffix free set of strings formed by the paths from the leaves to the root.

To specify a process we need in addition a set of CPFs$\{Q(a|s) : s \in \mathcal{T}\}$, that has $Q(a|s) > 0$ for all $a \in A$ and $s \in \mathcal{T}$. We will define the state map $\sigma(x_1^n) = s$ iff $s \prec x_1^n$ (since $\mathcal{T}$ is a suffix free set, s is unique), or $\sigma(x_1^n) = x_1^n$ if there is no such s (for small values of n). We define the process generated by $\langle \mathcal{T}, Q \rangle$ by

$$Q(x^n) = \prod_{i=1}^{n} Q(x_i|\sigma(x^{i-1})), \quad n \geq 1.$$

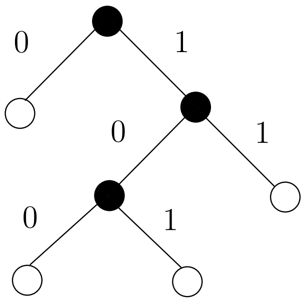


Figure 3.1: The context tree $T = \{0, 001, 101, 11\}$.

For this we need to know the CPFs for the states that are not leaves of the tree $\mathcal{T}$, since this states are transient the choice of their CPFs is unimportant, and they play a role similar to that of the initial distribution in a Markov chain. We can alternatively define the process without transient states in a simpler way described below.

Let $h = h(\mathcal{T}) = \max\{|s| : s \in \mathcal{T}\}$ be the maximum depth of the tree, and let $s_0 = 0^h$ denote the all-zero string of length. We define a new state function $\sigma'(x_1^n) = \sigma(0^h x_1^n)$ where $0^h x_1^n$ denotes the string x_1^n padded with h initial zeros. This guarantees that the state is a leaf for all $n > 0$, in fact $Q(x_1^n)$ using this σ' is equivalent to $Q(x_1^n|s_0)$.

Remark. In Chapter 2 we drop the completeness requirement, that each non-leaf node of the context tree has as many children as the alphabet size. If some strings have zero probability for the given process, these can not be contexts, and then the context tree need not be complete. However, the set of strings without a suffix in the tree has zero probability. The results hold then if we force the tree to be complete.

3.3.1 Minimal context trees

A tree model $\langle \mathcal{T}, Q \rangle$ is said to be *minimal* if no other tree model $\langle \mathcal{T}', Q' \rangle$ generates the same process and has a smaller number of permanent states.

We will prove the following characterization of minimality, $\langle \mathcal{T}, Q \rangle$ is minimal if for every node u in T with all its successors vu as leaves, there exist a, b, and c in $\mathcal{A}$ satisfying $P(a|bu) \neq P(a|cu)$. Clearly, if for some such node u the distributions $P(\cdot|bu)$ are equal for all b, we could lump the successors into u and have a smaller complete tree representing the same process.

In a minimal tree representation, the state $\sigma(x_1^n)$ for $x_1^n \in \mathcal{A}^n$ is determined by the smallest integer $\ell(x_1^n)$ such that $P(\cdot|ux_{n-\ell(x_1^n)+1}^n)$ is independent of u, with $Q(ux_{n-\ell(x_1^n)+1}^n) > 0$. Sets of "sibling" leaves $\{ab_{m-1} \dots b_2 b_1 | a \in A\}$ sharing the same CPF in the original model can be merged into one state (leaf), represented by the parent node $b_{m-1} \dots b_2 b_1$. The merging is repeated recursively whenever possible, seeking the shortest possible context that determines the CPF, until any set of $|A|$ sibling leaves contains at least two leaves with different associated CPFs.

We say that a tree $\mathcal{T}'$ is an *extension* of a tree $\mathcal{T}$ if it contains all the nodes of $\mathcal{T}$.

Proposition 3.1. *A complete tree model $\langle \mathcal{T}, Q \rangle$ such that $Q(s) > 0$ for $s \in A^*$ is minimal if and only if there is no set of $|A|$ sibling leaves of $\mathcal{T}$ sharing the same* CPF*. Moreover, if $\langle \mathcal{T}, Q \rangle$ is minimal, and $\langle \mathcal{T}', Q' \rangle$ generates the same process, where $\mathcal{T}'$ is also full, then $\mathcal{T}'$ is an extension of $\mathcal{T}$.*

Proof. The necessity of the minimality condition is straightforward, since sets of sibling leaves with identical CPFs can always be merged, reducing the number of states.

Assume the condition holds, and $\langle \mathcal{T}', Q' \rangle$ generates the same process as $\langle \mathcal{T}, Q \rangle$, with $\mathcal{T}'$ complete. Assume u is a node in $\mathcal{T} \backslash \mathcal{T}'$. Then, there is a leaf $u' \in \mathcal{T}'$ such that $u' \prec u$, and there is a complete set of sibling leaves of $\mathcal{T}$ that descend from u'. But, since $u' \in \mathcal{T}'$ is a leaf, $Q(s) > 0$ for all $s \in A^*$, and both tree models generate the same process, these leaves of $\mathcal{T}$ must be associated with the same CPF that is associated with $u' \in \mathcal{T}'$, contradicting the assumed condition. Thus, we must have $\mathcal{T} \subset \mathcal{T}'$, which also establishes the minimality of $\mathcal{T}$. □

3.4 Sparse Models

We are going to propose a new type of Markov models with sparse dependencies. We choose a fixed symbol $\phi \notin A$, and we denote by A_ϕ the *expanded alphabet* $A_\phi = A \cup \{\phi\}$. Strings over A_ϕ will be referred to as *patterns*, and patterns terminating in a symbol from A as *proper patterns*.

The set of proper patterns of length $m \geq 1$ will be denoted

$$\bar{A}_\phi^m = A_\phi^{m-1} A = \{\, w_1^m \in A_\phi^m \,|\, w_m \in A \,\}.$$

We will interpret ϕ as a wildcard. Given a pattern $w_1^m \in A_\phi^m$, sequences in the set

$$\mathcal{C}(w) = \{\, u_1^m \in A^m \,|\, u_i = w_i \ \text{ whenever } w_i \neq \phi \,\}$$

are said to be *consistent* or *conformal* with w_1^m.

We say a pattern v is a *ϕ-suffix* of w, denoted by $v \prec_\phi w$, when $l(v) \leq l(w)$ and $v_{n-i} = w_{m-i}$ for all $i \leq l(v)$ such that $v_{n-i} \neq \phi$, if $w \neq v$ we say that v is a proper ϕ-suffix. This implies that every $s \in \mathcal{C}(w)$ has a suffix in $\mathcal{C}(v)$.

3.4.1 Sparse Context Models

Let $\Gamma = \mathcal{P}(\{1 \ldots k\})$ denote the power set, we will identify Γ with the set of *Sparse Context Models* (SCMs). Given $\gamma \in \Gamma$ we define the model associated to γ as

$$M(\gamma) = \{w \in A_\phi^k : w_{k-i} = \phi \quad \forall i \in \gamma\}.$$

Define $\sigma_\gamma : A^* \to M(\gamma)$ the context function associated to γ given by $\sigma_\gamma(x_1^n) = w \in M(\gamma)$ if and only if $x_{n-l(w)+1}^n \in \mathcal{C}(w)$.

Definition 3.2. Given a SCM $\gamma = \{i_1, \ldots, i_d\} \in \Gamma$ we say that a process $\{X_n\}_{n \in \mathbb{N}}$ is *γ-adapted* if whenever $\sigma_\gamma(x_1^n) = w$ then

$$\mathbf{P}(X_{n+1} = a | X_1^n = x_1^n) = \mathbf{P}(X_{n+1} = a | \sigma_\gamma(X_1^n) = w).$$

Most of the properties of this class of Models are similar to those of the more general class of Sparse Tree Models (STMs) described below. We will concentrate on this bigger class of Models; however, the class of SCMs exhibit some of the sparseness properties of STMs, and are interesting representations of processes on their own. In a sense, SCMs are to STMs as fixed order Markov models are to Tree Models.

3.4.2 Sparse Tree Models

A set $\mathcal{T}$ of patterns is called a *sparse context tree* if no $w \in \mathcal{T}$ is a ϕ-suffix of any other $v \in \mathcal{T}$ (we say that $\mathcal{T}$ is a ϕ-suffix free set). We define the height of the tree $\mathcal{T}$ as $h(\mathcal{T}) = \max\{l(u) : u \in \mathcal{T}\}$. Moreover a sparse tree is said to be *complete* when every sequence x_1^n with $n > h(\mathcal{T})$ has a ϕ-suffix in the tree, i.e. if $h(\mathcal{T}) \leq k$ we have that $\{\mathcal{C}(w) : w \in \mathcal{T}\}$ are disjoint and their union is A^K. Denote $\mathcal{M}_k$ the set of complete sparse trees of height bounded by k. Moreover a sparse tree is said to be *irreducible* when no pattern can be replaced by a proper ϕ-suffix without violating the ϕ-suffix property, this notion generalizes the concept of a complete tree.

We can represent a sparse tree as a graph, where each pattern is visualized as a path from the root to a leaf, see Figure 3.4.2. The obtained structure is a unary,$|A|$-ary tree with labeled edges, in which every node has either $|A|$ children each one labeled with a different letter, or only 1 child labeled with the ϕ symbol.

Define $\sigma_T : A^* \to \mathcal{T}$ the *state function* associated to a sparse tree $\mathcal{T}$, by $\sigma_\mathcal{T}(x_1^n) = w \in \mathcal{T}$ if and only if $x_{n-l(w)+1}^n \in \mathcal{C}(w)$ i.e. $w \prec_\phi x_1^n$ (since $\mathcal{T}$ is a ϕ-suffix free set, w is unique).

