Two Remarks on Torsion-Point Attacks in Isogeny-Based Cryptography

Francesco Sica[0000−0002−6027−2548]

Florida Atlantic University*
sicaf@fau.edu

Abstract. We fix an omission in [8] on torsion point attacks of isogeny-based cryptosystems akin to SIDH, also reprised in [2, 4]. In these works, their authors represent certain integers using a norm equation to derive a secret isogeny. However, this derivation uses as a crucial ingredient [8, Section 4.3, Lemma 6], which we show to be incorrect. We then state sufficient conditions allowing to prove a modified version this lemma.
A further idea of parametrizing solutions of the norm equation will show that these conditions can be fulfilled under the same heuristics of these previous works. Our contribution is a theoretical one. It doesn't invalidate the attack, which works as well in practice, but gives a correct mathematical justification for it.
We also simplify the argument of [2, Theorem 3] to show that the requirement that m be small is unnecessary.

Keywords: Post-quantum cryptography, elliptic curve cryptography, isogenies.

1 Introduction

The recent attacks of Castryck-Decru [1], Maino-Martindale [6] and Robert [9] have dealt a mortal blow to the supersingular isogeny problem with torsion (see Problem 1 below). They are the culmination of a line of attacks begun with Petit [8] and developed by de Quehen, Kutas, Leonardi, Martindale, Panny, Petit and Stange [2] and Fouotsa and Petit [4]. However, there are recent attempts at repairing SIDH [3] (Supersingular Isogeny Diffie-Hellman), where masking the image of torsion points or the degree of the isogeny would thwart the higher dimensional attacks but would not shield against the earlier works mentioned above, in particular [4]. Consequently, investigations in earlier publications are still of interest.

In this work, we would like to correct some inaccuracies in [2, 8] ([4] also uses these results), in a way which we hope will also make them pedagogically more accessible. The contribution of this work is

1. to point out an error in a crucial technical lemma of [8],
2. a corresponding nontrivial fix, deployed in two steps (Theorems 1 and 3);

* The present work was supported by a university startup grant.

3. to point out that the work [2] is also inconsistent in one point and that in their Theorem 3 they don't need the hypothesis that d be coprime to B and consequently that m be small.

2 Background on elliptic curves and isogenies

We start by explaining the main terms of this work: elliptic curves and isogenies. We will only give the necessary explanations for the rest of the work. Further details can be found in Silverman's classic [11]. In this work, the finite fields involved will be of large prime characteristic p. We will denote them by $\mathbb{F}_q$, where q is the number of elements in the field, a power of p.

An elliptic curve E over a field $\mathbb{F}_q$ is a nonsingular plane cubic with an $\mathbb{F}_q$-rational point. Without loss of generality, the equation of curve can be written $y^2 = x^3 + ax + b$, where the right-hand side is a polynomial with coefficients in $\mathbb{F}_q$ and distinct roots in an algebraic closure $\overline{\mathbb{F}}_q$ of $\mathbb{F}_q$.

The set of $\mathbb{F}_q$-rational points on E is endowed with a (finite) abelian group structure, whose identity is the point at infinity of E. This abelian group is denoted $E(\mathbb{F}_q)$.

Let E_0, E be two elliptic curves defined over $\mathbb{F}_q$. An isogeny $\phi\colon E_0 \to E$ defined over $\mathbb{F}_q$ is a homomorphism $E_0(\overline{\mathbb{F}}_q) \to E(\overline{\mathbb{F}}_q)$ given by rational functions in $\mathbb{F}_q(x, y)$. It follows that $\ker\phi$ is finite, unless it is all of $E_0(\overline{\mathbb{F}}_q)$ (trivial isogeny). It is a theorem of Tate [12] that such a nontrivial isogeny exists if and only if $|E_0(\mathbb{F}_q)| = |E(\mathbb{F}_q)|$, in which case we say that E_0 and E are isogenous over $\mathbb{F}_q$.

If E_0 and E are isomorphic (consider that $E_0 = E$) then the set of $\overline{\mathbb{F}}_q$ isogenies $\phi\colon E \to E$ (called endomorphisms) has a natural ring structure (where addition comes from the group structure on E, and multiplication is isogeny composition). This ring is denoted $\mathrm{End}(E)$. For every $m \in \mathbb{Z}$, The map $[m]\colon E \to E$ defined by $P \mapsto [m]P$ is a $\mathbb{F}_q$-endomorphism of E and in this way we get an embedding $\mathbb{Z} \subset \mathrm{End}(E)$.

To any isogeny ϕ one can define an integer $\deg\phi \geq 0$ called the degree of ϕ, which is zero if and only if $\phi = 0$ (the constant isogeny). It satisfies the property that $\deg(\psi\phi) = \deg\psi \deg\phi$. Also, for any $m \in \mathbb{Z}$, $\deg[m] = m^2$.

Given $m > 0$ and integer, the m-torsion of E is $E[m] = \ker[m]$. As a group, we have, for $k > 0$ integer and ℓ prime different from p,

$$E[\ell^k] \cong \mathbb{Z}/\ell^k \times \mathbb{Z}/\ell^k \ .$$

On the other hand, either $E[p^k] = 0$ for all $k > 0$ or $E[p^k] \cong \mathbb{Z}/p^k$ for all $k > 0$. In the first case we say that E is supersingular, in the second one that E is ordinary.

A nonzero isogeny ϕ is called separable if $\deg\phi = |\ker\phi|$. Since this is always the case if[1] $(\deg\phi, p) = 1$, all our isogenies will be separable. An isogeny is called cyclic if its kernel is cyclic.

[1] We write (a, b) for the gcd of a and b.

Given an isogeny $\phi\colon E_0 \to E$, there exists a (unique) isogeny $\hat{\phi}\colon E \to E_0$ with the property that $\hat{\phi}\phi = \phi\hat{\phi} = [\deg\phi]$. From the multiplicativity of the degree and the definition, it follows that $\deg\hat{\phi} = \deg\phi$ and that $\hat{\hat{\phi}} = \phi$. We also have that for ϕ, ψ isogenies (with domain and codomain such that the following makes sense): $\widehat{\phi+\psi} = \hat{\phi} + \hat{\psi}$ and $\widehat{\psi\phi} = \hat{\phi}\hat{\psi}$.

