involve
a journal of mathematics

Numerical secondary terms in a Cohen–Lenstra conjecture on real quadratic fields

Codie Lewis and Cassandra Williams

msp

2019 vol. 12, no. 2

INVOLVE 12:2 (2019)
dx.doi.org/10.2140/involve.2019.12.221

Numerical secondary terms in a Cohen–Lenstra conjecture on real quadratic fields

Codie Lewis and Cassandra Williams

(Communicated by Yves-François Pétermann)

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 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 1980s it was noted that certain finite abelian groups occur much less frequently than others as class groups. In their classic paper, Cohen and Lenstra [1984] 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 [Jacobson 1998; Jacobson et al. 2006; Mollin and Williams 1992; te Riele and Williams 2003] gave new or improved algorithms for computing class groups of quadratic fields, followed by a data comparison to

MSC2010: primary 11R29; secondary 11R11, 11Y35.
Keywords: Cohen–Lenstra, real quadratic field, secondary term.

conjectures from [Cohen and Lenstra 1984]. Both Jacobson [1998] and te Riele and Williams [2003] constructed real quadratic fields and gave numerical tables to support various conjectures for small primes. Jacobson [1998] computed 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, te Riele and Williams [2003] considered 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.

There have been recent attempts to define such secondary terms analytically. For example, Hough [2016] 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 [2013] and Bhargava, Shankar, and Tsimerman [Bhargava et al. 2013] each proved the secondary term for the number of cubic number fields conjectured in [Roberts 2001]. 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 3-part of the class group of a real quadratic field.

Unfortunately, for many of the original Cohen–Lenstra conjectures the methods in [Bhargava et al. 2013; Hough 2016; Taniguchi and Thorne 2013] 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 nonisomorphic real quadratic fields with fundamental discriminant less than or equal to the positive integer X. Then*

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

In [Jia 1993], the author proves that, assuming the Riemann hypothesis, the error term for $Q(X)$ can be reduced to $O(X^{17/54+\varepsilon})$. However, in this work, we will choose not to assume the Riemann hypothesis, and so will only assume the classical error term for $Q(X)$ given in Lemma 3.

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 cardinality. 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 [Cohen and Lenstra 1984] 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 [Cohen and Lenstra 1984]. *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*

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

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

[Stein et al. 2016]. 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. The `class_number` method by default does not assume the generalized Riemann hypothesis [Stein et al. 2016]; therefore all of our class numbers are unconditionally correct although the data-generation process was significantly slowed by this choice. Additionally, the computation of class numbers is parallelizable, although we lacked the technology to do so. Therefore, a future investigation could generate a larger data set more quickly even if the default method for `class_number` was again used.

Jacobson [1998] 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 $X \to \infty$

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

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

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

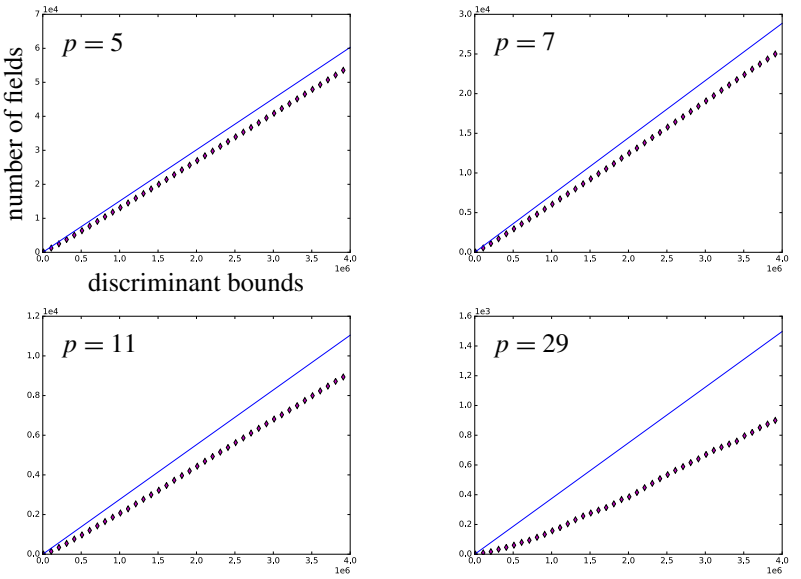


Figure 1. Plots of the actual number of fields (solid line) and predicted number of fields (diamonds) for $p = 5, 7, 11$, and 29.

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.