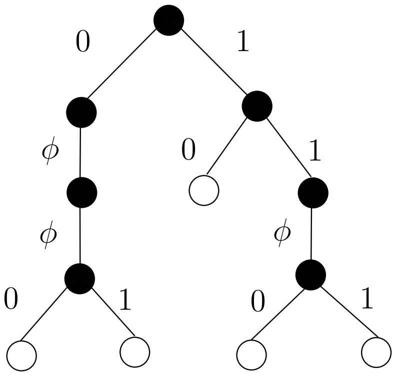


Figure 3.2: The sparse tree $\mathcal{T} = \{0\phi\phi0, 1\phi\phi0, 01, 0\phi11, 1\phi11\}$.

Consider a stationary Markov process $\{X_n\}_{n\in\mathbb{N}}$ taking values on A. Given a pattern w of length k we denote the probability of the pattern w by

$$Q(w) = \mathbf{P}(X_{n-k+1}^n \in \mathcal{C}(w)).$$

$$\text{If } Q(w) > 0, \quad Q(a|w) = \mathbf{P}(X_{n+1} = a | X_{n-k+1}^n \in \mathcal{C}(w)),$$

will stand for the conditional probability.

Definition 3.3. A *sparse tree model* over A is an ordered pair $\langle \mathcal{T}, Q\rangle$ such that: $\mathcal{T}$ is an complete sparse tree, and $Q = \{Q(\cdot|w) : w \in \mathcal{T}\}$ is a family of conditional probabilities over A.

Given a sparse context tree $\mathcal{T}$ we say that a process $\{X_n\}$ is $\mathcal{T}$*-adapted* if $\forall n \in \mathbb{N}$,$\forall a \in A$, and $\forall x_1^n$ such that $w \prec x_1^n$ we have that

$$\mathbf{P}(X_{n+1} = a|X_1^n = x_1^n) = Q(a|w)$$

Using the state function we can rewrite the above property as

$$\mathbf{P}(X_{n+1} = a|X_1^n = x_1^n) = \mathbf{P}(X_{n+1} = a|\sigma_T(X^n) = w).$$

Definition 3.4. A pattern w of length k is a *sparse context* for the process X if $Q(w) > 0$ and for any sequence x_1^n such that $x_{n-k+1}^n \in \mathcal{C}(w)$ we have that

$$\mathbf{P}(X_{n+1} = a|X_{n-k+1}^n = x_{n-k+1}^n) = Q(a|w) \quad \forall a \in \mathcal{A},$$

and no ϕ-prefix of w satisfies this equation. The set of contexts of a process form a sparse tree T, and it is clear that the process is T-adapted.

3.4.3 Entropy Rate

The entropy $\mathbf{H}(X)$ of a random variable X is defined by the formula

$$\mathbf{H}(X) = -\sum_{a\in A} Q(a)\log Q(a) = \mathbf{E}_P(\log Q(a)),$$

where Q is the probability distribution of X, i.e. $Q(a) = \mathbf{P}(X = a)$, $a \in A$; base two logarithms are used and $0 \log 0$ is defined to be 0.

If Y is another random variable, the joint entropy is defined to be the entropy of the pair (X, Y) as a random variable taking values on $A \times A$, the conditional entropy $\mathbf{H}(X|Y)$ is defined as the average, with respect to the distribution of Y, of the entropy of the conditional distribution of X, given $Y = y$.

Given a process $\mathbf{X} = \{X_n\}$ we define its *entropy rate* as the following limit provided that exists

$$\mathbf{H}(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n}\mathbf{H}(X_1,\ldots,X_n) = \sum_{a_1^n\in A^n} Q(a_1^n)\log Q(a_1^n),$$

where $Q(a_1^n) = \mathbf{P}(X_1^n = a_1^n)$ is the joint distribution of the n first variables.

Theorem 3.2. *For a stationary process $\mathbf{X} = \{X_n\}$ the entropy rate limit exists and it holds that*

$$\mathbf{H}(\mathbf{X}) = \lim_{n\to\infty} \mathbf{H}(X_n|X_{n-1},\ldots,X_1)$$

Proof. The limit exists since conditioning reduces the entropy and stationarity implies that

$$\begin{aligned}\mathbf{H}(X_{n+1}|X_n,\ldots,X_1) &\le \mathbf{H}(X_{n+1}|X_n,\ldots,X_2)\\ &\le \mathbf{H}(X_n|X_{n-1},\ldots,X_1),\end{aligned}$$

Hence the sequence $\mathbf{H}(X_n|X_{n-1},\ldots,X_1)$ is non-increasing and therefore it has a limit, we denote it by H.

Consider then

$$\begin{aligned}\frac{1}{n}\mathbf{H}(X_1,\ldots,X_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(X_i|X_{i-1},\ldots,X_1)\\ &= H + \frac{1}{n}\sum_{i=1}^{n}[\mathbf{H}(X_i|X_{i-1},\ldots,X_1) - H].\end{aligned}$$

The difference within the brackets is not greater than ε for $i \geq N_\varepsilon$, which implies that

$$\frac{1}{n}\mathbf{H}(X_1,\ldots,X_n) \geq H + \frac{1}{n}\sum_{i=1}^{N_\varepsilon}[H(X_i|X_{i-1},\ldots,X_1) - H] + \frac{n-N_\varepsilon}{n}\varepsilon.$$

The second term goes to zero with increasing n and since the last term is less than ε, we certainly have

$$\mathbf{H}(\mathbf{X}) \leq H + 2\varepsilon.$$

Since we can take ε as small as we like, $\mathbf{H}(\mathbf{X}) \leq H$. To get the opposite inequality, notice that by stationarity and the fact that increasing the amount of conditioning cannot increase the entropy $\mathbf{H}(X_i|X_{i-1},\ldots,X_1) - H \geq 0$. Hence $\frac{1}{n}\mathbf{H}(X_1,\ldots,X_n) \geq H$, and so is the limit $\mathbf{H}(\mathbf{X})$. □

3.4.4 Faithful representations

Given a process $\mathbf{X} = \{X_n\}$ we denote its entropy rate by $\mathbf{H}(\mathbf{X})$. The $\mathcal{T}$*-entropy* of a process is given by the function

$$\mathbf{H}_{\mathcal{T}}(\mathbf{X}) = -\sum_{w\in\mathcal{T}} Q(w)\sum_{a\in A} Q(a|w)\log Q(a|w).$$

Proposition 3.3. *If* $\mathbf{X}$ *is* $\mathcal{T}$*-adapted then* $\mathbf{H}(\mathbf{X}) = \mathbf{H}_{\mathcal{T}}(\mathbf{X})$.

Proof. Let $f(t) = t\log t$ and k be the memory of $\{X_n\}$:

$$\begin{aligned}
\mathbf{H}(\mathbf{X}) &= \lim_{n\to\infty}\mathbf{H}(X_{n+1}|X_1^n) = \mathbf{H}(X_{k+1}|X_1^k) = \\
&= -\sum_{u\in A^k}\mathbf{P}(X_1^k = u)\sum_{a\in A} f(\mathbf{P}(X_{k+1} = a|X_1^k = u)) \\
&= -\sum_{w\in\mathcal{T}}\sum_{u\in\mathcal{C}(w)}\mathbf{P}(X_1^k = u)\sum_{a\in A} f(\mathbf{P}(X_{k+1} = a|X_1^k = u)) \\
&= -\sum_{w\in\mathcal{T}} Q(w)\sum_{a\in A} Q(a|w)\log Q(a|w) = \mathbf{H}_{\mathcal{T}}(\mathbf{X})
\end{aligned}$$

□

Given $\mathbf{X}$ we define the set of *faithful representations*

$$\mathcal{F}(\mathbf{X}) = \{\mathcal{T}\in\mathcal{M} : \ \mathbf{X} \text{ is } \mathcal{T}\text{-adapted}\},$$

and the *minimizing sparse tree*

$$\mathcal{T}_0(\mathbf{X}) = \arg\min\{|\mathcal{T}| : \mathcal{T}\in\mathcal{F}(X)\}.$$

Proposition 3.4. *Given a process* $\mathbf{X} = \{X_n\}$ *we claim that* $\mathbf{H}_{\mathcal{T}}(\mathbf{X})$ *is minimal, if and only if* $\mathbf{X}$ *is* $\mathcal{T}$*-adapted. In particular it holds that*

$$\mathcal{F}(\mathbf{X}) = \{\mathcal{T} \in \mathcal{M} : \ \mathbf{H}_{\mathcal{T}}(\mathbf{X}) \leq \mathbf{H}_{\mathcal{T}'}(\mathbf{X}) \quad \forall \mathcal{T}' \in \mathcal{M}\}.$$

