NUMERICAL SECONDARY TERMS IN A COHEN-LENSTRA CONJECTURE ON REAL QUADRATIC FIELDS

CODIE LEWIS AND CASSANDRA WILLIAMS

Abstract. In 1984, Cohen and Lenstra made a number of conjectures regarding the class groups of quadratic fields. In particular, they predicted the proportion of real quadratic fields with class number divisible by an odd prime. We numerically investigate the difference between reality and these predictions. Using 4 million data points, we perform a curve fitting of the difference with a monomial term and demonstrate that there is reason to believe the term can be effectively approximated within the scope of our data set for several odd primes less than 30. We use cross-validation to show that including our monomial term as a secondary term to the original conjecture reduces the overall error.

1. Introduction

Though class groups of number fields have been studied by the number theory community since the latter half of the 19th century, it was not until the rise of modern computing that it was possible to compute a large set of examples. In the early 1980's it was noted that certain finite abelian groups occur much less frequently than others as class groups. In their classic 1984 paper, Cohen and Lenstra [3] gave the theoretical basis for a heuristic to explain these experimental observations on the frequency with which groups occur as the class group of a number field. Cohen and Lenstra then used their heuristic to generate a set of 12 conjectures about various attributes (such as size or group structure) of class groups of imaginary and real quadratic fields.

With advances in both technology and the efficiency of algorithms for computing class groups, various authors produced larger and larger data sets of class groups, often framing their numerical results as support for the conjectures of Cohen and Lenstra. For example, each of [6, 7, 8, 11] gave new or improved algorithms for computing class groups of quadratic fields, followed by a data comparison to conjectures from [3]. Both Jacobson in 1998 [6] and te Riele and Williams in 2003 [11] construct real quadratic fields and give numerical tables to support various conjectures for small primes. In [6], the author computes the density of fields of odd discriminant less than 10^9 and with a class number having a given prime divisor. On the other hand, the authors of [11] consider the actual density of fields with prime discriminant less than $2 \cdot 10^{11}$ and a given odd class number. In each case, the actual densities approach those of the conjecture.

However, it also appears that the convergence of the data to the conjectured densities is quite slow in many cases. (See, for example, Figure 1.) This implies that we may be able to refine the original conjectures via secondary terms.

Date: June 15, 2017.

There have been recent attempts to define such secondary terms analytically. For example, Hough [5] conjectured a negative secondary term for the mean size of the k-part of the class group of an imaginary quadratic field. Taniguchi and Thorne [10], and Bhargava, Shankar, and Tsimerman [1] each proved the secondary term for the number of cubic number fields conjectured by Roberts in 2001 [9]. The two papers used very different methods: Taniguchi and Thorne used the Shintani zeta function, while Bhargava, Shankar, and Tsimerman gave a geometric argument. The result on cubic number fields can be reformulated to instead give a secondary term for the size of the three-part of the class group of a real quadratic field.

Unfortunately, for many of the original Cohen-Lenstra conjectures the methods in [1], [5], and [10] do not apply. In the present work, we focus on predicting secondary terms for one of the Cohen-Lenstra conjectures for real quadratic fields using strictly numerical methods.

2. Real quadratic fields and the Cohen-Lenstra heuristics

Let d be a positive square-free integer so that $\mathbb{Q}(\sqrt{d})$ is a real quadratic field with fundamental discriminant D. We collect in this section some classical results on real quadratic fields and their class groups that will be useful in the sequel.

Lemma 1. *Let d be a square-free integer. Then the discriminant D of the quadratic field $\mathbb{Q}(\sqrt{d})$ is also a fundamental discriminant and is given by*

$$D = \begin{cases} d & \text{if } d \equiv 1 \bmod 4, \\ 4d & \text{if } d \equiv 2, 3 \bmod 4. \end{cases}$$

Remark 2. A corollary to Lemma 1 is that quadratic fields with the same fundamental discriminant are isomorphic, and so counting fields by fundamental discriminant ensures that we have only counted unique fields.

The next lemma counts such fields, and is a standard result in analytic number theory.

Lemma 3. *Let $Q(X)$ be the number of non-isomorphic real quadratic fields with fundamental discriminant less than or equal to the positive integer X. Then*

$$Q(X) = \frac{3}{\pi^2}X + O(X^{1/2}).$$

The *class group* of a number field K is a finite abelian group constructed as the quotient of the fractional ideals of K modulo the principal fractional ideals of K, and the *class number* is its size. If the ring of integers of K has unique factorization, the class group will be trivial and the class number will be 1. The class group (and thus the class number) can be interpreted as a measure of the extent to which unique factorization fails in the ring of integers of K.

Many of the conjectures in [3] are stated as the probability of a class group having a given attribute. Conjecture C7 concerns the probability of an odd prime dividing the class number.