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

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

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

Applying the logarithm to (3-2), we have

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

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 10,000-th statistic) into five 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 is 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, (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) with their curves of best fit determined by (3-2) for four primes. (For the plots of the difference (3-1) 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

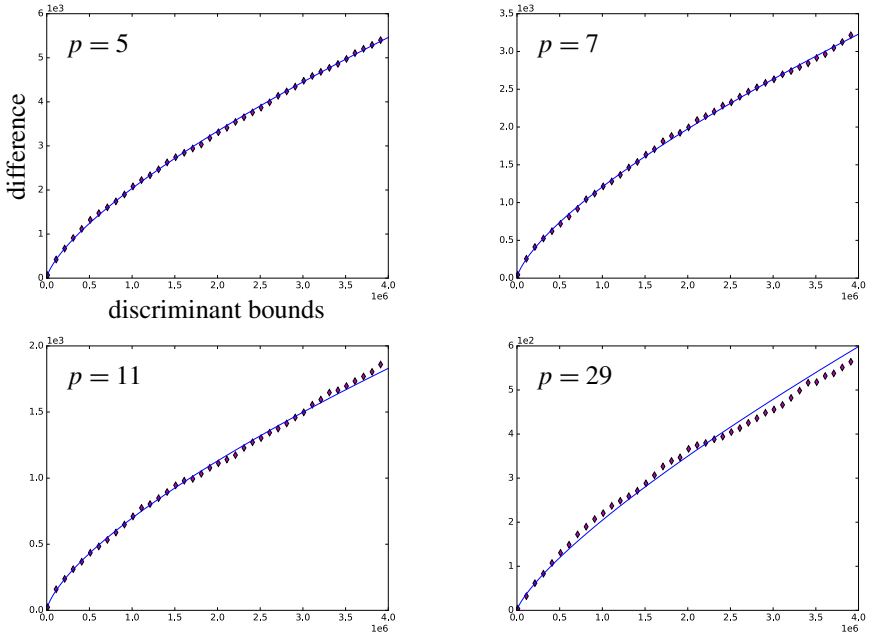


Figure 2. Plots of difference (3-1) with fitted curve from (3-2) for $p = 5, 7, 11$, and 29.

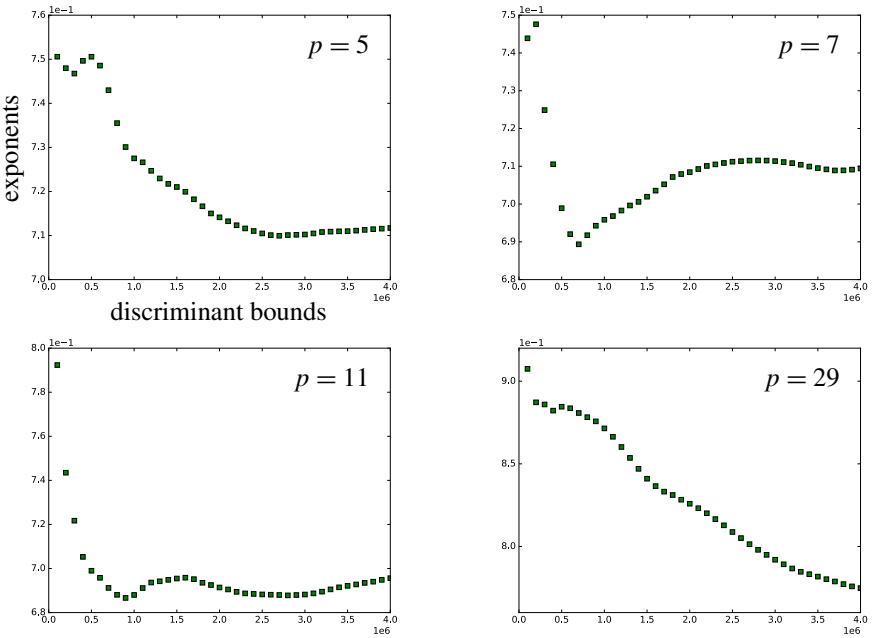


Figure 3. Plots of exponents s_p for varying discriminant bound and $p = 5, 7, 11, 29$.

