ALGEBRAIC AND GEOMETRIC ASPECTS OF LOG-LINEAR MODELS

ADWAY S. WADEKAR

Abstract. This is an expository thesis on algebraic aspects of log-linear models, conducted over two independent studies under the Prof. Ezra Miller towards graduation with distinction in mathematics. We describe the statistical intuition behind concepts arising in the algebraic statistics of discrete log-linear models. We show that in essence, a log-linear model relates to a multinomial distribution over colors, which induces a distribution on experimental outcomes. In the second half of the thesis, we explore two extensions of the classical algebraic and geometric notions of log-linear models. The first is the Bayesian setting, where prior belief/information and observed data are combined to make inferences. The second relates to uncertainty quantification and how algebra can be used to estimate the uncertainty of a point estimate for a parameter without appealing to asymptotic arguments used in classical statistics.

Contents

1. Introduction

In recent years, connections between classical algebraic geometry and log-linear models have been described by a number of authors [YAK+21, Sul23, PS05]. Most often, when using log-linear models, data analysts are concerned with the maximum likelihood estimate (MLE), which is the maximizer of the likelihood function conditional on observed data. In log-linear models, a vector p of probabilities, which encodes a distribution over a set of *experimental outcomes* is given by a parametrization in terms of probabilities on *states* on

Date: April 2025.

components of each experiment. In algebraic terms, this is related to a monomial map and an element in the d-dimensional algebraic torus' action on the all-ones vector.

To a statistician, what is observed is data concerning these experimental outcomes: in particular, a vector of counts of each experimental outcome. The goal is to estimate the probabilties of each experimental outcome using the data that are those counts. In [YAK+21], polyhedral conditions are given for the existence of the MLE in a log-linear model, which are related to the stability of the action on the all-ones vector of the algebraic torus. These polyhedral conditions align with the intuitive solution for the MLE when each experimental outcome is observed at least once (i.e. when each entry in the vector of counts is nonzero). The MLE, in this case, is simply the vector of counts divided by the number of experimental trials.

The key difficulty, as we explain, is that sometimes the observed data have counts of zero for certain experimental outcomes. This is particularly prevalent in so-called "small data" settings, where it is prohibitive to run an experiment a large number of times. In this setting, One cannot simply allow for the standard approach to constructing the MLE, since only distributions over the experimental outcomes containing all positive probabilities are allowed in the log-linear models. Therefore, there must be some alternative vector of possible counts that maximizes the likelihood function. The first half of this thesis is spent on providing the statistical intuition behind these polyhedral conditions.

In this thesis, we also explore two second-order extensions of MLE analysis in an algebraic context. The first is how the algebra and geometry of log-linear models relates to the Bayesian framework. The Bayesian framework is useful for incorporating prior information into statistical inferences. We show that for the classical Bayesian point estimate for a parameter of interest, the Bayesian setup "factors through" and that the same polyhedral formula for the existence of this point estimate holds. The second extension is about uncertainty quantification. We explore a particular method of uncertainty quantification that relies on sampling from the sample space conditioned on a given value of a sufficient statistic. This method relies on finding the generating set of an ideal in a polynomial ring. Computing ranges of parameters from this sort of sample is more appropriate for "small data" settings, where asymptotic arguments do not hold.

2. Preliminaries

2.1. **Statistical Preliminaries.**

Statistics is concerned with gleaning insight about parameters of a generating process from data. At a basic level, this means that given a set of data and a parametrizable generating process (model), the goal is to estimate the parameters, with high confidence, that are likely to have generated the data. A map from the data to the parameter space is called an estimator. Statistics, in a sense, is probability flipped on its head. Whereas in probability, the goal is to estimate the odds of seeing a particular data set given known parameters, in statistics, the data are known and the parameters are unknown. This is perhaps best illustrated through the trial and error experiment process, where the outcomes of several experiments are recorded.

Example 2.1 (Binomial distribution)**.** Consider an experiment in which someone flips a possibly weighted coin n times and sees k heads. The coin in question has a certain probability θ of landing on heads. We seek to estimate θ from the data. Recall that the distribution

function for k success in n independent trials under θ, the probability of a head, is

$$p(k;\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Some estimators are more useful than others. For example, mapping the data to a constant is a valid estimator but provides no real information about the parameter. Maximum likelihood estimation is one method for estimating the parameter θ from the observed data and has several nice asymptotic properties (i.e. consistency and asymptotic normality). In maximum likelihood estimation, we view the distribution function with the data as constants and the parameters as variables. Taking the MLE is to find the parameters that maximize the value of the density or mass function as follows, we have

$$\begin{aligned}
\log p(k;\theta) &= \log\binom{n}{k} + k\log\theta + (n-k)\log(1-\theta) \\
\frac{\partial \log p(k;\theta)}{\partial\theta} &= \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \\
k(1-\theta) &= (n-k)\theta \\
k - k\theta &= n\theta - k\theta \\
k &= n\theta
\end{aligned}$$

which implies that

$$\hat{\theta} = \frac{k}{n}.$$

The maximum likelihood estimate $\hat{\theta}$ is the estimator for the parameter θ.

2.2. **Algebraic Preliminaries.** In this section, we will discuss the algebraic and geometric structures and setup to which connections to toric actions and varieties are made in discrete statistics. For a discussion about continuous models, see the material in [AKRS21], which discusses. We will soon provide an algebraic definition for the main object of study in the paper on toric varieties for log-linear models: the log-linear model itself. However, first, we must provide an algebraic and geometric definition of what is considered to be a (discrete) statistical model, that we will work with from the algebraic statistics perspective.

Consider the $(m-1)$-dimensional probability simplex (denoted as having $m-1$ dimensions because there are are $m-1$ free parameters):

$$\Delta_{m-1} = \left\{ p \in \mathbb{R}^m \mid p_j \geq 0 \text{ for all } j \text{ and } \sum_{j=1}^{m} = 1 \right\}.$$

Each point in this simplex corresponds to a probability mass function (i.e. a discrete distribution) as each of m possible states is assigned a non-negative theoretical probability. A statistical model $\mathcal{M}$ is a collection of distributions, and therefore a subset of the $m-1$-dimensional simplex. The MLE procedure described in Section 2.1 is to observe counts of these m states and then find the point $p \in \mathcal{M} \subseteq \Delta_{m-1}$ that maximizes the likelihood (or probability) of observing those counts, up to swapping of the order in which the states occur. As such, the likelihood, defined in Section 2.1 is given by

$$f(u;p) = p_1^{u_1} \dots p_m^{u_m},$$

where u is a vector of observed counts of the states and p is the vector of the theoretical probabilities of the states.

The idea of characterizing a statistical model as a collection of points in a simplex is a very geometric one, and in reality, a statistician would conceive of a (discrete) model to have much more structure than merely a collection of possible points where the individual entries lie in the positive orthant and sum to 1. In the discrete world, statisticians often deal with counts of data, and there are many common models for such data: binomial, Poisson, multinomial, negative binomial, etc. Indeed, the paper [YAK+21] puts a form of algebraic structure on the model that is considered. This structure agrees with what is essentially the multinomial distribution for count data, as I will describe in Section 3. But for now, I will continue to focus on the algebraic characterizations of the statistical model presented and its properties which connect it to the algebraic torus. In their paper, [YAK+21] describe what they call a "log-linear model." We first provide a slightly more general definition of this model after [Sul23], and then examine its connection to algebraic structure.

Definition 2.2 (Log-affine model)**.** Fix $A \in \mathbb{Z}^{d\times m}$ as a matrix of integers, and let $h \in \mathbb{R}^m_{>0}$. The *log-affine* model formed from this matrix is the set of probability distributions (or mass functions) that satisfy

$$\mathcal{M}_{A,h} := \{p \in \Delta_{m-1} : \log p \in \log h + \text{rowspan}(A)\}.$$

When $h = \mathbf{1}$, the all 1's vector, then we denote $\mathcal{M}_{A,h} := \mathcal{M}_A$ and call the model *log-linear*, which is the specific model used in [YAK+21]. From now on, we will take $h = \mathbf{1}$.