Conjecture C7 ([3])**.** *Let d be a positive squarefree integer, let p be an odd prime, and let h be the size of the odd part of the class group of $\mathbb{Q}(\sqrt{d})$. Then the probability that p divides h is*

$$1 - \prod_{k\geq 2}(1 - p^{-k}).$$

In what follows, we denote this probability by ξ_p.

For our investigation it is more useful to consider this conjecture as an asymptotic density statement in terms of a discriminant bound X. We restate Conjecture C7 in this context below.

Conjecture C7*. *Let d be a positive squarefree integer so that $\mathbb{Q}(\sqrt{d})$ is a real quadratic field with fundamental discriminant D. Let p be an odd prime, and let h be the size of the odd part of the class group of $\mathbb{Q}(\sqrt{d})$. Then*

$$\lim_{X\to\infty} \frac{\#\{\mathbb{Q}(\sqrt{d}) \mid p|h \textit{ and } D < X\}}{Q(X)} = 1 - \prod_{k\geq 2}(1-p^{-k}) = \xi_p.$$

In the next section we will use data to empirically investigate the discrepancy between this conjectured value and the actual density of such real quadratic fields.

3. Methods

In order to calculate the actual statistics for Conjecture C7*, we first generated the class numbers of a large set of real quadratic fields. Computations were done in Sage [[4]]. Utilizing the `class_number` method for the `quadratic_field` class, we computed the class numbers of all real quadratic fields $K = \mathbb{Q}(\sqrt{d})$ for square-free integers $0 < d < 4 \cdot 10^6$ (about $2.4 \cdot 10^6$ fields). Each of these fields is unique (see Remark 2), and we used these class numbers, ordered by the field discriminant, to complete the following computations of statistics related to Conjecture C7*. Fields with fundamental discriminant $D > 4 \cdot 10^6$ were not used in our calculations.

Even using this class and function in Sage, computing the list of class numbers was the most computationally expensive process. Although we lacked the technology to do so, the computation of class numbers is parallelizable so a future investigation could generate a larger data set more quickly. Jacobson [6] also investigated Conjecture C7 for real quadratic fields, but presented data only for fields of odd discriminant less than 10^9. Odd discriminants account for (asymptotically) one third of all fundamental discriminants. Thus, while our discriminant bound is lower than theirs, using all fundamental discriminants below that bound gives us a denser set of data points from which to work.

Conjecture C7* is stated in terms of the density of fields with class number divisible by an odd prime. The calculation of the actual density of such fields is also parallelizable although such a consideration is not necessary since a pattern can be discerned from calculating the statistics at fixed intervals instead of at every valid fundamental discriminant. In what follows, we compute any statistics for discriminant bounds X at intervals of 10,000 and for all odd primes less than 100.

Our script counted the number of fields of discriminant $D < X$ with class number divisible by the chosen prime p and then divided by the number of fields with discriminant $D < X$. This actual density is denoted $\sigma_p(X)$ in the following equations. That is, as $X \to \infty$

$$\#\{\mathbb{Q}(\sqrt{d}) \mid p|h \text{ and } D < X\} = \sigma_p(X)\frac{3X}{\pi^2}.$$

Apply Lemma 3 to Conjecture C7* and rearrange the terms. Then for X sufficiently large, the number of distinct quadratic fields with discriminant $D < X$ which have class number divisible by the odd prime p is approximately the product of $Q(X)$ and ξ_p. More concisely we have that as $X \to \infty$, Cohen and Lenstra's conjecture predicts that

$$\#\{\mathbb{Q}(\sqrt{d}) \mid p|h \text{ and } D < X\} \sim \xi_p \, \frac{3X}{\pi^2}.$$

Plots of the actual and predicted number of fields satisfying Conjecture C7 at each discriminant bound X show that there is a discrepancy between these values. In particular, as the discriminant bound X grows, the predicted value overestimates the actual value fairly dramatically, as seen in Figure 1.

