Counting Gene Finding Functions

Heemal Dhanjee, Shaowei Lin

May 8, 2007

Abstract

In biology, a gene-finding function is one which takes a DNA sequence and returns the most likely underlying intron/exon structure. Generalizing to graphical models, maps from observed states to their most probable hidden state are called inference functions. While the problem of counting inference functions for a given model is not easy, it has been shown that the number of inference functions grows at most polynomially. In this report, we study the Neyman hidden Markov model of length n, and show that the number of inference functions is exactly $4(n - \lceil n/4 \rceil) + 2$.

1 Gene-finding in Biology

In all organisms, there is a process where a genomic DNA sequence is transcribed into an RNA sequence, after which the RNA sequence can ultimately go on to become some kind of functional unit. The sequences in the genome that undergo this process are what we refer to as genes, while in between genes in the genome are intergenic sequences that do not code for any functional unit. Very often, the above process also involves the translation of the RNA sequence into an amino acid sequence via a set of rules in which three RNA bases code for a particular amino acid. This sequence of amino acids thus constitutes a protein.

Genomes, however, typically contain both coding and non-coding regions. The human genome for example, has only about 5% coding sequences, by which I mean that only 5% of the genome seem to be genes, while 95% of the genome is non-coding and commonly referred to as genomic 'junk'. So, when looking at gene homology between organisms, say a human and a mouse, we only want to be looking at the coding regions of the genome. The problem of finding alignments in coding regions between genomes of different organisms can be made simpler if we know where the coding regions are, since looking for gene homology in junk sequences is senseless and causes much inefficiency.

Unfortunately, the problem of finding coding regions in eukaryotes is made more difficult because the process is much more complicated. In eukaryotes, genes from the genome get transcribed into a pre-mRNA sequences (an 'immature' RNA sequence) which consists introns and exons. What happens next is that the pre-mRNA sequence undergoes a splicing process in which the introns are spliced out, and the exons are annealed together to form the mature mRNA sequence. So, in eukaryotic organisms, the only parts of the gene DNA sequence that become functional units are the exons. Furthermore, there are other properties that characterize gene structures, such as signal sequences. For example, the nucleotide sequence ATG in a gene is the site at which translation into amino acid sequences originates, and likewise there are the known stop sequences TAA, TAG, and TGA which signal for translation to stop. There are also signal sequences that are observed almost without

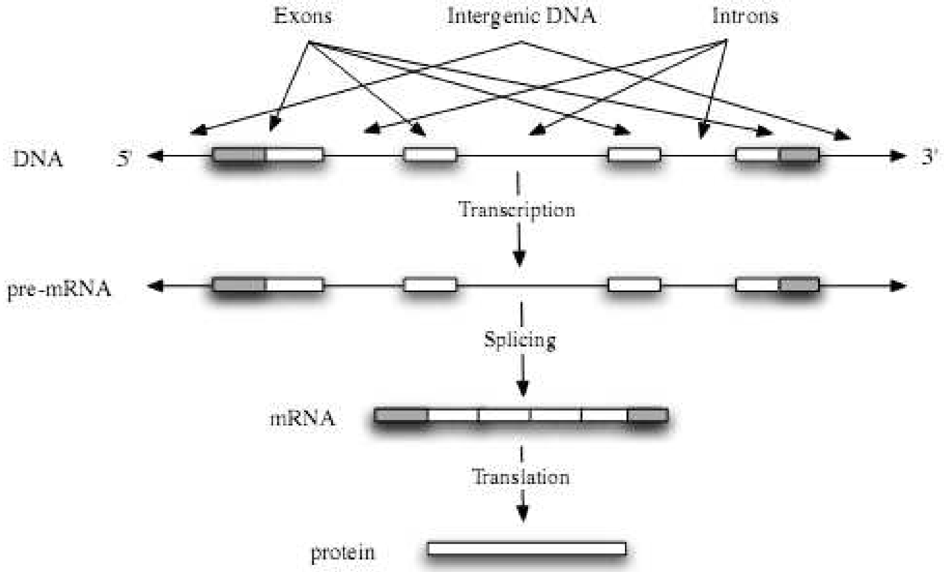

Figure 1: Gene Structure, from [7]

exception at the boundaries between introns and exons. These sequences are GT at the 5' end of an intron and AG at the 3' end.

Thus, the problem is this: given a genomic DNA sequence from an organism where the above splicing occurs, which parts of the sequence are exons, which parts are the introns, and which the intergenic sequences?

To tackle this problem, we simplify matters by assuming that there are only introns and exons, and use a homogenous hidden Markov model. Suppose we have a sequence of n nucleotides. In this model we have n hidden random variables $X_1, X_2, \ldots, X_n$ and n observed random variables $Y_1, Y_2, \ldots, Y_n$. The hidden variables take values from the alphabet $\Sigma = \{\mathtt{e}, \mathtt{i}\}$ corresponding to introns and exons, while the observed variables take values from the alphabet $\Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ corresponding to the nucleotides. As shown in Figure 2, this model is also characterized by transition matrices

$$S = \begin{pmatrix} s_{\mathtt{ee}} & s_{\mathtt{ei}} \\ s_{\mathtt{ie}} & s_{\mathtt{ii}} \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} t_{\mathtt{eA}} & t_{\mathtt{eC}} & t_{\mathtt{eG}} & t_{\mathtt{eT}} \\ t_{\mathtt{iA}} & t_{\mathtt{iC}} & t_{\mathtt{iG}} & t_{\mathtt{iT}} \end{pmatrix}.$$

which are determined by 12 probability parameters. For each position i, the transition matrix S gives us the probabilities that X_{i+1} is an intron or exon given that we have an intron/exon at X_i. Likewise, our transition matrix T denotes the probabilities that we observe each of the nucleotides $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ at Y_i in our DNA sequence given an intron/exon at X_i.

Thus, our model is a map

$$f \colon \mathbb{R}^{12} \to \mathbb{R}^{4^n}$$

from the parameter space into the probability space, where n is the length of the observed nucleotide sequence. For example, suppose $n = 3$. Then we have a map from our parameter

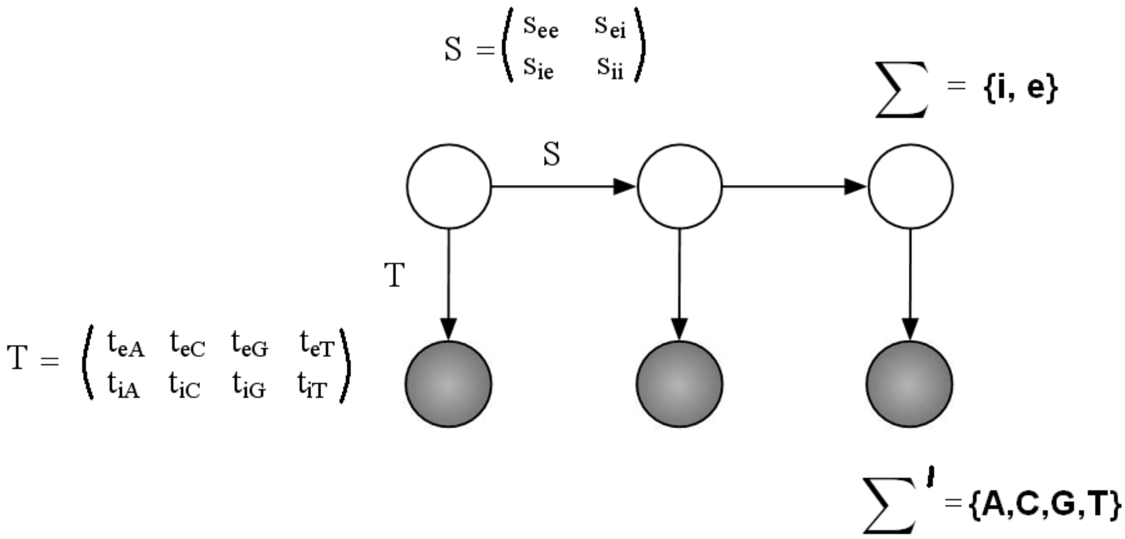
S = (s_ee s_ei / s_ie s_ii)
Σ = {i, e}
S
T
T = (t_eA t_eC t_eG t_eT / t_iA t_iC t_iG t_iT)
Σ' = {A,C,G,T}

