

## Lecture 2 — January 29

Lecturer: Lior Pachter

Scribe/ Editor: Allen Chen/ Anne Shiu

## 2.1 Log linear models (toric models)

We will discuss toric models based on a Laminar family  $H$ . An alphabet is a finite set we denote by  $\Sigma$ . For example,  $\Sigma$  could be  $\{0, 1\}$  or  $\{A, C, G, T\}$ . The first type of model we consider are those in which the state space is  $\Sigma^{|X|}$ . For example, consider the alphabet  $\Sigma = \{0, 1\}$ . In the last lecture we looked at a model with state space consisting of  $(0,0)$ ,  $(0,1)$ ,  $(1,0)$  and  $(1,1)$ . Here  $|X| = 2$ ,  $X = \{1, 2\}$ , and  $H = \{\{1\}, \{2\}\}$ . Next we parametrize the model by the function

$$P : \Theta \rightarrow \Delta_{\xi}$$

$$(\theta_1, \theta_2, \theta_3, \theta_4) \mapsto (\theta_1\theta_3, \theta_1\theta_4, \theta_2\theta_3, \theta_2\theta_4) = (p_{00}, p_{01}, p_{10}, p_{11}).$$

Here the parameter vectors  $\Theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in \mathbb{R}^4$  must satisfy  $\theta_1 + \theta_2 = 1$ ,  $\theta_3 + \theta_4 = 1$ , and  $\theta_i \geq 0$ . Note that we have the relation  $p_{00}p_{11} = \theta_1\theta_2\theta_3\theta_4 = p_{01}p_{10}$ . In fact the model is the set  $\{(p_{00}, p_{01}, p_{10}, p_{11}) : p_{00}p_{11} - p_{01}p_{10} = 0, p_{00} + p_{01} + p_{10} + p_{11} - 1 = 0, p_{ij} \geq 0\}$ .

**Definition 2.1.** A log-linear model is a model of the form

$$P : \mathbb{R}^d \rightarrow \mathbb{R}^m$$

$$(\theta_1, \dots, \theta_d) \mapsto \frac{1}{\sum_{j=1}^m \theta^{a_j}} (\theta^{a_1}, \dots, \theta^{a_m}),$$

where  $A$  is a matrix  $A = (a_{ij})_{j=1, \dots, m}^{i=1, \dots, d}$  with all column sums the same. Note that we use multi-index exponent notation  $\theta^{a_j} = \prod_{i=1}^d \theta^{a_{ij}}$ .

For example, the  $A$  from the previous example is the following matrix, which realizes this model as a toric model:

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Data for this model take the form  $i_1, \dots, i_N$  where  $i_j \in \Sigma^{|X|}$ . The number  $N$  denotes the sample size. If the data are independent and identically distributed (i.i.d.), then we summarize the data with the vector  $(u_1, \dots, u_m)$ , where  $u_k =$  number of integers  $j$  such that  $i_j = k$ . Note that  $u_1 + \dots + u_m = N$ .

## 2.2 The maximum likelihood problem

(ref Pachter and Sturmfels pp.8-13)

**Definition 2.2.** *Maximum likelihood problem:* Given a model, find the parameters  $\theta \in \Theta$  that maximizes the likelihood function

$$\mathbf{L}(\theta) = p_1(\theta)^{u_1} p_2(\theta)^{u_2} \cdots p_m(\theta)^{u_m} \quad .(*)$$

and  $\mathbf{L}$  maps the low dimensional parameter  $\theta$  into a high-dimensional space with data  $u_i$

$$\theta \rightarrow (p_1, \cdots, p_m).$$

A statistic is a function of the data. A *sufficient statistics* is a function whose image is the least amount of information the likelihood depends on.

## 2.3 Toric models

For *log-linear models*, also known as *toric model*, maximum likelihood is “nice”, because the problem (\*) reduces to maximizing  $\theta^b$  such that  $\theta \in \Theta$  where  $b = Au$ . Note that  $\theta^{Au}$  has the form

$$\theta^{Au} = \prod_{i=1}^d \theta_i^{a_{i1}u_1 + \cdots + a_{im}u_m}.$$

**Theorem 2.3.** (see Pachter and Sturmfels p13 Proposition 1.9)

Let  $A$  be a log-linear (toric) model with data  $u \in \mathbb{N}^m$ , sample size  $N$  and sufficient statistic  $b = Au$ . Let  $\theta$  be any local maximum for the maximum likelihood (ML) problem. Then

$$\text{if } \hat{p} = \mathbf{P}(\hat{\theta}), \quad \text{then } A \cdot \hat{p} = \frac{1}{N} \cdot b = \frac{1}{N} \cdot Au.$$

The ML estimate will be  $\mathbf{L}(\theta^*)$  where  $\theta^*$  solves the ML problem.

*Homework 2:* Review and prove the Lagrange multiplier theorem.

**Theorem 2.4.** Every local maxima is a critical point of the following function in  $d + 1$  unknowns  $\theta_1, \dots, \theta_d, \lambda$ :

$$\theta^b + \lambda \left( 1 - \sum_{j=1}^m \theta^{a_j} \right).$$

We give an outline here (see also Pachter and Sturmfels p.13). If you look at the system of equations forming the partial derivatives, then you get the condition that vector  $A \cdot \hat{p}$  is a scalar multiple of the vector  $b = Au$ .