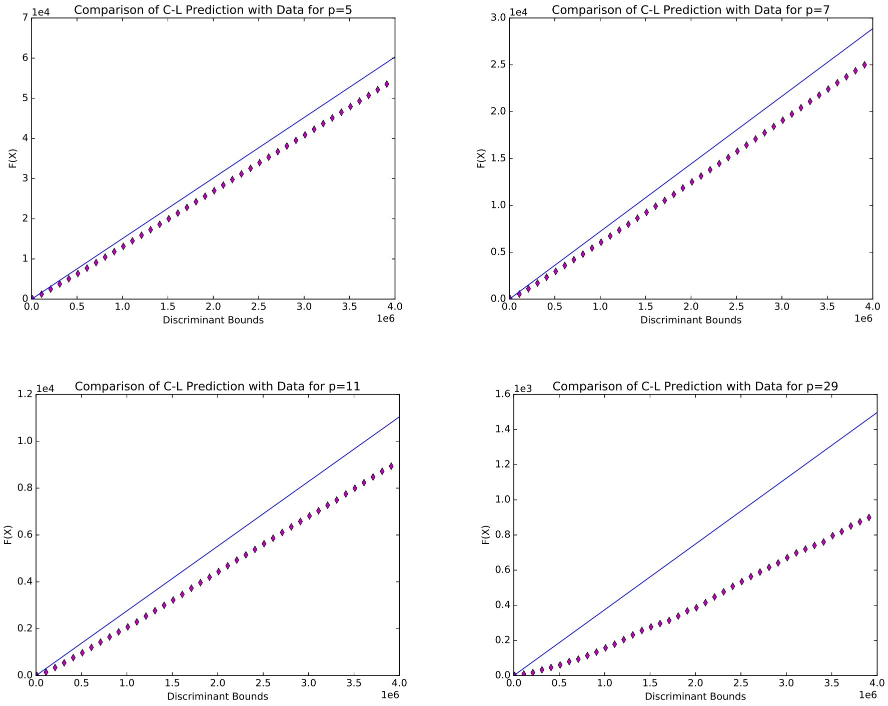


FIGURE 1. Plots of the actual number of fields (solid line) and predicted number of fields (diamonds) for $p = 5, 11, 17$, and 29.

Consider the difference between the predicted and actual field counts,

$$\frac{3X}{\pi^2}[\xi_p - \sigma_p(X)]. \tag{3.1}$$

Fitting a curve to this difference will yield a function which could be used as a secondary term to modify the original Conjecture C7. The plot of this difference is concave down and increasing for each p (as seen in Figure 2), so we will model the error as a monomial of form CX^s with $0 < s < 1$. (A logarithm model for the error was attempted but failed to produce a sufficient fit to the data.)

Then, as $X \to \infty$, we predict

$$\frac{3X}{\pi^2}[\xi_p - \sigma_p(X)] = C_p X^{s_p} \tag{3.2}$$

and thus

$$\#\{\mathbb{Q}(\sqrt{d}) \mid p|h \text{ and } D < X\} = \sigma_p(X)\frac{3X}{\pi^2} = \xi_p\frac{3X}{\pi^2} - C_pX^{s_p}.$$

It is the function $C_pX^{s_p}$ that we will analyze for the remainder of the paper.

Applying the logarithm to (3.2), we have

$$\log\left(\frac{3X}{\pi^2}[\xi_p - \sigma_p(X)]\right) = \log(C_pX^{s_p}) = \log(C_p) + s_p\log(X). \tag{3.3}$$

We apply to the log of the data points a standard linear fit by least squares to find the coefficient and the exponent for the secondary term for each prime p and each discriminant bound X (again, in intervals of 10,000). The fitted curve was then compared to the actual difference (the left hand side of (3.2)). Additionally, we looked for patterns in the coefficients and exponents for each prime as we increased the discriminant bound.

Finally, we calculated the error between our fit and the actual difference using cross-validation. For this, we divided our data (every 10,000th statistic) into 5 bins for a total of 80 data points per bin. Then we computed the fit model excluding one bin. After the fit model was determined, we calculated the fit error for the excluded bin. After repeating this process five times, once for each bin, we then averaged the five errors into one fit error for the prime p. This was done for each odd prime up to 29, and is called "CV Error" in what follows.

Because the number of fields satisfying the conjecture for a given prime divisor are dramatically different between primes, we scaled the error from the cross-validation so that we could compare these errors between primes. We chose to scale by the Cohen-Lenstra prediction ξ_p, which is equivalent to scaling by the predicted number of fields.

4. Results

Although the differences and curves of best fit (equations (3.1) and (3.2), respectively) were computed for all odd p less than 100, the number of fields in our sample with large odd prime factors in their class number is too small to confidently identify any patterns in the exponents or coefficients of those curves. Therefore we present results only for odd primes less than 30 because they exemplify the patterns we found while also including values of p for which there were not enough data points to suggest convergence of s_p or C_p as X increased.

In all plots, the discriminants on the x-axis are the discriminant bounds X. For example, a point above X=100,000 represents the value using all real quadratic fields with fundamental discriminant less than 100,000. Some markers are omitted in the plots to prevent marker overlap.

Figure 2 below gives plots of the difference (3.1) (labeled on the y-axis as $G(X)$) with their curves of best fit determined by equation (3.2) for four primes. (For the plots of $G(X)$ and the curves of best fit for other primes, please see Appendix A.)