The first immediate property that can be seen is that when $\mathbf{1} \in \text{rowspan}(A)$ and $h = \mathbf{1}$, the *uniform distribution* is contained within the model $\mathcal{M}_A$. This is because one can normalize, dividing p entry-wise by m to get a point with all states having equal probability in the simplex. Discrete log-linear models are known commonly as toric models to algebraic statisticians. To show connections to toric varieties, we first provide another definition of a monomial map associated to a log-linear model.

Definition 2.3 (Monomial map associated to log-linear model)**.** From the previous definition take $A \in \mathbb{Z}^{d\times m}$ with entries denoted a_{ij}, set $h = \mathbf{1}$, and consider $\mathcal{M}_A$. Also assume that the uniform distribution is contained within $\mathcal{M}_A$. We have a monomial map of the following form associated with the model $\mathcal{M}_A$:

$$\phi^A : \mathbb{R}^d \to \mathbb{R}^m$$

where

$$\theta \mapsto \left(\prod_{i=1}^{d} \theta_i^{a_{ij}}\right)_{1\le j\le m}.$$

Some authors choose to write a normalization constant of $Z(\theta)$ in the front of this monomial map, which ensures that the map sends θ into the simplex Δ_{m-1}.

We note that this parametrization of a log-linear model can be contextualized in terms of real-world applications, which we will do in the next section, but for the sake of self-containment, we will move on now to a proposition that describes the toric ideal that is connected to this model.

Definition 2.4 (Toric ideal)**.** Let $A \in \mathbb{Z}^{d\times m}$ as before and put $p \in \mathbb{R}^m$. The toric ideal

$$I_A := I(\phi^A(\mathbb{R}^d)) \subseteq \mathbb{R}[p_1, \dots p_m]$$

is the vanishing ideal generated by the set of possible distributions on the states of the model $\mathcal{M}_A$.

For the reader unfamiliar with the notation $I(\cdot)$, for $W \subseteq \mathbb{K}^r$ for a field $\mathbb{K}$,

$$I(\cdot) := \{f \in \mathbb{K}[p] : f(a) = 0 \text{ for all } a \in W\}.$$

Therefore, I_A is the set of all polynomials $f \in \mathbb{R}[p]$ such that $f(x) = 0$ for $x \in \operatorname{im}\phi$. We now present the proof of a statement that appears in [YAK+21] and is taken for granted. It is a standard result in [Sul23], but one that is not directly intuitive (at least to me) so I provide its proof, spelling out all of the details.

Proposition 2.5. *The toric ideal for a log-linear model associated with $A \in \mathbb{Z}^{d\times m}$ is a binomial ideal and*

$$I_A = \langle p^u - p^v : u, v \in \mathbb{N}^m \text{ and } Au = Av\rangle.$$

Proof. First, consider a binomial $p^u - p^v$ such that $Au = Av$. Then, we have that this polynomial is in I_A. To show this, pick $x \in \operatorname{im}\phi$. Recall that $p_j = \prod_{i=1}^d \theta_i^{a_{ij}}$. Therefore,

$$p_j^{u_j} = (\theta_1^{a_{1j}} \cdots \theta_d^{a_{dj}})^{u_j} = \theta_1^{a_{1j}u_j} \cdots \theta_d^{a_{dj}u_j}.$$

This means that

$$p^u = \prod_{j=1}^m \theta_1^{a_{1j}u_j} \cdots \theta_d^{a_{dj}u_j} = \theta_1^{\sum_{j=1}^m a_{1j}u_j} \cdots \theta_d^{\sum_{j=1}^m a_{dj}u_j}.$$

Notice that p^v takes the same form with the u_j's replaced with v_j's. But in the exponents, the sums are simply the products of the rows of A with u and v respectively, which are known to be equal. Factorizing therefore yields the claim that $p^u - p^v$ is in the ideal. This statement has an intuitive statistical rationale that will be explained in the next section.

To complete the proof, we must show that all such polynomials that vanish when evaluated at a point $p \in \operatorname{im}\phi$ are generated by the set of these binomials. Pick such a polynomial $f \in I_A$ and take $c_u p^u$ to be a monomial in this polynomial that has nonzero coefficient. As the polynomial must vanish for any point p, for any θ in the source of ϕ (i.e. $\mathbb{R}^d$), the polynomial must vanish as a function of θ in the form $f(\phi(\theta))$. This means that $c_u p^u$ must have a paired term that allows it to cancel. From above, note that $p^u = \theta_1^{\sum_{j=1}^m a_{1j}u_j} \cdots \theta_d^{\sum_{j=1}^m a_{dj}u_j}$, so in order to cancel for any choice of θ, there must be a term $c_v p^v$ that allows for this cancellation. This means that the dot products must align, implying that the restriction on u and v is that $Au = Av$. Note that we must have this "corresponding negative term" for any such $c_u p^u$ in the polynomial f, which means that f can and must be represented as a sum of binomials of the form $c_{u,v}(p^u - p^v)$ where $Au = Av$. We have shown that a generating set for this ideal is those very binomials. □

3. A Statistical Interpretation of Toric Models

This discussion first rephrases terms related to the algebraic characterizations of log-linear models presented in [YAK+21] and in the previous section in ways that I feel are more natural to a frequentist statistician who thinks in terms of a repeated experiments framework. In particular, we can map each of the ideas presented in Section 2.2 to a setting

where experiments are performed by repeatedly drawing balls of different colors from a bag. Moreover, the paper starts with a discussion of the algebra and the geometry that is used, and then moves into the statistics. I will do the opposite: first, I will describe the statistical framework and experimental design, and then, I will show how this connects to algebraic structures.

Example 3.1. Consider an experiment in which there are balls of d different colors, each with a certain probability θ_i of being chosen from the bag, such that $\sum_{i=1}^{d} = 1$. Suppose the person conducting the experiment draws with replacement from the bag n times. Put the number of times that a ball of color $i \in [d]$ is drawn as n_i, with $\sum_{i=1}^{d} n_i = n$. Now, the probability mass function of drawing the $(n_i)_{i=1}^{d}$ balls is

$$\begin{aligned} p(n_1, \ldots, n_d) &= \frac{n!}{n_1! \cdots n_d!} \theta_1^{n_1} \times \cdots \times \theta_d^{n_d} \\ &\propto \theta_1^{n_1} \times \cdots \times \theta_d^{n_d}. \end{aligned}$$

Already, one can see that this probability mass function, which is the probability mass function for the multinomial distribution, is beginning to take the form of the monomial map in Definition 2.3. Moreover, notice that here, θ lies in the Δ_{d-1} dimensional simplex, a fact that we will come back to in Section 4.

To extend this example further so that it agrees with the log-linear model setup concerning a matrix A and a probability simplex with m "states," consider an extended version of the experiment in Example 3.1. In particular, suppose that the person conducting the experiment wishes to conduct the experiment *multiple times*. That is, they draw from the bag n times and record the empirical distribution of the colors as just one experiment, and suppose they conduct k experiments.

First, notice that the number of colors of balls in the bag and the number of times the person conducting the experiment draws from the bag induces a probability distribution on the *possible outcome* of any given experiment. Indeed, the probability of observing any particular possible experiment is the probability of observing the colors that comprise that possible experiment. Therefore, if there are m possible outcomes for an experiment, where each of these outcomes representing a "binning" of colors, then we have a distribution over m states (i.e. $p \in \Delta_{m-1}$), where each state is a possible experimental outcome.

Example 3.2. Consider an example where each experiment contains three draws with a bag of two colors a and b. The possible outcomes for the number of color a and the number of color b, or in other words, the possible outcomes for any given experiment, are $(3, 0)$, $(2, 1)$, $(1, 2)$ and $(0, 3)$. Then, suppose the person conducting the experiment does so twice with one outcome resulting in $(3, 0)$ and another resulting in $(1, 2)$. What is the probability of this occurring? We have

$$p((3,0) \text{ and } (1,2)) \propto \binom{2}{1} \theta_a^3 \theta_b^0 \theta_a^1 \theta_b^2 \propto \theta_a^4 \theta_b^2.$$

Notice that this example directly aligns with the example of the twisted cubic in [Sul23, Example 6.2.5]. More generally, the probability of observing a certain count of experimental outcomes is simply the probability of observing the total number of colors summed over all the observed experiments, up to permutation (i.e. disregarding the factorial terms in front, which swap the order of the experiments and in the case of the singular experiment, swap the order of the draws).