Proof. ($\Uparrow$) Let $\mathbf{X}$ be $\mathcal{T}$-adapted we shall prove that $\mathbf{H}_{\mathcal{T}}(\mathbf{X})$ is minimum. Let $\mathcal{T}'$ be another tree, then for $v \in \mathcal{T}'$ if we define $R(w|v) = \mathbf{P}(X_1^k \in \mathcal{C}(w)|X_1^k \in \mathcal{C}(v))$ we have that $\sum_{w \in T} R(w|v) = 1$ and

$$Q(a|v) = \sum_{w \in T} Q(a|w)R(w|v).$$

Then applying Jensen inequality to the function $f(x) = -x \log x$ we obtain

$$\begin{aligned}
\mathbf{H}_{\mathcal{T}'}(\mathbf{X}) &= -\sum_{v \in \mathcal{T}'} Q(v) \sum_{a \in A} Q(a|v) \log Q(a|v) \\
&\geq -\sum_{v \in \mathcal{T}'} Q(v) \sum_{a \in A} \sum_{w \in T} R(w|v) Q(a|w) \log Q(a|w) \\
&\geq -\sum_{w \in \mathcal{T}} \sum_{v \in T'} Q(v) R(w|v) \sum_{a \in A} Q(a|w) \log Q(a|w) \\
&\geq -\sum_{w \in \mathcal{T}} Q(w) \sum_{a \in A} Q(a|w) \log Q(a|w) = \mathbf{H}_{\mathcal{T}}(\mathbf{X})
\end{aligned}$$

($\Downarrow$) Now let $\mathcal{T}$ be such that $\mathbf{H}_{\mathcal{T}}(\mathbf{X})$ is minimum. Consider $\mathcal{T}' = A^k$ the complete context tree. It is clear that $\mathbf{H}_{\mathcal{T}'}(\mathbf{X})$ is minimum because $\mathbf{X}$ is $\mathcal{T}'$-adapted. Using the argument written above we have that $\mathbf{H}_{\mathcal{T}}(\mathbf{X}) \geq \mathbf{H}_{\mathcal{T}'}(\mathbf{X})$, but since $\mathbf{H}_{\mathcal{T}}(\mathbf{X})$ is minimum they must be equal. Recall that for all $w \in \mathcal{T}$ we have $Q(a|w) = \sum_{u \in \mathcal{T}'} Q(a|u)R(u|w)$. Since we have the equality in Jensen theorem and $f(x) = x \log x$ is strictly convex, it must be $Q(a|w) = Q(a|u)$ for all u such that $R(u|w) > 0$. This exactly means that

$$\mathbf{P}(X_{k+1} = a|X_1^k = u) = \mathbf{P}(X_{k+1} = a|X_1^k \in \mathcal{C}(w))$$

because $Q(u|w) > 0$ when $w = \sigma_{\mathcal{T}}(v)$ which is equivalent to the statement that $\mathbf{X}$ is $\mathcal{T}$-adapted. □

3.5 Consistency of MDL estimators

Given a realization $x_1^n \in A^n$ of a process of order k consider the counters,

$$N_n(w) = N_n(w|x_1^n) = \#\{i : k \leq i \leq n, x_{i-l(w)}^{i-1} \in C(w)\}.$$

Let $N_n(w,a)$ denote the number of occurrences of the pattern $w \in A_\phi^{l(w)}$ followed by the letter $a \in A$ in the sample x_1^n then

$$N_n(w,a) = N_n(w,a|x_1^n) = \#\{i : k \leq i \leq n, x_{i-l(w)}^{i-1} \in C(w), x_i = a\}.$$

For any symbol $a \in A$, the empirical transition probability is defined by

$$\hat{P}_n(a|w) = \hat{P}(a|w)(x^n) = \frac{N_n(w,a)}{N_n(w)}.$$

Therefore the ML probability of a sequence x_1^n is given by the product

$$P_{\mathrm{ML},\mathcal{T}}(x_1^n) = \prod_{w\in\mathcal{T}} \prod_{a\in A} \hat{P}_n(a|w)^{N_n(w,a)}.$$

We define the ML expected code length as

$$\mathrm{ML}_{\mathcal{T}}(x^n) = -\log P_{\mathrm{ML},\mathcal{T}}(x^n) = -\sum_{w\in\mathcal{T}} \sum_{a\in A} N_n(w,a) \log \hat{P}_n(a|w).$$

In a similar way we define the KT expected code length as

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = -\log P_{\mathrm{KT},\mathcal{T}}(x_1^n),$$

where $P_{\mathrm{KT},\mathcal{T}}(x_1^n)$ is the KT probability of x_1^n corresponding to $\mathcal{T}$ given by

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^K} \prod_{w\in\mathcal{T}} P_{\mathrm{KT},w}(x_1^n), \text{ where}$$

$$P_{\mathrm{KT},w}(x_1^n) = \prod_{a\in A} \frac{(N_n(w,a)-\frac{1}{2})(N_n(w,a)-\frac{3}{2})\ldots\frac{1}{2}}{(N_n(w)-1+\frac{|A|}{2})(N_n(w)-2+\frac{|A|}{2})\ldots\frac{|A|}{2}}, \tag{3.2}$$

is the KT-factor corresponding to w.

3.5.1 Preliminary lemmas

We will need some results on typicality and a bound of the difference between KT and ML probabilities based on Stirling's approximation of the Gamma function.

Lemma 3.5. *Given a process Q, to any $\delta > 0$ there exists $\kappa > 0$ such that, eventually almost surely as $n \to \infty$*

$$\left|\frac{N_n(w,a)}{N_n(w)} - Q(a|w)\right| \leq \sqrt{\frac{\delta \log n}{N_n(w)}}$$

simultaneously for all patterns w with $l(s) \leq \kappa \log n$ and $N_n(w) \geq 1$ which have a ϕ-suffix in the sparse context tree of Q.

Proof. By a generalization of Theorem 2.11 for the case of sparse tree models, we have that for $\xi > (\log |A|)/2$ there exist $\eta > 0$ and $c > 0$ such that, eventually almost surely,

$$\left| \frac{N_n(w,a)}{N_n(w)} - Q(a|w) \right| \leq \sqrt{\frac{\max\{\xi l(s), \eta \log\log n\}}{N_n(w)}} \tag{3.3}$$

simultaneously for all patterns w with $N_n(w) \geq cl(w)$ which have a ϕ-suffix in the sparse context tree of Q. While Theorem 2.11 is stated for Tree models only, the proof relies upon the martingale property of the sequence Z_n defined in equation (2.23), and $Z_n = N_n(w,a)Q(a|w)N_n(w)$ defines a martingale whenever w has a ϕ-suffix in the sparse context tree of the process Q. Thus, the mentioned proof applies literally.

Then the choice $\kappa = \delta / \max\{\xi, c/4\}$ is suitable for this Lemma. Indeed, if $N_n(w) \geq cl(w)$, the bound (3.3) holds and gives the assertion, because

$$\begin{aligned} \eta \log\log N_n(w) \leq \eta \log\log n \leq \delta \log n \quad &\text{for } n \geq n_0 \\ \xi l(s) \leq \xi\kappa \log n \leq \delta \log n \end{aligned}$$

While in the opposite case $N_n(w) < cl(w) \leq c\kappa \log n$ we have

$$\sqrt{(\delta \log n)/N_n(s)} \geq \sqrt{\delta/(c\kappa)} \geq 2$$

and the assertion holds trivially. □

Remark. We only need the result for fixed w, a so a simpler proof could be obtained using the Law of the iterated logarithm (see Neveu 1975) for the martingale Z_n. However we prefer this proof since all the difficulties have already been introduced for the VLMC case.

Lemma 3.6. *There is a constant C depending only on the alphabet size $|A|$ such that for every $n \geq 1$ and $x_1^n \in A^n$,*

$$\left| \log P_{\mathrm{KT}}(x_1^n) - \sum_{a \in A} N_n(a) \log \frac{N_n(a)}{n} + \frac{|A|-1}{2} \log n \right| \leq C \tag{3.4}$$

and, for every $\mathcal{T}$ sparse tree it holds

$$\left| \log P_{\mathrm{KT},\mathcal{T}}(x_1^n) - \log P_{\mathrm{ML},\mathcal{T}}(x_1^n) + \frac{|A|-1}{2} \sum_{w \in \mathcal{T}} \log N_n(w) \right| \leq C' \tag{3.5}$$

Proof. Recall the formula for the KT probability

$$P_{\mathrm{KT}}(x_1^n) = \prod_{a\in A} \frac{(N_n(a)-\frac{1}{2})(N_n(a)-\frac{3}{2})\dots\frac{1}{2}}{(n-1+\frac{|A|}{2})(n-2+\frac{|A|}{2})\dots\frac{|A|}{2}}, \tag{3.6}$$

using the Γ function this can be written in the form

$$P_{\mathrm{KT}}(x_1^n) = \frac{\Gamma(\frac{|A|}{2})}{\Gamma(n+\frac{|A|}{2})} \prod_{a\in A} \frac{\Gamma(N_n(a)+\frac{1}{2})}{\Gamma(\frac{1}{2})},$$

taking logarithms we obtain

$$\log P_{\mathrm{KT}}(x_1^n) = \log\Gamma(\tfrac{|A|}{2}) - \log\Gamma(n+\tfrac{|A|}{2}) + \sum_{a\in A} \log\Gamma(N_n(a)+\tfrac{1}{2}) - \log\Gamma(\tfrac{1}{2}).$$

Then the following bound holds

$$\left| \log P_{\mathrm{KT}}(x_1^n) + \log\Gamma(n+\tfrac{|A|}{2}) - \sum_{a\in A} \log\Gamma(N_n(a)+\tfrac{1}{2}) \right| \le K_1 \tag{3.7}$$

where $K_1 = \log\Gamma(\frac{|A|}{2}) + |A|\log\Gamma(\frac{1}{2})$.