We now consider an example defined by

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}.$$

In other words, the model is parametrized by

$$(\theta_1, \theta_2) \mapsto \frac{1}{\theta_1^2 + \theta_1 \theta_2 + \theta_2^2} (\theta_1^2, \theta_1 \theta_2, \theta_2^2).$$

Suppose the data vector  $u = (11, 17, 23)$ , with  $N = 11 + 17 + 23 = 51$ .

The ML problem is to maximize the likelihood

$$\mathbf{L}(\theta) = \theta_1^{39} \theta_2^{63} = (\theta_1^2)^{11} (\theta_1 \theta_2)^{17} (\theta_2^2)^{23}.$$

$$A \cdot \hat{p} = \frac{1}{N} Au \Rightarrow$$

$$\begin{pmatrix} 2\hat{\theta}_1^2 + \hat{\theta}_1 \hat{\theta}_2 \\ \hat{\theta}_1 \hat{\theta}_2 + 2\hat{\theta}_2^2 \end{pmatrix} = \frac{1}{51} \begin{pmatrix} 39 \\ 63 \end{pmatrix} \Rightarrow$$

$$\hat{\theta}_1 = \frac{1}{51} \sqrt{1428 - 51\sqrt{277}} = 0.472$$

$$\hat{\theta}_2 = \frac{1}{51} \sqrt{2040 - 51\sqrt{277}} = 0.677.$$

The probability distribution corresponding to these parameter values is

$$\hat{p} = (\hat{p}_1, \hat{p}_2, \hat{p}_3) = (\hat{\theta}_1^2, \hat{\theta}_1 \hat{\theta}_2, \hat{\theta}_2^2) = (0.2227, 0.3193, 0.4580).$$

## 2.4 Birch's theorem

Given the matrix  $A \in \mathbb{R}^{d \times m}$ , we consider the set

$$P_A(b) = \{p \in \mathbb{R}^m : A \cdot p = \frac{1}{N} \cdot b \quad \text{and} \quad p_j > 0, \forall j\}.$$

This is a relatively open *polytope*. We are interested in  $P_A(b) \cap P(\Theta)$ , which consists of at most one point.

**Theorem 2.5.** (Birch's Theorem) *Fix a toric model  $A$  and let  $u \in \mathbb{N}_{>0}^m$  be a strictly positive data vector with sufficient statistic  $b = Au$ . The intersection of the polytope  $P_A(b)$  with the toric model  $\mathbf{f}(\mathbb{R}_{>0}^d)$  consists of precisely one point. That point is the maximum likelihood estimate  $\hat{p}$  for the data  $u$ .*

Comment: For the complete proof, see Pachter and Sturmfels p. 14. The key idea is look at the *entropy function*

$$\begin{aligned} H : \mathbb{R}_{\geq 0}^m &\rightarrow \mathbb{R}_{\geq 0} \\ (p_1, \dots, p_m) &\mapsto - \sum_{i=1}^m p_i \log(p_i) \end{aligned}$$

This function  $H$  is strictly concave. This means for  $0 < \lambda < 1, p \neq q$

$$H(\lambda p + (1 - \lambda)q) > \lambda H(p) + (1 - \lambda)H(q).$$

And  $H$  attains its maximum at some point  $p^*$  in  $P_A(b)$ . Now we must show that  $p^*$  actually solves the ML problem.

## 2.5 Markov chains

Let

$$H = \{\{1\}, \{1, 2\}, \{1, 2, 3\}, \dots, \{1, 2, \dots, n\}\}$$

and

$$X = \{1, 2, \dots, n\}$$

The *Markov Chain (MC)* is a model on  $\sum^{|H|} = \sum^n$  which has parameter set  $\Theta$  that consists of  $l \times l$  matrices where  $|\sum| = l$ .

An example, toric Markov chains from Pachter and Sturmfels pp 25-27:  
Let us consider words of length  $n = 4$  over binary alphabet

$$\sum = \{0, 1\},$$

so that  $l = 2, d = 4$  and  $m = 16$ . The matrix  $A$  is the following  $4 \times 16$  matrix:

$$A = \begin{pmatrix} 3 & 2 & 1 & 1 & 1 & 1 & 0 & 0 & 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 2 & 3 \end{pmatrix}.$$

The toric Markov chain of length  $n = 4$  for the binary alphabet  $l = 2$  is the image of  $\Theta$  under the monomial map:

$$(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}) \mapsto \frac{1}{\sum_{ijkl} p_{ijkl}} \cdot (p_{0000}, p_{0001}, p_{0010}, \dots, p_{1111}).$$

with 4 rows and 16 columns. where  $p_{ijkl}$  has the following form:

$$p_{ijkl} = \theta_{ij}\theta_{jk}\theta_{kl}.$$

The toric Markov chain model  $\mathbf{f}_{2,4}(\Theta)$  is a three-dimensional object inside the 15-dimensional simplex.

Algebraically, the simplex is specified by *model invariants*. A *model invariant* is an algebraic relation that holds for all probability distributions in the model. The simplest model invariant is:

$$p_{0000} + \dots + p_{1111} = 1,$$

where the  $p_{ijkl}$  are unknowns which represent the probabilities of the 16 states.

The other linear invariants come from the fact that the matrix  $A$  has some repeated columns:

$$p_{0110} = p_{1011} = p_{1101},$$

$$p_{0010} = p_{0100} = p_{1001}.$$

For remaining invariants, see text on p. 26.