We now have the intuition to reconstruct the setup in [YAK+21] from the viewpoint of a distribution on the colors. Instead of constructing a distribution on the possible experimental outcomes first and relating it retrospectively to the distribution on the colors via a monomial map parametrization and a torus action, we will take the more intuitive approach, building up from the distribution on the colors, showing how a torus action naturally relates the distribution on the colors to the distribution on the experimental outcomes. Then, having observed many experiments and knowing the binning of colors that make up each potential experimental outcome, we will show that whether or not it is possible to find the maximum likelihood estimate for the distribution of experimental outcomes is related to the properties of the torus action that maps between the distribution on the colors to the distribution on the experimental outcomes.

To do so, for a set of m experimental outcomes determined by n draws per experiment with d colors, we would like to construct a map $\psi : \Delta_{d-1} \to \Delta_{m-1}$ (up to normalizing constants), which takes a distribution $\theta \in \Delta_{d-1}$ on the colors and takes it to a distribution p on the possible experimental outcomes. For a given experiment $j \in [m]$, let us observe $n_{1j}, \ldots, n_{dj}$ colors, where $\sum_{i=1}^{d} n_{im} = n$. We have

$$\psi : \Delta_{d-1} \to \Delta_{m-1}$$

with

$$\theta \mapsto \left(\theta_1^{n_{1j}} \times \cdots \times \theta_d^{n_{dj}}\right)_{1 \le j \le m}.$$

Notice that we have arrived back at the very monomial map associated to the log-linear model for the potential experiments that we stated in Definition 2.3, where each component of p is a mass from the multinomial distribution, up to ordering of the draws for each experiment. We can now identify $(n_{ij})_{1 \le i, \le d, 1 \le j \le m}$ with the matrix $A \in \mathbb{Z}^{d \times m}$. That is, the columns of A are identified with the various counts of the d colors in the bag for each experimental outcome. We can take this further to show that this map is the action of θ as an element of the d-dimensional algebaric torus.

Definition 3.3 (Torus action)**.** Consider the d dimensional complex torus GT_d. The action of GT_d on $\mathbb{P}^{m-1}_{\mathbb{C}}$ under the matrix $A \in \mathbb{Z}^{d \times m}$ is given by the map that first sends a torus element $\lambda = (\lambda_1, \ldots, \lambda_d)$ to the matrix

$$\begin{bmatrix} \lambda_1^{a_{11}} \cdots \lambda_d^{a_{d1}} & & & & \\ & \lambda_1^{a_{12}} \cdots \lambda_d^{a_{d2}} & & & \\ & & \lambda_1^{a_{13}} \cdots \lambda_d^{a_{d3}} & & \\ & & & \ddots & \\ & & & & \lambda_1^{a_{1m}} \cdots \lambda_d^{a_{dm}} \end{bmatrix}.$$

The torus element then acts on a vector $v \in \mathbb{P}^m_{\mathbb{C}}$ through right multiplication by the above matrix.

To show how $p \in \Delta_{m-1}$ arises from a torus action, we make the (trivial) statement that the probability of observing an experimental outcome is the probability of observing *one time* in a repeated experiments framework. This statement allows us to encode the distribution p as an action of the torus element θ on $\mathbf{1}$, the all 1's vector. Indeed, let $\lambda = \theta$ as an element in the d-dimensional algebraic torus, and identify $(n_{ij})_{1 \le i, \le d, 1 \le j \le m}$ with A as before, and let θ act on $\mathbf{1}$.

Revisiting the first statement of Proposition 2.5, that binomials of the form $p^u - p^v$ are in the toric ideal when $Au = Av$, also now has a very intuitive rationale. Recall that the columns of A are identified with the counts of the d different colors, which means the row vectors identify the numbers of color $i \in [d]$ observed across the possible experimental outcomes. Therefore if $Au = Av$, that means that u and v are two count vectors for experimental outcomes that yield the same count of each of the d colors. As the vector p is simply a function of the probability of the colors for fixed theoretical experimental outcomes, the binomial *must* evaluate to zero because the counts of the colors is the same.

3.1. **Maximum Likelihood Estimation.** Ultimately, the goal of [YAK+21] is to use the general toric geometry of GT_d to find the maximum likelihood estimate for the true distribution $p \in \Delta_{m-1}$ over the experimental outcomes from a vector of counts of these outcomes taken over n experiments. There are certain conditions that must be met by the torus action of GT_d under a certain *linearization* for the MLE to even exist, relating to its stability of the $\mathbf{1}$ vector under the action of elements in the torus (i.e. possible distributions of colors).

3.1.1. *Geometric Background Related to Torus Actions.* A linearization of the action of GT_d is a corresponding action that takes the matrix A and subtracts a vector $b \in \mathbb{Z}^d$ from each column of A. We will now describe certain notions related to the stability of a group action, and then a torus action. For a vector v define its capacity $\mathrm{cap}(v) := \inf_{g \in G} \|g \cdot v\|$ where g an element in a group G. We have the following definition for different forms of stability under the group action, that will be related to geometry for a torus action, noting that the algebraic torus is a group.

Definition 3.4 (Notions for stability)**.** Let $v \in \mathbb{C}^m$. Denote the orbit of v by $G \cdot v$, and the orbit closure in the Euclidean sense by $\overline{G \cdot v}$. Let the stabilizer be $G_v = \{g \in G : g \cdot v = g\}$. For those unfamiliar with group theory, the "orbit" of a vector is the possible locations the vector can be sent by the group action. The stabilizer of a vector are the members of the group that map the vector to itself. We call v

(a) *unstable* if $0 \in \overline{G \cdot v}$. If 0 is in the orbit closure, then the capacity of v is 0.
(b) *semistable* if $0 \notin \overline{G \cdot v}$. If 0 is not in the orbit closure, then the capacity of v is greater than 0 as the closure of the orbit includes the infimum over all possible destination points.
(c) *polystable* if $v \neq 0$ and the orbit is closed.
(d) *stable* if v is polystable and the stabilizer is finite.

Unstable points form the *null cone* of the action.

Denote the convex hull of the columns of A as a polytope

$$P(A) := \mathrm{conv}\{a_1, \dots a_j\}.$$

Points in $P(A)$ are equivalently represented by Au, where $u \in \Delta_{m-1}$ by the definition of a convex hull. A subpolytope in this convex hull for an index set $J \subseteq [m]$ is denoted by

$$P_J(A) := \mathrm{conv}\{a_j | j \in J\}.$$

Finally, for a vector v, denote its support $\mathrm{supp}(v) := \{j \mid v_j \neq 0\}$. In other words, $\mathrm{supp}(v)$ forms an index set $J \in [m]$ that we can use to make subpolytopes. Denote

$$P_v(A) := \mathrm{conv}\{a_j \mid j \in \mathrm{supp}(v)\}.$$

For a given polytope $P \subseteq \mathbb{R}^d$, we will denote the interior by $\text{int}(P)$ and its relative interior by $\text{relint}(P)$. The following theorem relates Definition 3.4, the various notions of stability, to the geometry of polytopes constructed from the columns of A, when the group action is given by a torus.

Theorem 3.5 (Hilbert-Mumford criterion for a torus)**.** *Let $v \in C^m$ and take the action of GT_d given by a matrix $A \in \mathbb{Z}^{d\times m}$ with a linearization $b \in \mathbb{Z}^d$. We have,*

(a) v is unstable if and only if $b \notin P_v(A)$
(b) v is semistable if and only if $b \in P_v(A)$.
(c) v is polystable if and only if $b \in \text{relint}(P_v(A))$
(d) v is stable if and only if $b \in \text{int}(P_v(A))$.

The proof of this theorem is provided elsewhere [YAK$^+$21, Appendix A]. We now have enough to describe how the existence of the MLE for the distribution of experimental outcomes given an observed count of each of these experimental outcomes, stored in a vector u. We first remark that this vector u must have its components sum to the total number of experiments undertaken (i.e. $\sum_{i=1}^m u_i = n$). If we divide u by n elementwise, we have an empirical, or observed distribution over the possible experiments. Denote this empirical distribution as $\bar{u}$.