If $\phi \in \mathrm{End}(E)$, the endomorphism $\phi + \hat{\phi}$ equals $[t]$ for some $t \in \mathbb{Z}$ called the trace of ϕ and denoted $\mathrm{tr}\,\phi$.

Every endomorphism ϕ satisfies the quadratic polynomial $\phi^2 - [\mathrm{tr}\,\phi]\phi + [\deg\phi] = 0$. This polynomial is irreducible if ϕ is not in $\mathbb{Z}$. It is then called the characteristic polynomial of ϕ and defines an imaginary quadratic extension of $\mathbb{Q}$.

Finally, given a finite subgroup G of E_0, there exist an elliptic curve E and a (separable) isogeny $\phi\colon E_0 \to E$ such that $G = \ker\phi$ (see [11, Chapter III, Prop. 4.12]). Also, if $H \leq G$, then an isogeny $\phi\colon E_0 \to E$ with kernel G factors through an isogeny ϕ' with kernel H: there exist an elliptic curve E' and isogenies $\phi'\colon E_0 \to E'$ and $\psi\colon E' \to E$ such that $\phi = \psi\phi'$ and $\ker\phi' = H$. Moreover, $\ker\psi \cong G/H$. Note that if $|G| = n$ and d is a positive divisor of n, since G is abelian, we can always find $H \leq G$ of order d to make this decomposition work.

Keeping the same notations as in the previous paragraph, if ϕ (i.e. G) is cyclic, in the decomposition above, for $H \leq G$, then both ϕ' and ψ are cyclic, as subgroups and quotient groups of cyclic groups are cyclic. In short, if a composition of (any number of) isogenies is cyclic, then all the factors are cyclic.

The effective computation of an isogeny ϕ passes through Vélu's formulas [13]. These are efficient as long as $|\ker\phi|$ is small. The previous paragraph allows to split an isogeny ϕ of degree ℓ^k into a product of k isogenies of degree ℓ, allowing for a polynomial-time (in $\log(\deg\phi)$) evaluation of the isogeny when its degree is divisible only by small fixed primes. In the following, for all practical purposes, the problem of computing an isogeny will therefore be equivalent to the problem of describing its kernel. The following result then shows that computing ϕ or $\hat{\phi}$ are equivalent.

Lemma 1. *Let $\phi\colon E_0 \to E$ be an isogeny of degree A. Then $\ker\hat{\phi} = \phi(E_0[A])$. Moreover, if ϕ is cyclic, then so is $\hat{\phi}$.*

Proof. We prove the lemma in the case when $\phi, \hat{\phi}$ are separable, which is the only one relevant to this work. Note that $\hat{\phi}\phi = [A]$ implies that $\ker\phi \subseteq E_0[A]$ and that $\ker\hat{\phi} \supseteq \phi(E_0[A])$. Therefore

$$E_0[A]/\ker\phi \cong \phi(E_0[A]) \subseteq \ker\hat{\phi}$$

Comparing the cardinalities of both members ($= A$), we deduce that we have equality.

For the second statement, we assume that $(A, p) = 1$ for simplicity, which is the case of interest to us (p large and A only divisible by small primes). Then $E_0[A] \cong \mathbb{Z}/A \times \mathbb{Z}/A$ and we can complete $\ker\phi \cong \mathbb{Z}/A$ to a $\mathbb{Z}/A-$basis of $E_0[A]$. Then $\ker\hat{\phi} \cong E_0[A]/\ker\phi \cong \mathbb{Z}/A$. □

The subject of our investigations is the following computational problem, called supersingular isogeny problem with torsion (SSI-T).

Problem 1 (SSI-T). Let $\phi\colon E_0 \to E$ be a cyclic isogeny between two supersingular elliptic curves with $\deg\phi = A$. Let $B > 0$ an integer coprime to A. The *supersingular isogeny problem with torsion* asks to recover ϕ (i.e. $\ker\phi$), from the knowledge of E_0, E and $\phi|_{E_0[B]}$.

In the case of SIDH, ϕ is Alice's secret key (say), while $\phi|_{E_0[B]}$ is public. Hence the solution of SSI-T is equivalent to a complete key break of SIDH.

3 Petit's torsion point attack

In [8], the following strategy is presented to solve the SSI-T problem in some cases, where B is significantly larger than A. We follow notations as in Problem 1. The idea is to look for $\tau \in \mathrm{End}(E)$ of the form

$$\tau = \phi\theta\hat{\phi} + [d]$$

for some known $\theta \in \mathrm{End}(E_0)$ and $d \in \mathbb{Z}$, such that

$$\deg\tau = \deg(\phi\theta\hat{\phi} + [d]) = Be \tag{1}$$

where e is a small positive integer. The previous equation is called the norm equation. We can then factor $\tau = \tau_e\tau_B$, with $\deg\tau_e = e$ and $\deg\tau_B = B$. Since $\ker\tau_B \subseteq \ker\tau \cap E[B]$ is a subgroup of order B, by looking at $\tau|_{E[B]}$ one can recover $\ker\tau_B$ since e is small[2]. Also, if $(e, B) = 1$, then $\ker\tau_B = \ker\tau \cap E[B]$, so the action of τ on $E[B]$ immediately gives $\ker\tau_B$. Note that the way τ is constructed and using the torsion knowledge from Problem 1, it is possible to compute $\tau|_{E[B]}$. Indeed, since B is coprime to A and $\ker\phi \subseteq E_0[A]$, $\phi|_{E_0[B]}$ is invertible and therefore

$$\hat{\phi}\Big|_{E[B]} = A\left(\phi|_{E_0[B]}\right)^{-1} .$$

Once τ_B is found, one does a clever exhaustive search (meet-in-the-middle guess) on isogenies of degree e, which is feasible if e is small. Now that τ is known, it is possible to calculate

$$(\tau - [d])|_{E[A]} = \phi\theta\hat{\phi}\Big|_{E[A]}$$

From the knowledge of $G = \ker\phi\theta\hat{\phi}\Big|_{E[A]} = \ker\phi\theta\hat{\phi} \cap E[A]$, it is shown how to recover $\ker\phi \subseteq E_0[A]$ and therefore ϕ, see [8, Section 4.3]. Note that G can be easily calculated by a Pohlig-Hellman type approach (akin to a two-dimensional

[2] It is assumed that B is smooth, in order to derive $\ker\tau \cap E[B]$ from the action of τ on a basis of $E[B]$.