Figure 2: Hidden Markov Model for Gene-finding Functions

space (S,T) into a probability space represented by $4^3 = 64$ coordinate polynomials.

$$(S,T) \mapsto \begin{pmatrix} p_{\texttt{AAA}} \\ p_{\texttt{AAC}} \\ p_{\texttt{AAG}} \\ \vdots \\ p_{\texttt{TTT}} \end{pmatrix}$$

For each observed state $\tau \in \Sigma'^n$, its corresponding coordinate polynomial p_τ consists of the sum of the probabilities $\mathrm{Prob}(X = \sigma, Y = \tau)$ where σ varies over all hidden states in Σ^n. As an example, take the observation ACG of length 3. Then our coordinate polynomial will have $2^3 = 8$ terms, corresponding to the 8 hidden states shown below.

$$p_{\texttt{ACG}} = \frac{1}{2}\begin{pmatrix} t_{\texttt{eA}}s_{\texttt{ee}}t_{\texttt{eC}}s_{\texttt{ee}}t_{\texttt{eG}} \\ + \; t_{\texttt{eA}}s_{\texttt{ee}}t_{\texttt{eC}}s_{\texttt{ei}}t_{\texttt{iG}} \\ + \; t_{\texttt{eA}}s_{\texttt{ei}}t_{\texttt{iC}}s_{\texttt{ie}}t_{\texttt{eG}} \\ \vdots \\ + \; t_{\texttt{iA}}s_{\texttt{ii}}t_{\texttt{iC}}s_{\texttt{ii}}t_{\texttt{iG}} \end{pmatrix} \begin{matrix} \texttt{eee} \\ \texttt{eei} \\ \texttt{eie} \\ \\ \texttt{iii} \end{matrix}$$

Now, suppose we pick a point $(S,T) \in \mathbb{R}^{12}$ in the parameter space. We define the *explanation* for an observation $\tau \in \Sigma'^n$ to be the hidden state $\sigma \in \Sigma^n$ that maximizes the probability $\mathrm{Prob}(X = \sigma, Y = \tau)$ of seeing our observation.

Definition. *A gene-finding function is a map* $g\colon \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}^n \to \{\texttt{e}, \texttt{i}\}^n$ *that takes each observation* $\tau \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}^n$ *to its explanation* $\sigma \in \{\texttt{e}, \texttt{i}\}^n$.

In general, there could be multiple explanations maximizing the probability of seeing our observation. For simplification, we will assign in advance a tie-breaking rule that decides which of them to pick. Any total ordering, such as the lexicographic ordering, on the set of hidden states will work as a tie-breaking rule. In practice, a gene-finding function is a map that, given a specific observation of length n, tells you what the most likely break up of that sequence is in terms of introns and exons. Given a gene region on a genomic sequence, a gene finding function could tell us the most likely functional parts of that gene region.

What we have described is a very basic method which could be used in gene-finding. In reality, the methods used in gene prediction are usually more sophisticated. Gene prediction is an area that has accrued many different algorithms over the years. It originated with the study of different forms of *ab initio* methods and later, alignment-based methods. More recently there has been a trend in using a hybrid form of these methods to produce more efficient algorithms. See [1], [2], [3] for more information on these developments.

The rest of this report is organized as follows. In Section 2, we define inference functions for general graphical models and introduce the Few Inference Function Theorem. Section 3 describes the Neyman hidden Markov model and proves a formula for the number of inference functions on the model. Section 4 closes with a discussion of the lessons learnt from studying the Neyman model and lists some open questions.

2 Inference Functions for Graphical Models

The above discussion on hidden Markov models can be generalized to graphical models. We refer the reader to Section 1.5 of [4] for the definition of a graphical model. Suppose we have a graphical model with n observed random variables $Y_1, Y_2, \ldots, Y_n$ and q hidden random variables $X_1, X_2, \ldots, X_q$ whose values are taken from finite alphabets Σ' and Σ respectively. Figure 3 illustrates an example of what the underlying graph of a graphical model might look like, where the shaded vertices represent observed nodes while the unshaded ones represent hidden nodes. Associated with each edge of the underlying graph is a transition matrix. Let the transition matrices be determined by d parameters $\theta_1, \theta_2, \ldots, \theta_d$. Then, the model is a polynomial map

$$\begin{array}{rcl} f\colon \mathbb{R}^d & \longrightarrow & \mathbb{R}^{|\Sigma'^n|} \\ (\theta_1, \ldots, \theta_d) & \longmapsto & (p_\tau)_{\tau \in \Sigma'^n} \end{array}$$

For a given observation $\tau \in \Sigma'^n$, the corresponding coordinate polynomial p_τ is a sum of probabilities $\mathrm{Prob}(X = \sigma, Y = \tau)$ as σ varies over all hidden states Σ^q. Each probability in the sum is a monomial in the parameters $\theta_1, \ldots, \theta_d$. Now, fix a point $(\theta_1, \ldots, \theta_d)$ in the parameter space $\mathbb{R}^d$. Then, an explanation for an observation $\tau \in \Sigma'^n$ is a hidden state $\sigma \in \Sigma^q$ that maximizes the probability $\mathrm{Prob}(X = \sigma, Y = \tau)$.

Definition. *An inference function is a map $g\colon \Sigma'^n \to \Sigma^q$ that takes each observation $\tau \in \Sigma'^n$ to its explanation in Σ^q.*

In general, a given observation may have several explanations. Once again, we will resolve this minor issue by deciding on a tie-breaking rule in advance, so that our inference functions are well-defined. Thus, a gene-finding function is just an inference function for the hidden Markov model with observed and hidden alphabets $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $\{\mathtt{e}, \mathtt{i}\}$ respectively. So far, we have seen how each choice of parameters allows us to construct some inference function. As we change our parameters, the associated inference function may change. We say that an inference function g is *generic* if there exists some open set $U \subset \mathbb{R}^d$ such that the inference function constructed from each point in U is g.

The question at the heart of this project is this: given a graphical model, how many generic inference functions are there?

Naively, one could estimate the number of generic inference functions by counting all possible maps from Σ'^n to Σ^q. The total number, $|\Sigma^q|^{|\Sigma'^n|}$, gives an upper bound on the number of inference functions and is itself doubly exponential in n. However, in [5], Elizalde proved a remarkable result, called the Few Inference Functions Theorem:

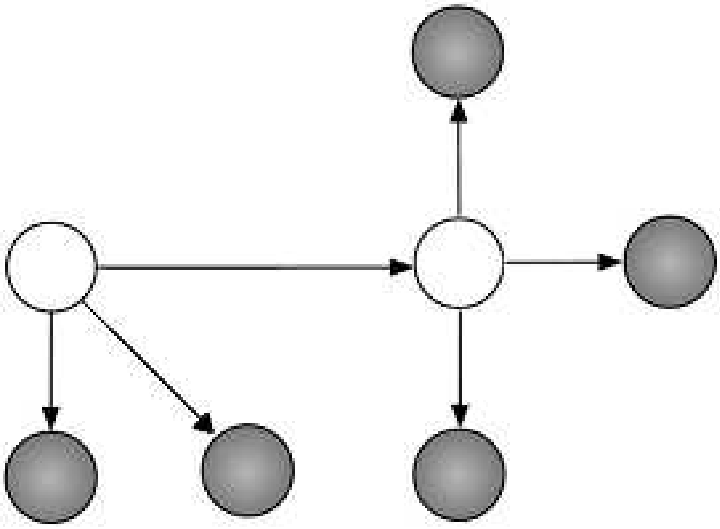

Figure 3: The Underlying Graph of a Graphical Model.

Theorem 2.1. *Let d be a fixed positive integer. Consider a graphical model with d parameters, and let E be the number of edges of the underlying graph. Then, the number of inference functions of the model is at most $O(E^{d(d-1)})$.*

In other words, the number of inference functions grows polynomially, not double exponentially. This means that majority of the maps from Σ'^n to Σ^q are not inference functions.