An immediate possible conclusion is that $\bar{u}$ is the MLE for p, and indeed this may be a possible MLE. But are there other possible MLE's? The answer is a possible "yes," as similar to the first statement in Proposition 2.5, different observed counts of experiments could lead to the same number of observed colors, and so even though we observed one particular experimental distribution of experiments, we could just as easily have observed a different distribution of experiments. This idea is what encapsulates the notion of a *sufficient statistic* – that the statistic contains all there is to know about the model parameters. In other words, knowing a count of observed experiments does not tell you everything about the true probabilities of observing a particular experiment, but knowing the counts of the colors does. This intuition is encapsulated in the idea that the MLE for p is a vector q that satisfies

$$Aq = A\bar{u},$$

which is a fact that is shown in [Sul23, Corollary, 7.3.9] as a result of a theorem about sufficient statistics in exponential families.

There is an issue, however, with this approach, and that is what happens if we observe a count of zero for one experimental outcome? It is certainly possible that this could be the case experimentally, but any p containing zero lies on the boundary of $\mathcal{M}_A$. This is due to the logarithmic condition that defines $\mathcal{M}_A$: we cannot include outcome probabilities of zero in the model as the logarithm is not defined; however, we can get arbitrarily close to zero.

So, in order to find an MLE, there must be some other count vector that yields the same number of *colors* as that of u, which defines the empirical distribution $\frac{u}{n} = \bar{u}$ with one of the coordinates zero. In essence, if we cannot do this, the MLE does not exist. However, in this case, if we extend $\mathcal{M}_A$ to its closure in the Euclidean topology (i.e., to include distributions where one or more of the experimental outcomes has probability zero), we can always find an MLE, as the likelihood is continuous and $\overline{\mathcal{M}_A}$ is a compact set. We will call the MLE on this set the *extended MLE*. The next theorem will provide geometric conditions related to the stability of the torus action, which shows when we can find a true MLE and when we must default to the extended MLE.

Theorem 3.6 (MLE existence)**.** *Let $u \in \mathbb{Z}^m_{\geq 0}$ be a vector of counts of the observed experimental outcomes that sum to n. Let $A \in \mathbb{Z}^{d\times m}$ encode these experimental outcomes, and let $\mathcal{M}_A$ be the associated log-linear model such that the uniform distribution exists in the model. Then, the stability under the torus GT_d with matrix nA and linearization $b = Au$ is related to the following:*

(a) $\mathbf{1}$ *unstable, which does not happen*
(b) $\mathbf{1}$ *semistable if and only if the extended MLE exists*
(c) $\mathbf{1}$ *polystable if and only if the MLE exists*
(d) $\mathbf{1}$ *stable, which does not happen*

Proof. First, observe that $\mathbf{1}$ unstable does not happen. $\mathbf{1}$ would be unstable if and only if $b \notin P_{\mathbf{1}}(nA) = P(nA)$. But notice that $Au = b$, so we divide $u/n = \bar{u}$, and multiply A by n, we get a vector such that the entires sum to 1, we get $nA\bar{u} = b$, so $b \in P(nA)$. Now, we will show that $\mathbf{1}$ is never stable under the action. To do so, we will show that the interior of this polytope is empty. By assumption $\mathbf{1}$ is in the rowspace of A (so it is in the rowspace of nA). Therefore, all columns of A lie on the same hyperplane defined by $r_1x_1 + \cdots + r_dx_d = \mathbf{1}$. As such, any affine combination of the vectors that lie in this hyperplane will also like in the hyperplane, since we're taking a weighted sum of the constant that defines the hyperplane.

Note that $\mathbf{1}$ is now either going to be semistable or polystable. If we show that the MLE exists if and only if $\mathbf{1}$ is polystable, then we will have shown also shown (b), as the extended MLE always exists. Suppose $\mathbf{1}$ is polystable. This is identified with $b = Au \in \mathrm{relint}(P_{\mathbf{1}}(nA)) = \mathrm{relint}(P(nA))$. We will show that this implies MLE existence. Polystable means that there is some convex combination of the columns of nA such that $Au = nAv$ with the entries of v strictly positive that sum to 1 [Bro12, Chapter 3]. Recall that an MLE is a solution to the equation $Aq = A\bar{u}$, where q lies in $\mathcal{M}_A$ (i.e. requires all positive entries). Dividing through, we have that $A\frac{u}{n} = Av$, and so the v that ensures that Au is in the relative interior of $P(nA)$ is the q that satisfies the MLE equation.

Finally, suppose the MLE does exist. In this case, we have a strictly positive solution to the equation $Aq = A\bar{u}$. Then, clearly, $nAq = Au$, which means that $Au \in \mathrm{relint}P(nA)$ and the action on 1 is polystable. □

Note that in the case when u contains zeros, but the action is still polystable with respect to the linearization Au, it is possible to find another observed experiment count that results in the same counts of colors as that experiment count with zeros for some of the potential experiment. And, by the convexity of the likelihood function, this MLE will be unique.

It is possible to give another geometric condition for MLE existence, this time in terms of the semistability of the vector of counts, which is more intuitive than using a vector of all 1's.

Theorem 3.7. *Let $u \in \mathbb{Z}^m_{\geq 0}$ be a vector of counts of the observed experimental outcomes that sum to n. Let $A \in \mathbb{Z}^{d\times m}$ encode these experimental outcomes. Then, the MLE exists if and only if there is some $b \in \mathbb{Z}^d$ where $b = Av$ for some $v \in \mathbb{R}^m_{>0}$, such that u is semistable for the torus action given by nA with linearization b.*

Proof. Assume the MLE exists. First of all, we know that $Au = b$ lies in the polytope $P_u(nA)$. This is because we are only considering columns that have a corresponding nonzero entry in u itself. Au is obviously in the span of only those columns, and $P_u(nA)$ is an affine combination of those columns. Shifting the n to multiply the weights of the affine

combination matches up the two terms. Using criterion (b) in Theorem 3.5, u is semistable with respect to the linearization Au, as $Au \in P_u(nA)$. We must now show that u has all positive real entries and meets the criteria for v. We may use the previous theorem, noting that Au is in the relative interior of $P(nA)$, so shifting the n to the affine weights yields the claim. Therefore, if the MLE exists, u is semistable with respect to the torus action given by nA with linearization $b = Au$.

Now, suppose that u is semistable for the torus action given by nA for some linearization $b = Av$ with $v \in \mathbb{R}^d_{>0}$. That is, there is some $Av \in P_u(nA)$, which in turn means that there is some u^* with support matching the indices where u is nonzero, and with the sum of entries equalling 1, such that $Av = nAu^*$. We are free to set this $u^* = \frac{u}{n} = \bar{u}$ as $\frac{u}{n}$ meets both criteria for u^*. Doing so provides the solution to the MLE equation $Aq = A\bar{u}$, where $q = \frac{v}{n}$. □

This means that the MLE exists if and only if there exists some completely positive "count" vector for experimental outcomes such that when you subtract the induced color counts from the color counts of the original experimental outcomes, and take the "difference probability" of the experimental outcomes as follows,

$$\begin{bmatrix} \lambda_1^{a_{11}-b_1}\cdots\lambda_d^{a_{d1}-b_d} & & & & \\ & \lambda_1^{a_{12}-b_1}\cdots\lambda_d^{a_{d2}-b_d} & & & \\ & & \lambda_1^{a_{13}-b_1}\cdots\lambda_d^{a_{d3}-b_d} & & \\ & & & \ddots & \\ & & & & \lambda_1^{a_{1m}-b_1}\cdots\lambda_d^{a_{dm}-b_d} \end{bmatrix},$$

the original observed experimental outcomes under these difference in color counts probabilities, for any such color probabilities, have a nonzero maximum lower bound probability. Though interesting, I am not sure (yet) if this provides me with any additional statistical intuition.