Hensel lifting), if A is the product of small primes to large powers, as is typically the case in SIDH, where A is a power of 2 or a power of 3. If G is cyclic, then $\ker \hat{\phi} \subseteq G$ is its only subgroup of order A. The problem arises when there is a large rank 2 subgroup inside G, the extreme case being (if for instance θ is a scalar) when $G = E[A]$ and then we don't get any information on ϕ whatsoever.

To avoid this case, the following lemma is needed [8, Section 4.3, Lemma 6].

Lemma 2 (Incorrect!). *Let M be a divisor of A and κ the number of distinct prime factors of M. Then there are most 2^κ cyclic subgroups H of order M in $E_0[M]$ such that $\theta(H) = H$.*

Counterexample for $\kappa = 1$: Let E an elliptic curve defined over a field of characteristic different from 2. Using an isomorphism $E[4] \cong \mathbb{Z}/4 \times \mathbb{Z}/4$, suppose that some endomorphism θ is described on $E[4]$ by a matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$; notice that θ is not a scalar and that $\deg \theta \equiv -1 \pmod 4$. Then the following four distinct order 4 subgroups of $\mathbb{Z}/4 \times \mathbb{Z}/4$ are fixed by θ (we write the column vectors as row vectors here):

$$\langle (1,0) \rangle \ , \quad \langle (0,1) \rangle \ , \quad \langle (1,2) \rangle \ , \quad \langle (2,1) \rangle \ .$$

The case of a supersingular elliptic curve where this is happening is the following: consider the curve E defined by $y^2 = x^3 - x$ over $\mathbb{F}_p$ with $p \equiv 7 \pmod{16}$. This is a special case of the the curve E_0 which appears in Section 5. In particular, it is supersingular, and its 2-torsion is all contained in $E(\mathbb{F}_p)$ since $x^3 - x = x(x-1)(x+1)$ splits into linear factors over $\mathbb{F}_p$. The cardinality of $E(\mathbb{F}_p)$ is $p+1$ which is divisible exactly by 8. This means that $E[4]$ has a basis consisting of one point $P_1 \in E(\mathbb{F}_p)$ and another point P_2 defined over a (quadratic) extension of $\mathbb{F}_p$. Taking as θ the p-Frobenius endomorphism, its matrix, when restricted to $E[4]$, is therefore $\left(\begin{smallmatrix} 1 & c \\ 0 & -1 \end{smallmatrix}\right)$, because $\operatorname{tr} \theta = 0$ (since E is supersingular). Also, since θ acts like the identity on $E[2]$, we must have $c = 0, 2$. If $c = 0$, we are exactly in the situation described above. The case when $c = 2$ reduces to the former one in the new basis $\{P_1, P_1 + P_2\}$.

Remark 1. The assumption of θ not being a scalar as curve endomorphism (not explicitly mentioned in the statement of Lemma 2), is nevertheless not enough. Even if θ is not a scalar, it could be a scalar on the M-torsion and then everything would be invariant. One clearly needs some additional restriction on θ, which is the subject of Theorem 1 in the next section.

4 Fixing Lemma 2

In [8, Section 4.3, Lemma 6], it is incorrectly stated that the number of invariant subgroups or order M of a nonscalar endomorphism θ restricted to $E_0[M]$ is bounded by 2^k where k is the number of prime factors of M. Furthermore, in that section's analysis, it is implicitly assumed that $\deg \theta$ is coprime to A. However, later analysis of the norm algorithm skips this coprimality check, which is important and will be discussed in Section 5. For the remainder of the section, we will write E for E_0, since we will work on one curve only.

Theorem 1. *Let q be a power of a prime p and E an elliptic curve defined over $\mathbb{F}_q$. Let $\theta \in \mathrm{End}(E)$ be a nonzero endomorphism such that $\mathrm{tr}\,\theta = 0$ and $(\deg\theta, M) = 1$, where $M > 1$ is an integer with κ distinct prime factors. Then the number of cyclic subgroups $H \subseteq E[M]$ of order M such that $\theta(H) = H$ is at most $2^{\kappa+1}$.*

Proof. The proof uses a Hensel lifting procedure. Note that if $\ell \neq p$ is prime and $k \geq 1$ is a positive integer, then $E[\ell^k] \cong \mathbb{Z}/\ell^k \times \mathbb{Z}/\ell^k$, while $E[p^k] \cong 0, \mathbb{Z}/p^k$, depending on whether the curve is supersingular or ordinary [11, Chapter III, Corollary 6.4]. By the Chinese Remainder Theorem (CRT),

$$E[M] \cong \prod_{\ell^k \| M} E[\ell^k] \ ,$$

where the subscript notation means that the product is taken over all prime powers $\ell^k \mid M$ with $\ell^{k+1} \nmid M$. Therefore counting the number of fixed cyclic subgroups of order M of $E[M]$ is equivalent to counting the number of fixed cyclic subgroups of order ℓ^k of $E[\ell^k]$ for $\ell^k \| M$ and multiplying the results together. When $\ell = p$ this number is at most one. Otherwise we proceed as in [8].

Fix a basis $\{P, Q\}$ of $E[\ell^k]$ and let $H = \langle \alpha P + \beta Q \rangle$ be of order ℓ^k, which is to say that ℓ does not divide both variables $1 \leq \alpha, \beta \leq \ell^k$. The action of θ on $E[\ell^k]$ is described by a matrix $\left(\begin{smallmatrix} a & -b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{Z}/\ell^k)$ (the reason for the minus sign will become evident below).

As shown, the condition that $\theta(H) = H$ is equivalent to

$$c\beta^2 + (a-d)\alpha\beta + b\alpha^2 = c\beta^2 + 2a\alpha\beta + b\alpha^2 \equiv 0 \pmod{\ell^k} \ , \tag{2}$$

in view of the fact that $\mathrm{tr}(\theta) = a + d \equiv 0 \pmod{\ell^k}$.