Elizalde proved his result by exploiting the relationship between inference functions and polytopes. We refer the reader to Section 2.3 of [4] for a more in-depth discussion on polytopes. Meanwhile, we will reproduce some notations here because we will be using them in the coming sections. Given a polytope $P \subset \mathbb{R}^d$ and a vector $w \in \mathbb{R}^d$, define

$$\text{face}_w(P) = \{x \in P \mid x \cdot w \leqslant y \cdot w \text{ for all } y \in P\}$$

Conversely, given a face F of P, we define the normal cone of P at F to be

$$N_P(F) = \{w \in \mathbb{R}^d \mid \text{face}_w(P) = F\}$$

The collection of all normal cones of P as F varies over all faces of P is called the normal fan of P, and is represented by $\mathcal{N}(P)$. The normal fan can thus be viewed as a partition of $\mathbb{R}^d$ into cones. Given two normal fans $\mathcal{N}_1$ and $\mathcal{N}_2$, we can consider the collection of cones obtained by intersecting a cone from each fan. This collection of cones, $\mathcal{N}_1 \wedge \mathcal{N}_2$, is known as the common refinement of the normal fans. Now, given two polytopes P and Q, their sum is the convex hull of the union of P and Q, and is denoted by $P \oplus Q$. Their product, also called the Minkowski sum, is given by

$$P \odot Q = \{p + q \in \mathbb{R}^d \mid p \in P, q \in Q\} \tag{1}$$

The following lemma [6] describes the relationship between Minkowski sums and normal fans:

Lemma 2.2. *Let P_1, ..., P_k be polytopes. Then,*

$$\mathcal{N}(P_1 \odot \cdots \odot P_k) = \mathcal{N}(P_1) \wedge \cdots \wedge \mathcal{N}(P_k)$$

Given a polynomial h on d indeterminates $\theta_1, \ldots, \theta_d$, we denote its Newton polytope by $\mathcal{NP}(h)$. Suppose that we have a graphical model $f : \mathbb{R}^d \to \mathbb{R}^{|\Sigma'^n|}$ as described at the start of this section. We define the Newton polytope $\mathcal{NP}(f)$ of f to be the Minkowski sum of the Newton polytopes of its coordinate polynomials, i.e.

$$\mathcal{NP}(f) = \bigodot_{\tau \in \Sigma'^n} \mathcal{NP}(p_\tau)$$

It turns out that the generic inference functions are in one-to-one correspondence with the vertices of $\mathcal{NP}(f)$. In [5], Elizalde details the reason behind this correspondence. The reader is encouraged to check this out before proceeding, even though knowledge of it is not required for the remaining sections.

3 Neyman Hidden Markov Model

Elizalde's theorem significantly improved bounds on the number of inference functions, but his estimate is often still far from the actual number. For instance, suppose we have the homogeneous binary hidden Markov model of length n, where the alphabets for the hidden and observed random variables are both $\{0, 1\}$. The matrices for the transitions between the hidden nodes are all equal to some 2×2 matrix S, while the matrices for the transitions between the hidden and observed nodes are all equal to some 2×2 matrix T. Hence, this model has a total of 8 parameters. Consider the case $n = 6$. The naive guess for the number of inference functions is $64^{64} \approx 3.9 \times 10^{115}$. Meanwhile, Elizalde's method gives us 7.2×10^{72}. The actual number, however, is 17354.

To study the dynamics behind computing the exact number of inference functions, we decided to start with a simple hidden Markov model. We call it the homogeneous binary Neyman hidden Markov model of length n, where $n \geqslant 1$. In this model, the alphabet for all the hidden and observed nodes is $\{0, 1\}$. The transition matrices between the hidden nodes are equal to S, while the transition matrices between the hidden and observed nodes are equal to T, where S and T are given by

$$S = \begin{pmatrix} s_{00} & s_{01} \\ s_{10} & s_{11} \end{pmatrix} = \begin{pmatrix} s & \tilde{s} \\ \tilde{s} & s \end{pmatrix}, \quad T = \begin{pmatrix} t_{00} & t_{01} \\ t_{10} & t_{11} \end{pmatrix} = \begin{pmatrix} t & \tilde{t} \\ \tilde{t} & t \end{pmatrix}$$

for some $s, \tilde{s}, t, \tilde{t} \in \mathbb{R}$. Notes that s and t are the probabilities that the two random variables in a transition are equal while $\tilde{s}$ and $\tilde{t}$ are the probabilities that they are different. Usually, we also require that $s + \tilde{s} = 1$ and $t + \tilde{t} = 1$, but this condition is not crucial to our analysis and will be ignored. Our model is named this way, because the transition matrices S and T have the same structure as those in the phylogenetical Neyman model. Denote the number of generic inference functions on this model by $G(n)$.

We start by developing some useful notations. Given $\tau \in \{0, 1\}$, let $\bar{\tau}$ represent the flipped symbol $1 - \tau \in \{0, 1\}$. When we say that we have the configuration

$$\begin{matrix} \sigma_1 & \sigma_2 & \ldots & \sigma_n \\ \tau_1 & \tau_2 & \ldots & \tau_n \end{matrix}$$

it means that the hidden and observed sequences are $X = \sigma_1\sigma_2 \ldots \sigma_n$, $Y = \tau_1\tau_2 \ldots \tau_n$. We may also say that we have the configuration $(\sigma_1\sigma_2 \ldots \sigma_n, \tau_1\tau_2 \ldots \tau_n)$ to mean the same thing. Any equality $\sigma_i = \tau_i$ will be called a vertical equality, while any equality $\sigma_i = \sigma_{i+1}$ will be called a horizontal equality. Vertical and horizontal inequalities are defined in a similar way.

We will also refer to these equalities/inequalities as the horizontal and vertical relations of the configuration, or of $\sigma_1\sigma_2\ldots\sigma_n$ when it is clear what $\tau_1\tau_2\ldots\tau_n$ is.

Given some observation $\tau \in \{0,1\}^n$, the coordinate polynomial p_τ is the sum of monomials of the form $s^x\tilde{s}^{\tilde{x}}t^y\tilde{t}^{\tilde{y}}$. Note that the Newton polytope $\mathcal{NP}(p_\tau)$ of p_τ lies on the two dimensional subspace of $\mathbb{R}^4$ cut out by the equations $x+\tilde{x}=n-1$ and $y+\tilde{y}=n$. Thus, by projecting $\mathcal{NP}(p_\tau)$ onto $\mathbb{R}^2$ via the map $\psi : \mathbb{R}^4 \to \mathbb{R}^2, (x,\tilde{x},y,\tilde{y}) \mapsto (x,y)$, we get a polytope $\mathcal{P}_\tau$ that is combinatorially isomorphic to our original Newton polytope. We call $\mathcal{P}_\tau$ the coordinate polytope of τ. If $\mathcal{LP}_\tau$ be the set of lattice points $\{(x,y)|s^x\tilde{s}^{\tilde{x}}t^y\tilde{t}^{\tilde{y}}$ is a monomial in $p_\tau\}$, then $\mathcal{P}_\tau$ is the convex hull of $\mathcal{LP}_\tau$. Note that if (x,y) corresponds to some monomial $s^x\tilde{s}^{\tilde{x}}t^y\tilde{t}^{\tilde{y}}$ in p_τ coming from $\mathrm{Prob}(X=\sigma, Y=\tau)$, then x is the number of horizontal equalities and y the number of vertical equalities in the configuration

$$\begin{matrix}\sigma_1 & \sigma_2 & \ldots & \sigma_n \\ \tau_1 & \tau_2 & \ldots & \tau_n\end{matrix}$$

We say that the configuration *gives rise* to the point (x,y), and that (x,y) *arises* from the configuration. Note that each configuration gives rise to exactly one point, but a given point may arise from many different configurations.