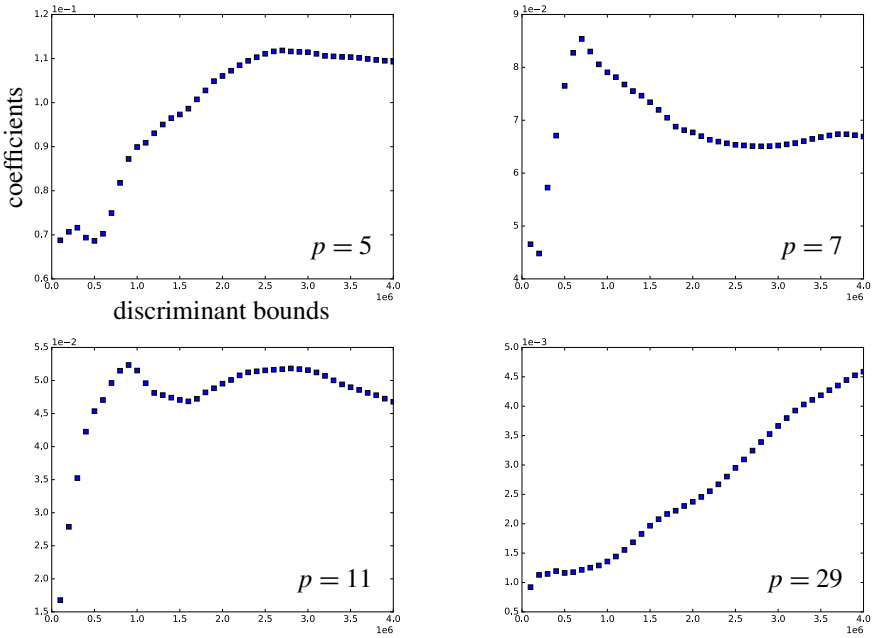


Figure 4. Plots of coefficients C_p for varying discriminant bound and $p = 5, 7, 11, 29$

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 our data for some primes, there does seem to be some predictability to the monomial term given by (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 10,000-th 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.

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

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

prime (p)	CV error	scaled CV error
3	23,408	146,473
5	7914	159,625
7	3869	163,020
11	1623	178,704
13	1129	176,258
17	661	180,072
19	495	169,324
23	356	180,490
29	185	150,771

Table 2. Quality of fit.

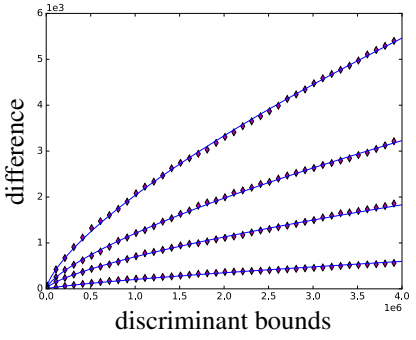


Figure 5. Comparison of the data points and fit curves for $p = 5, 7, 11, 29$

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 counterintuitive looking at the comparison of

the fit curves and plots given above in Figure 2. In Figure 5, we plot the fit curves 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 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 B). However, it might be 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 could give rise to significant improvement in Conjecture C7 of [Cohen and Lenstra 1984].

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. Under our assumption of a monomial model for the error, a modification of the conjecture might be of the form

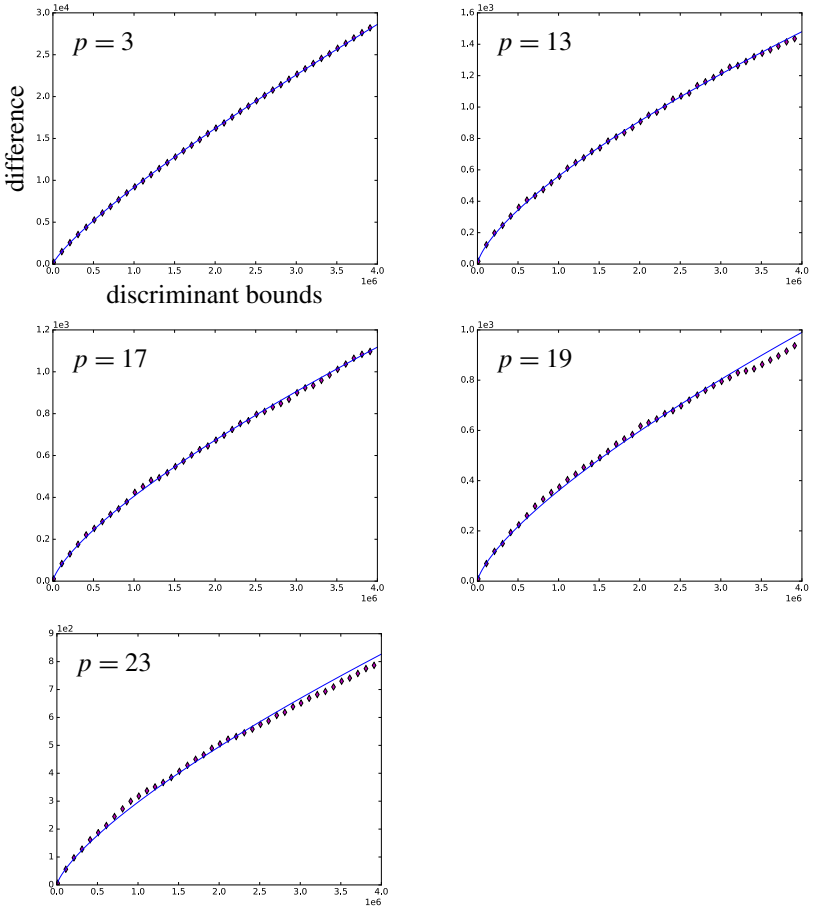
$$\#\{\mathbb{Q}(\sqrt{d}) : p \mid h \text{ and } D < X\} \sim \xi_p \frac{3X}{\pi^2} - C_p X^{s_p},$$