4. Bayesian Formulations

We have thus far assumed that the value of p, and therefore the values of θ that induce p via the monomial map are fixed, but unknown quantities. But what if they were random variables themselves and followed some sort of distribution that we had prior knowledge about? How would observing counts of experimental outcomes change our intuition about our prior beliefs of the distribution of these experimental outcomes? This is the central question of Bayesian statistics, where parameters of generating processes are treated as variable in and of themselves. In this section, I give a brief introduction to Bayesian formulations in classical statistics and present some results regarding the algebra of Bayesian log-linear models.

4.1. Bayesian preliminaries.

We will introduce Bayesian updating through the example of a binomial distribution with parameter θ that is a random variable. Consider the setup in Example 2.1, the binomial distribution with parameter θ, corresponding to a series of coin flips. The likelihood of the data, given the parameter θ is

$$p(k \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

and the goal is to learn θ, the probability of landing on "heads."

Let us assume that we have a prior belief about θ in that it follows some distribution $p(\theta)$. Because θ is not fixed and follows a distribution, we do what's called "turning the Bayesian crank" to find a posterior distribution for the parameter. The posterior distribution is the distribution of the parameter conditional on the observed data. By Bayes' theorem, we have

$$p(\theta \mid k) = \frac{p(k \mid \theta)p(\theta)}{p(k)}.$$

Note that $p(k) = \int p(k \mid \theta)p(\theta)d\theta$, as this expression marginalizes out θ. Oftentimes, this integral is a tedious and possibly intractable computational task, and indeed, there are many statistical tools for doing so, as well as connections found to algebra and geometry [Lin11]. But often, there exist models that can represent prior beliefs in the data that result in the posterior belonging to the same family of models. This motivates the notion of a *conjugate prior*.

Definition 4.1 (Conjugate prior)**.** If, given a likelihood $p(x \mid \theta)$, the posterior distribution $p(\theta \mid x)$ is of the same family as the prior $p(\theta)$, the prior $p(\theta)$ is said to be conjugate with the likelihood.

Importantly, as they are chosen for algebraic convenience, conjugate priors allow us to disregard the integral in the denominator of Bayes theorem as a known constant because we can read off the posterior based on the kernel of the likelihood and the kernel of the conjugate prior. In the case of the previous example, the prior family that's conjugate with the binomial likelihood is the beta prior, resulting in the beta-binomial model. We show the conjugate update for the beta-binomial model below.

Example 4.2 (Beta-binomial update)**.** We have a prior on θ as

$$p(\theta) = \text{dbeta}(\alpha, \beta)$$

and the likelihood for the trials is

$$p(k \mid \theta) = \text{dbinom}(\theta).$$

We have that

$$\begin{aligned} p(\theta \mid k) &\propto p(k \mid \theta)p(\theta) \\ &\propto \theta^k(1-\theta)^{n-k}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \\ &\propto \text{dbeta}(\alpha+k, \beta+n-k). \end{aligned}$$

This is the posterior distribution for θ given the prior, the sampling model and the observed data.

4.2. **Conjugate setup in log-linear models.** A Bayesian formulation, in a geometric sense, is about putting a probability measure over points in a simplex which contains the acceptable parameters for a model. In the case of the log-linear model, this can be done both at the level of colors (i.e. a probability measure over points in Δ_{d-1}) or at the level of experiments (i.e. a probability measure over points in Δ_{m-1}). The setup in [YAK+21] concerning models and estimation is at the level of experimental outcomes, the distributions on which are said to be parametrized by a distribution on the colors. As such, we seek to put a tractable, conjugate prior over $p \in \mathcal{M}_A$, which is to say, we seek to put a prior distribution

over the experimental outcomes. While it certainly possible to do this, in statistics, it is much more natural to put a prior over the distribution of the colors. This is because a statistician would interpret a distribution over the colors as the parameters of the model. Indeed, in [YAK+21], p is described as being parameterized by θ via the monomial map. We wish first to show the conditions under which these setups are equivalent given the standard map that takes a given distribution over colors to its induced distribution on the experimental outcomes. In a statistical sense, this reduces to when a (possibly empirical) distribution colors is identifiable from a (possibly empirical) distribution over the experimental outcomes. In an algebraic sense, given a distribution $p \in \Delta_{m-1}$, we should have $\psi^{-1}(p)$ the fibers contain only one element. In other words, the map

$$\psi : \Delta_{d-1} \to \Delta_{m-1}$$

should be injective. We provide the conditions for this in the following claim. Recall from [Wad24] that moving between a distribution on colors to a distribution on experimental outcomes is given by the torus element $\theta = (\theta_1, \dots, \theta_d)$ acting on the all-ones vector $\mathbf{1}$ with matrix A and the trivial linearization.

Proposition 4.3. *As in Definition 3.3, let A be a matrix in $\mathbb{Z}^{d\times m}$ and θ be an element in Δ_{d-1}^+. Then, the map $\psi : \Delta_{d-1}^+ \to \Delta_{m-1}$ under the action*

$$\theta \mapsto \begin{bmatrix} \theta_1^{a_{11}} \cdots \theta_d^{a_{d1}} & & & & \\ & \theta_1^{a_{12}} \cdots \theta_d^{a_{d2}} & & & \\ & & \theta_1^{a_{13}} \cdots \theta_d^{a_{d3}} & & \\ & & & \ddots & \\ & & & & \theta_1^{a_{1m}} \cdots \theta_d^{a_{dm}} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

is injective if and only if A is of rank d.

Proof. Suppose ψ is injective. Then, for $\psi(\theta) = \psi(\theta')$, we must have $\theta = \theta'$. Towards a contradiction, suppose that the rank of A is not d. Then, consider for some $\psi(\theta) = \psi(\theta')$, we have

$$\begin{bmatrix} \theta^{a_1} - \theta'^{a_1} & & & & \\ & \theta^{a_1} - \theta'^{a_1} & & & \\ & & \theta^{a_1} - \theta'^{a_1} & & \\ & & & \ddots & \\ & & & & \theta^{a_1} - \theta'^{a_1} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 0$$

where $\theta = (\theta_1, \dots, \theta_d)$ and $\theta' = (\theta'_1, \dots, \theta'_d)$. This implies that each $\theta^{a_i} = \theta'^{a_i}$ for each $i \in [d]$. Further, we have $\log \theta^{a_i} = \log \theta'^{a_i}$ which is unique, as θ, θ' are in the positive orthant. This can be rewritten as

$$a_{i1}(\log \theta_1 - \log \theta'_1) + \cdots + a_{id}(\log \theta_d - \log \theta'_d) = 0. \tag{1}$$

Now, suppose that the rank of A is not d. Then, there is some nonzero $c = (c_1, \dots c_d)$ such that $ca_{i1} + \cdots + ca_{id} = 0$. Equipped with c, we can find θ and θ' that are not equal, where the images are equal with (1) holding, violating injectivity. Thus, A must have rank d. Suppose now that A has rank d and towards a contradiction, suppose that the map is not injective. Let θ and θ' be nonequal but let their images $\psi(\theta)$ and $\psi(\theta')$ be equal. Then, (1) must hold. But because the rank of A is d, the coefficients must be 0, implying that $\theta = \theta'$. So, ϕ is injective. □

Just as in the coin flipping case, where the beta family allows us to tractably represent different beliefs in the values of θ and by extension $1-\theta$ by choosing α and β, the Dirichlet prior allows us to do the same for $\gamma = (\gamma, \dots, \gamma_k)$ such that $\sum_{i=1}^{k} \gamma_k = 1$. In other words, the Dirichlet distribution given parameters $\alpha_1, \dots, \alpha_k > 0$ is a conjugate prior distribution over the probability simplex

$$\Delta_{k-1} = \left\{ \gamma \in \mathbb{R}^k \mid \gamma_i \geq 0 \text{ for all } j \text{ and } \sum_{i=1}^{d} \gamma_i = 1 \right\}.$$

Algebraically speaking, there prior distribution on Δ_{k-1} is a map

$$\xi : \Delta_{k-1} \to \mathbb{R}_{>0}$$

where by the Dirichlet distribution,

$$\gamma \mapsto \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{k} \gamma_i^{\alpha_i - 1}.$$

For a likelihood over k states, the posterior distribution $p(\gamma \mid u)$ can be written as

$$p(\gamma, u) \propto \prod_{i=1}^{k} \gamma_i^{\alpha_i + u_i - 1}. \tag{2}$$

4.3. **Bayesian algebraic statistics setup.** We will now present a brief and simple calculation to show that the Bayesian setup "factors through," in the sense that the maximum a posteriori estimate (MAP), which is the maximizer of the posterior distribution in the parameter, exists under shifted conditions as the MLE. Moreover, under a conjugate prior for the toric model, we show that the Bayesian analogue of the MLE, the MAP always exists when any sort of informative prior is assumed on and is analogous to the regular MLE. Recall that the issue with simply turning the Bayesian crank as in (2) is that in the log-linear model, data are given in the form of $u \in \mathbb{Z}^m$ but the parameters are given in the form of $\theta \in \mathbb{R}^d$.