Define $\mathcal{P}_{(n)}$ to be the Minkowski sum of the coordinate polytopes $\mathcal{P}_\tau$ where τ varies over $\{0,1\}^n$. We say that $\mathcal{P}_{(n)}$ is the model polytope of the Neyman model of length n. Observe that $\mathcal{P}_{(n)}$ is combinatorially isomorphic to the Newton polytope $\mathcal{NP}(f)$ of the Neyman model $\boldsymbol{f} : R^4 \to R^{|\{0,1\}^n|}$, since it is the image of $\mathcal{NP}(f)$ under the projection ψ. Since the number $G(n)$ of inference functions is equal to the number of vertices in $\mathcal{NP}(f)$, it follows that

$$G(n) = |\{\text{vertices in } \mathcal{P}_{(n)}\}|$$

Let P be any two-dimensional lattice polytope, and let F be some edge of P. Then, the normal cone $\mathcal{N}_P(F)$ may be written as $\{\lambda(u,v) \mid \lambda > 0\}$ for some lattice vector $(u,v) \in \mathbb{Z}^2$. In particular, if we require that $\gcd(u,v)=1$, then (u,v) is determined uniquely. Let $\mathcal{LV}(P)$ be the collection of all such lattice vectors (u,v) as F varies over all the edges of P, i.e.

$$\mathcal{LV}(P) = \{(u,v) \in \mathbb{Z}^2 \mid \gcd(u,v)=1, \mathrm{face}_{(u,v)}P \text{ is an edge}\}$$

Since the number of vertices in P equals the number of edges in P which are in one-to-one correspondence with vectors in $\mathcal{LV}(P)$, we have

$$|\{\text{vertices in } \mathcal{P}_{(n)}\}| = |\mathcal{LV}(P)|$$

The following proposition explores the relationship between $\mathcal{LV}$ and Minkowski sums:

Proposition 3.1. *Let $P_1, \ldots, P_k$ be two-dimensional lattice polytopes. Then,*

$$\mathcal{LV}(P_1 \odot \cdots \odot P_k) = \mathcal{LV}(P_1) \cup \cdots \cup \mathcal{LV}(P_k)$$

Proof. This follows immediately from Lemma 2.2, and from the fact that in $\mathbb{R}^2$, the collection of one-dimensional cones of a refinement of normal fans is the union of the collection of one-dimensional cones for each normal fan. □

From the above discussion, we see that $G(n) = |\{\text{vertices in } \mathcal{P}_{(n)}\}| = |\mathcal{LV}(\mathcal{P}_{(n)})|$. Thus, to count our inference functions, it would be useful to understand the structure of $\mathcal{LV}(\mathcal{P}_{(n)})$. Figure 4 shows each $\mathcal{LV}(\mathcal{P}_{(n)})$ for small values of n. Some patterns are easily recognized from the figure. The following two lemmas list some of these patterns, and allow us to characterize $\mathcal{LV}(\mathcal{P}_{(n)})$ completely.

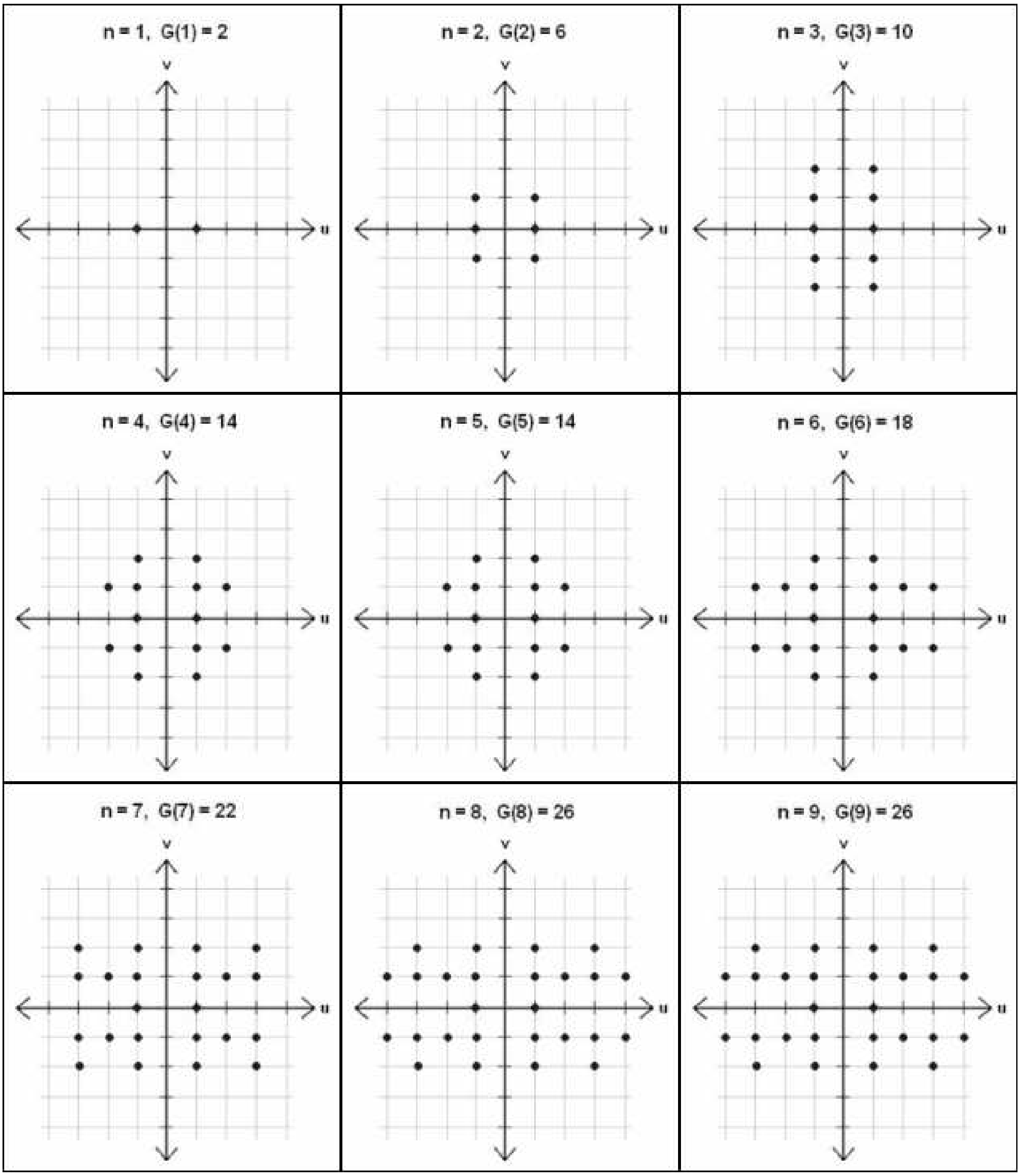

Figure 4: $\mathcal{LV}(\mathcal{P}_{(n)})$ for the Neyman Hidden Markov Model of Length n

Lemma 3.2. *If $n \geqslant 2$ and $(u,v) \in \mathcal{LV}(\mathcal{P}_{(n)})$, then*

i. $(-u,v) \in \mathcal{LV}(\mathcal{P}_{(n)})$.

ii. $(u,-v) \in \mathcal{LV}(\mathcal{P}_{(n)})$.

iii. $|u| \leqslant \lfloor \frac{n}{2} \rfloor$.

iv. $|v| \leqslant 2$.

Proof. Before proving the statements, we apply Proposition 3.1 to get

$$\mathcal{LV}(\mathcal{P}_{(n)}) = \bigcup_{\tau \in \{0,1\}^n} \mathcal{LV}(\mathcal{P}_\tau) \tag{2}$$

i. We claim that given any $\tau \in \{0,1\}^n$, there exists some $\tau' \in \{0,1\}^n$ such that $\mathcal{P}_{\tau'}$ is the reflection of $\mathcal{P}_\tau$ about the line $x = \frac{n-1}{2}$. Then, statement (i) follows from this claim because given any $(u,v) \in \mathcal{LV}(\mathcal{P}_{(n)})$, equation (2) tells us that $(u,v) \in \mathcal{LV}(\mathcal{P}_\tau)$ for some $\tau \in \{0,1\}^n$. The claim assures us of some $\mathcal{P}_{\tau'}$ that is the horizontal reflection of $\mathcal{P}_\tau$, therefore $(-u,v) \in \mathcal{LV}(\mathcal{P}_{\tau'}) \subset \mathcal{LV}(\mathcal{P}_{(n)})$.