We computed the coefficients C_p and exponents s_p as the discriminant bound X increased for each p with the goal of determining whether these coefficients and exponents showed convergent behavior within or across primes. As X increased, many of the primes' coefficients and exponents demonstrated seemingly stable behavior while others varied too much to support any conjectures without more data (see Figures 3 and 4). Overall, despite the values not stabilizing within the reach of

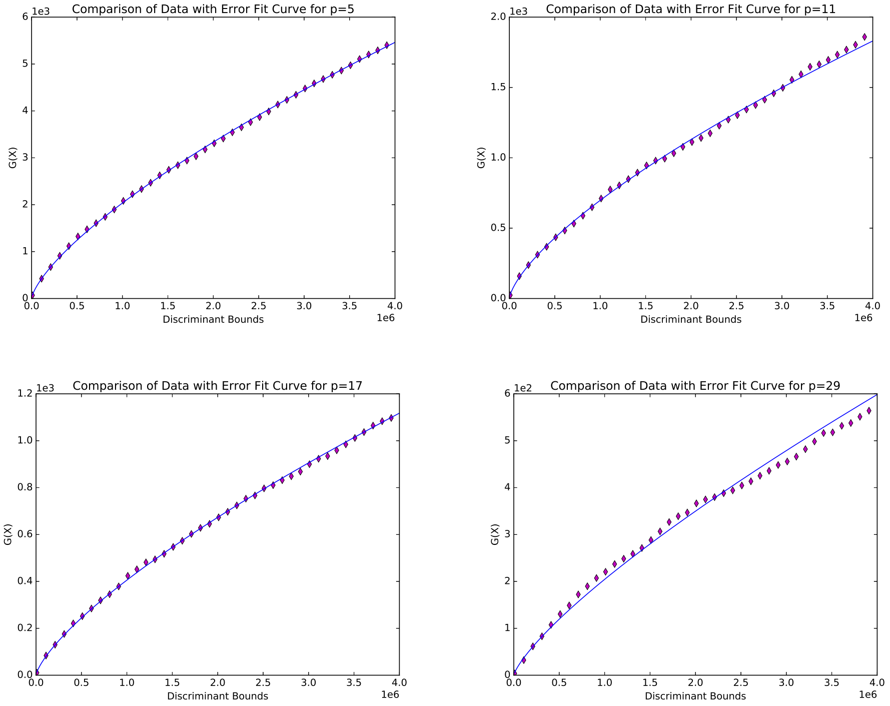


Figure 2. Plots of equation (3.2) with fitted curve from equation (3.3) for $p = 5, 11, 17$, and 29

our data for some primes, there does seem to be some predictability to the monomial term given by equation (3.3). We view this as evidence that the assumption of a monomial secondary term is valid. Further discussion of these values is in the next section.

Table 1 contains the parameters for the error function when we use every 10,000th statistic over our full data set. We also include the root-mean-square error of the fit for each prime as a measure of how much variability should be expected when more data points are calculated.

We found small proportions of error when applying the cross-validation calculation to our models for each prime. That is, we computed C_p and s_p using a subset of our data, then computed the error between our predicted fit and the remaining data. In Table 2, the cross-validation error (CV Error) is the average of the absolute errors given by the five trials in the cross-validation method (measured in number of fields) and the Scaled CV Error gives that error scaled by ξ_p in order to produce values that can be compared between primes. Both errors are truncated to an integer value.

Notice that while the absolute error is very different between primes, the scaled error is comparable between all primes.

Table 2 shows that the fit curve for $p = 5$ is a worse fit for the data than the fit curve for $p = 29$. This may seem counter-intuitive looking at the comparison of the fit curves and plots given above in Figure 2. In Figure 5, we plot the fit curves