But using Proposition 4.3, we can directly assume that there is a prior distribution over $\mathcal{M}_A$, which will immediately be identifiable as a prior distribution over Δ_{d-1}, as long as we assume that A is of full rank. That is, let each point $p \in \mathcal{M}_A$ be distributed $p \sim \pi_0$, and assume that $\pi_0 \sim \text{Dirichlet}(\alpha_1 \dots \alpha_m)$.

Then, the posterior of p, with p now being treated as the "parameters of interest," is

$$p(p \mid u) = p^u p^\alpha.$$

One can identify this function as a likelihood with the data $L(u + \alpha; p)$, where

$$L(u + \alpha; p) = \theta_1^{a_1 \cdot (u+\alpha)} \theta_d^{a_d \cdot (u+\alpha)},$$

where $a_1, \dots a_d$ are the rows of the matrix A. Recall that the maximizer is available in the variable p, where this maximizing value of p is contained in the model $\mathcal{M}_A$ by the conditions in Theorem 3.6. The only difference is that instead of the torus acting on $\mathbf{1}$ through the matrix nA with linearization $Au = b$, to determine the existence of the MLE, the torus acts on $\mathbf{1}$ with linearization $A(u + \alpha) = b'$. Then, note that if α is nonzero, which corresponds to an informative prior, there is no question that a MAP can always be found, as 1 will always be polystable.

The proof is as follows. Suppose that $\mathbf{1}$ is not polystable under the action. Then,

$$b' = A(u + \alpha) \notin \text{relint}(P_{\mathbf{1}}(nA)).$$

However, α is nonzero and $u \in \mathbb{N}^M$, so clearly there is a positive solution v to $b' = nAv$. This solution is $v = \frac{1}{n}(u + \alpha)$, which is the MAP when an informative prior is placed.

5. Confidence Bounds: Sampling from conditional distributions

Beyond the MLE, which in statistics is known as a *point estimate* of a parameter, what is also of interest is determining ranges where the true parameter could lie. These ranges, in the frequentist paradigm, are known as confidence intervals. The procedure for their creation yields that with a certain probability (i.e. before observing the data) the true parameter is contained within the interval. Therefore, after observing the data and using the procedure, an analyst can say with a certain degree of confidence that a computed bound contains the true parameter that is to be learned. In many cases, computing such bounds relies on asymptotic results concerning the Central Limit Theorem. For example, the fact that a sampling distribution of a sequence of i.i.d. random variables converges after scaling and shifting to a standard normal distribution, allows for the construction of confidence intervals for a parameter that coincides with the expectation of one of these random variables in many large N settings.

But what happens when it is inappropriate to appeal to the Central Limit Theorem or another such asymptotic result? This is often the case in so-called small N settings, such as that of clinical trials and associated experiments. Often, these settings correspond to multi-way contingency tables, the sampling distribution for which is often assumed to be a multinomial distribution, which of course are log-linear models.

In this section, we provide an exposition of a foundational paper in algebraic statistics, which presents the backbone of a new approach to produce accurate confidence intervals, particularly in small N settings. The approach to construct confidence intervals is as follows. For every distribution in a discrete exponential family, there exists a sufficient statistic which can be computed from the data. One can sample from the original distribution conditioned on the sufficient statistic, which combines a notion of variability with that of what is observed through the data. Sampling from this exact distribution can be done a large number of times with the construction of a Markov chain; computing estimates of parameters from each sample can yield credible intervals for the parameter of interest. In [DS98], methods for constructing the Markov chains are presented and are tied to computational algebraic geometry. To construct the Markov chains, what is needed is a *Markov basis*. Finding the Markov basis is equivalent to finding the generators for an ideal in a ring of polynomials, which is where the connection to algebra arises.

5.1. General discrete exponential families.

Following [DS98], consider a general discrete exponential family over a finite set $\mathcal{X} \subset \mathbb{R}^m$,

$$P_\theta(x) = Z(\theta) \exp\{\theta \cdot T(x)\}$$

where $\theta \in \mathbb{R}^d$ and T, a sufficient statistic, is a map from $\mathbb{R}^m$ to $\mathbb{R}^d$. It is immediate that the standard log-linear model $\mathcal{M}_A$ falls into such a discrete exponential family. To see this, recall the monomial map given in Definition 2.3, and further note that the likelihood for

log-linear model can be written as

$$L(p;u) = p^u = \theta_1^{\sum_{j=1}^m a_{1j}u_j} \cdots \theta_d^{\sum_{j=1}^m a_{dj}u_j} = \theta_1^{a_1 \cdot u} \cdots \theta_d^{a_d \cdot u}$$

up to the normalizing constant $Z(\theta)$ for $u \in \mathbb{R}^m$. Reparametrizing $\theta_1, \ldots, \theta_d$ with the exponential operation and identifying T with the action of A on u by matrix mutiplication yields this claim. Statistically speaking, u is a vector of counts of each experimental outcome, and Au is a vector of total counts of states across all observed experimental outcomes. Reframing the goal of this section, just as the objective of [YAK+21] was to understand when it is possible to find vector(s) of counts of experimental outcomes that agree on total counts of states, the goal of [DS98] is to walk around in the distribution of such vectors.

To do so, consider the set $\mathcal{Y}_t = \{(x_1, \ldots x_n) \in \mathcal{X}^N : T(x_1) + \cdots + T(x_n) = t\}$. This is the set of possible samples from N runs of the experiment that yield a particular value t of the sufficient statistic. In an exponential family, sampling from this conditional distribution is uniform, but it is difficult to enumerate all of $\mathcal{Y}_t$. Another representation of t that will prove useful for the construction of the Markov chain over $\mathcal{Y}_t$ is

$$t = \sum_x G(x)T(x) = \sum_{i=1}^N T(x_i), \tag{3}$$

where $G(x)$ is $\#\{i : X_i = x\}$. By representing t as a sum over the entire experiment space $\mathcal{X}$, where $G(x)$ is a function that maps each possible experiment to the number of times it is observed in a given sample, one can walk over the space of functions that satisfy Equation 3. These equations are equivalent to different counts of experimental outcomes that yield the value t of the statistic T over N samples. Therefore, let

$$\mathcal{F}_t = \{f : \mathcal{X} \to \mathbb{N} : \sum_x f(x)T(x) = t\}.$$

The image of the uniform distribution on $\mathcal{Y}_t$ under the map from $\mathcal{Y}_t$ to $\mathcal{F}_t$ is the hypergeometric distribution

$$H_t(f) = \frac{N!}{|\mathcal{Y}_t|} \prod_x (f(x)!)^{-1}.$$

To sample from this distribution using a Markov chain, we use a *Markov basis*, for which we now provide a definition.

Definition 5.1 (Markov basis)**.** A Markov basis is a set of functions $f_1, \ldots, f_L : \mathcal{X} \to \mathbb{Z}$ such that

$$\sum_x f_i(x)T(x) = 0.$$

Moreover, for any t and $f, f' \in \mathcal{F}_t$, there exist $(\epsilon_1, f_{i_1}), \ldots, (\epsilon_A, f_{i_A})$, where $\epsilon_i \in \{-1, 1\}$, such that

$$f' = f + \sum_{j=1}^A \epsilon_i f_{i_j} \text{ and } f + \sum_{j=1}^a \epsilon_j f_{i_j} \geq 0 \text{ for } 1 \leq a \leq A. \tag{4}$$

To sample from $\mathcal{F}_t$, we use the basis as follows. From a starting $f \in \mathcal{F}_t$, pick a number l uniformly at random in $\{1, \ldots, L\}$, and $\epsilon \in \{-1, 1\}$ in the same way. Then, form $f + \epsilon f_l$. If $f + \epsilon f_l$ is nonnegative, the chain moves there. Otherwise, the chain stays at f. Note that due to the first condition, the chain stays in $\mathcal{F}_t$, because one can factor out the ϵ and recognize

that $\sum_x f_l(x)T(x) = 0$, thereby not changing the value of the sufficient statistic. In order to devote more time to the algebra associated with finding this Markov basis, we leave out details concerning the fact that this algorithm results in an irreducible, aperiodic Markov chain with stationary distribution $H_t(f)$, corresponding to the uniform distribution over $\mathcal{Y}_t$.