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. In future work, we hope to recreate this numerical investigation on a larger data set to refine a possible conjecture.

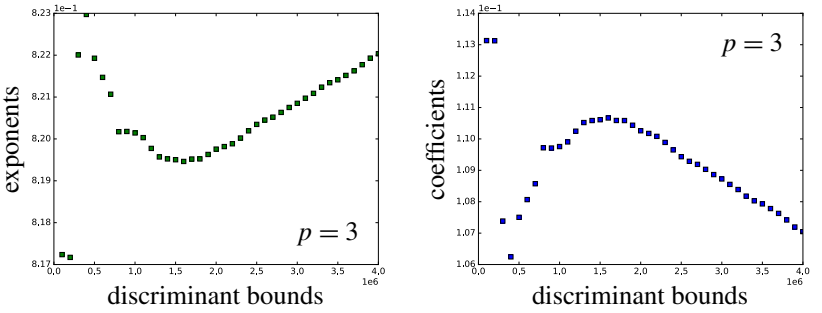
We note that there are actually two sources of error that contribute to the secondary term as we have estimated it here; one is the actual error in the Cohen–Lenstra conjecture, while the other is the error in approximating $Q(X)$ by only $3X/\pi^2$. Although we chose not to control for the error arising from $Q(X)$, we believe it would be possible in future work to separate these two sources and identify an error term for each of them. In that case, one could consider the effect on error of assuming the Riemann hypothesis and using the smaller error term on $Q(X)$ given in [Jia 1993]. (See the discussion after Lemma 3.)

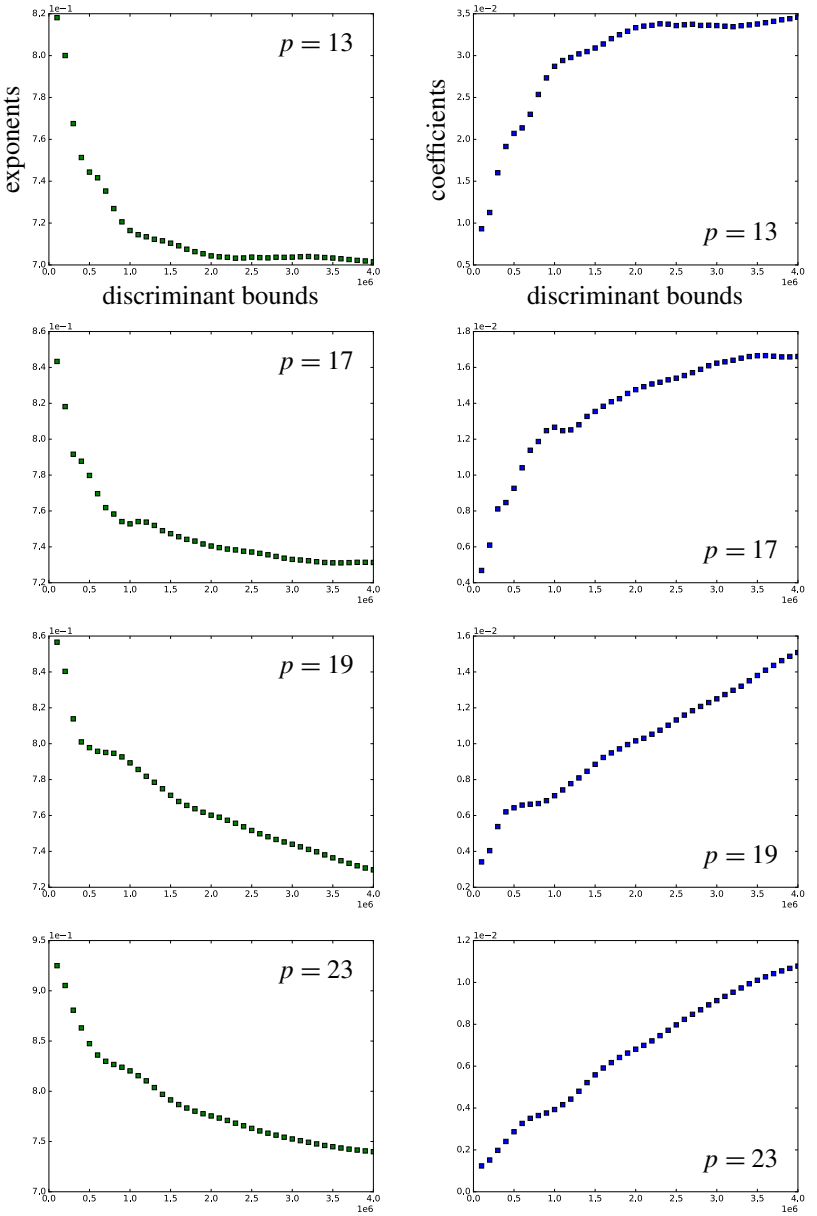
At the time of this writing, we are not aware of any analytic approach to finding a secondary term for Conjecture C7, even for particular primes. We would be interested to see such a method and compare to our numerical results.