Now, $\mathcal{P}_\tau$ and $\mathcal{P}_{\tau'}$ are convex hulls of the lattice points $\mathcal{LP}_\tau$ and $\mathcal{LP}_{\tau'}$ respectively. Thus, to prove the claim, it suffices to show that given $\mathcal{LP}_\tau$, there exists some $\mathcal{LP}_{\tau'}$ which is the reflection of $\mathcal{LP}_\tau$ about the line $x = \frac{n-1}{2}$. Indeed, let τ' be $\tau_1\bar{\tau}_2\tau_3\bar{\tau}_4 \ldots \in \{0,1\}^n$, the string where every even-positioned letter in τ is flipped. Suppose that (x,y) is a lattice point in $\mathcal{LP}_\tau$ and that the configuration

$$\begin{matrix} \sigma_1 & \sigma_2 & \ldots & \sigma_n \\ \tau_1 & \tau_2 & \ldots & \tau_n \end{matrix}$$

gives rise to the point (x,y), i.e. x is the number of horizontal equalities and y the number of vertical equalities in the configuration. Then, the configuration

$$\begin{matrix} \sigma_1 & \bar{\sigma}_2 & \sigma_3 & \bar{\sigma}_4 & \ldots \\ \tau_1 & \bar{\tau}_2 & \tau_3 & \bar{\tau}_4 & \ldots \end{matrix} \tag{3}$$

gives rise to the point $(n-1-x,y) \in \mathcal{LP}_{\tau'}$, since none of the vertical relations are changed while all of the horizontal relations are flipped. This point $(n-1-x,y)$ is the reflection of (x,y) about the line $x = \frac{n-1}{2}$. Conversely, each point in $\mathcal{LP}_{\tau'}$ is the reflection of some point in $\mathcal{LP}_\tau$ by the above reasoning, because flipping the even-positioned letters of τ' gives us τ again. This finishes the proof of our claim, and statement (i) follows.

To prove (ii), (iii) and (iv), because of equation (2), it suffices to show that they are true for each $\mathcal{P}_\tau$, i.e. given $(u,v) \in \mathcal{LV}(\mathcal{P}_\tau)$, we have $(u,-v) \in \mathcal{LV}(\mathcal{P}_\tau)$, $|u| \leqslant \lfloor \frac{n}{2} \rfloor$ and $|v| \leqslant 2$. In the following arguments, we will fix $\tau \in \{0,1\}^n$.

ii. To prove this, it is enough to show that $\mathcal{P}_\tau$ is symmetrical about the line $y = \frac{n}{2}$. In fact, it is sufficient to prove it for the lattice points $\mathcal{LP}_\tau$. Let (x,y) be any lattice point in $\mathcal{LP}_\tau$. Suppose the configuration

$$\begin{matrix} \sigma_1 & \sigma_2 & \ldots & \sigma_n \\ \tau_1 & \tau_2 & \ldots & \tau_n \end{matrix}$$

gives rise to the point (x,y). Then, the configuration

$$\begin{matrix} \bar{\sigma}_1 & \bar{\sigma}_2 & \ldots & \bar{\sigma}_n \\ \tau_1 & \tau_2 & \ldots & \tau_n \end{matrix}$$

gives rise to the point $(x, n-y) \in \mathcal{LP}_\tau$, since none of the horizontal relations are changed while all of the vertical relations are flipped. Note that $(x, n-y)$ is the reflection of (x, y) about the line $y = \frac{n}{2}$. Thus, $\mathcal{LP}_\tau$ is symmetrical about the line $y = \frac{n}{2}$, as required.

iii. Given some $(u, v) \in \mathcal{LV}(\mathcal{P}_\tau)$, consider the edge $F = \text{face}_{(u,v)}(\mathcal{P}_\tau)$. Let the vertices on the ends of F be (x_1, y_1) and (x_2, y_2). Then, $u(x_2 - x_1) + v(y_2 - y_1) = 0$. Since the quantities involved are integers and $\gcd(u, v) = 1$, we have $|x_2 - x_1| = c|v|$ and $|y_2 - y_1| = c|u|$ for some positive integer c. Due to the symmetry of $\mathcal{P}_\tau$ from statement (ii), the edge F lies either completely above the line $y = \frac{n}{2}$, completely below that line, or crosses the line perpendicularly. In the first two cases, note that the ends of F may touch the line of symmetry. Also, each y_i represents the number of vertical equalities in some configuration, so $0 \leqslant y_i \leqslant n$. Hence, either $0 \leqslant y_1, y_2 \leqslant \frac{n}{2}$ or $\frac{n}{2} \leqslant y_1, y_2 \leqslant n$. In both scenarios, $|y_2 - y_1| \leqslant \frac{n}{2}$, and because the LHS of the inequality is integer, we have $|y_2 - y_1| \leqslant \lfloor \frac{n}{2} \rfloor$. Therefore, $|u| = \frac{1}{c}|y_2 - y_1| \leqslant \lfloor \frac{n}{2} \rfloor$ as required. In the third case, $v = \frac{1}{c}(x_1 - x_2) = 0$. Because $\gcd(u, v) = 1$, we have $|u| = 1 \leqslant \lfloor \frac{n}{2} \rfloor$ so statement (iii) is trivially satisfied in this case.

iv. Given some $(u, v) \in \mathcal{LV}(\mathcal{P}_\tau)$, we want to show that $|v| \leqslant 2$. Again, consider the edge $F = \text{face}_{(u,v)}(\mathcal{P}_\tau)$. Let $(x_1, y_1), (x_2, y_2) \in \mathcal{LP}_\tau$ be two distinct lattice points on F. Then, $(x_2 - x_1, y_2 - y_1) \cdot (u, v) = 0$. Also, $|x_2 - x_1| = c|v|$ and $|y_2 - y_1| = c|u|$ for some positive integer c. Let

$$\begin{matrix} \sigma'_1 & \sigma'_2 & \dots & \sigma'_n \\ \tau_1 & \tau_2 & \dots & \tau_n \end{matrix}$$

be a configuration that gives rise to the point (x_1, y_1) and let

$$\begin{matrix} \sigma_1 & \sigma_2 & \dots & \sigma_n \\ \tau_1 & \tau_2 & \dots & \tau_n \end{matrix}$$

be a configuration that gives rise to (x_2, y_2). Then, for each $1 \leqslant i \leqslant n$, either $\sigma'_i = \sigma_i$ or $\sigma'_i = \bar{\sigma}_i$. We partition the sequence σ' into blocks as follows. Let $r = 0$ if $\sigma'_1 = \sigma_1$, and $r = 1$ otherwise. Also, define $0 = b_0 < b_1 < \ldots < b_k < b_{k+1} = n$ to be the unique sequence of integers such that for each $0 \leqslant j \leqslant k$, we have

$$\sigma'_i = \begin{cases} \sigma_i & \text{if } j \equiv r \pmod 2 \\ \bar{\sigma}_i & \text{otherwise.} \end{cases} \qquad \text{for } b_j + 1 \leqslant i \leqslant b_{j+1}$$

For example, suppose $n = 10$ and

$$\sigma' = \bar{\sigma}_1\bar{\sigma}_2\sigma_3\sigma_4\sigma_5\sigma_6\bar{\sigma}_7\bar{\sigma}_8\bar{\sigma}_9\sigma_{10}. \tag{4}$$

Note that the sequence is partitioned into blocks which are alternatingly flipped and not flipped: the 1st - 2nd letters are flipped, the 4th - 6th letters are not flipped, the 7th - 9th letters are flipped, and the 10th letter is not flipped. We represent this with the integer sequence $(b_1, b_2, b_3) = (2, 6, 9)$ and $k = 3$. The i-th block is defined to be the letters in the $(b_i + 1)$-th to (b_{i+1})-th positions, for $0 \leqslant i \leqslant k$.

Now, we want to transform σ' to σ by a sequence of *moves*. Each move involves the flipping of a contiguous block of letters. For each $0 \leqslant i \leqslant k$, define $M_i : \{0, 1\}^n \to \{0, 1\}^n$ to be the move which flips the i-th block of letters. Furthermore, among these moves, we want to consider only those which flip letters of σ' that are different from those in σ, i.e. M_i such that $\sigma'_{b_i+1} = \bar{\sigma}_{b_i+1}$. We let $\mathcal{M}$ be the set of such moves. Returning to the example in (4), we see that $\mathcal{M} = \{M_0, M_2\}$. In general, either $\mathcal{M} = \{M_i \mid i \text{ is even}\}$ or $\mathcal{M} = \{M_i \mid i \text{ is odd}\}$. Note that applying all the moves in $\mathcal{M}$ to σ' gives σ.

We can think of each move as a map between configurations, by applying the move to the hidden state of the configuration. Since each configuration gives rise to some lattice point in $\mathcal{LP}_\tau$, the moves can also be viewed as hops from one lattice point to another in $\mathcal{LP}_\tau$. Call the resulting change in lattice coordinates the *hop vector*. In general, the hop vector of a move depends on the configuration that the move is applied to. In our case, we are applying the moves from $\mathcal{M}$ in some order to σ'. We claim that in this case, the hop vector of each $M_i \in \mathcal{M}$ does not depend on the order of application. Indeed, M_i changes exactly:

(a) the horizontal relation between X_j and X_{j+1}, for $j \in \{b_i, b_{i+1}\} \setminus \{0, n\}$

(b) the vertical relation between X_j and Y_j, for $b_i < j \leqslant b_{i+1}$.