We will analyze (2) in its dehomogenized forms

$$cx^2 + 2ax + b \equiv 0 \pmod{\ell^k} \ , \tag{3a}$$

where $x = \beta\alpha^{-1} \pmod{\ell^k}$, or

$$bx^2 + 2ax + c \equiv 0 \pmod{\ell^k} \ , \tag{3b}$$

where $x = \alpha\beta^{-1} \pmod{\ell^k}$.

The symmetry of (3a) and (3b) is now visible and will allow a single treatment of both equations. The slight complication is that since α, β are unknown, we may have to resort to both these dehomogenizations, leading to double counting cases when $\ell \nmid \alpha\beta$.

Remark that the effect of the previous dehomogenizations is equivalent to the selection of one particular generator of H, namely one of the form $P + xQ$ (Equation (3a)) or $xP + Q$ (Equation (3b)). In particular, solutions for $x \pmod{\ell^k}$ will correspond bijectively to different H with generators of that form.

The only time (3a) (resp. (3b)) cannot be used is to find subgroups H such that $\ell \mid \alpha$ (resp. $\ell \mid \beta$), and this can happen if and only if $\ell \mid c$ (resp. $\ell \mid b$). In that case, one has to resort to the other dehomogenization.

Both equations (3a) and (3b) have discriminant $\Delta = 4(a^2 - bc) \equiv -4 \deg \theta \pmod{\ell^k}$. This is the analysis of [8], where the conclusion would follow by invoking Hensel's lemma. However, applying this lemma requires a more careful analysis, in particular regarding the case $\ell = 2$, where the roots are multiple.

With this setup, let's consider what happens for various primes ℓ. The easiest case is the next one.

4.1 $\ell > 2$

By assumption, the discriminant Δ is nonzero modulo ℓ, and any (simple) solution of (3) modulo ℓ lifts uniquely to a solution modulo ℓ^k. Therefore, for each case (3a) or (3b), we get at most 2 solutions.

A separate analysis is needed when the two dehomogenizations are needed, which as seen above happens if and only if $b \equiv c \equiv 0 \pmod{\ell}$. In this case the each equation (3a) and (3b) is in fact linear mod ℓ (but not necessarily mod ℓ^k). Hensel's lemma still applies, giving rise to at most 2 solutions mod ℓ^k again, one for each dehomogenization.

A more complicated analysis is needed for $\ell = 2$, since in this case, Hensel's lemma doesn't apply in its simple form.

4.2 $\ell = 2$

We will need here the following stronger form of Hensel's lemma, which is rarely taught in elementary number theory courses. See [10, Ch II, Section 2.2], with a proof.

Lemma 3 (Hensel's lemma for multiple roots). *Let $f \in \mathbb{Z}[x]$, ℓ a prime, $j, r \geq 1$ positive integers such that $2j + 1 \leq r$. Let $x_0 \in \mathbb{Z}$ satisfy $f(x_0) \equiv 0 \pmod{\ell^r}$, with $\ell^j \parallel f'(x_0)$. Then*

1. *If $x_1 \in \mathbb{Z}$ and $x_1 \equiv x_0 \pmod{\ell^{r-j}}$, we have $f(x_1) \equiv f(x_0) \equiv 0 \pmod{\ell^r}$ and $\ell^j \parallel f'(x_1)$.*
2. *There is exactly one $t \bmod \ell$ such that $f(x_0 + t\ell^{r-j}) \equiv 0 \pmod{\ell^{r+1}}$.*
3. *The same conclusions hold with any $k \geq r$ in substitution of r.*

This lemma shows that from a certain point onward, solutions lift (to ℓ solutions each at every step) for j steps, but only one (mod ℓ^{k+1-j}) will continue lifting to the next step (mod ℓ^{k+1}). We illustrate first this lemma with the simplest case when $j = 1$, i.e. when the root is double but "barely so". This is also the case which will be of concern to us.

Example 1. Consider the equation $x^2 \equiv 1 \pmod{2^k}$. In this case the equation $f(x) = x^2 - 1$ has the single solution $x = 1 \bmod 2$. Since $f'(x) = 2x$, we have $2 \parallel f'(1)$. In this case, to start the Hensel lifting procedure, we need $k \geq 3$. Solutions mod 8 are $1, 5$ and $3, 7$. The first two are lifts of 1 mod 4, the latter two of -1 mod 4.

We can conclude from Hensel's lemma that exactly one lift of 1 mod 4 will lift to a solution mod 16, and similarly one lift of -1 mod 4 will lift to a solution mod 16. Of course these lifts are respectively 1 and -1, which lift respectively to $1, 9$ and $-1, -9$.

We can continue like this ad infinitum by repeated applications of the lemma and come to the conclusion that the only solutions of $x^2 \equiv 1 \pmod{2^k}$ for $k \geq 3$ are $1, 1+2^{k-1}, -1, -1-2^{k-1}$. At each step, one of the first two solutions (in this case 1) lifts to two solutions mod 2^{k+1} while the other one $(1+2^{k-1})$ dies. Similarly for the latter two.

Proof (Hensel's lemma). We reproduce the proof here for completeness. Consider the Taylor expansion (with integer coefficients)

$$\begin{aligned} f(x_0 + t\ell^{r-j}) &= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} t^n \ell^{nr-nj} \\ &\equiv f(x_0) + t\ell^{r-j} f'(x_0) \pmod{\ell^{2r-2j}} \\ &\equiv f(x_0) + t\ell^{r-j} f'(x_0) \pmod{\ell^{r+1}} \end{aligned}$$

since $r+1 \leq 2r-2j$ by hypothesis. The last line is $\equiv f(x_0) \equiv 0 \pmod{\ell^r}$, which shows that, calling $x_1 = x_0 + t\ell^{r-j}$, we have $f(x_1) \equiv 0 \pmod{\ell^r}$. Moreover, since $x_1 \equiv x_0 \pmod{\ell^{r-j}}$, and writing $f'(x_0) = \ell^j u$ with $\ell \nmid u$ we obtain

$$f'(x_1) \equiv f'(x_0) \pmod{\ell^{r-j}} \Rightarrow f'(x_1) \equiv f'(x_0) = \ell^j u \pmod{\ell^{j+1}}$$

which shows that $\ell^j \parallel f'(x_1)$. This wraps up the first point.

