ARITHMETIC STATISTICS COURSE NOTES

ROBERT J. LEMKE OLIVER

1. Integer polynomials and Hilbert irreducibility

We begin with a very loose statement of Hilbert's irreducibility theorem[1]. Unpacking how to turn this loose statement into a rigorous one will motivate some of the key themes of this course. And of course, its proof is good too!

Theorem 1.1 (Hilbert irreducibility; loose version)**.** *Asymptotically,* 100% *of integer polynomials of degree n are irreducible and have Galois group S_n.*

It turns out this theorem is true in a very robust sense (meaning it's applicable in many settings and variations, with respect to many notions of complexity; that's why it's a template for an entire class of theorem), but let's consider what this is saying relative to the notion of complexity we considered in the previous section, the largest absolute value of the coefficients. What this theorem is considering, in words, is the following:

- First, count the number of integer polynomials with complexity at most X, i.e. the size of the set
$$\mathcal{P}_n(X) := \{f \in \mathbb{Z}[x] : f(x) = x^n + a_1x^{n-1} + \cdots + a_n, |a_i| \leq X \text{ for } 1 \leq i \leq n\}.$$
- Second, count the subset of those that are irreducible and have Galois group S_n, i.e. the size of the set
$$\mathcal{P}_n(X; S_n) := \{f \in \mathcal{P}_n(X) : f \text{ is irreducible, and } \mathrm{Gal}(f) \simeq S_n\}.$$
- Then, compute the proportion of those polynomials that are irreducible with Galois group S_n, and take the limit as $X \to \infty$, i.e.

$$\lim_{X\to\infty} \frac{\#\mathcal{P}_n(X; S_n)}{\#\mathcal{P}_n(X)}. \tag{1.1}$$

The Hilbert irreducibility theorem, in its rigorous form, asserts that this limit in (1.1) is equal to 1:

Theorem 1.2 (Hilbert irreducbility; rigorous version)**.** *With notation as above, for any integer $n \geq 2$,*
$$\lim_{X\to\infty} \frac{\#\mathcal{P}_n(X; S_n)}{\#\mathcal{P}_n(X)} = 1.$$

Date: January 20, 2022.

[1] The Hilbert irreducibility theorem is now regarded as a template for a much broader class of theorem, some of which look almost nothing like this one. This one is, I believe, close to Hilbert's original version.

1.1. **The counting problem.** We will take a look at actual data in some small cases in just a second to get a better feel of what this theorem is asserting, but before we do that, it's convenient to first consider the counting problem. In this case, we're just trying to estimate $\#\mathcal{P}_n(X)$. If X is a positive integer, then there are exactly $2X+1$ integers in the interval $[-X,X]$, and hence exactly $(2X+1)^n$ choices for the n different coefficients of a polynomial $f\in\mathcal{P}_n(X)$. It's not a problem to assume that X is an integer, but it's convenient for the proof to come (and as a means of introducing some useful notation) to consider what happens when X is not an integer. We could still get an exact formula by replacing X by its floor – the number of integers in $[-X,X]$ is exactly $2\lfloor X\rfloor+1$ – but instead, it's more motivating to consider how wrong our "simple" estimate $(2X+1)^n$ can be. In particular, most counting problems in arithmetic statistics don't admit an exact formula (for example, we're not going to get an exact answer for $\#\mathcal{P}_n(X;S_n)$) and so it's useful to understand how to write rigorous but inexact formulas.