To construct the Markov basis, we need the following further notation. For each $x \in \mathcal{X}$, introduce an indeterminate also denoted x in the ring of polynomials $k[\mathcal{X}]$ over any field k. A function $g(x)$ will be represented by monomials $\prod_x x^{g(x)}$ in this polynomial ring and will be denoted $\mathcal{X}^g$. The sufficient statistic T will be further represented by a homomorphism of polynomial rings

$$\phi_T : k[\mathcal{X}] \to k[t_1, \dots t_d]$$

where

$$x \mapsto t_1^{T(x)_1} t_2^{T(x)_2} \cdots t_d^{T(x)_d}.$$

Recall that $T : \mathcal{X} \to \mathbb{R}^d$. In this notation, $T(x)_i$ corresponds to the ith coordinate of the sufficient statistic $T(x)$. A key ingredient is that any function $f : \mathcal{X} \to \mathbb{Z}$ can be written as $f^+ - f^-$,where $f^+ = \max(f(x), 0)$ and $f^- = \max(-f(x), 0)$. Finally, let $\mathcal{F}_T = \{p \in k[\mathcal{X}] : \phi_T(p) = 0\}$ (note the capital T in the subscript). Now, observe that $\sum_x f(x)T(x) = 0$ holds if and only if the difference $\mathcal{X}^{f+} - \mathcal{X}^{f-} \in \mathcal{F}_T$, where $\mathcal{F}_T$ is represented an ideal in $k[\mathcal{X}]$. To see this, first pick f such that the sum condition holds and split it into $f^+ - f^-$. Recall, for the monomial $\prod_x x^{g(x)}$, we have that $T : \mathcal{X} \to \mathbb{N}^d$ is represented by the map ϕ_T, which takes

$$\prod_x x^{g(x)} \mapsto \prod_x (t_1^{T(x)_1} \cdots t_d^{T(x)_d})^{g(x)} = t_1^{\sum_x T(x)_1 g(x)} \cdots t_d^{\sum_x T(x)_d g(x)}$$

by the homomorphism rules. So, we have,

(5)

$$T\left(\prod_x x^{f^+(x)} - \prod_x x^{f^-(x)}\right) = t_1^{\sum_x T(x)_1 f^+(x)} \cdots t_d^{\sum_x T(x)_d f^+(x)} - t_1^{\sum_x T(x)_1 f^-(x)} \cdots t_d^{\sum_x T(x)_d f^-(x)}.$$

In order for $\sum_x f(x)T(x) = \sum_x (f^+(x) - f^-(x))T(x) = 0$, the terms $\sum_x f^+(x)T(x)$ and $\sum_x f^-(x)T(x)$ must agree on every coordinate of $T(x)$. Given this fact, we can write the above as

$$t_1^{a_1} \cdots t_d^{a_d} - t_1^{a_1} \cdots t_d^{a_d} = 0,$$

which means $\mathcal{X}^{f^+} - \mathcal{X}^{f^-} = 0$. Having written the first direction out, the second is straightforward. Suppose there is a function f such that Equation 5 equals zero. Recognizing that the sums must be the same value for this to happen, coordinate-wise, one can rearrange and use the distributive property to yield the claim.

Now, consider a collection of functions $f_1, \dots f_L$ and the monomial differences $\mathcal{X}^{f_i^+(x)} - \mathcal{X}^{f_i^-(x)}$. A stronger statement about these monomial differences encapsulates when $f_1, \dots, f_L$ is a Markov basis.

Theorem 5.2. *The collection of functions $f_1, \dots, f_L$ constitutes a Markov basis if and only if the set*

$$\mathcal{B} = \{\mathcal{X}^{f_i^+} - \mathcal{X}^{f_i^-} : 1 \le i \le L\}$$

generates the ideal $\mathcal{F}_T$.

To prove Theorem 5.2, we first need a lemma.

Lemma 5.3. *Consider $\mathcal{F}'$, the ideal generated by the monomial differences for functions f where $\sum_x f(x)T(x) = 0$,*

$$\mathcal{F}' = \langle \mathcal{X}^{f^+} - \mathcal{X}^{f^-} : \sum_x f(x)T(x) = 0 \rangle.$$

We have that $\mathcal{F}' = \mathcal{F}_T$.

Proof. Observe from above that $\mathcal{F}' \subseteq \mathcal{F}_T$. To show that $\mathcal{F}_T \subseteq \mathcal{F}'$, put a total order on the set of monomials by ordering the variables and then saying that one monomial is larger than the other by first checking the degree. If the degrees are equal, the monomial that has the higher power on the first variable on which the two monomials disagree is said to be larger.

Towards a contradiction, suppose that $\mathcal{F}_T$ is not contained in $\mathcal{F}'$. Then, there exists $p \in \mathcal{F}_T - \mathcal{F}'$. However, because $p \in \mathcal{F}_T$, for the largest monomial $\mathcal{X}^\alpha$ of p, there exists a corresponding monomial $\mathcal{X}^\beta$ such that $\phi_T(\mathcal{X}^\alpha) = \phi(\mathcal{X}^\beta)$, by definition of $\mathcal{F}_T$. Factoring out by the common variables of $\mathcal{X}^\alpha$ and $\mathcal{X}^\beta$ yields $\mathcal{X}^\gamma(\mathcal{X}^{\alpha'} - \mathcal{X}^{\beta'})$. Now, under the map ϕ_T, $\mathcal{X}^{\alpha'} - \mathcal{X}^{\beta'} = 0$. As such, letting $f(x) = \alpha'(x) - \beta'(x)$ yields a function where $\sum_x f(x)T(x) = 0$. We note that $\alpha'(x)$ takes the role of $f^+(x)$ and $\beta'(x)$ takes the role of $f^-(x)$ because after factoring $\mathcal{X}^\gamma$, the two remaining terms have disjoint support. Moreover, subtracting a multiple of $\mathcal{X}^\alpha - \mathcal{X}^\beta$ from p yields a polynomial in fewer terms and a smaller leading monomial. Repeating this process yields the claim that $\mathcal{F}' = \mathcal{F}_T$. □

We note that the techniques used in the proof of Lemma 5.3 is very similar in spirit to the proof for Proposition 2.5. In fact, the proof of Proposition 2.5 is almost a special case of the proof of the above lemma.

Proof of Theorem 5.2. To show that $\mathcal{B}$ is a Markov basis if and only if it generates $\mathcal{F}_T$, observe first that the first condition $\sum_x f_i(x)T(x) = 0$ is satisfied if and only if $\mathcal{B} \subset \mathcal{F}_T$. Then, it remains to be shown that the connectivity properties in (4) hold if and only if $\mathcal{B}$ as a set of monomial differences in $k[\mathcal{X}]$ generates $\mathcal{F}_T$ as an ideal in $k[\mathcal{X}]$. Suppose (4) holds. Then, using the fact that $\mathcal{F}' = \mathcal{F}_T$ from the first part of the proof, it suffices to show that $\mathcal{B}$ generates the generators of $\mathcal{F}'$. In other words, pick a function f such that $\sum_x f(x)T(x) = 0$. Then, the associated monomial difference $\mathcal{X}^{f^+} - \mathcal{X}^{f^-}$ must be in the ideal generated by $\mathcal{B}$. Once this is true, any element in the ideal $\mathcal{F}' = \mathcal{F}_T$ can be constructed by the elements in the set of monomial differences $\mathcal{B}$.