For example, the diagram below indicates using $*$ the relations changed by M_2 in (4).

$$\begin{array}{ccccccccccc}
\cdots & \sigma_6 & * & \bar{\sigma}_7 & & \bar{\sigma}_8 & & \bar{\sigma}_9 & * & \sigma_{10} & \cdots \\
 & & & * & & * & & * & & & \\
\cdots & \tau_6 & & \tau_7 & & \tau_8 & & \tau_9 & & \tau_{10} & \cdots
\end{array}$$

Since either $\mathcal{M} = \{M_i \mid i \text{ is even}\}$ or $\mathcal{M} = \{M_i \mid i \text{ is odd}\}$, the set of relations changed by each $M_i \in \mathcal{M}$ is disjoint from the other set of relations. Thus, the hop vectors which come from changes in the number of horizontal and vertical equalities does not depend the order of application of the moves on σ'. We can simply compute the hop vectors by considering the action of each move on (x_1, y_1), the lattice point for $X = \sigma', Y = \tau$. In addition, the sum of all the hop vectors represents the total action of all the moves in $\mathcal{M}$ on σ'. Hence, this sum is exactly $(x_2 - x_1, y_2 - y_1)$.

We now come to the crux of the proof of statement (iv). Suppose otherwise that $|v| \geqslant 3$. Then, $|x_2 - x_1| = c|v| \geqslant 3$, so the difference in the number of horizontal equalities of σ and of σ' is at least 3. This implies that there are at least 3 differences in horizontal relations between σ and σ'. We claim that there is no way to transform σ' to σ by exactly one move, i.e. there does not exist a contiguous block of letters in σ' which, when flipped, gives σ. Indeed, if such a block exists, then flipping the block changes at most two horizontal relations, contradicting the lower bound on the number of differences in horizontal relations.

Given some $M_i \in \mathcal{M}$, let M_i take (x_1, y_1) to $(x_1 + \Delta x, y_1 + \Delta y)$ so that its hop vector is $(\Delta x, \Delta y)$. The above claim implies that $(x_1 + \Delta x, y_1 + \Delta y)$ cannot be another lattice point on F, otherwise there are two points on F with exactly one move between them. Hence, $(\Delta x, \Delta y)$ cannot satisfy $(\Delta x, \Delta y) \cdot (u, v) = 0$. Also, since (x_1, y_1) minimizes the equation $(x, y) \cdot (u, v)$, we must have $(x_1 + \Delta x, y_1 + \Delta y) \cdot (u, v) \geqslant (x_1, y_1) \cdot (u, v)$, giving us $(\Delta x, \Delta y) \cdot (u, v) > 0$. Therefore, the sum $\sum(\Delta x, \Delta y)$ of hop vectors over all moves in $\mathcal{M}$ must also satisfy $\sum(\Delta x, \Delta y) \cdot (u, v) > 0$. But $\sum(\Delta x, \Delta y) \cdot (u, v) = (x_2 - x_1, y_2 - y_1) \cdot (u, v) = 0$, a contradiction. Thus, $|v| \leqslant 2$, as required.

□

Lemma 3.3. *Let $m = \lfloor \frac{n}{2} \rfloor$. Then,*

i. $(u, 1) \in \mathcal{LV}(\mathcal{P}_{(n)})$ *for all integers* $u \in [1, m]$.

ii. $(u, 2) \in \mathcal{LV}(\mathcal{P}_{(n)})$ *for all odd integers* $u \in [1, m-1]$.

iii. $(m, 2) \in \mathcal{LV}(\mathcal{P}_{(n)})$ *if and only if* $n \equiv 3$ *(mod* 4*).*

iv. $(1, 0) \in \mathcal{LV}(\mathcal{P}_{(n)})$ *and* $(0, 1) \notin \mathcal{LV}(\mathcal{P}_{(n)})$*.*

Proof. In this proof, we will frequently use constructed sequences. For convenience, define

$\mathcal{S}_0(i)$	$= 0 \dots 0$	a sequence of i zeroes
$\mathcal{S}_1(i)$	$= 1 \dots 1$	a sequence of i ones
$\mathcal{S}_{01}(i, j)$	$= 0 \dots 01 \dots 1$	a sequence of i zeroes and j ones
$\mathcal{S}_{10}(i, j)$	$= 1 \dots 10 \dots 0$	a sequence of i ones and j zeroes
$\mathcal{S}_{010}(i, j, k)$	$= 0 \dots 01 \dots 10 \dots 0$	a sequence of i zeroes, j ones and k zeroes
$\mathcal{S}_{101}(i, j, k)$	$= 1 \dots 10 \dots 01 \dots 1$	a sequence of i ones, j zeroes and k ones

i. Given an integer $u \in [1, m]$, define sequences $\tau = \sigma' = \mathcal{S}_{10}(u, n-u)$ and $\sigma = \mathcal{S}_0(n)$. Then, the configuration (σ, τ) gives rise to the point $(n-1, n-u) \in \mathcal{LP}(\mathcal{P}_\tau)$ while the configuration (σ', τ) gives rise to the point $(n-2, n)$. If these two points minimize the dot-product $(x, y) \cdot (-u, -1)$, then $(-u, -1) \in \mathcal{LV}(\mathcal{P}_\tau)$, so by Proposition (3.1) and Lemma 3.2, we have $(u, 1) \in \mathcal{LV}(\mathcal{P}_{(n)})$.

   Suppose otherwise that there exist some point $(x_0, y_0) \in \mathcal{LP}_\tau$ arising from a different configuration (σ'', τ) such that

$$(n-2, n) \cdot (-u, -1) > (x_0, y_0) \cdot (-u, -1).$$

   The coordinates x_0, y_0 trivially satisfy $0 \leqslant x_0 \leqslant n-1$ and $0 \leqslant y_0 \leqslant n$, so we get

$$\begin{array}{lclcl} -u(n-2) - n & > & -ux_0 - y_0 & \geqslant & -u(n-1) - y \\ -u(n-2) - n & > & -ux_0 - y_0 & \geqslant & -ux_0 - n \end{array}$$

   which imply $x_0 = n-1$ and $y_0 > n-u$. Since $x_0 = n-1$, all the horizontal relations in σ'' must be equalities, so either $\sigma'' = \mathcal{S}_0(n) = \sigma$ or $\sigma'' = \mathcal{S}_1(n)$. We disregard the first case because $\sigma'' \neq \sigma$ by definition. In the second case, it is easy to see that $y_0 = u$, but $u \leqslant n-u$ as $u \in [1, m]$, contradicting $y_0 > n-u$. Hence, the points $(n-1, n-u)$ and $(n-2, n)$ minimize the dot-product $(x, y) \cdot (-u, -1)$, and statement (i) follows.

Before we prove (ii) and (iii), we need the following claim: let $b \in [1, m]$ be an odd integer, and let $\tau = \mathcal{S}_{010}(a, b, n-a-b)$. Then, $(b, 2) \in \mathcal{LV}(\mathcal{P}_\tau)$ if and only if

$$\frac{b}{2} \leqslant a \leqslant n - \frac{3b}{2} \tag{5}$$

Indeed, let $\sigma = \mathcal{S}_0(n)$, and $\sigma' = \tau$. If the points $(n-1, n-b)$ and $(n-3, n)$ arising from the configurations (σ, τ) and (σ', τ) minimize the dot-product $(x, y) \cdot (-b, -2)$, then $(b, 2) \in \mathcal{LV}(\mathcal{P}_\tau)$. Conversely, suppose $(b, 2) \in \mathcal{LV}(\mathcal{P}_\tau)$. Pick points $(x_1, y_1), (x_2, y_2) \in \mathcal{LP}_\tau$ lying on the edge $F = \mathrm{face}_{(-b,-2)}(\mathcal{P}_\tau)$. Now, suppose for each i, $(n-3, n) \cdot (-b, -2) > (x_i, y_i) \cdot (-b, -2)$. Plugging in $0 \leqslant y_i \leqslant n$, we get $x_i \in \{n-2, n-1\}$, and so $|x_2 - x_1| \leqslant 1$. But $(x_2 - x_1, y_2 - y_1) \cdot (-b, -2) = 0$ and $\gcd(b, 2) = 1$ implies that for some positive integer c, $|x_2 - x_1| = 2c \geqslant 2$, a contradiction. We have just shown that $(b, 2) \in \mathcal{LV}(\mathcal{P}_\tau)$ if and only if $(n-1, n-b)$ and $(n-3, n)$ minimize the dot-product $(x, y) \cdot (-b, -2)$.