Define f the following auxiliary function by the expression

$$\begin{aligned}
f(x_1^n) =& \sum_{a\in A} \log\Gamma(N_n(a)+\tfrac{1}{2}) - \sum_{a\in A} N_n(a)\log N_n(a) \\
& - \log\Gamma(n+\tfrac{|A|}{2}) + \left(\sum_{a\in A} N_n(a)\right)\log n + \tfrac{|A|-1}{2}\log n \\
=& \sum_{a\in A} \log\Gamma(N_n(a)+\tfrac{1}{2}) - N_n(a)\log N_n(a) \\
& - \log\Gamma(n+\tfrac{|A|}{2}) + n\log n + \tfrac{|A|-1}{2}\log n \\
& + \left(n+\tfrac{|A|}{2} - \sum_{a\in A}(N_n(a)+\tfrac{1}{2})\right)\log e \\
=& \sum_{a\in A} \log\Gamma(N_n(a)+\tfrac{1}{2}) - N_n(a)\log N_n(a) - (N_n(a)+\tfrac{1}{2})\log e \\
& - \log\Gamma(n+\tfrac{|A|}{2}) + (n+\tfrac{|A|-1}{2})\log n + (n+\tfrac{|A|}{2})\log e
\end{aligned}$$

using Stirling's formula for Γ-functions

$$|\log\Gamma(z) - (z-\tfrac{1}{2})\log z - z\log e| \le M, \qquad z \ge \tfrac{1}{2},$$

we have that

$$\Big| \log\Gamma(n+\tfrac{|A|}{2}) - (n+\tfrac{|A|-1}{2})\log(n+\tfrac{|A|}{2}) - (n+\tfrac{|A|}{2})\log e \Big| \le M$$

$$\Big| \log\Gamma(N_n(a)+\tfrac{1}{2}) - N_n(a)\log(N_n(a)+\tfrac{1}{2}) - (N_n(a)+\tfrac{1}{2})\log e \Big| \le M$$

applying the mean value theorem to the log function we obtain that

$$\log(a+b) - \log a \le b\,\tfrac{1}{a},$$

then in particular

$$\begin{array}{ll} \log(n+\tfrac{|A|}{2}) - \log n \le \tfrac{|A|}{2}\tfrac{1}{n} & \text{which implies} \\ (n+\tfrac{|A|-1}{2})\log(n+\tfrac{|A|}{2}) - (n+\tfrac{|A|-1}{2})\log n \le (n+\tfrac{|A|-1}{2})\tfrac{|A|}{2}\tfrac{1}{n} \le L & \end{array}$$

and

$$\begin{array}{ll} \log(N_n(a)+\tfrac{1}{2}) - \log N_n(a) \le \tfrac{1}{2}\tfrac{1}{N_n(a)} & \text{which implies} \\ N_n(a)\log(N_n(a)+\tfrac{1}{2}) - N_n(a)\log N_n(a) \le N_n(a)\tfrac{1}{2}\tfrac{1}{N_n(a)} \le \tfrac{1}{2} & \end{array}$$

then the following bound holds

$$\begin{aligned} \Bigg| & -\log\Gamma(n+\tfrac{|A|}{2}) + \sum_{a\in A}\log\Gamma(N_n(a)+\tfrac{1}{2}) \\ & \qquad - \sum_{a\in A} N_n(a)\log\frac{N_n(a)}{n} + \frac{|A|-1}{2}\log n \Bigg| \le K_2 \end{aligned} \tag{3.8}$$

where $K_2 = |A|(M+\frac{1}{2}) + M + L$.

By combining bounds (3.7) and (3.8) we obtain the desired result,

$$\Bigg| \log P_{\mathrm{KT}}(x_1^n) - \sum_{a\in A} N_n(a)\log\frac{N_n(a)}{n} + \frac{|A|-1}{2}\log n \Bigg| \le K_1 + K_2.$$

For the second inequality we shall see that the factors in the definition (3.2) of $P_{\mathrm{KT},\mathcal{T}}$ are of the form of the definition (3.6) of P_{KT}, with $N_n(w)$ and $N_n(w,a)$ in the role of n and $N_n(a)$, respectively. Applying (3.4) to each of these factors we obtain

$$\Bigg| \log P_{\mathrm{KT},w}(x_1^n) - \sum_{a\in A} N_n(w,a)\log\frac{N_n(w,a)}{N_n(a)} + \frac{|A|-1}{2}\log N_n(w) \Bigg| \le C$$

then we can write

$$\sum_{w\in\mathcal{T}}\left|\log P_{\mathrm{KT},w}(x_1^n)-\sum_{a\in A}N_n(w,a)\log\frac{N_n(w,a)}{N_n(a)}+\frac{|A|-1}{2}\log N_n(w)\right|\le|\mathcal{T}|C$$

recalling that

$$\log P_{\mathrm{ML},\mathcal{T}}=\sum_{w\in\mathcal{T}}\sum_{a\in A}N_n(w,a)\log\frac{N_n(w,a)}{N_n(a)},$$

and

$$\log P_{\mathrm{KT},\mathcal{T}}=\sum_{w\in\mathcal{T}}\sum_{a\in A}N_n(w,a)\log\frac{N_n(w,a)}{N_n(a)},$$

we obtain that

$$\left|\log P_{\mathrm{KT},\mathcal{T}}(x_1^n)-\log P_{\mathrm{ML},\mathcal{T}}(x_1^n)+\frac{|A|-1}{2}\sum_{w\in\mathcal{T}}\log N_n(w)\right|\le|\mathcal{T}|C+K_3$$

where K_3 is a constant bigger than $|\mathcal{T}|C$ in order to take care of the $k\log|A|$ term. □

3.5.2 Proof of the theorem

Theorem 3.7. *Let $\{X_n\}$ be a process with with minimal sparse tree $\mathcal{T}_0\in\mathcal{M}_k$ then the* MDL *like estimator based on* KT *probability assignment*

$$\widehat{\mathcal{T}}_n=\widehat{\mathcal{T}}(x_1^n)=\arg\min_{\mathcal{T}\in\mathcal{M}_k}\mathrm{KT}_{\mathcal{T}}(x_1^n)$$

satisfies $\widehat{\mathcal{T}}_n=\mathcal{T}_0$ eventually almost surely.

Proof. Since $\mathcal{M}_k$ is finite it suffices to prove that for any single tree we have that eventually almost surely

$$\mathrm{KT}_{\mathcal{T}_0}(x_1^n)<\mathrm{KT}_{\mathcal{T}}(x_1^n).$$

The proof separates in two cases, the overestimation and underestimation events, whether $\{X_n\}$ is $\mathcal{T}$-adapted or not. We shall use the bound from Lemma 3.6 a couple of times:

$$\left|\mathrm{KT}_{\mathcal{T}}(x_1^n)-\mathrm{ML}_{\mathcal{T}}(x_1^n)-\frac{|A|-1}{2}\sum_{w\in\mathcal{T}}\log N_n(w)\right|<C. \tag{3.9}$$

Underestimation: If $\{X_n\}$ is not $\mathcal{T}$-adapted then by Proposition 3.4 we have that $\mathbf{H}_{\mathcal{T}}(\mathbf{X})$ is strictly bigger that $\mathbf{H}_{\mathcal{T}_0}(\mathbf{X})$. We shall prove that $\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{KT}_{\mathcal{T}_0}(x_1^n) > 0$. We can write

$$\begin{aligned}\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{KT}_{\mathcal{T}_0}(x_1^n) &= \mathrm{KT}_{\mathcal{T}}(x_1^n) - n\mathbf{H}_{\mathcal{T}}(\mathbf{X}) &(3.10)\\ &\quad + n\mathbf{H}_{\mathcal{T}}(\mathbf{X}) - n\mathbf{H}_{\mathcal{T}_0}(\mathbf{X}) &(3.11)\\ &\quad + n\mathbf{H}_{\mathcal{T}_0}(\mathbf{X}) - \mathrm{KT}_{\mathcal{T}_0}(x_1^n). &(3.12)\end{aligned}$$

When we divide everything by n we see that the middle term (3.11) is equal to $\mathbf{H}_{\mathcal{T}}(\mathbf{X}) - \mathbf{H}_{\mathcal{T}_0}(\mathbf{X})$ which is strictly positive, and the external terms (3.10) and (3.12) converge to zero almost surely. Effectively, our claim holds from

$$\begin{aligned}\left|\frac{1}{n}\,\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathbf{H}_{\mathcal{T}}(\mathbf{X})\right| &\le \frac{1}{n}\left|\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n)\right| &(3.13)\\ &\quad + \left|\frac{1}{n}\,\mathrm{ML}_{\mathcal{T}}(x_1^n) - \mathbf{H}_{\mathcal{T}}(\mathbf{X})\right|, &(3.14)\end{aligned}$$

the convergence to zero of (3.13) follows from (3.9), because

$$\begin{aligned}\left|\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n)\right| &\le C + \frac{|A|-1}{2}\sum_{w\in\mathcal{T}}\log N_n(w)\\ \frac{1}{n}\left|\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n)\right| &\le \frac{C}{n} + \frac{|A|-1}{2n}\sum_{w\in\mathcal{T}}\log N_n(w)\\ &\le \frac{C}{n} + \frac{(|A|-1)|\mathcal{T}|\log n}{2n} \to 0\end{aligned}$$

while the term (3.14) converges to zero since

$$\hat{P}_n(a|w) \to Q(a|w) \qquad \text{and} \qquad \hat{P}_n(w) \to Q(w)$$

almost surely by the ergodic theorem and

$$\begin{aligned}\frac{1}{n}\,\mathrm{ML}_{\mathcal{T}}(x_1^n) &= -\frac{1}{n}\log P_{\mathrm{ML},\mathcal{T}}(x_1^n)\\ &= -\frac{1}{n}\sum_{w\in\mathcal{T}}\sum_{a\in A} N_n(w,a)\log\frac{N_n(w,a)}{N_n(a)}\\ &= -\sum_{w\in\mathcal{T}}\frac{N_n(w)}{n}\sum_{a\in A}\frac{N_n(w,a)}{N_n(w)}\log\frac{N_n(w,a)}{N_n(a)}\\ \mathbf{H}_{\mathcal{T}}(\mathbf{X}) &= \sum_{w\in\mathcal{T}} Q(w)\sum_{a\in A} Q(a|w)\log Q(a|w)\end{aligned}$$

the convergence to zero follows since $\mathcal{T}$ is finite and $x \log x$ is continuous. For the case of $\mathcal{T}_0$ the proof is the same. Then $\frac{1}{n}\big(\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{KT}_{\mathcal{T}_0}(x_1^n)\big) > 0$ eventually almost surely.