In the case of counting integers in the interval $[-X,X]$, the formula $2X+1$ is always at least as large as the right answer, $2\lfloor X\rfloor+1$, and it can only off by just less than 2 at the worst, coming from when X is just below an integer (e.g., $X=99.999$). That means the number of integers in $[-X,X]$ is always between $2X-1$ and $2X+1$, so we can write it as $2X+\theta$ for some $\theta\in[-1,1]$. Implicitly, θ is a function of X, but the content of this statement is that it's a *bounded* function. Back to polynomials, again by considering the choices for the n different coefficients, this means $\#\mathcal{P}_n(X)=(2X+\theta)^n$. Using the binomial theorem, we rewrite this in the temporarily cumbersome form,

$$\#\mathcal{P}_n(X)=(2X)^n+\binom{n}{1}(2X)^{n-1}\theta+\binom{n}{2}(2X)^{n-2}\theta^2+\cdots+\theta^n.$$

What's important to absorb here is not the exact form – again, we're ultimately not shooting for an exact formula – so much as the order of magnitude of the various terms, remembering that we'll ultimately be taking the limit as $X\to\infty$. In particular, every subsequent term on the right-hand side is of a smaller order of magnitude than the first term, with the second term being the largest of these subsequent terms. Thus, we expect the contribution of all subsequent terms to have order of magnitude X^{n-1}, which we make rigorous by noting that the function

$$\frac{\binom{n}{1}(2X)^{n-1}\theta+\binom{n}{2}(2X)^{n-2}\theta^2+\cdots+\theta^n}{X^{n-1}} \tag{1.2}$$

is bounded as $X\to\infty$. We thus introduce some notation that will enable us to write $\#\mathcal{P}_n(X)=(2X)^n+O(X^{n-1})$, where the big-oh term $O(X^{n-1})$ is keeping track of the order of magnitude of the hidden parts of this formula.

Notation (Big-oh)**.** Given two functions $f(X)$ and $g(X)$ with $g(X)$ strictly positive, we say that $f(x)=O(g(X))$ if the ratio $f(X)/g(X)$ is bounded as $X\to\infty$. If we write $f(X)=g(X)+O(h(X))$, then we mean that the difference $f(X)-g(X)$ is $O(h(X))$.

Remark. If the bound implicit in a big-oh statement depends on a parameter in the problem, we sometimes denote that with a subscript. For example, as noted earlier, the function in (1.2) is bounded as $X\to\infty$, but the bound depends on the parameter n. We therefore might more specifically say it's $O_n(X^{n-1})$ and write $\#\mathcal{P}_n(X)=(2X)^n+O_n(X^{n-1})$. By contrast, in the formula $2X+\theta$ for the number of integers in $[-X,X]$, the parameter θ is between -1 and 1, neither of which depends on n in any way. We'd therefore write this formula as $2X + O(1)$, without any subscript. I'll usually include subscripts when necessary to be precise, but you may pretend they're not there without losing anything.

We summarize this lengthy discussion of the (fairly easy!) counting problem in the following:

Proposition 1.3. *For any $n \geq 1$ and $X \geq 1$, $\#\mathcal{P}_n(X) = 2^n X^n + O_n(X^{n-1})$. If X is an integer, then we have further $\#\mathcal{P}_n(X) = (2X+1)^n$.*

1.2. **Actual data.** Let's now actually look at some data for polynomials. For any specific values of n and X, we can compute all polynomials in $\mathcal{P}_n(X)$, see whether they factor, and (particularly with the aid of a computer) their Galois group, assuming that they're irreducible. I've done this only for some very small degrees and values of X, but this data is representative of the general picture.

1.2.1. *Cubic polynomials.* If $n = 3$, then any polynomial $f \in \mathcal{P}_3(X)$ has at most three distinct roots in $\mathbb{C}$. If f is irreducible, these three roots must be distinct, and the Galois group must rearrange or permute them in some way. In particular, if f is irreducible, then $\text{Gal}(f)$ will be a subgroup of the symmetric group S_3 (the group of all permutations of $\{1, 2, 3\}$), and it turns out that it's either all of S_3 or the alternating subgroup A_3, which (for $n = 3$ only!) is the same as the cyclic subgroup $C_3 = \langle(1, 2, 3)\rangle$. [2]

For the values $X = 5$, 10, and 20, we now compute (by computer) how many polynomials in $\mathcal{P}_3(X)$ are irreducible vs. reducible, and of those that are irreducible, how many have Galois group S_3 vs. A_3.