By definition, $(n-1, n-b)$ and $(n-3, n)$ minimize $(x, y) \cdot (-b, -2)$ if and only if there does not exist $(x_0, y_0) \in \mathcal{LP}_\tau$ such that

$$-b(n-3) - 2n > -bx_0 - 2y_0. \tag{6}$$

Suppose such a point (x_0, y_0) exists, arising from some configuration (σ'', τ). Using $0 \leqslant x_i \leqslant n-1$ and $0 \leqslant y_i \leqslant n$, we get $x_i \in \{n-2, n-1\}$ and $y_i > n-b$. If $x_i = n-1$, then all the horizontal relations in σ'' are equalities, so either $\sigma'' = \mathcal{S}_0(n) = \sigma$ or $\sigma'' = \mathcal{S}_1(n)$. Again, disregard the first case because $\sigma'' \neq \sigma$. In the second case, $y_0 = b$, but $b \leqslant n-b$ as $b \in [1, m]$, contradicting $y_0 > n-b$. Hence, $x_i = n-2$. This means that $y_0 > n - \frac{b}{2}$, which comes from substituting $x_0 = n-2$ into (6). It also means that exactly one of the horizontal relations in σ'' is an inequality. We have two scenarios: either $\sigma'' = \mathcal{S}_{01}(d, n-d)$ or $\sigma'' = \mathcal{S}_{10}(d, n-d)$, for some integer d. In the first scenario, a quick analysis gives:

$$n - \frac{b}{2} < y_0 = \begin{cases} b+d & \leqslant b+a \qquad \text{if } 0 < d \leqslant a \\ 2a+b-d & \leqslant b+a \qquad \text{if } a < d \leqslant a+b \\ d-b & \qquad\qquad\quad\ \text{if } a+b < d \leqslant n \end{cases}$$

The third case above is impossible since $d - b \leqslant n - b < n - \frac{b}{2}$. Thus, $n - \frac{b}{2} < a + b$, or in other words, $a > n - \frac{3b}{2}$. The second scenario is symmetrical to the first scenario: we reflect the concerned sequences back to front and use new parameters $a' = n-a-b$, $b' = b$ and $d' = n-d$. Hence, we arrive at $a' > n - \frac{3b'}{2}$, or equivalently, $a < \frac{b}{2}$.

Conversely, if $a > n - \frac{3b}{2}$, we consider the configuration $(\mathcal{S}_{01}(a, n-a), \tau)$ which gives rise to the point $(x_0, y_0) = (n-2, a+b)$. Then,

$$-bx_0 - 2y_0 = -b(n-2) - 2(a+b) < -b(n-3) - 2n,$$

so there exists $(x_0, y_0) \in \mathcal{LP}_\tau$ satisfying (6). If $a < \frac{b}{2}$, pick the configuration $(\mathcal{S}_{10}(a+b, n-a-b), \tau)$ which gives rise to the point $(x_0, y_0) = (n-2, n-a)$. Then,

$$-bx_0 - 2y_0 = -b(n-2) - 2(n-a) < -b(n-3) - 2n,$$

thus satisfying (6). Therefore, there does not exist $(x_0, y_0) \in \mathcal{LP}_\tau$ satisfying (6) if and only if $\frac{b}{2} \leqslant a \leqslant N - \frac{3b}{2}$. This completes the proof of the claim.

ii. Given an odd integer $u \in [1, m-1]$, set $a = \frac{u+1}{2}$. Let $\tau = \mathcal{S}_{010}(a, u, n-a-u)$. Since

$$\frac{u}{2} < \frac{u+1}{2} = a \leqslant \frac{n}{4} = n - \frac{3(n/2)}{2} \leqslant n - \frac{3u}{2},$$

the claim in (5) tells us that $(u, 2) \in \mathcal{LV}(\mathcal{P}_\tau) \subseteq \mathcal{LV}(\mathcal{P}_{(n)})$.

iii. If $n \equiv 0, 1 \pmod 4$, then $m = \lfloor \frac{n}{2} \rfloor$ is even and $\gcd(m, 2) \neq 1$, so trivially $(m, 2)$ is not in $\mathcal{LV}(\mathcal{P}_{(n)})$. If $n \equiv 3 \pmod 4$, then $m = \frac{n-1}{2}$. Set $a = \frac{m+1}{2}$. Let $\tau = \mathcal{S}_{010}(a, m, n-a-m)$. Then, the claim in (5) tells us that $(m, 2) \in \mathcal{LV}(\mathcal{P}_\tau) \subseteq \mathcal{LV}(\mathcal{P}_{(n)})$ because

$$\frac{m}{2} < \frac{m+1}{2} = a = \frac{n+1}{4} \leqslant \frac{n+3}{4} = n - \frac{3m}{2},$$

If $n \equiv 2 \pmod 4$, suppose on the contrary that $(m, 2) \in \mathcal{LV}(\mathcal{P}_\tau)$ for some $\mathcal{P}_\tau$. Then, there exists some edge $F \in \mathcal{P}_\tau$ with end-vertices (x_1, y_1) and (x_2, y_2) satisfying $|x_2 - x_1| = 2c$ and $|y_2 - y_1| = mc$ for some positive integer c. Due to the symmetry of $\mathcal{P}_\tau$ from Lemma 3.2.ii, the edge F lies either completely above the line $y = \frac{n}{2}$, completely below that line, or crosses the line perpendicularly. The third case is impossible because the normal vector to F is $(m, 2)$. Using Lemma 3.2.ii again, we can assume that F lies above the line. Then, $|y_2 - y_1| \leqslant \frac{n}{2} = m$. But $|y_2 - y_1| = mc \geqslant m$. Hence, $|y_2 - y_1| = m = \frac{n}{2}$. This means that one of the ends of F, say (x_1, y_1), lies on the line $y = \frac{n}{2}$, while the other end, (x_2, y_2), lies on the line $y = n$. It also means that $c = 1$, so $|x_2 - x_1| = 2$.

The next step is to show that $x_1 \in \{0, n-1\}$. Visually, this seems obvious but we shall proceed cautiously. We start by observing that the edge (or vertex, if the following two points are the same) joining the points arising from the configurations $(\mathcal{S}_0(n), \tau)$ and $(\mathcal{S}_1(n), \tau)$ intersects the line $y = \frac{n}{2}$ at $x = n-1$. Similarly, the edge (or vertex) joining the points arising from the configurations $(010\ldots01, \tau)$ and $(101\ldots10, \tau)$ (the hidden sequences are alternating zeroes and ones) intersects the line $y = \frac{n}{2}$ at $x = 0$. Hence, $\mathcal{P}_\tau \cap \{y = \frac{n}{2}\}$ is the line segment $[0, n-1] \times \{\frac{n}{2}\} \subseteq \mathbb{R}^2$. Since $(x_1, y_1) = (x_1, \frac{n}{2})$ is a boundary point of $\mathcal{P}_\tau$, either $x_1 = 0$ or $x_1 = n-1$. By Lemma 3.2.i, we can pick τ so that $x_1 = n-1$. Therefore, we now have $(x_1, y_1) = (n-1, \frac{n}{2})$ and $(x_2, y_2) = (n-3, n)$.