We also have

$$f(x_0 + t\ell^{r-j}) \equiv f(x_0) + tu\ell^r \pmod{\ell^{r+1}}$$

and note that since $\ell^r \mid f(x_0)$, $f(x_0 + t\ell^{r-j}) \equiv 0 \pmod{\ell^{r+1}}$ is equivalent to

$$\frac{f(x_0)}{\ell^r} + tu \equiv 0 \pmod{\ell} \ ,$$

which has exactly one solution mod ℓ. This shows the second point. The final point follows by induction on $k \geq r$.

□

Let's go back to Equation (2). Since $\Delta = -4 \deg \theta \equiv 0 \pmod 2$, roots are multiple. However our hypothesis again allows us to limit the extent to which the derivative is divisible by powers of 2.

Note that since $\deg \theta = bc - a^2 \equiv 1 \pmod 2$, only the following situations can happen:

$2 \mid a$ and $2 \nmid bc$. Here (2) can be transformed into either (3a) or (3b). Fixing the first choice, let[3] $f(x) = cx^2 + 2ax + b$ so that we are looking for solutions of

[3] By abuse of notation, we consider $f \in \mathbb{Z}[x]$, for instance by choosing $0 \leq a, b, c < 2^k$.

$f(x) \equiv 0 \pmod{2^k}$. Then $f(x) \equiv x^2 + 1 \pmod 2$ hence any solution $r \bmod 2^k$ is congruent to $1 \bmod 2$ and $f'(r) = 2cr + 2a \equiv 2 \pmod 4$. In other terms, $2 \parallel f'(r)$, so we can apply Hensel's lemma with $j = 1$, after analyzing solutions mod 8. Since there is one solution mod 2, there are at most two solutions mod 4 and at most four solutions mod 8. Applying Hensel's lemma as in the example, we see that there will be at most four solutions modulo 2^k for all $k \geq 1$. For $k \geq 3$, one solution dies while the other one lifts.

$2 \nmid a$ and 2 divides precisely one of b or c. Without loss of generality, we can suppose that $2 \mid b$ and $2 \nmid ac$. It means that only one dehomogenization (3a) (with respect to α) is needed to enumerate the subgroups H. One gets as before $f(x) = cx^2 + 2ax + b \equiv 0 \pmod{2^k}$. Here $f(x) \equiv x^2 \pmod 2$, so any solution $r \bmod 2^k$ is congruent to $0 \bmod 2$ but then again $f'(r) = 2cr + 2a \equiv 2 \pmod 4$. As before, we conclude that there will be at most four solutions modulo 2^k for all $k \geq 1$.

$2 \nmid a$, $2 \mid b$ and $2 \mid c$. This is the most arduous case, in that we have to study concurrently both (3a) and (3b) to account for all possible H. Note that whatever dehomogenization we are using, $f(x)$ is identically 0 mod 2. This means we could have up to $3 \cdot 2^{k-1}$ subgroups $\bmod 2^k$ (since when using (3a) or (3b), a nonzero solution will relate to an H common to both dehomogenizations). We will show that in fact we have at most four subgroups for all $k \geq 1$.

Consider (3a) first and let $f(x) = cx^2 + 2ax + b$. Remark that any solution $r \bmod 2^k$ will satisfy $f'(r) = 2cr + 2a \equiv 2 \pmod 4$, therefore as in the previous cases, we have $2 \parallel f'(r)$. The same considerations apply to (3b) with $g(x) = bx^2 + 2ax + c$. By Hensel's lemma, this observation already caps the number of possible H to 12 for all $k \geq 1$, but we can do better.

1. Modulo 2: As noted, $f(x) \equiv g(x) \equiv 0$, hence $x = 0, 1$ are solutions for both f and g, leading to three subgroups.
2. Modulo 4: $x = 0$ lifts (to 0 and 2) for f (resp. for g) if and only if $b \equiv 0 \pmod 4$ (resp. $c \equiv 0 \pmod 4$). Also, $x = 1$ lifts (to 1 and -1) if and only if $b + c \equiv 2 \pmod 4$. At most two of these three conditions can hold, giving four subgroups.
3. Modulo 8: For f, 0 lifts iff $b \equiv 0 \pmod 8$, while 2 lifts iff $b \equiv 4 \pmod 8$. Therefore one of the lifts dies. Similarly for g with c in place of b. Furthermore 1 lifts iff $b + c + 2a \equiv 0 \pmod 8$ whereas -1 lifts iff $b + c - 2a \equiv 0 \pmod 8$. Again both cannot hold since $4a \equiv 4 \pmod 8$. The conclusion is that out of the four possible lifts mod 4, half will die while the other half will lift to four solutions mod 8.
4. Modulo $2^t, 3 < t \leq k$. Applying Hensel's lemma again with $j = 1$, of the four solutions $\bmod 2^{t-1}$, two die while two lift to four solutions $\bmod 2^t$.

□

There is an easier proof of the previous theorem, when M is squarefree. We state:

Theorem 2. *Let q be a power of a prime p and E an elliptic curve defined over $\mathbb{F}_q$. Let $\theta \in \mathrm{End}(E)$ be a nonzero endomorphism such that $\operatorname{tr}\theta = 0$ and $(\deg\theta, M) = 1$, where $M > 1$ is a squarefree integer with κ distinct prime factors. Then the number of cyclic subgroups $H \subseteq E[M]$ of order M such that $\theta(H) = H$ is at most $3 \cdot 2^{\kappa-1}$.*

Proof. As previously, using the CRT, we are reduced to the case when $M = \ell \neq p$ is prime. Then $E[\ell] \cong \mathbb{F}_\ell^2$ is a $\mathbb{F}_\ell$-vector space of dimension 2. Fixing a basis, we identify as before θ as a matrix in $\mathrm{GL}_2(\mathbb{F}_\ell)$. The invariant subspaces H are then eigenspaces of θ. The crucial observation is that since $\det\theta \equiv \deg\theta \not\equiv 0 \pmod{\ell}$ and $\operatorname{tr}\theta \equiv 0 \pmod{\ell}$, θ cannot be a multiple of the identity, except possibly when $\ell = 2$. In this case, we have at most three subspaces. Otherwise, the Jordan normal form (over an algebraic closure) of θ is one of the following:

1. $\begin{pmatrix}\lambda & 0\\ 0 & \mu\end{pmatrix}$ with $\lambda \not\equiv \mu \pmod{p}$. We then have two or zero invariant subspaces, depending on whether the eigenvalues λ, μ are in $\mathbb{F}_p$ or are conjugate in a quadratic extension.
2. $\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix}$. In this case, $\lambda \in \mathbb{F}_p$ has geometric multiplicity 1, and there is only one invariant subspace.

