Statistical Intuition Behind Algebraic Structure in Log-Linear Models with Views Toward Bayesian Formulations

ADWAY S. WADEKAR

Abstract. This is an expository article on what I studied in a reading/research independent study under the direction of Prof. Ezra Miller towards graduation with distinction in mathematics. We describe the statistical intuition behind concepts arising in an algebraic statistics paper on 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. As such, given the existence of a conjugate prior for the multinomial distribution, we show that the algebraic and geometric findings in the log-linear model setup are ripe for consideration from a Bayesian viewpoint, particularly from the perspective of inducing a probability measure over a group orbit.

1. Introduction

Algebraic statistics is a nascent field that lies at the intersection of algebraic geometry and statistics. The fundamental premise that spurred its creation is that many statistical problems can be formulated as algebraic ones. In recent years, algebraic statistics has grown tremendously, with a full Graduate Studies in Mathematics textbook having being published in 2018 [Sul23]. In addition, there have been connections made between statistics, algebra and evolutionary biology, which are most unexpected [PS05]. While the field of algebraic statistics is new, *connections* with algebra from the world of (discrete) statistics have been made since the mid-1970s.

For one, there are connections between the design of experiments and algebra. More recently, however, several objects in algebraic geometry have been found to have connections with objects in statistics. These connections range from direct correspondence to a rough correspondence. The goal of this paper is to elucidate some of the connections between torus actions and maximum likelihood estimation in log-linear models presented in a recent article. I learned the latter two terms in introductory statistics classes, and before this semester, I was unfamiliar with the former. The published paper emphasizes the geometry and the algebra, and as I was reading it, I felt I could do with some statistical intuition. Here, I provide the intuition I would have liked to have when reading this paper. I conclude my discussion with opportunities for future directions to explore the connections between Bayesian formulations of log-linear models and geometry.

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

Date: December 2024.

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 [AKRS21a], 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 [AKRS21b] 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, [AKRS21b] 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

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

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 [AKRS21b]. From now on, we will take $h = \mathbf{1}$.

The first immediate property that can be seen is that when $\mathbf{1} \in \operatorname{rowspan}(A)$ 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 [AKRS21b] 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 [AKRS21b] 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 [AKRS21b] 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 $(n_{ij})_{1\leq i,\leq d, 1\leq j\leq m}$ 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 [AKRS21b] 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

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

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

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

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 [AKRS21b, 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]. 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. Towards Bayesian Formulations

Throughout this paper, we have 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. Bayesian formulations of log-linear models is what I hope to study from an algebraic and geometric perspective in the following semester.

In this section, I give a brief introduction to Bayesian formulations in classical statistics and present some possible directions for future work.

4.1. **Bayesian Statistics: A Primer.** 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. The likelihood of the data, given the parameter θ is

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

Now, let's 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.

In the Bayesian regime, the analogue to the MLE is the maximum a posteriori estimate (MAP), which is the value of the parameter that maximizes the posterior density or probability of the parameter. This is the quantity we seek to study in an algebraic and geometric setting. MAPs have previously been studied in [Sul23, Chapter 18], for the special case of a hidden Markov model. The authors of [Sul23, PS05] have further connected MAP estimates to tropical geometry.

4.2. **Bayesian formulations in log-linear models.** 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 may be possible to do this (in fact, I suspect that one could naturally use the Dirichlet prior as I'm about to describe below), in statistics, it is much more natural to put a prior over the distribution of the colors.

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 $(\theta_1, \dots, \theta_d)$ such that $\sum_{i=1}^{d} \theta_i = 1$. In other words, the Dirichlet distribution given parameters $\alpha_1, \dots, \alpha_d > 0$ is a conjugate prior distribution over the probability simplex

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

for the colors. Algebraically speaking, there is a map

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

where by the Dirichlet distribution,

$$\theta \mapsto \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{d} \theta_i^{\alpha_i - 1}.$$

Now observe that if the colors have a prior distribution (i.e. every point in the $(d-1)$-dimensional simplex is associated with some prior probability), given a matrix $A \in \mathbb{Z}^{d \times m}$, we can induce a prior probability on the experimental outcomes directly. Recall that the probability of an experimental outcome j encoded by the column a_j of A given a specific distribution on the colors is

$$p(a_j \mid \theta) \propto \theta_1^{a_{1j}} \dots \theta_d^{a_{dj}}.$$

In the above context, the induced (fixed) probability distribution of the experiments is given by the action of the torus element θ under the matrix $A \in \mathbb{Z}^{d \times m}$ on $\mathbf{1}$, the all 1's vector.

But now, it may be somewhat bold to claim this, but we have a probability distribution over θ, which means we have a probability distribution over the torus element chosen. As the probability distribution for the experimental outcomes is equivalent to the appropriate point in the orbit of $\mathbf{1}$ for a given torus element acting, the prior distribution for the experimental states may actually be a distribution over the (possibly restricted) orbit of $\mathbf{1}$. These are the sorts of characterizations I hope to study next semester, along with discovering how to relate such geometry to MAP estimates.

References

[AKRS21a] 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.

[AKRS21b] 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.

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

[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.