Overestimation: If $\{X_n\}$ is $\mathcal{T}$-adapted we have that $\mathbf{H}_{\mathcal{T}}(\mathbf{X}) = \mathbf{H}_{\mathcal{T}_0}(\mathbf{X})$ so the argument above fails. In order to prove that $\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{KT}_{\mathcal{T}_0}(x_1^n) > 0$ we now write

$$\begin{aligned}\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{KT}_{\mathcal{T}_0}(x_1^n) &= \mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n)\\ &\quad - (-\mathrm{ML}_{\mathcal{T}}(x_1^n) - \log Q(x_1^n))\\ &\quad - (\log Q(x_1^n) + \mathrm{KT}_{\mathcal{T}_0}(x_1^n)),\end{aligned}$$

where $Q(x_1^n)$ is the probability of the observed sequence. In this case we will show that for any $\varepsilon > 0$ eventually almost surely we have

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n) > \frac{|A|-1}{2}|\mathcal{T}|(1-\varepsilon)\log n - C, \tag{3.15}$$

$$-\mathrm{ML}_{\mathcal{T}}(x_1^n) - \log Q(x_1^n) < |\mathcal{T}|\varepsilon \log n, \tag{3.16}$$

$$\log Q(x_1^n) + \mathrm{KT}_{\mathcal{T}_0}(x_1^n) < \frac{|A|-1}{2}|\mathcal{T}_0|\log n + C. \tag{3.17}$$

Since $|\mathcal{T}| > |\mathcal{T}_0|$ by choosing ε small enough we have that

$$\frac{|A|-1}{2}\big[\,|\mathcal{T}|(1-\varepsilon) - |\mathcal{T}_0|\,\big] - |\mathcal{T}|\varepsilon > 0,$$

which completes the proof.

From (3.9) we have that

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n) - \frac{|A|-1}{2}\sum_{w\in\mathcal{T}}\log N_n(w) > -C,$$

then inequality (3.15) follows using the fact that $N_n(w) = \sum_{u\in\mathcal{C}(w)} N_n(u) > n^{1-\varepsilon}$ eventually almost surely

$$\begin{aligned}\mathrm{KT}_{\mathcal{T}}(x_1^n) - \mathrm{ML}_{\mathcal{T}}(x_1^n) &> \frac{|A|-1}{2}\sum_{w\in\mathcal{T}}\log N_n(w) - C\\ &> \frac{|A|-1}{2}\sum_{w\in\mathcal{T}}(1-\varepsilon)\log n - C\\ &> \frac{|A|-1}{2}|\mathcal{T}|(1-\varepsilon)\log n - C.\end{aligned}$$

We can decompose the probability of the sequence x_1^n in

$$Q(x_1^n) = Q(x_1^k) \prod_{w \in \mathcal{T}} \prod_{a \in A} Q(a|w)^{N_n(w,a)}$$

then we can write the left hand side of (3.16) in the following way

$$\begin{aligned}
-\operatorname{ML}_{\mathcal{T}}(x_1^n) - \log Q(x_1^n) &= \log P_{\mathrm{ML},\mathcal{T}}(x_1^n) - \log Q(x_1^n) \\
&= -\log Q(x_1^k) + \sum_{w \in \mathcal{T}} \sum_{a \in A} N_n(w,a) \log \frac{N_n(a|w)/N_n(w)}{Q(a|w)} \\
&= -\log Q(x_1^k) + \sum_{w \in \mathcal{T}} N_n(w) \sum_{a \in A} \frac{N_n(w,a)}{N_n(w)} \log \frac{N_n(a|w)/N_n(w)}{Q(a|w)} \\
&= -\log Q(x_1^k) + \sum_{w \in \mathcal{T}} N_n(w) \mathbf{D}(\hat{P}_n(\cdot|w) \,\|\, Q(\cdot|w))
\end{aligned}$$

Now combining Lemmas 2.6 and 3.5 we can write eventually almost surely

$$\begin{aligned}
\mathbf{D}(\hat{P}_n(\cdot|w) \,\|\, Q(\cdot|w)) &\le \sum_{a \in A} \frac{(P(a|w) - Q(a|w))^2}{Q(a|w)} \\
&\le \frac{1}{\rho} \sum_{a \in A} (P(a|w) - Q(a|w))^2 \\
&\le \frac{1}{\rho} \sum_{a \in A} \frac{\delta \log n}{N_n(w)} \le \frac{1}{\rho} |A| \frac{\delta \log n}{N_n(w)},
\end{aligned}$$

where ρ denotes the smallest element of the transition probabilities $Q(a|w)$, with $a \in A$ and $w \in \mathcal{T}$. From here we can prove the bound

$$\sum_{w \in \mathcal{T}} N_n(w) \mathbf{D}(\hat{P}_n(\cdot|w) \,\|\, Q(\cdot|w)) < |\mathcal{T}| \frac{1}{\rho} |A| \delta \log n < \varepsilon |\mathcal{T}| \log n$$

by choosing δ small enough, and this implies inequality (3.16).

Now for the last inequality (3.17) we use

$$\log Q(x_1^n) + \operatorname{KT}_{\mathcal{T}_0}(x_1^n) \le \log P_{\mathrm{ML},\mathcal{T}_0}(x_1^n) - \log P_{\mathrm{KT},\mathcal{T}_0}(x_1^n),$$

which holds because $Q(x_1^n) \leq P_{\mathrm{ML},\mathcal{T}_0}(x_1^n)$. From (3.9) we have that

$$\begin{aligned} P_{\mathrm{ML},\mathcal{T}_0}(x_1^n) - \log P_{\mathrm{KT},\mathcal{T}_0}(x_1^n) - \frac{|A|-1}{2} \sum_{w \in \mathcal{T}_0} \log N_n(w) &< C \\ P_{\mathrm{ML},\mathcal{T}_0}(x_1^n) - \log P_{\mathrm{KT},\mathcal{T}_0}(x_1^n) &< \frac{|A|-1}{2} \sum_{w \in \mathcal{T}_0} \log N_n(w) + C \\ &< \frac{|A|-1}{2} \sum_{w \in \mathcal{T}_0} \log n + C \\ &< \frac{|A|-1}{2} |\mathcal{T}_0| \log n + C \end{aligned}$$

using the fact that $N_n(w) \leq n$. □

Remark. Strictly speaking, the MDL principle would require to minimize the code length $L(\mathcal{T}, x^n) = \mathrm{KT}_{\mathcal{T}}(x^n) + L(\mathcal{T})$, where $L(\mathcal{T})$ is the cost of $\mathcal{T}$, i.e. the length of a code describing $\mathcal{T}$. This additional term is omitted, since this does not affect the consistency result.

3.5.3 Computation of the estimator

Definition 3.5. Given a pattern w with $0 \leq l(w) \leq k$, we assign recursively, starting from the patterns of the full template tree $(A_\phi)^k$, the score value

$$S_w(x_1^n) = \begin{cases} \min\{S_{\phi w}(x_1^n), \sum_{a \in A} S_{aw}(x_1^n)\}, & \text{if } |w| < k \\ -\log P_{\mathrm{KT},w}(x_1^n), & \text{if } |w| = k \end{cases}$$

Given a pattern w we define $\mathcal{M}(w)$ the set of sparse subtrees of w, that is $\mathcal{M}(w) = \{\mathcal{T} \in \mathcal{M}_k : \forall u \in \mathcal{T} \quad w \prec_\phi u\}$.