Therefore, when $\ell \neq 2$, we have at most two invariant H. This concludes the proof.

□

Remark 2. In [8, Section 4.3, Lemma 7] it is written the expected number k of invariant subgroups is $\leq 2\log\log A$ when A is powersmooth (i.e. with all prime factors less than $\log A$). In fact we don't need a powersmooth A, since in the crucial step, we can just use

$$\sum_{\substack{\ell \leq A \\ \ell \text{ prime}}} \frac{2}{\ell} = 2\log\log A + c + O\left(\frac{1}{\log A}\right) \ ,$$

for some $c \in \mathbb{R}$, a formula due to Mertens [7].

5 A new analysis of the norm equation

We now want to prove that it is possible to solve the norm equation (1) (heuristically) satisfying the conditions enumerated in Theorem 1, namely that $\operatorname{tr}\theta = 0$ (as done before us) and $(\deg\theta, A) = 1$ (our second contribution), for the curve $E_0/\mathbb{F}_p$ with $j(E_0) = 1728$. The resulting modification of the original algorithm appears as Algorithm 1.

This curve has equation $y^2 = x^3 - Dx$ for some $D \in \mathbb{F}_p$ and is supersingular if and only if $p \equiv 3 \pmod{4}$. The endomorphism ring of this curve contains a $\mathbb{Z}$-module of index four generated by $1, \iota, \pi, \iota\pi$, where $\iota(x, y) = (-x, iy)$ with $i^2 \equiv -1 \pmod{p}$ and $\pi(x, y) = (x^p, y^p)$.

Algorithm 1 Solving norm equation (5).

input : SIDH parameters p, A, B. **output** : A solution (a, b, c, d, e) to (9).

$s \leftarrow 0$
if A is even **then**
    $A \leftarrow A/2$
    $s \leftarrow 1$
Set $e \leftarrow 1$.
while $Be - A^2$ is a quadratic non-residue mod p **do**
    $e \leftarrow e + 1$
    while $(e, A) > 1$ **do**
        $e \leftarrow e + 1$
    Compute c as the smallest positive integer such that $c^2 \equiv Be - A^2 \pmod{p}$.
    if $Be > A^2 + c^2$ **then**
        if $\frac{Be-A^2-c^2}{p}$ is prime **then**
            if $\frac{Be-A^2-c^2}{p} \equiv 1 \pmod 4$ **then**
                Find $a, b \in \mathbb{N}$ such that $a^2 + b^2 = \frac{Be-A^2-c^2}{p}$
                $d \leftarrow p(a^2 + b^2) + c^2 - A^2$
                if $s = 0$ **then**
                    $(a, b, c) \leftarrow (2a, 2b, 2c)$
                return [break] (a, b, c, d, e)

Note that $\iota^2 = -1$ and $\pi^2 = -p$ (the trace of Frobenius is divisible by p for supersingular curves and at most $2\sqrt{p}$ in absolute value, hence is zero). Moreover, since $i \notin \mathbb{F}_p$, we have

$$\pi\iota(x, y) = \pi(-x, iy) = (-x^p, -iy^p) = -\iota\pi(x, y) \ ,$$

therefore $\pi\iota = -\iota\pi$. Then

$$(\pi\iota)^2 = \pi\iota\pi\iota = -\pi^2\iota^2 = -p \ .$$

From the characteristic equations of all these endomorphisms, we deduce that

$$\operatorname{tr}\iota = \operatorname{tr}\pi = \operatorname{tr}(\iota\pi) = 0 \ . \tag{4}$$

Choosing $\theta = a\pi + b\pi\iota + c\iota$ with $a, b, c \in \mathbb{Z}$, using (4) and the linearity of the trace we get $\operatorname{tr}\theta = 0$. Similarly, since

$$\phi\theta\hat{\phi} + \widehat{\phi\theta\hat{\phi}} = \phi\theta\hat{\phi} + \phi\hat{\theta}\hat{\phi} = \phi(\theta + \hat{\theta})\hat{\phi} = 0$$

we derive $\operatorname{tr}(\phi\theta\hat{\phi}) = 0$. Letting $\tau = \phi\theta\hat{\phi} + [d]$ as in Section 3, the norm equation $\deg\tau = B^2e^2$ (where we give the variant in the improvement of [2], with B^2 instead of B and write e^2 for e to get a homogeneous equation) can be explicitly written in terms of a, b, c, d, e, after noticing that

$$\theta\hat{\theta} = (a\pi + b\pi\iota + c\iota)(a\hat{\pi} + b\hat{\iota}\hat{\pi} + c\hat{\iota}) = [pa^2 + pb^2 + c^2]$$

which implies that $\deg\theta = pa^2+pb^2+c^2$. A similar calculation with $\phi\theta\hat{\phi}$ of trace zero, shows that the modified norm equation $\deg\tau = B^2e^2$ reads

$$\deg\tau = A^2(pa^2+pb^2+c^2)+d^2 = B^2e^2 \ . \tag{5}$$

We will now show how to ensure that $\deg\theta = pa^2+pb^2+c^2$ be coprime to A, which is the main contribution of this section. The idea is to view (5) as the equation of a (real) projective quadric in $\mathbb{RP}^4$. Since the quadric has the rational point in projective coordinates $[a,b,c,d,e] = [0,0,0,B,1]$, it can be parametrized by polynomial functions (with rational coefficients). The idea is a familiar one, which is first seen in the parameterization of the (projective) unit circle $x^2+y^2=z^2$ as $[x,y,z] = [2st, t^2-s^2, t^2+s^2]$ (Pythagorean triples).