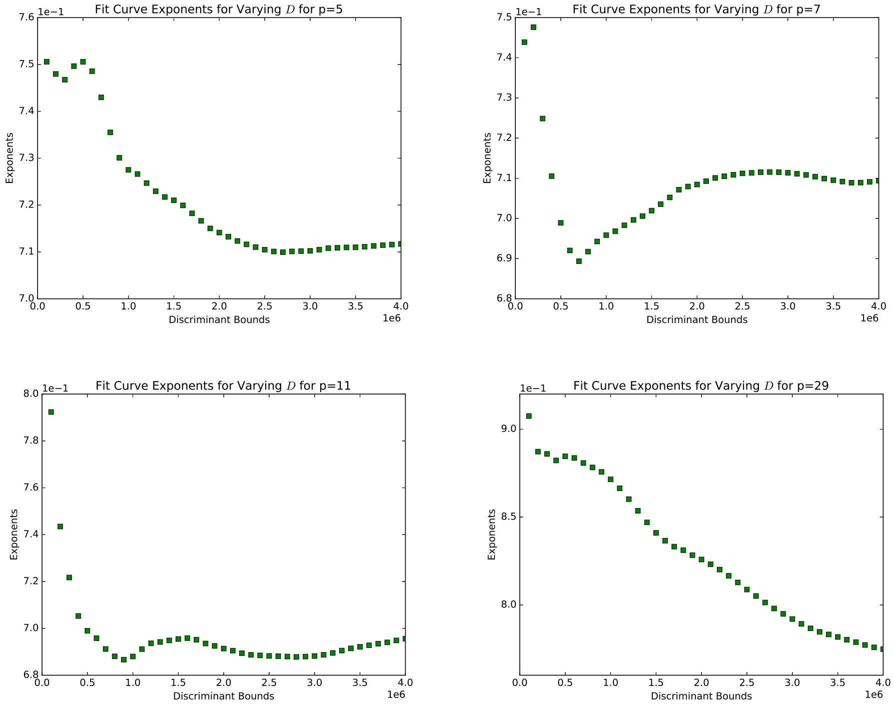


FIGURE 3. Plots of exponents s_p for varying discriminant bound and $p = 5, 7, 11, 29$

Prime	Exponent (s_p)	Coefficient (C_p)	RMSE
$p = 3$	0.822	0.107	50.425
$p = 5$	0.712	0.109	31.896
$p = 7$	0.709	0.067	22.967
$p = 11$	0.696	0.047	22.500
$p = 13$	0.701	0.035	9.494
$p = 17$	0.731	0.017	7.906
$p = 19$	0.730	0.015	18.370
$p = 23$	0.740	0.011	14.982
$p = 29$	0.775	0.005	15.599

TABLE 1. Parameters and root-mean-square error for the error function up to discriminant bound $X = 4 \cdot 10^6$.

for $p = 5, 7, 11$, and 29 on the same axes to avoid the effect of scale on the visual representation of error.

5. Discussion

Some interesting patterns emerge in the coefficients and exponents. First, as the discriminant bound X increases the exponents s_p seem to converge for each odd

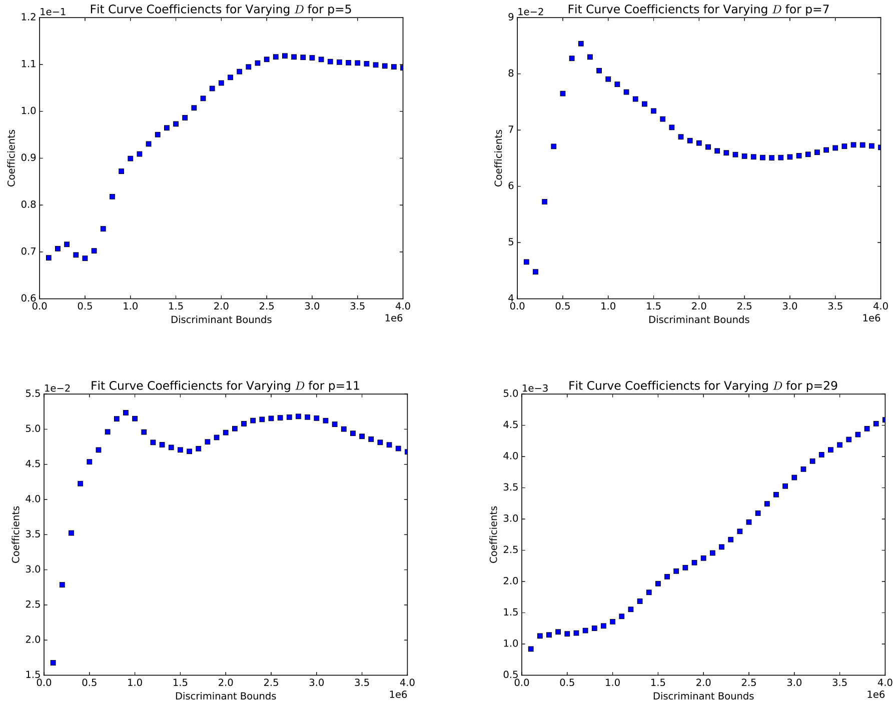


FIGURE 4. Plots of coefficients C_p for varying discriminant bound and $p = 5, 7, 11, 29$

Prime	CV Error	Scaled CV Error
$p = 3$	23,408	146,473
$p = 5$	7914	159,625
$p = 7$	3869	163,020
$p = 11$	1623	178,704
$p = 13$	1129	176,258
$p = 17$	661	180,072
$p = 19$	495	169,324
$p = 23$	356	180,490
$p = 29$	185	150,771

TABLE 2. Quality of fit

prime p. Moreover, the exponents s_p approach similar limit values (between 0.7 and 0.8) for all p less than 30. On the other hand, the coefficients C_p seem to vary depending on the value of p, but do appear to approach a limit for constant p and increasing discriminant bound X.