Theorem 3.8. *For any w with $0 \leq l(w) \leq k$ it holds that*

$$S_w(x_1^n) = \min_{\mathcal{T} \in \mathcal{M}(w)} \sum_{u \in \mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n)$$

Proof. If $l(w) = k$ the proof is obvious since the only possible subtree of w is $\mathcal{T} = \{w\}$, therefore,

$$S_w(x_1^n) = -\log P_{\mathrm{KT},w}(x_1^n) = \min_{\mathcal{T} \in \mathcal{M}(w)} \sum_{u \in \mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n).$$

The proof goes by induction on the length of the pattern w. Suppose the assertion holds for all patterns of length m, and let w with $l(w) = m - 1$. Define

$$\begin{aligned}\mathcal{M}_A(w) &= \{\mathcal{T} \in \mathcal{M}(w) : w \text{ has } |A| \text{ children in } \mathcal{T}\},\\ \mathcal{M}_\phi(w) &= \{\mathcal{T} \in \mathcal{M}(w) : w \text{ has a } \phi \text{ child in } \mathcal{T}\}.\end{aligned}$$

It holds that $\mathcal{M}(w) = \mathcal{M}_\mathcal{A}(w) \cup \mathcal{M}_\phi(w)$. It follows from the inductive hypothesis that:

$$\begin{aligned}\sum_{a\in A} S_{aw}(x_1^n) &= \sum_{a\in A}\left(\min_{\mathcal{T}_a\in\mathcal{M}(aw)} \sum_{u\in\mathcal{T}_a} -\log P_{\mathrm{KT},u}(x_1^n)\right)\\ &= \min_{\mathcal{T}\in\mathcal{M}_\mathcal{A}(w)} \sum_{u\in\mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n), \qquad (3.18)\end{aligned}$$

where the second equality is a consequence of considering $\mathcal{T} = \cup_{a\in A}\mathcal{T}_a$.

$$\begin{aligned}S_{\phi w}(x_1^n) &= \min_{\mathcal{T}_\phi\in\mathcal{M}(\phi w)} \sum_{u\in\mathcal{T}_\phi} -\log P_{\mathrm{KT},u}(x_1^n)\\ &= \min_{\mathcal{T}\in\mathcal{M}_\phi(w)} \sum_{u\in\mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n). \qquad (3.19)\end{aligned}$$

Then, combining (3.18) and (3.19) with the fact that $\mathcal{M}(w) = \mathcal{M}_\mathcal{A}(w) \cup \mathcal{M}_\phi(w)$ we have that

$$\begin{aligned}S_w(x_1^n) &= \min\left\{S_{\phi w}(x_1^n), \sum_{a\in A} S_{aw}(x_1^n)\right\}\\ &= \min_{\mathcal{T}\in\mathcal{M}(w)} \sum_{u\in\mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n).\end{aligned}$$

□

Remark. In particular the result holds for the empty pattern λ, and since $\mathcal{M}(\lambda) = \mathcal{M}$ we have that

$$\begin{aligned}S_\lambda(x_1^n) &= \min_{\mathcal{T}\in\mathcal{M}} \sum_{u\in\mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n)\\ &= \min_{\mathcal{T}\in\mathcal{M}} \sum_{u\in\mathcal{T}} -\log P_{\mathrm{KT},u}(x_1^n) - K\log|A|\\ &= \min_{\mathcal{T}\in\mathcal{M}} -\log P_{\mathrm{KT},\mathcal{T}}(x_1^n).\end{aligned}$$

Therefore the MDL model can be obtained by tracking the recursive choices in order to obtain the optimal tree. This describes an algorithm to determine the optimal tree model. For this we consider the complete tree of depth k over A_ϕ. We recursively calculate the score $S_w(x_1^n)$ for each node of this tree starting from the leaves, and at each node we prune its children in order to obtain its optimal subtree. Once at the root, the tree obtained is $\widehat{\mathcal{T}}(x_1^n)$.

CHAPTER 4

Applications

4.1 Protein classification

The primary structure of a protein is represented by a sequence of 20 different symbols called aminoacids. Proteins can be composed of one or more functional regions, called domains; the identification of domains that occur within proteins can provide insights into their function. For this reason biologists classify protein domains into families and care about the reliability of the classification (Stein, 2001). On the other hand, it is also important the coverage of all the proteins encoded by a genome, called proteome coverage.

The Pfam database is a large collection of protein domain families (Finn et al., 2006). In its last release of July 2007, the Pfam database comprises 9318 families (Pfam-A), covering 73.23% of all proteins in Pfamseq, a database based on UniProt 9.7 (Wu et al., 2006). Another 13% are covered by Pfam-B families, an un-annotated and of lower quality set of families, generated automatically from another database called ProDom (release 2005.1, Servant et al. 2002).

A protein sequence can be thought as a realization of a discrete time stochastic process having as state space the set A of 20 aminoacids. This is the basic idea in the modeling of protein domains by HMMs or VLMCs.

Proteins consist of sequences of amino acids codified with the alphabet

$$\mathcal{A} = \{G, A, V, L, I, P, M, F, Y, W, S, T, C, N, Q, R, K, H, D, E\}.$$

Each Pfam-A family consist of two parts: a manually curated set of protein domains called *seed* and a set of automatically detected protein domains using

G	Gly	Glycine
A	Ala	Alanine
V	Val	Valine
L	Leu	Leucine
I	Ile	Isoleucine
P	Pro	Proline
M	Met	Methionine
F	Phe	Phenylalanine
Y	Tyr	Tyrosine
W	Trp	Tryptophan
S	Ser	Serine
T	Thr	Threonine
C	Cys	Cysteine
N	Asn	Asparagine
Q	Gln	Glutamine
R	Arg	Arginine
K	Lys	Lysine
H	His	Histidine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid

Figure 4.1: The alphabet of Aminoacids.

a *profile hidden Markov model* (profile HMM), whose parameters are estimated from the seed of the family.

VLMC have been successfully applied to model and classify protein sequences (Bejerano and Yona, 2001). As in the case of profile HMM in the construction of the Pfam families, the VLMC approach of Bejerano and Yona takes, for each family, a set of already classified protein domains and estimates a VLMC model, i.e. a pair (t, p). Then, the estimated VLMC model is used to classify other protein sequences into the family.

The context trees of sequences of a given family are assumed to be random samples of a distribution associated to the family; this distribution is used as a *signature* of the family (Galves et al. 2004; Leonardi et al. 2008). For each sequence they construct the estimated context tree using the PST algorithm introduced by Ron, Singer, and Tishby (1996) and implemented by Bejerano (2004). Then they proceed to classify the proteins into families.

As mentioned in Devroye, Györfi, and Lugosi (1996), "pattern recognition or classification is about guessing or predicting the unknown nature of an observation, a discrete quantity such as black or white, one or zero, sick or healthy, real or fake. An observation is a collection of numerical measurements such as an image (which is a sequence of bits, one per pixel), a vector of weather data, an electrocardiogram, or a signature on a check suitably digitized.” In the problem under consideration the observations will be the VLMC context trees.

Given a finite set $\{1,\dots,m\}$ and an arbitrary space E, an observation is a pair $(x,y)\in E\times\{1,\dots,m\}$, where x is known and y is called a class or label that denotes the unknown nature of the observation. A mapping $g:E\to\{1,\dots,m\}$ is called a classifier and represents our guess of the class y given it’s associated vector $x\in E$. The classification is wrong if given an observation (x,y), $g(x)\neq y$.

Let $(X,Y)\in E\times\{1,\dots,m\}$ be a random pair. Since an error occurs if $g(X)\neq Y$, the probability of misclassification for g is

$$L(g)=P[g(X)\neq Y]. \tag{4.1}$$

Then the best possible classifier is the function g^* that minimizes $L(g)$ (4.1). The minimal minimum error probability (the Bayes error) is denoted by $L^*=L(g^*)$.

In order to obtain g^*, the distribution of (X,Y) should be known, but this does not happen often. One must build up a classifier based on a training sample of independent pairs $\{(X_i,Y_i);1\le i\le n\}$, with the same distribution as the pair (X,Y) and known $Y_1,\dots,Y_n$ values. Then the classifier based on the training sample $\{(X_i,Y_i):1\le i\le n\}$ is a function

$$g_n(\,\cdot\,;X_1,Y_1,\dots,X_n,Y_n):E\times(E\times\{1,\dots,m\})^n\to\{1,\dots,m\}.$$

The performance of g_n is measured by the conditional error probability

$$L_n(g_n)=P\left[g_n(X;X_1,Y_1,\dots,X_n,Y_n)\neq Y|X_1,Y_1,\dots,X_n,Y_n\right].$$

An individual mapping

$$g_n:E\times(E\times\{1,\dots,m\})^n\to\{1,...,m\}$$

is called a classifier and a sequence of classifiers $\{g_n;n\ge 1\}$ is called a rule. A rule is consistent when $\lim_n L_n(g_n)=L^*$. Asymptotic results about consistency of discrimination rules can be found in Devroye, Györfi, and Lugosi (1996).

The most simple and used classification rules are the k-nearest neighbor rules which only depends on the distances between individuals, and can be applied in general metric spaces. Formally, if $m=2$, we define the k–NN rule by

$$g_n(x)=\begin{cases}1 & \text{if } \displaystyle\sum_{i=1}^n w_{ni}I_{\{Y_i=1\}}>\sum_{i=1}^n w_{ni}I_{\{Y_i=0\}}\\ 0 & \text{otherwise,}\end{cases}$$

where

$$w_{ni} = \begin{cases} 1 & \text{if } X_i \text{ is among the } k \text{ nearest neighbors of } x \\ 0 & \text{otherwise.} \end{cases}$$

X_i is said to be the k-th nearest neighbor of x when the distance $||X_i - x||$ is the k-th smallest among $||X_1 - x||, \dots, ||X_n - x||$.

Devroye, Györfi, and Lugosi (1996) proved, under mild conditions, universal consistency for the case of non random nearest neighbor k, if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ (see, for instance,theorem 6.4, pp. 101). They also considered the case of automatic nearest neighbor rules, where the parameter K is random and depends on the data sequence $(X_1, Y_1), \dots, (X_n, Y_n)$. In this case, the same result is true if the sequence K_n satisfies that $K_n \to \infty$ and $K_n/n \to 0$ with probability one, as $n \to \infty$ (theorem 26.1, pp. 451)

There are several approaches to estimate context trees and transition probabilities of VLMC's. The context algorithm of Rissanen (1983a) and some variations proposed by Ron, Singer, and Tishby (1996), Buhlmann and Wyner (1998) and Galves, Maume-Deschamps, and Schmitt (2008). Csiszár and Talata (2006) proposed the use of the Bayesian Information Criterion (BIC) and the Minimum Description Length Principle (MDL). These algorithms converge to the true parameters when the sample size increases.