For instance, to get these expressions, we parametrize the line passing through $[0,0,0,B,1]$ and a generic point $[c_0,c_1,c_2,c_3,0]$ and look for the other intersection point with the quadric. The coordinates of this point are expressed as rational functions of the coefficients of the quadratic equation which defines the other intersection point of this line with the quadric. After a tedious computation, one finds the following parameterization:

$$\begin{aligned}(a,b,c) &= 2Bc_3(c_0,c_1,c_2) \ , \\ d &= Bp(c_0^2+c_1^2)+Bc_2^2-BA^2c_3^2 \ , \\ e &= p(c_0^2+c_1^2)+c_2^2+A^2c_3^2 \ .\end{aligned} \tag{6}$$

Unfortunately, this parameterization still suffers from large coefficients. For instance, $e \geq A^2$. To be able to find small solutions in e, we have to use some number theoretical argument. We will therefore look for e of the form $e = Be'$ and a small positive integer $e' = O(\log p)$, since all other coordinates have B as a common factor (we are taking $c_0, c_1, c_2, c_3 \in \mathbb{N}$). If we can simplify this large common factor B, we will be left with a solution of the kind sought by [2] and [8].

By the same heuristics as in [2], we find that, when $c_3 = 1$, there is a solution in integers to

$$Be' = p(c_0^2+c_1^2)+c_2^2+A^2 \tag{7}$$

provided $B > \max(p^2, A^2)$. Starting from $e' = 1$ (testing with increments of 1 if the requirements below fail), one attempts to solve the equation

$$Be' - A^2 \equiv c_2^2 \pmod{p}$$

This is solvable with probability $1/2$. For such a solution we test whether the positive quantity

$$\frac{Be' - A^2 - c_2^2}{p} = \alpha$$

is a prime congruent to $1 \pmod 4$. This should happen with probability bounded by $O(1/\log p)$, since $\alpha < Be' < p^h$ for some $h > 0$. In this case, express $\alpha =$

$c_0^2 + c_1^2$ by Cornacchia's algorithm, for instance. Therefore we expect to find a solution to (7) in $O(\log p)$ steps and consequently (5) will be solved with

$$\begin{aligned} (a,b,c) &= 2(c_0,c_1,c_2) \ , \\ d &= p(c_0^2+c_1^2)+c_2^2-A^2 \ , \\ e &= e' = \frac{p(c_0^2+c_1^2)+c_2^2+A^2}{B} \ . \end{aligned} \tag{8}$$

This method has some advantages and some disadvantages over the [2] and [8] method. The disadvantage is that the range of parameters to which it applies is worse if $p \gg A$ or $p \ll A$ and similar only if $p \approx A$. On the plus side however, there is no requirement that $A < p$ nor that it have $O(\log\log p)$ prime factors. Another advantage is that, although the final size of e^2 is $O(\log^2 p)$ as in [2], since we're only sieving through squares, the expected time before finding a solution is faster, $O(\log p)$ instead of $O(\log^2 p)$.

There is a final advantage of our method, which goes in hand with our attempt at fixing [8, Lemma 6]. Recall that we need $\deg\theta = pa^2+pb^2+c^2$ to be coprime to A. This is almost achieved at the current stage. It suffices to impose when solving (7) that e' be coprime to A. These e' are then sampled from a set of density at least

$$\prod_{\substack{\ell \le \log A \\ \ell \text{ prime}}} \left(1-\frac{1}{\ell}\right) = O\left(\frac{1}{\log\log A}\right) = O\left(\frac{1}{\log\log p}\right)$$

again by the Mertens theorem quoted above, since an integer A has at most $O(\log A)$ prime divisors, by a theorem dating back to Tchebychev. This will heuristically increase the runtime of the algorithm by a multiplicative factor $O(\log\log p)$.

Then by (7), $p(c_0^2+c_1^2)+c_2^2$ will be coprime to A. But by (8)

$$\deg\theta = pa^2+pb^2+c^2 = 4\big(p(c_0^2+c_1^2)+c_2^2\big)$$

so that $(\deg\theta, A) \le 4$. To find a fully coprime degree (which is now necessary only if A is even), solve the norm equation (5) in our fashion (7) with $A/2$ instead of A and assume that $A/2$ is also even (otherwise as in the proof of Theorem 2 for $\ell = 2$, we have at most three invariant subspaces mod 2). In that case, choosing again e' coprime to A will give

$$(p(c_0^2+c_1^2)+c_2^2, A/2) = (p(c_0^2+c_1^2)+c_2^2, A) = 1 \ .$$

The corresponding solution to (5) reads

$$\left(\frac{A}{2}\right)^2 (pa^2+pb^2+c^2)+d^2 = B^2e^2 \iff A^2(pc_0^2+pc_1^2+c_2^2)+d^2 = B^2e^2 \ ,$$

and now $\deg\theta = pc_0^2+pc_1^2+c_2^2$ is coprime to A.

We summarize our findings in the following

Theorem 3. *Let A, B be two coprime integers and p be prime. Assume $B > \max(p^2, A^2)$. The norm equation*

$$A^2(pa^2 + pb^2 + c^2) + d^2 = B^2e^2 \tag{9}$$

has an integer solution (a, b, c, d, e) with $abc \neq 0$, $(pa^2 + pb^2 + c^2, A) = 1$ and $e = O(\log p \log\log p)$, which can be found in heuristic time $O(\log p \log\log p)$ if B is at most polynomial in A.

6 Remarks on the Crypto 2021 work

The authors of [2] improve the norm equation $A^2(pa^2 + pb^2 + c^2) + d^2 = Be$ to

$$\deg\tau = A^2(pa^2 + pb^2 + c^2) + d^2 = B^2e \ ,$$

with $\tau = \phi\theta\hat{\phi} + [d]$ as before. They then prove the following result.