Appendix A: Cohen–Lenstra error fitting for other $p < 30$



Appendix B: Exponent and coefficient plots for other $p < 30$





Acknowledgments

We thank the referee for insightful comments and suggestions.

References

[Bhargava et al. 2013] M. Bhargava, A. Shankar, and J. Tsimerman, "On the Davenport–Heilbronn theorems and second order terms", *Invent. Math.* **193**:2 (2013), 439–499. MR Zbl

[Cohen and Lenstra 1984] H. Cohen and H. W. Lenstra, Jr., "Heuristics on class groups of number fields", pp. 33–62 in *Number theory* (Noordwijkerhout, 1983), edited by H. Jager, Lecture Notes in Math. **1068**, Springer, 1984. MR Zbl

[Hough 2016] B. Hough, "Equidistribution of bounded torsion CM points", preprint, 2016. To appear in *J. Anal. Math.* arXiv

[Jacobson 1998] M. J. Jacobson, Jr., "Experimental results on class groups of real quadratic fields (extended abstract)", pp. 463–474 in *Algorithmic number theory* (Portland, OR, 1998), edited by J. P. Buhler, Lecture Notes in Comput. Sci. **1423**, Springer, 1998. MR Zbl

[Jacobson et al. 2006] M. J. Jacobson, Jr., S. Ramachandran, and H. C. Williams, "Numerical results on class groups of imaginary quadratic fields", pp. 87–101 in *Algorithmic number theory*, edited by F. Hess et al., Lecture Notes in Comput. Sci. **4076**, Springer, 2006. MR Zbl

[Jia 1993] C. H. Jia, "The distribution of square-free numbers", *Sci. China Ser. A* **36**:2 (1993), 154–169. MR Zbl

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

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

[Roberts 2001] D. P. Roberts, "Density of cubic field discriminants", *Math. Comp.* **70**:236 (2001), 1699–1705. MR Zbl

[Stein et al. 2016] W. A. Stein et al., *Sage mathematics software*, Version 7.3, Sage Development Team, 2016, available at http://www.sagemath.org.

[Taniguchi and Thorne 2013] T. Taniguchi and F. Thorne, "Secondary terms in counting functions for cubic fields", *Duke Math. J.* **162**:13 (2013), 2451–2508. MR Zbl

Received: 2017-06-15 Revised: 2018-03-08 Accepted: 2018-06-11

lewis@math.colostate.edu *Department of Mathematics, Colorado State University, Fort Collins, CO, United States*

willi5cl@jmu.edu *Department of Mathematics and Statistics, James Madison University, Harrisonburg, VA, United States*

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2019 is US $195/year for the electronic version, and $260/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFlow® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2019 vol. 12 no. 2

1944-4176(2019)12:2;1-3