X	$\#\mathcal{P}_3(X)$	# Irred.	# Red.	# Gal(f) $\simeq S_3$	# Gal(f) $\simeq A_3$
5	1,331	1,002	329	976	26
10	9,261	7,878	1,383	7,760	118
20	68,921	63,274	5,647	62,906	368

Table 1. Statistics of integer cubic polynomials in $\mathcal{P}_3(X)$, i.e. those $f(x) = x^3 + a_1x^2 + a_2x + a_3$ with each $|a_i| \leq X$.

Theorem 1.2 asserts that the S_3 column should make up a larger and larger proportion of $\mathcal{P}_3(X)$ as X tends to infinity, eventually making up essentially 100%. For the data points we have, the percentages are about 73%, 84%, and 91%, which, considering that we've only taken $X = 20$, is not shabby!

1.2.2. *Quartic polynomials.* We now consider the analogous computation if $n = 4$, though with fewer values of X owing to the increased size of the problem. It turns out that there are five choices for the Galois group of an irreducible quartic polynomial: the full symmetric group S_4, the alternating group A_4, the dihedral group D_4, the Klein four group V_4, and the cyclic group C_4. We find the following:

As before, we can compute the percentage of polynomials that are irreducible with Galois group S_4; we find them to be about 71% and 84%. This is exhibiting the same general trend as the degree 3 case.

[2] The operative fact here is that the Galois group of an irreducible polynomial of degree n is a *transitive* subgroup G of S_n, which means that for every $i \leq n$, there is some element of G that sends 1 to i. For example, the cyclic subgroup $\langle(1, 2)\rangle \subseteq S_3$ is *not* transitive, since it has no element that sends 1 to 3.

X	$\#\mathcal{P}_4(X)$	# Irred.	# Red.	# S_4	# A_4	# D_4	# V_4	# C_4
5	14,641	11,246	3,395	10,382	16	774	46	28
10	194,481	169,214	25,267	163,588	182	5,118	218	108

Table 2. Statistics of integer quartic polynomials in $\mathcal{P}_4(X)$, i.e. those $f(x) = x^4 + a_1x^3 + a_2x^2 + a_3x + a_4$ with each $|a_i| \le X$.

1.2.3. *Van der Waerden's conjecture.* Another observation we can make from this data is that it appears in the process of going from (all polynomials) to (irreducible polynomials) to (irreducible polynomials with Galois group S_n), it's the first step that removes the most. In other words, it appears that there are more reducible polynomials than there are irreducible polynomials with Galois group not equal to S_n. This was also noticed by van der Waerden in 1936, and he conjectured that this phenomenon should hold for any degree n. There are always at least $(2X - 1)^{n-1}$ reducible polynomials – choose the constant term equal to 0 – and in fact, this is essentially the right order of magnitude: the number of reducible polynomials in $\mathcal{P}_n(X)$ turns out to be $O_n(X^{n-1})$. (We'll prove a slighly weaker version of this below.) Van der Waerden's conjecture is that this is also at most the order of magnitude of the non-S_n irreducible polynomials:

Conjecture 1.4 (Van der Waerden; 1936). *For any n, the number of irreducible polynomials in $\mathcal{P}_n(X)$ with Galois group not equal to S_n is $O_n(X^{n-1})$. Equivalently, $\#\mathcal{P}_n(X; S_n) = 2^nX^n + O_n(X^{n-1})$.*[3]

After 85 years, van der Waerden's conjecture was just proven by Bhargava in Summer 2021:

Theorem 1.5 (Bhargava; 2021). *Van der Waerden's conjecture is true.*

We won't discuss this theorem more today, but I am hoping to at least present the ideas behind its proof later in the semester.

1.3. **A soft proof of Hilbert irreducibility.**

[3] This equivalence comes from noting that any polynomial not in $\mathcal{P}_n(X; S_n)$ is either irreducible or has small Galois group, and using the above claimed bound on reducible polynomials.