Theorem 4 ([2, Theorem 3]). *Suppose we are given an instance of SSI-T where A has $O(\log\log p)$ distinct prime factors. Assume we are given the restriction of a trace-zero endomorphism $\theta \in \mathrm{End}(E_0)$ to $E_0[B]$, **an integer** d **coprime to** B, and a smooth integer e such that $\deg\tau = B^2e$. Then we can compute ϕ in time $O(\sqrt{e} \cdot \mathrm{polylog}(p))$.*

The new norm equation allows as explained in Section 2 to factor τ as $\tilde{\psi}'\tilde{\eta}\tilde{\psi}$, where $\deg\tilde{\psi}' = \deg\tilde{\psi} = B$ and $\deg\tilde{\eta} = e$. In fact, by the same argument, if $E[m'] \subseteq \ker\tau$ is largest, we can write $m' = mh$ where $m \mid B$ is largest (hence $h^2 \mid e$).

Since $\ker\tau/E[m']$ is cyclic of order $(B/m)^2(e/h^2)$, we can decompose $\tau = \psi'\eta\psi \circ [m']$ where now $\psi'\eta\psi$ is cyclic and hence so are the single isogenies ψ', η and ψ of degree respectively $B/m, e/h^2, B/m$. Note that m can be easily found as $E[m] \subseteq \ker\tau \cap E[B]$ is largest; we know $\tau|_{E[B]}$, we can therefore test $\tau|_{E[\ell^k]}$ for all prime powers $\ell^k \mid B$ and combine the results via CRT.

Calling $B' = B/m$ and $\epsilon = e/h^2$, after composing with $[m']$, we end up with a *cyclic* $\psi'\eta\psi\colon E \to E$ of degree $B'^2\epsilon$. Therefore $\ker\psi$ will be the only subgroup of order B' of $\ker(\psi'\eta\psi) \cap E[B']$. Similarly, considering the dual (cyclic by Lemma 1) isogeny $\hat{\psi}\hat{\eta}\hat{\psi}'\colon E \to E$ we find $\ker\hat{\psi}'$ as the only subgroup of order B' of $\ker(\hat{\psi}\hat{\eta}\hat{\psi}') \cap E[B']$. Having recovered ψ and $\hat{\psi}'$, hence ψ', one then finds $\tilde{\psi}' = \psi' \circ [m]$ and $\tilde{\psi} = \psi \circ [m]$ and successively $\tilde{\eta}$ by a meet-in-the-middle guess, as explained in [2]. Once τ is found, one recovers ϕ à la Petit, as explained in Section 3.

Therefore it is not needed to suppose that m be small, and no condition on d in the norm equation is necessary in Theorem 4. We also note that this condition was in any case not subsequently checked in any of the algorithms of [2].

7 Conclusion

We hope to have put the attack in [8] on a mathematically satisfactory level. Several other points remain to be addressed to a similar degree of rigor. They include the Frobenius isogeny attack of [2] or the heuristic argument to justify the runtime. The Frobenius isogeny attack is harder to deal with, because the smaller number of variables makes it harder to find a suitable parameterization. This is an example of how in lower dimensions sometimes the smaller degree of freedom complicates the approach.

However, regarding the rigorous justification of the runtime, one promising avenue could be to use theta functions to study via the circle method à la Hardy [5] the (asymptotic or exact) number of representations of a large enough integer in terms of a diagonal quadratic form. We leave these investigations to a separate paper, as they are of a completely different nature.

Acknowledgements. I am thankful to Koblandy Idrissov for the interesting online discussions.

References

1. W. Castryck and T. Decru. An efficient key recovery attack on SIDH (preliminary version). Cryptology ePrint Archive, Paper 2022/975, 2022. https://eprint.iacr.org/2022/975.
2. V. de Quehen, P. Kutas, C. Leonardi, C. Martindale, L. Panny, C. Petit, and K. E. Stange. Improved torsion-point attacks on SIDH variants. In Tal Malkin and Chris Peikert, editors, *Advances in Cryptology - CRYPTO 2021 - 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16-20, 2021, Proceedings, Part III*, volume 12827 of *Lecture Notes in Computer Science*, pages 432–470. Springer, 2021.
3. T. B. Fouotsa, T. Moriya, and P. Petit. M-SIDH and MD-SIDH: countering SIDH attacks by masking information. In Carmit Hazay and Martijn Stam, editors, *Advances in Cryptology - EUROCRYPT 2023 - 42nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Lyon, France, April 23-27, 2023, Proceedings, Part V*, volume 14008 of *Lecture Notes in Computer Science*, pages 282–309. Springer, 2023.
4. T. B. Fouotsa and C. Petit. A new adaptive attack on SIDH. Cryptology ePrint Archive, Paper 2021/1322, 2021. https://eprint.iacr.org/2021/1322.
5. G. H. Hardy. On the representation of a number as the sum of any number of squares, and in particular of five. *Trans. Amer. Math. Soc.*, 21:255–284, 1920.
6. L. Maino and C. Martindale. An attack on SIDH with arbitrary starting curve. Cryptology ePrint Archive, Paper 2022/1026, 2022. https://eprint.iacr.org/2022/1026.
7. F. Mertens. Ein Beitrag zur analytischen Zahlentheorie. *J. reine angew. Math*, 78:46–62, 1874.
8. C. Petit. Faster algorithms for isogeny problems using torsion point images. In Tsuyoshi Takagi and Thomas Peyrin, editors, *Advances in Cryptology - ASIACRYPT 2017 - 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017,*

Proceedings, Part II, volume 10625 of *Lecture Notes in Computer Science*, pages 330–353. Springer, 2017.
9. D. Robert. Breaking SIDH in polynomial time. Cryptology ePrint Archive, Paper 2022/1038, 2022. https://eprint.iacr.org/2022/1038.
10. J-P. Serre. *A course in arithmetic*, volume 7 of *Graduate Texts in Mathematics*. Springer, 1973.
11. J. H. Silverman. *The Arithmetic of Elliptic Curves*, volume 106 of *Graduate Texts in Mathematics*. Springer, 1992.
12. J. Tate. Endomorphisms of abelian varieties over finite fields. *Invent. Math.*, 2:134–144, 1966.
13. J. Vélu. Isogénies entre courbes elliptiques. *C.R. Acad. Sc. Paris. Série A*, 273:238–241, 1971.