The $p = 3$ case defies both of these general trends. For this prime, there is an approximately linear change of the exponent and coefficient values for increasing X greater than 1.5 million (for plots, see Appendix A). However, it might be

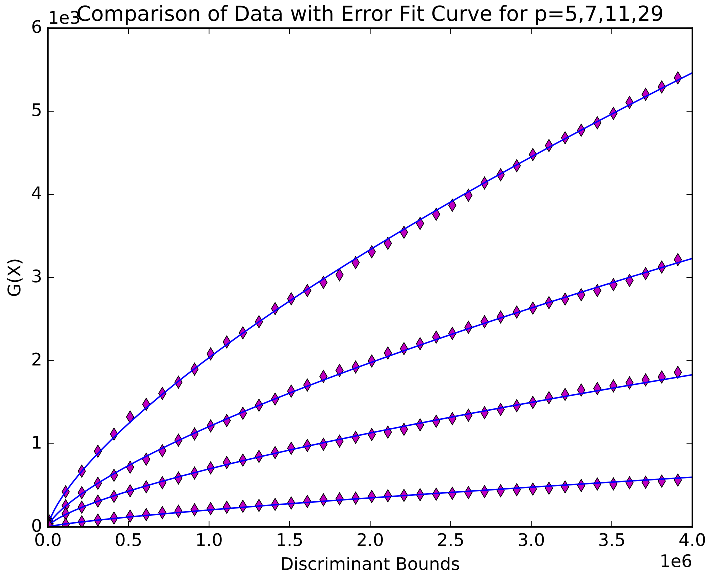


FIGURE 5. Comparison of the data points and fit curves for $p = 5, 7, 11, 29$

reasonable to suspect that since there are so many more fields, especially with smaller discriminant, for which $p = 3$ divides its class number, the exponents and coefficients may not fit the overall pattern as well as those for larger primes.

Analysis of our two measures of error in the secondary term suggest that the coefficients and exponents we obtained were a reasonable fit for the data and therefore we believe that a single monomial secondary term gives rise to significant improvement in Conjecture C7 of Cohen and Lenstra [3].

A proper investigation of odd primes greater than 30 would require generation of far more data than could be efficiently constructed using the computing hardware available to us at the time of data generation.

Though these results are experimental and represent a small portion of the possible data, they do lend support to the existence of a secondary term for Conjecture C7 of [3]. Under our assumption of a monomial model for the error, a modification of the conjecture might be of the form

$$\#\{\mathbb{Q}(\sqrt{d}) \mid p|h \text{ and } D < X\} \sim \xi_p \frac{3X}{\pi^2} - C_p X^{s_p}$$

where C_p depends on p, and s_p may be coherent for odd primes and may have value(s) between 0.7 and 0.8. At the time of this writing, we are not aware of any analytic approach to finding a secondary term for Conjecture C7 even for particular primes. We would be interested to see such a method and compare to our numerical results.

References

[1] Manjul Bhargava, Arul Shankar, and Jacob Tsimerman. On the Davenport-Heilbronn theorems and second order terms. *Invent. Math.*, 193(2):439–499, 2013.

[2] H. Cohen and H. W. Lenstra, Jr. Heuristics on class groups. In *Number theory (New York, 1982)*, volume 1052 of *Lecture Notes in Math.*, pages 26–36. Springer, Berlin, 1984.

[3] Henri Cohen and Hendrik W. Lenstra, Jr. Heuristics on class groups of number fields. In *Number theory, Noordwijkerhout 1983 (Noordwijkerhout, 1983)*, volume 1068 of *Lecture Notes in Math.*, pages 33–62. Springer, Berlin, 1984.

[4] The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 7.3)*, 2016. `http://www.sagemath.org`.

[5] B. Hough. Equidistribution of bounded torsion CM points. *ArXiv e-prints*, May 2016.

[6] Michael J. Jacobson, Jr. Experimental results on class groups of real quadratic fields (extended abstract). In *Algorithmic number theory (Portland, OR, 1998)*, volume 1423 of *Lecture Notes in Comput. Sci.*, pages 463–474. Springer, Berlin, 1998.

[7] Michael J. Jacobson, Jr., Shantha Ramachandran, and Hugh C. Williams. Numerical results on class groups of imaginary quadratic fields. In *Algorithmic number theory*, volume 4076 of *Lecture Notes in Comput. Sci.*, pages 87–101. Springer, Berlin, 2006.

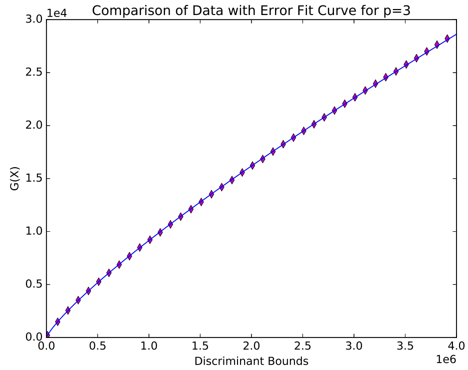
[8] R. A. Mollin and H. C. Williams. Computation of the class number of a real quadratic field. *Utilitas Math.*, 41:259–308, 1992.

