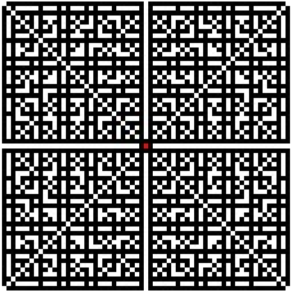
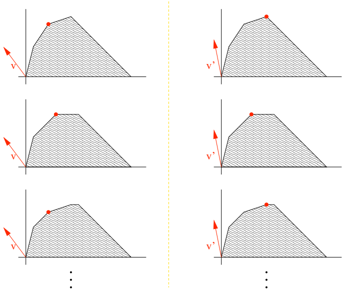
Primitive Sets and Inference Functions: Pure and Applied Combinatorics

Kevin Woods, Oberlin College
(joint work with Sergi Elizalde, Dartmouth)





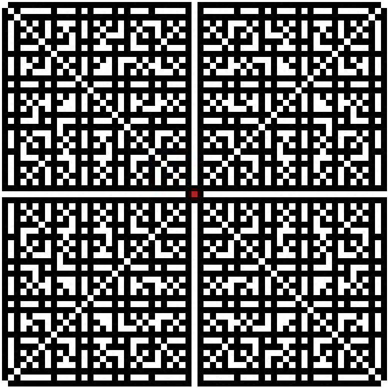
Two Stories

Pure Story: Geometry of Numbers

Applied Story: Inference for Bayesian networks

The Pure Story

Question: What proportion of $(a, b) \in \mathbb{Z}^2$ are visible from the origin?



i.e., a and b relatively prime
i.e., (a, b) is a basis for the lattice $\text{span}_{\mathbb{R}}(a, b) \cap \mathbb{Z}^2$.

Moral Proof

$$\begin{aligned}
\text{Probability} &= \left(1-\frac{1}{4}\right)\left(1-\frac{1}{9}\right)\left(1-\frac{1}{25}\right)\cdots \\
&= \frac{1}{\prod_{p \text{ prime}} 1/(1-p^{-2})} \\
&= \frac{1}{\sum_{i=1}^{\infty} i^{-2}} \\
&= \frac{1}{\zeta(2)}.
\end{aligned}$$

Immoral proof is not too bad either.

Moral Proof

$$\begin{aligned}\text{Probability} &= \left(1-\frac{1}{4}\right)\left(1-\frac{1}{9}\right)\left(1-\frac{1}{25}\right)\cdots \\ &= \frac{1}{\prod_{p \text{ prime}} 1/(1-p^{-2})} \\ &= \frac{1}{\sum_{i=1}^{\infty} i^{-2}} \\ &= \frac{1}{\zeta(2)}.\end{aligned}$$

Immoral proof is not too bad either.

Moral Proof

$$\begin{aligned}
\text{Probability} &= \left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{9}\right)\left(1 - \frac{1}{25}\right)\cdots \\
&= \frac{1}{\prod_{p \text{ prime}} 1/(1 - p^{-2})} \\
&= \frac{1}{\sum_{i=1}^{\infty} i^{-2}} \\
&= \frac{1}{\zeta(2)}.
\end{aligned}$$

Immoral proof is not too bad either.

Moral Proof

$$\begin{aligned}\text{Probability} &= \left(1-\frac{1}{4}\right)\left(1-\frac{1}{9}\right)\left(1-\frac{1}{25}\right)\cdots \\ &= \frac{1}{\prod_{p \text{ prime}} 1/(1-p^{-2})} \\ &= \frac{1}{\sum_{i=1}^{\infty} i^{-2}} \\ &= \frac{1}{\zeta(2)}.\end{aligned}$$

Immoral proof is not too bad either.

Generalizing

The probability that a point in $\mathbb{Z}^d$ is visible from the origin is $1/\zeta(d)$. [Nymann, 1974]

$S = \{s_1, s_2, \ldots, s_m\} \subseteq \mathbb{Z}^d$ is primitive if it is a basis for

$$\mathrm{span}_{\mathbb{R}}(S) \cap \mathbb{Z}^d$$

The probability that S is primitive is

$$\frac{1}{\zeta(d)\zeta(d-1)\cdots\zeta(d-m+1)}.$$

[Elizalde, W]

Moral proof

Write S as rows of a matrix:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

Column operations (over $\mathbb{Z}$) don't change primitivity.

$$\begin{bmatrix} \gcd(a,b,c) & 0 & 0 \\ d' & e' & f' \end{bmatrix}$$

Must have $\gcd(a,b,c) = 1$ (probability $1/\zeta(3)$).

$$\begin{bmatrix} \gcd(a,b,c) & 0 & 0 \\ d' & \gcd(e',f') & 0 \end{bmatrix}$$

Must have $\gcd(e',f') = 1$ (probability $1/\zeta(2)$).

Moral proof

Write S as rows of a matrix:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

Column operations (over $\mathbb{Z}$) don't change primitivity.

$$\begin{bmatrix} \gcd(a,b,c) & 0 & 0 \\ d' & e' & f' \end{bmatrix}$$

Must have $\gcd(a,b,c) = 1$ (probability $1/\zeta(3)$).

$$\begin{bmatrix} \gcd(a,b,c) & 0 & 0 \\ d' & \gcd(e',f') & 0 \end{bmatrix}$$

Must have $\gcd(e',f') = 1$ (probability $1/\zeta(2)$).