We describe here the PST *algorithm* of Ron, Singer, and Tishby (1996), see also Bejerano (2004), which have been used to classify the protein families.

Suppose $x_1, \dots, x_l$ is a sample of a VLMC over A specified by the pair (t, p) (in our setting $x_1, \dots, x_l$ represents a protein over the alphabet of 20 aminoacids). For any sequence $a_1^j \in A^j$ define the counters

$$N(a_1^j) = \sum_{i=0}^{l-j} \mathbf{1}\{x_{i+1}^{i+j} = a_1^j\},$$

where the function $\mathbf{1}$ takes value 1 if $x_{i+1}^{i+j} = a_1^j$ and 0 otherwise. For any sequence $a_1^j \in A^j$ such that $N(a_1^j) \geq 1$ and any symbol $a \in A$ we define the empirical transition probabilities $\hat{p}(a|a_1^j)$ as

$$\hat{p}(a|a_1^j) = \frac{N(a_1^j a)}{\sum_{b \in A} N(a_1^j b)}, \tag{4.2}$$

where $a_1^j a$ denotes the sequence a_1^j concatenated with the symbol a.

To estimate the context tree associated to the sequence two parameters are fixed: L, the maximal depth of the estimated tree $\hat{t}$ and $r > 1$, a threshold value.

The PST algorithm defines the *context tree estimator* $\hat{t}$ as the tree containing all the sequences a_1^j, with $j \leq L$, $N(a_1^j) \geq 1$, such that there exists a symbol $a \in A$ satisfying

$$|\log \hat{p}(a|a_2^j) - \log \hat{p}(a|a_1^j)| \geq \log r\,. \tag{4.3}$$

That is, the node a_1^j is a node of $\hat{t}$ if the conditional probabilities $\hat{p}(\cdot|a_1^j)$ and $\hat{p}(\cdot|a_2^j)$ are sufficiently far in the sense of (4.3). To guarantee that $\hat{t}$ is a tree, include also all suffixes of included nodes; that is, $a_1^j \in \hat{t}$ implies $a_2^j \in \hat{t}$.

The PST algorithm uses other parameters to *smooth* the estimated transition probabilities given by (4.2). This smoothing is useful to avoid null estimated probabilities that can damage the prediction step in classification of new sequences. We refer the interested reader to Ron, Singer, and Tishby (1996), Bejerano (2003) and Bejerano (2004) for a full explanation of the PST algorithm, its implementation and some basic examples.

Leonardi (2006) considers a functionality family as a random variable assuming values in the space of context trees, conjecturing that different families induce different random variables. The context tree of a protein is then seen as a realization of the random variable corresponding to its family. This approach permits the use of traditional distance based pattern recognition methods to classify proteins, make them emerge from the bulk as families and check the validity of known families.

An alternative way to deal with this problem is to model the proteins with sparse Markov models and apply the Minimum Description Length Principle (MDL) we propose to associate to each protein a sparse tree. Then perform the classification rule on the space of sparse trees.

4.2 Universal source coding

An arithmetic code for a distribution Q_n on A^n has the length function given by $\lceil -\log Q_n(x_1^n) \rceil$ and produces expected length within 1 bit of the entropy lower bound $\mathbf{H}(Q_n)$; it therefore provides an almost optimal method for coding if it is known that the data x_1^n is governed by Q_n. In practice, however, the distribution governing the data is usually not known, though it may be reasonable to assume that the data are coming from an unknown member of a known class $\mathcal{P}$ of processes, such as the i.i.d. or Markov or stationary processes. Then it is desirable to use "universal" codes that perform well no matter which member of is the true process.

Arithmetic coding

Arithmetic coding is a procedure introduced in Rissanen (1976), and developed further in Pasco (1976) and Rissanen and Langdon (1981). The objective of the arithmetic coding algorithm is to represent a sequence of random variables by a subinterval in $[0, 1)$.

Let $B = \{0, 1\}$ denote the binary alphabet. An arithmetic code is a map $C : A^* \to B^*$, defined as follows. Let Q_n, $n = 1, 2, \ldots$ be probability distributions on the sets A^n satisfying the consistency conditions

$$Q_n(x_1^n) = \sum_{a \in \mathcal{A}} Q_{n+1}(x_1^n a),$$

these are necessary and sufficient for the distributions Q_n to be the marginal distributions of a process.

For each n, partition the unit interval $[0, 1)$ into subintervals $J(x_1^n) = [\ell(x_1^n), r(x_1^n))$ of length $r(x_1^n) - \ell(x_1^n) = Q_n(x_1^n)$ in a nested manner, i.e., such that $\{J(x_1^n a) : a \in A\}$ is a partitioning of $J(x_1^n)$, for each $x_1^n \in A^n$.

If the endpoints of $J(x_1^n)$ have binary expansions

$$\ell(x_1^n) = 0.z_1 z_2 \ldots z_m 0 \ldots, \quad r(x_1^n) = 0.z_1 z_2 \ldots z_m 1 \ldots,$$

we define $C(x_1^n) = z_1^m$, so the length function satisfy $L_C(x_1^n) < \lceil -\log Q_n(x_1^n) \rceil$, and the mapping $C|_{A^n}$ is one-to-one (since the intervals $J(x_1^n)$ are disjoint). Moreover, C has the feature of that $C(x_1^n)$ is always a prefix of $C(x_1^{n+1})$.

In order to determine the codeword $C(x_1^n)$, the nested partitions above need not be actually computed, it suffices to find the interval $J(x_1^n)$. This can be done in steps, the i-th step is to partition the interval $J(x_1^{i-1})$ into $|A|$ subintervals of length proportional to the conditional probabilities

$$Q(a|x_1^{i-1}) = \frac{Q_i(x_1^{i-1} a)}{Q_{i-1}(x_1^{i-1})}, \quad a \in A.$$

Arithmetic codes reduce the lossless compression problem to one of finding the best probability assignment for the given data x_1^n, that which will provide the shortest ideal code length.

Redundancy

For many practical situations, however, the probability distribution underlying the data may be unknown. Instead, all we know is a class of distributions. In yet other cases, there is no probability distribution underlying the data, all we are given is an individual sequence of outcomes. Examples of such data sources include text and music. We can then ask the question: How well can we compress the sequence? If we do not put any restrictions on the class of algorithms, we get a meaningless answer - there always exists a function that compresses a particular sequence to one bit while leaving every other sequence uncompressed. This function is clearly “overfitted” to the data. However, if we compare our performance to that achievable by optimal codes with respect to i.i.d. or kth-order Markov processes, we obtain more interesting answers.

The ideal codelength of a message $x_1^n \in A^n$ coming from a process with distribution P is defined as $-\log P(x_1^n)$. For an arbitrary code $C : A^* \to B^*$, the difference of its length function from the “ideal” will be called the *redundancy function* $\mathcal{R} = \mathcal{R}_{P,C}$:

$$\mathcal{R}_{P,C}(x_1^n) = L_C(x_1^n) + \log P(x_1^n).$$

Moreover, given a class $\mathcal{P}$ of processes, we define the worst case expected and maximum redundancy

$$\overline{\mathcal{R}}_C = \sup_{P\in\mathcal{P}} \mathbf{E}_P(\mathcal{R}_{P,C}) \qquad \text{and} \qquad \mathcal{R}_C^* = \sup_{P\in\mathcal{P}} \max_{x^n\in A^n} \mathcal{R}_{P,C}(x_1^n).$$

We say that a code is *strongly universal* for a class $\mathcal{P}$ of processes if either

$$\frac{1}{n}\overline{\mathcal{R}}_C \to 0 \qquad \text{or} \qquad \frac{1}{n}\mathcal{R}_C^* \to 0.$$

For any class $\mathcal{P}$ of processes with alphabet A, the least possible value of $\overline{\mathcal{R}}_C$ or $\mathcal{R}_C^*$ for prefix codes $C : A^* \to B^*$ “almost” equals

$$\overline{\mathcal{R}}_n = \min_Q \sup_{P\in\mathcal{P}} \sum_{x_1^n\in A^n} P(x^n) \log \frac{P(x_1^n)}{Q(x_1^n)} \qquad \text{or} \qquad \mathcal{R}_n^* = \min_Q \sup_{P\in\mathcal{P}} \max_{x_1^n\in A^n} \log \frac{P(x_1^n)}{Q(x_1^n)}$$

In particular, for the subclass of kth order Markov chains with alphabet A and sparse context tree $\mathcal{T}$ it can be proved that

$$\frac{|\mathcal{T}|(|A|-1)}{2} \log n - K_1 \le \overline{\mathcal{R}}_n \le \mathcal{R}_n^* \le \frac{|\mathcal{T}|(|A|-1)}{2} \log n - K_2,$$

with suitable constants K_1 and K_2, this shows that there exists universal codes for a fixed structure tree and unknown parameters.

Relation with model selection techniques

In the classical setting of universal coding it is assumed that, although the exact process is unknown, it is still known to belong to a given class $\mathcal{P}$, e.g., i.i.d., first-order Markov processes, and so on. The performance of a universal code is measured in terms of the excess compression ratio beyond the entropy, namely, the redundancy rate $\mathcal{R}_{P,C}$, which depends on the code length function $L_C(\cdot)$, the process P, and the data string x_1^n.

We can extend the scope of universal coding theory to deal with hierarchies of classes. Consider a sequence of classes of process, $\{\Lambda_i : i \in \mathcal{I}_n\}$ with possibly different capacities $\mathcal{R}_n^*(\Lambda_i)$. The number of classes $|\mathcal{I}_n|$ may be finite and fixed, or growing with n, or even countably infinite for all n. We know that the active process P belongs to one of the classes Λ_i but we do not know i in advance. The challenge is to provide coding schemes with optimum "adaptation" capability in the sense that, first, the capacity of the active class $\mathcal{R}_n^*(\Lambda_i)$ is always approached, and moreover, the extra redundancy due to the lack of prior knowledge of i is minimum.