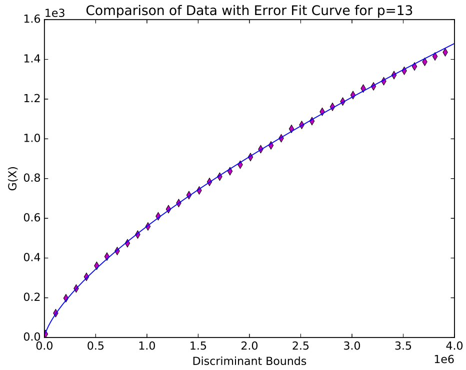
[9] David P. Roberts. Density of cubic field discriminants. *Math. Comp.*, 70(236):1699–1705 (electronic), 2001.

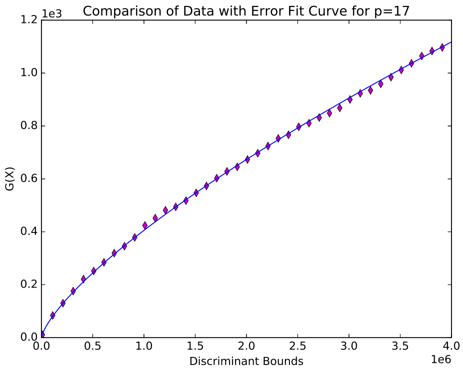
[10] Takashi Taniguchi and Frank Thorne. Secondary terms in counting functions for cubic fields. *Duke Math. J.*, 162(13):2451–2508, 2013.

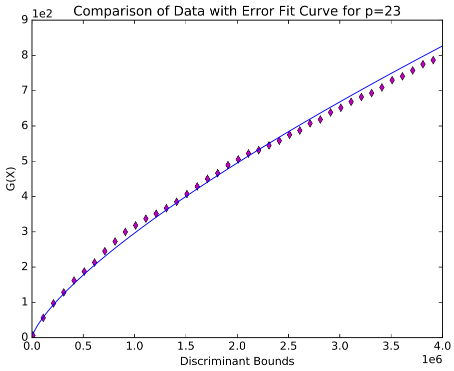
[11] Herman te Riele and Hugh Williams. New computations concerning the Cohen-Lenstra heuristics. *Experiment. Math.*, 12(1):99–113, 2003.

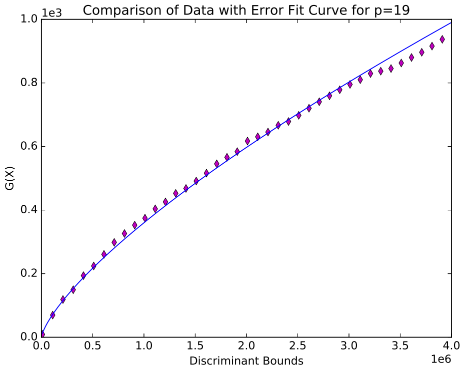
Appendix A. Cohen-Lenstra Error Fitting for other $p < 30$