Let (σ', τ) and (σ, τ) be configurations which give rise to (x_1, y_1) and (x_2, y_2) respectively. Since $x_2 = n-3$, there are exactly two inequalities among the horizontal relations in σ. As a result, the possibilities are $\sigma = \mathcal{S}_{010}(a, b, n-a-b)$ or $\sigma = \mathcal{S}_{101}(a, b, n-a-b)$ for some integers a, b. Since $y_2 = n$, all the vertical relations are equalities, so $\tau = \sigma$. Without loss of generality, we can assume $\tau = \sigma = \mathcal{S}_{010}(a, b, n-a-b)$. Now, $x_1 = n-1$, all the horizontal relations in σ' are equalities so $\sigma = \mathcal{S}_0(n)$ or $\sigma = \mathcal{S}_1(n)$. Since $y_1 = \frac{n}{2}$ is the number of vertical equalities, we see that in each case the number of ones in τ is $\frac{n}{2}$. Thus, $b = \frac{n}{2}$. By claim (5), $(b, 2) = (m, 2) \in \mathcal{LV}(\mathcal{P}_\tau)$ implies that $\frac{b}{2} \leqslant a \leqslant n - \frac{3b}{2}$. But $\frac{b}{2}$ is non-integer and $\frac{b}{2} = n - \frac{3b}{2}$, so there are no integers between the two bounds, a contradiction. Hence, $(m, 2) \notin \mathcal{LV}(\mathcal{P}_\tau)$, as required.

iv. The configurations $(\mathcal{S}_0(n), \mathcal{S}_0(n))$ and $(\mathcal{S}_1(n), \mathcal{S}_0(n))$ give rise to the points $(n-1, n)$ and $(n-1, 0)$, both of which minimizes the dot-product $(x, y) \cdot (-1, 0)$. Hence, $(1, 0) \in \mathcal{LV}(\mathcal{P}_{(n)})$. Now, suppose otherwise that $(0, 1) \in \mathcal{LV}(\mathcal{P}_\tau$ for some $\tau \in \{0, 1\}^n$. Then, there exists at least two points in $\mathcal{LP}_\tau$ minimizing the dot-product $(x, y) \cdot (0, -1) = -y$, i.e. maximizing y. But the configuration (τ, τ) gives rise to a point that attains the maximum $y = n$, and (τ, τ) is the only configuration attaining this maximum since $y = n$ implies that all vertical relations are equalities. This contradicts the earlier fact that at least two points maximize y. Therefore, $(0, 1) \notin \mathcal{LV}(\mathcal{P}_{(n)})$.

□

With the above two lemmas, we can now proceed to prove our main theorem.

Theorem 3.4. *$G(n) = 4(n - \lceil \frac{n}{4} \rceil) + 2$*

Proof. Let $\mathcal{V} = \mathcal{LV}(\mathcal{P}_{(n)})$. This theorem is equivalent to showing that $|\mathcal{V}| = 4(n - \lceil \frac{n}{4} \rceil) + 2$. Lemma 3.2.(iii-iv) bounds the location of the vectors in $\mathcal{V}$, while Lemma 3.2.(i-ii) highlights the symmetry of $\mathcal{V}$ so it is enough to count vectors in the first quadrant $\{(u, v) \mid u \geqslant 0, v \geqslant 0\}$. Meanwhile, keep in mind that if $(u, v) \in \mathcal{V}$, then $\gcd(u, v) = 1$.

On the axes, Lemma 3.3.iv indicates that $(1, 0), (-1, 0) \in \mathcal{V}$ while $(0, 1), (0, -1) \notin \mathcal{V}$. This accounts for the summand 2 in the desired formula. It remains for us to consider the region $\mathcal{R} = \{(u, v) \mid 0 < u \leqslant \lfloor \frac{n}{2} \rfloor, 0 < v \leqslant 2\}$. By Lemma 3.3.(i-iii), we have for some k,

$$|\mathcal{R} \cap \mathcal{V}| = \begin{cases} 2k + k & = n - \lceil \frac{n}{4} \rceil \quad \text{if } n = 4k \\ 2k + k & = n - \lceil \frac{n}{4} \rceil \quad \text{if } n = 4k+1 \\ (2k+1) + k & = n - \lceil \frac{n}{4} \rceil \quad \text{if } n = 4k+2 \\ (2k+1) + (k+1) & = n - \lceil \frac{n}{4} \rceil \quad \text{if } n = 4k+3 \end{cases}$$

Because of symmetry, the total number of vectors in $\mathcal{V}$ is given by $4|\mathcal{R} \cap \mathcal{V}| + 2 = 4(n - \lceil \frac{n}{4} \rceil) + 2$, which is the formula to be shown. □

4 Discussion and Conclusion

Applying Elizalde's theorem to the Neyman model of length n, we attain the upper bound $O(n^{d(d-1)}) = O(n^2)$, where we set d to the dimension of the Newton polytope rather than to the number of parameters. Hence, his theorem puts a quadratic bound on $G(n)$, while from Theorem 3.4, we see that $G(n)$ grows only linearly. A natural question to ask is: are there any reasons for this discrepancy?

It turns out that the culprit is Lemma 3.2.iv, which bounds the vectors $(u, v) \in \mathcal{LV}(\mathcal{P}_{(n)})$ by $|v| \leqslant 2$. Without this bound, $\mathcal{LV}(\mathcal{P}_{(n)})$ might be free to fill the plane, causing $G(n)$ to grow quadratically. A deeper look at the lemma shows that v is related to a special move which take configurations to configurations. Adjacent vertices on the coordinate polytopes must be linked by exactly one such move, and this fact gives the bound on v. This seems to suggest that in order to find better bounds on a certain graphical model, one should attempt to discover similar relationships between adjacent vertices, edges and other faces on the Newton polytopes of the coordinate polynomials.

In summary, we looked at the problem of gene-finding in biology and studied its relationship to graphical models. We developed the concept of an inference function for general graphical models, and saw that the number of such functions grows only polynomially through the Few Inference Functions Theorem. We studied the simple case of the Neyman hidden Markov model and was able to get an exact count of the inference functions for this model.

Having produced a formula for $G(n)$, the next goal would be to do the same for the homogeneous binary hidden Markov model described at the beginning of Section 3. Currently, the exact numbers [8] are only known up to $n = 7$, and it is exponentially expensive to compute it for higher n's. Also, as mentioned in [7], it is an open problem to give a combinatorial characterization of inference functions, i.e. given a map $\Sigma'^n \rightarrow \Sigma^q$, how do we tell if it is an inference function? Thus, one could try to characterize inference functions for the Neyman model or for the homogeneous binary hidden Markov model. By understanding these simpler cases, perhaps a universal method for attacking the general case will emerge.

5 Acknowledgements

We would like to thank Prof. Bernd Sturmfels for teaching the Math 127 course, Algebraic Statistics for Computational Biology, thus giving us the opportunity to carry our this project. The idea of considering the simpler homogenous binary hidden Markov model case was also suggested by Prof. Sturmfels. We also want to appreciate Dustin Cartwright, a consultant for the course, for his valuable discussions with us on gene-finding and his excellent advice on our presentation of the material. Last but not least, we thank Peter Huggins, a graduate student at the UC Berkeley Mathematics department, for sharing his insights on writing computer algorithms for calculating Minkowski sums of polytopes.

References

[1] I. Korf, P. Flicek, D. Duan and M. Brent, *Integrating Genomic Homology into Gene Structure Prediction*, Bioinformatics, vol. 17, no. 1, 2001, pp. S140-S148.

[2] J. Allen and S. Salzberg, *A Phylogenetic Generalized Hidden Markov Model for Predicting Alternatively Spliced Exons*, Algorithms for Molecular Biology, 2006, 1:14.

[3] W. Majoros, M. Pertea, A. Delcher and S. Salzberg, *Efficient Decoding Algorithms for Generalized Hidden Markov Model Gene Finders*, BMC Bioinformatics, 2005, 6:16.

[4] L. Pachter and B. Sturmfels, eds., *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge, UK, 2005.

[5] S. Elizalde and K. Woods, *Bounds on the Number of Inference Functions of a Graphical Model*, Statistica Sinica, to appear, arxiv:math.CO/0610233.

[6] P. Gritzmann and B. Sturmfels, *Minkowski Addition of Polytopes: Computational Complexity and Applications to Gröbner Bases*, SIAM Journal of Discrete Mathematics, 6:246-269, 1993.

[7] L. Pachter and B. Sturmfels, *The Mathematics of Phylogenomics*, SIAM Review, vol. 49, no. 1, 2007, pp. 3-31.

[8] B. Sturmfels, *Parametric Inference*, CMI Workshop (ASCB), Cambridge, MA, 2005. `http://www.claymath.org/programs/cmiworkshops/ascb/Sturmfels/Sturmfels.pdf`

[9] K. Fukuda and C. Weibel, *On f-vectors of Minkowski Additions of Convex Polytopes*, 2005. arxiv:math.CO/0510470.

[10] J. Wierer and N. Boston, *Newton Polytopes of Two-Dimensional Hidden Markov Models*, Experimental Mathematics, to appear, 2006.