If one views this problem just as universal coding with respect to the union of classes $\Lambda = \cup_i \Lambda_i$, then the redundancy would be the capacity $\mathcal{R}_n^*(\Lambda)$ associated with Λ. For example, if $\mathcal{I}_n = \{1, \ldots, M_n\}$ and Λ_i is the class of Markov models of order i, then $\mathcal{R}_n^*(\Lambda)$ is essentially the same as the redundancy associated with the maximum order M_n. Obviously, it is easy to do better than that as there are many ways to approach the capacity $\mathcal{R}_n^*(\Lambda_i)$ of the class corresponding to the active process.

This approach is known as *twice-universal or hierarchical coding*, we answer questions such as: should we model the data as i.i.d. or as Markov of order 1? A trade-off occurs when varying the model size because increasing the model size allows the model to better fit the data and thus the data description length decreases while the model description length increases.

One way to achieve this adaptation property is to apply a two-part code, where the first part is a code for the index $i \in \mathcal{I}_n$, and the second part implements an optimum universal code within each class. The value of i is chosen so as to minimize the total length of the code. This approach is known as semi-predictive, two-pass or model selection method. By doing this, one can achieve redundancy essentially as small as $\mathcal{R}_n^*(\Lambda_i) + L_C(i)$. This method, however, requires a comparison between competing codes for all $i \in \mathcal{I}_n$ or a good estimator for the true i.

The idea of the above construction is based on the *minimum description length* (MDL) *principle* of statistical inference, which states the simple idea that the best way to capture regular features in data is to construct a model in a certain class which permits the shortest description of the data and the model itself.

These topics, among others, are developed in depth in my Master thesis in Computer Science (Fraiman, 2008).

Bibliography

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math. 22*, 203–217.

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Akadémia Kiadó, Budapest, pp. 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr. 19*, 716–723.

Akaike, H. (1977). On entropy maximization principle. In *In Applications of Statistics (Proc. Sympos., Wright State Univ., Dayton, Ohio, 1976)*, Amsterdam, pp. 27–41. North-Holland.

Anderson, T. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist. 33*, 255–265.

Anderson, T. (1963). Determination of the order of dependence in normally distributed time series. In *Proc. Sympos. Time Series Analysis (Brown Univ., 1962)*, New York, pp. 425–446. Wiley.

Baron, D. and Y. Bresler (2004). An o(n) semipredictive universal encoder via the bwt. *IEEE Trans. Inform. Theory 50*, 928–937.

Baron, D., J. Rissanen, and B. Yu (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory 44*, 2743–2760.

Bejerano, G. (2003). *Automata learning and stochastic modeling for biosequence analysis.* Ph. D. thesis, Institute of Computer Science, Hebrew University.

Bejerano, G. (2004). Algorithms for variable length markov chain modeling. *Bioinformatics 20*, 788–789.

Bejerano, G. and G. Yona (2001). Variations on probabilistic suffix trees: statistical modeling and the prediction of protein families. *Bioinformatics 17*, 23–43.

Bourguignon, P. Y. and D. Robelin (2004). Modèles de markov parcimonieux : sélection de modèle et estimation. In *Proceedings of Journees Ouvertes Biologie Informatique.*

Buhlmann, P. and A. Wyner (1998). Variable length Markov chains. *Annals Math. Statist. 27*, 480–513.

Csiszár, I. (2002). Large-scale typicality of markov sample paths and consistency of mdl order estimators. *IEEE Trans. Inform. Theory 48*, 1616–1628.

Csiszár, I. and P. Shields (2000). The consistency of the bic Markov order estimator. *Annals Math. Statist. 28*(6), 1601–1619.

Csiszár, I. and Z. Talata (2006). Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Trans. Inform. Theory 52*, 1007–1016.

Davisson, L. (1965). Prediction error of stationary gaussian time series of unknown variance. *IEEE Trans. Inform. Theory 19*, 783–795.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* New York: Springer Verlag.

Eskin, E., W. Grundy, and Y. Singer (2000). Protein family classification using sparse markov transducers. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 134–145.

Finesso, L. (1992). Estimation of the order of a finite markov chain. In *Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing (Kobe, 1991)*, Tokyo, pp. 643–645. Mita.

Finn, R., J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. Eddy, E. Sonnhammer, and A. Bateman (2006). Pfam: clans, web tools and services. *Nucleic Acids Res. 34*, 47–51.

Forchhammer, S., X. Wu, and J. Andersen (2004). Optimal context quantization in lossless compression of image data sequences. *IEEE Trans. Image Proc. 13*(4), 509–517.

Fraiman, N. (2008). Universal souce coding via sparse tree models. Master's thesis, Instituto de Computación, Universidad de la República.

Galves, A., V. Maume-Deschamps, and B. Schmitt (2008). Exponential inequalities for vmlc empirical trees. *ESAIM Probab. Stat. 12*, 219–229.

Gerencsér, L. (1987). Order estimation of stationary gaussian arma processes using rissanen's complexity. Technical Report Computer and Automation Institute of the Hungarian Academy of Sciences.

Hamerly, E. and M. Davis (1989). Strong consistency of the pls criterion for order determination of autoregressive processes. *Annals Math. Statist. 17*, 941–946.

Hannan, E. (1980). The estimation of the order of an arma process. *Annals Math. Statist. 8*, 1071–1081.

Hannan, E. and B. Quinn (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B 41*, 190–195.

Haughton, D. (1988). On the choice of model to fit data from an exponential family. *Annals Math. Statist. 16*, 342–355.

Krichevsky, R. and V. Trofimov (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory 27*, 199–207.

Leonardi, F. (2006). A generalization of the pst algorithm: modeling the sparse nature of protein sequences. *Bioinformatics 22*(11), 1302–1307.

Leonardi, F., S. Matioli, H. Armelin, and A. Galves (2008). Detecting phylogenetic relations out from sparse context trees. Preprint.

Mallows, C. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the IMS, Manhattan, Kansas.

Mallows, C. (1973). Some comments on cp. *Technometrics 15*, 661–675.

Martín, A., G. Seroussi, and M. Weinberger (2004). Linear time universal coding and time reversal of tree sources via fsm closure. *IEEE Trans. Inform. Theory 50*, 1442–1468.

Neveu, J. (1975). *Discrete Parameter Martingales.* Elsevier Science.

Neyman, J. and E. Pearson (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika 20*, 263–294.

Nohre, R. (1994). *Some Topics in Descriptive Complexity.* Ph. D. thesis, Department of Computer Science, The Technical University of Linkoping, Sweden.

Pasco, R. (1976). *Source Coding Algorithms for Fast Data Compression.* Ph. D. thesis, Department of Electrical Engineering, Stanford University, California.

Rissanen, J. (1976). Generalized kraft inequality and arithmetic coding. *IBM J. Res. Devel. 20*, 198–203.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*, 465–471.

Rissanen, J. (1983a). A universal data compression system. *IEEE Trans. Inform. Theory 29*, 656–664.

Rissanen, J. (1983b). A universal prior for integers and estimation by minimum description length. *Annals Math. Statist. 11*, 416–431.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* River Edge, NJ: World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory 42*, 40–47.

Rissanen, J. and G. Langdon (1981). Universal modeling and coding. *IEEE Trans. Inform. Theory 27*, 12–23.

Ron, D., Y. Singer, and N. Tishby (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Mach. Learn. 25*, 117–149.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals Math. Statist. 6*, 461–464.

Servant, F., C. Bru, S. Carrère, E. Courcelle, J. Gouzy, D. Peyruc, and D. Kahn (2002). ProDom: Automated clustering of homologous domains. *Briefings in Bioinformatics 3*, 246–251.

Shibata, R. (1976). Selection of the order of an autoregressive model by akaike's information criterion. *Biometrika 63*, 117–126.

Shtarkov, Y. (1977). Coding of discrete sources with unknown statistics. In I. Csiszár and P. Elias (Eds.), *Topics in Information Theory*, Volume 23 of *Colloquia Math. Soc. J. Bolyai*, Amsterdam, pp. 559–574. North-Holland.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nat. Rev. Genet. 7*, 493–505.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B 36*, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J. Roy. Statist. Soc. Ser. B 39*, 44–47.

Weinberger, M., A. Lempel, and J. Ziv (1992). A sequentional algorithm for the universal coding of finite memory sources. *IEEE Trans. Inform. Theory 38*, 1002–1014.

Weinberger, M., J. Rissanen, and M. Feder (1995). A universal finite memory source. *IEEE Trans. Inform. Theory 41*, 643–652.

Willems, F. (1998). The context-tree weighting method: Extensions. *IEEE Trans. Inform. Theory 44*, 792–798.

Willems, F., Y. Shtarkov, and T. Tjalkens (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inform. Theory 41*, 653–664.

Willems, F., Y. Shtarkov, and T. Tjalkens (2000). Context-tree maximizing. In *Conference on Information Sciences and Systems*, Princeton, pp. 6–12.

Wu, C., R. Apweiler, A. Bairoch, D. Natale, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek (2006). The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic Acids Res. 34*, 187–191.

Zhao, X., H. Huang, and T. Speed (2004). Finding short dna motifs using permuted markov models. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, New York, USA, pp. 68–75. ACM.