Moral proof

Write S as rows of a matrix:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

Column operations (over $\mathbb{Z}$) don't change primitivity.

$$\begin{bmatrix} \gcd(a,b,c) & 0 & 0 \\ d' & e' & f' \end{bmatrix}$$

Must have $\gcd(a,b,c) = 1$ (probability $1/\zeta(3)$).

$$\begin{bmatrix} \gcd(a,b,c) & 0 & 0 \\ d' & \gcd(e',f') & 0 \end{bmatrix}$$

Must have $\gcd(e',f') = 1$ (probability $1/\zeta(2)$).

Immoral proof

Difficult, but interesting

- Triangulations
- Volumes of cross sections of d-cubes [Ball, 1989]
- Prime number theorem

Applied Story

This has it backwards. The applied story came first.

It uses combinatorial tools, but also inspired the previous combinatorial result.

Recombination

Given: Genomes of parent strains:

AAAAAA

CCCCCC

Observed: Child strain

ATACCC

Inference: Explanation of what recombination happened.

Recombination

Given: Genomes of parent strains:

AAAAAA

CCCCCC

Observed: Child strain

ATACCC

Inference: Explanation of what recombination happened.

Tradeoff between recombination and mutation.

Recombination

Given: Genomes of parent strains:

AAAAAA

CCCCCC

Observed: Child strain

ATACCC

Inference: Explanation of what recombination happened.

Tradeoff between recombination and mutation.

Recombination

Given R and M, the costs of a recombination event or a mutation.

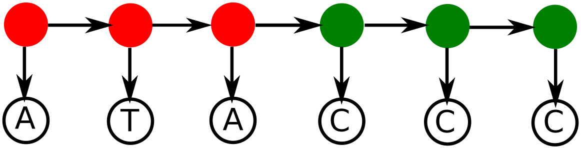
Minimize $R \cdot r + M \cdot m$ over all possible explanations
(r=number of recombinations, m=number of mutations).

Recombination

Given R and M, the costs of a recombination event or a mutation.

Minimize $R \cdot r + M \cdot m$ over all possible explanations
(r=number of recombinations, m=number of mutations).

This is one example of inference in a Bayesian network / graphical model.



Inference Functions

Given R and M and a length n,

Inference Function is a map
Input: Length n DNA sequence (the child)
Output: Best possible explanation

Different R and M may give different inference functions.
There seem to be

$$(2^n)^{4^n}$$

possible functions.

No Worries

Actually, there are only $O(n^2)$ inference functions.

In general, this is $O\left(n^{d(d-1)}\right)$, where d is the number of parameters. [Elizalde, W]

No Worries

Actually, there are only $O(n^2)$ inference functions.

In general, this is $O\left(n^{d(d-1)}\right)$, where d is the number of parameters. [Elizalde, W]

Example: For binary HMM's of length 5, there are

$$1461501637330902918203684832716283019655932542976$$

potential functions.

No Worries

Actually, there are only $O\left(n^2\right)$ inference functions.

In general, this is $O\left(n^{d(d-1)}\right)$, where d is the number of parameters. [Elizalde, W]

Example: For binary HMM's of length 5, there are

1461501637330902918203684832716283019655932542976

potential functions.

Only

5266

are actually inference functions. [Weibel]

Relation to Combinatorics

Translate to statement about Minkowski sums of polytopes:

The sum of a huge number of polytopes may have surprisingly few vertices. [Gritzmann, Sturmfels, 1993]

Relation to Combinatorics

Translate to statement about Minkowski sums of polytopes:

The sum of a huge number of polytopes may have surprisingly few vertices. [Gritzmann, Sturmfels, 1993]

In proving that the $O\left(n^{d(d-1)}\right)$ bound is tight, needed to know that a positive fraction of choices of

$$\{s_1, \cdots, s_m\} \subseteq \mathbb{Z}^d$$

are primitive.

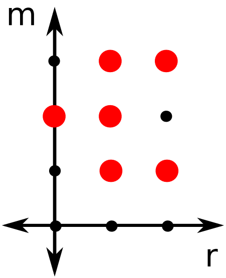
Translation to polytopes

Parents:

```
AAA
CCC
```

Given Child:

```
TAC
```



8 possible explanations.

Graph (r, m) for each explanation.

Translation to polytopes

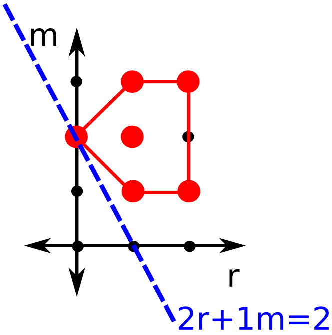
Example: $R = 2$, $M = 1$

Minimize

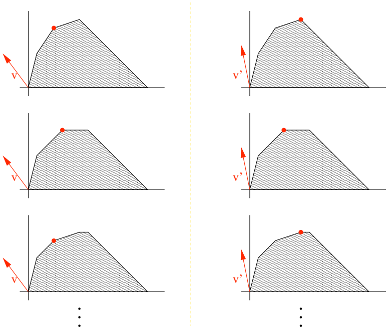
$$2r + 1m$$

over all points.

Linear Programming!



Translation to polytopes



Two different inference functions (for different R, M).

Inference Function = Vertex of Minkowski Sum

Translation to polytopes

Theorem (Gritzmann, Sturmfels)

Let $P_1, P_2, \ldots, P_k$ be polytopes in $\mathbb{R}^d$, and let m denote the number of non-parallel edges of $P_1, \ldots, P_k$. Then the number of vertices of $P_1 + \cdots + P_k$ is at most

$$2\sum_{j=0}^{d-1}\binom{m-1}{j}.$$

Note that this bound is independent of the number k of polytopes.