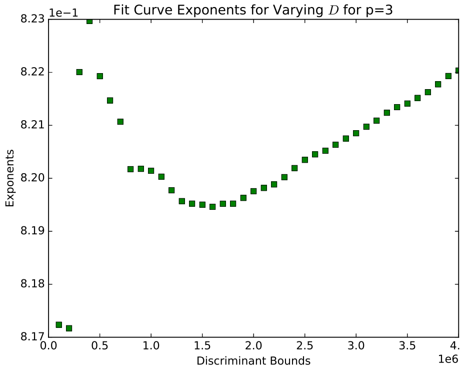


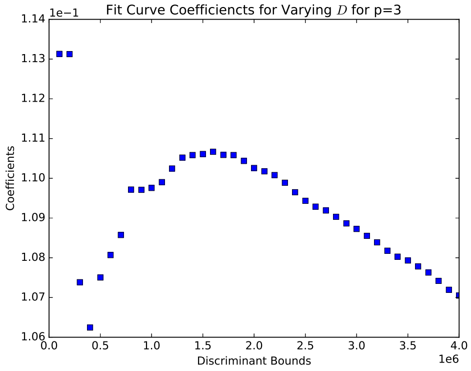


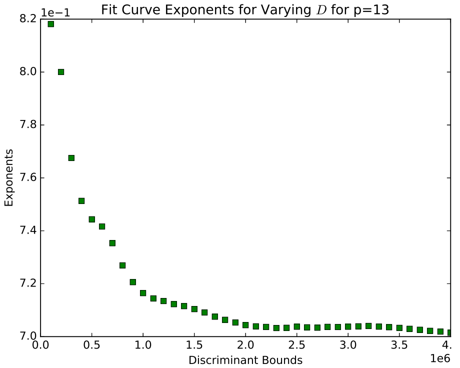


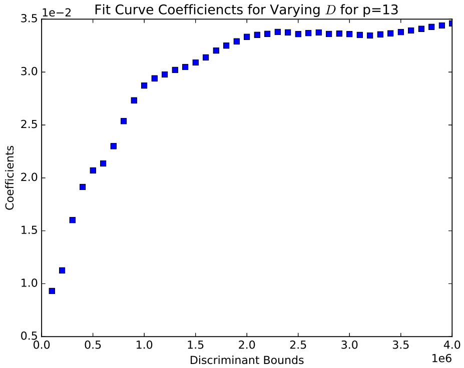


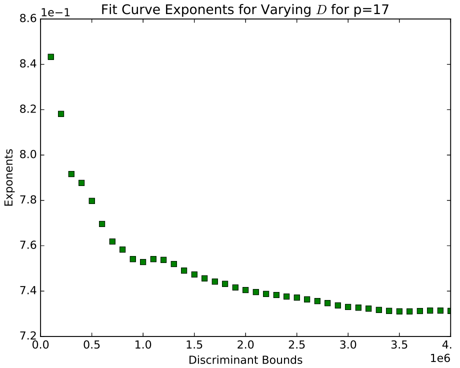
Appendix B. Exponent and Coefficient Plots for other $p < 30$

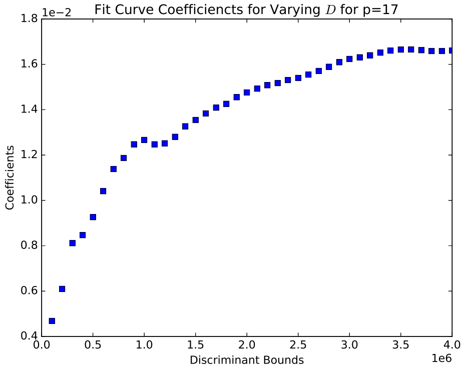


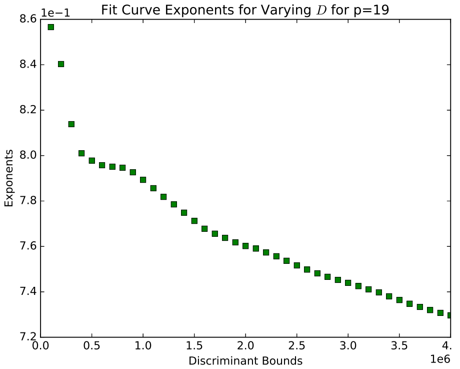


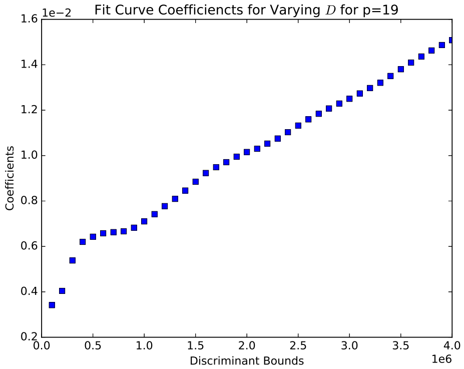

Fit Curve Exponents for Varying D for p=13

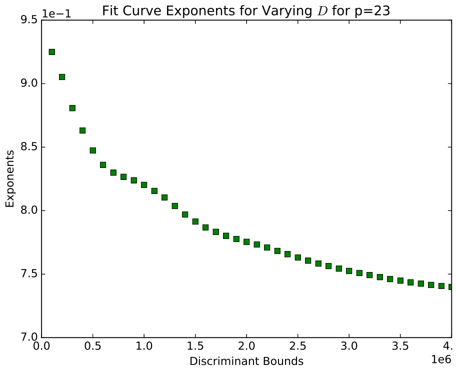

Fit Curve Coefficiencts for Varying D for p=13

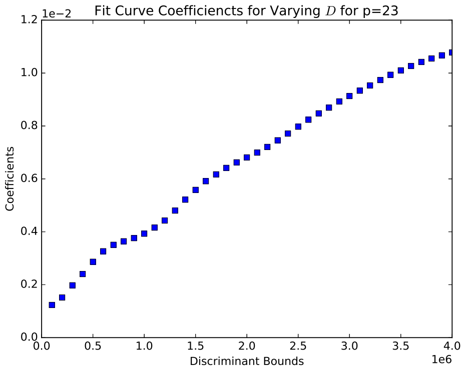

Fit Curve Exponents for Varying D for p=17


Fit Curve Coefficiencts for Varying D for p=17


Fit Curve Exponents for Varying D for p=19


Fit Curve Coefficiencts for Varying D for p=19





Colorado State University, Fort Collins, CO 80523-1874
E-mail address: `lewis@math.colostate.edu`

James Madison University, Harrisonburg, VA 22807
E-mail address: `willi5cl@jmu.edu`
URL: `http://educ.jmu.edu/~willi5cl`