Because (4) holds, pick an admissible f, set t to some value and let $g = f^+$ and $g' = f^-$, with the goal of the monomial differences in $\mathcal{B}$ generating. Note that this is admissible because $\sum_x f^+(x)T(x) - \sum_x f^-(x)T(x) = 0$, so the sums of the sufficient statistics over the counts of each experimental outcome must be the same for some value of t, and so for any f such that $\sum_x f(x)T(x) = 0$, f^+ and f^- lie in $\mathcal{F}_t$ for some t. Using (4), we have that

$$g' = g + \sum_{j=1}^{A} \epsilon_j f_{i_j} \text{ and } g + \sum_{j=1}^{a} \epsilon_i f_{i_j} \geq 0 \text{ for } 1 \leq a \leq A.$$

Suppose $A = 1$. If $\epsilon_1 = 1$, then $f^- = f^+ + f_{i_1}$, where $f_{i_1} \in f_1, \ldots, f_L$. Rearranging, $f^- - f^+ = f_{i_1} = f_{i_1}^+ - f_{i_1}^-$. Identifying each of these functions with their monomials and negating yields the claim for $A = 1$ and $\epsilon_1 = 1$ Further, a similar claim holds for $A = 1$ and $\epsilon_1 = -1$. For cases in which $A > 1$, one can always subtract the first term $\epsilon_j f_{i_j}$ from g'.

To show the other direction, suppose $\mathcal{B}$ as a set of monomial differences generates $\mathcal{F}_T$ as an ideal. Pick g, g' and some t. We must show that if $g, g' \in \mathcal{F}_t$, then it is possible to move between g and g' using $f_1, \ldots, f_L$ in the appropriate way of (4). To begin, note that if $g, g' \in \mathcal{F}_t$, then

$$\sum_x (g(x) - g'(x))T(x) = 0.$$

By a similar calculation as in (5), first represent $p = g - g'$ as $p = \mathcal{X}^g - \mathcal{X}^{g'}$ and recognize that this is in the kernel of ϕ_T, so $\mathcal{X}^g - \mathcal{X}^{g'} \in \mathcal{F}_T$. Then note that because $\mathcal{B}$ generates $\mathcal{F}_T$,

$$\mathcal{X}^g - \mathcal{X}^{g'} = \sum_{j=1}^{A} \epsilon_j \mathcal{X}^{h_j}(\mathcal{X}^{f_{i_j}^+} - \mathcal{X}^{f_{i_j}^-})$$

where $h_j : \mathcal{X} \to \mathbb{N}$. Note that we can put $\epsilon_j = \pm 1$ because one could always repeatedly sum for any coefficient of the outside monomial that is not 1. Now, note that $\mathcal{X}^g = \mathcal{X}^{h_r}\mathcal{X}^{f_{i_r}^-}$ for some r. Observe that $g - f_{i_r}^-$ is nonnegative, as $g = h_r + f_{i_r}^-$. Moreover, $g + f_{i_r} = g + f_{i_r}^+ - f_{i_r}^- \implies g + f_{i_r} = h_r + f_{i_r}^+$. Now, subtract the term $\mathcal{X}^{h_r}(\mathcal{X}^{f_{i_r}^+} - \mathcal{X}^{f_{i_r}^-})$ from both sides, this cancels the $\mathcal{X}^g$ term and results in a term $\mathcal{X}^{g+f_{i_r}}$ and an expression $\mathcal{X}^{g+f_{i_r}} - \mathcal{X}^{g'}$. The corresponding expression on the right hand side has 1 less term. If one keeps subtracting terms in this manner, it is possible to link $\mathcal{X}^{g+f_{i_r}}$. □

5.2. **Toric model interpretation.** Having stated and proven the theorem in a general sense, we now connect it back to toric models. In particular, suppose we have a toric model that satisfies the conditions of Proposition 2.5 for a matrix $A \in \mathbb{Z}^{d\times m}$. Then, $\mathcal{F}' = I_A$. To see this, pick a generator of $\mathcal{F}'$. That means, for the associated f^+ and f^-,

$$\begin{aligned}\sum_x (f^+(x) - f^-(x))T(x) &= \sum_x f^+(x)T(x) - \sum_x f^-(x)T(x) \\ &= A \cdot (f^+(x_1), \ldots, f^+(x_m)) - A \cdot (f^-(x_1), \ldots, f^-(x_m))\end{aligned}$$

by a vector representation of f^+ and f^- over the finite space $\mathcal{X}$. Observe that when treated as a vector, which is exactly what $\mathcal{X}^{f+}$ and $\mathcal{X}^{f-}$ is doing, we have found two vectors $u = f^+$ and $v = f^-$ such that $Au = Av$, where u and v are both in $\mathbb{N}^m$ by the definitions of f^+ and f^-. Therefore, $\mathcal{X}^{f^+} - \mathcal{X}^{f^-}$ is in I_A. For the other direction, pick $p^u - p^v$ such that $Au = Av$. We will show that $p^u - p^v$ lies in $\mathcal{F}_T$, which is equivalent to $\mathcal{F}'$. By a similar calculation as in (5), we get that $\phi_T(p) = 0$. As the members of each set of generators lies in the other ideal, we have shown the equality.

This correspondence allows us to use Theorem 5.2 as follows. Identifying $f_1, \ldots, f_L$ such that $\sum_x f_i(x)T(x) = 0$ in the setting of Theorem 5.2 with vectors in the kernel of A, we have that these vectors are a Markov basis if and only if the set

$$\{p^{f_i^+} - p^{f_i^-}\}$$

generates $I_A = \mathcal{F}_T = \mathcal{F}'$. I_A, of course, is generated by $p^u - p^v$ such that $Au = Av$. Therefore, finding a Markov basis is equivalent to finding vectors of counts of experimental outcomes where the suffficient statistics match for any value of the sufficient statistic. Doing so allows us to walk in the space of counts of experimental outcomes that yield a particular sufficient statistic.

6. Updating to Zero

In this section, we will revisit the recurring theme of "zeros" in count vectors of experimental outcomes in the log-linear model, and how algebra can help us bypass them. Somewhat philosophically speaking, the question then becomes: *should* we bypass the zeros? Or, in other words, what happens if it is true that a certain experimental outcome or cell in a contingency table has zero probability, either structurally or otherwise. As it turns out, detecting these cells was first proposed in an algebraic way, leveraging similar formulations as to those presented in Section 5. The basic idea, from my understanding, is to define a "maximal" toric model using the generators of the orthogonal space to basis of functions f such that $\sum_x f(x)T(x) = 0$. Models are then presented as *mixtures* of models with certain parameters set to zero, each with some prior probability, and Bayes factors are used to provide evidence in favor of a model with zero-probability cells or not.

Understanding these formulations with respect to both log-linear models (i.e. discrete models and contingency tables) as well as the broader problem of pushing posterior probabilities of events to zero when they "should be" is an area of possible future work.

References

[AKRS21] Carlos Améndola, Kathlén Kohn, Philipp Reichenbach, and Anna Seigal. Invariant theory and scaling algorithms for maximum likelihood estimation. *SIAM Journal on Applied Algebra and Geometry*, 5(2):304–337, 2021.

[Bro12] Arne Brondsted. *An introduction to convex polytopes*, volume 90. Springer Science & Business Media, 2012.

[DS98] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of statistics*, 26(1):363–397, 1998.

[Lin11] Shaowei Lin. *Algebraic methods for evaluating integrals in Bayesian statistics*. University of California, Berkeley, 2011.

[PS05] Lior Pachter and Bernd Sturmfels. *Algebraic statistics for computational biology*, volume 13. Cambridge university press, 2005.

[Sul23] Seth Sullivant. *Algebraic statistics*, volume 194. American Mathematical Society, 2023.

[Wad24] Adway Wadekar. Toric geometry of log-linear models. Unpublished manuscript, 2024.

[YAK+21] Mom Youg, Carlos Améndola, Kathlén Kohn, Philipp Reichenbach, and Anna Seigal. Toric invariant theory for maximum likelihood estimation in log-linear models. *Algebraic Statistics*, 12(2):187–211, 2021.

Departments of Mathematics and of Statistical Science, Duke University, Durham, NC, 27708