

A KLEPTOGRAPHIC ATTACK ON CSIDH

TREY LI

ABSTRACT. This note reports an attack scenario for CSIDH that does not appear to have been considered in the literature. A subverted device of Alice can leak a shared curve between Alice and Bob through Alice’s public curves in later CSIDH sessions. The attack adapts the Young–Yung kleptographic attack from classical Diffie–Hellman to CSIDH using the Goldreich–Levin theorem and rejection sampling. We prove that a CSIDH public curve $[\mathbf{a}] \star E_0$ remains computationally indistinguishable from an honest public curve even when its secret class $[\mathbf{a}]$ is rejection-sampled according to a hard-core bit of a hidden parallelization curve $[\mathbf{a}][\mathbf{c}] \star E_0$ formed by the same class $[\mathbf{a}]$ together with an independent curve $[\mathbf{c}] \star E_0$. The proofs rely only on the computational assumptions underlying CSIDH itself. We hope that this note provides a useful starting point for the study of algorithm-substitution attacks in isogeny-based cryptography, complementing existing work on side-channel and fault attacks.

1. INTRODUCTION

Kleptographic and algorithm-substitution attacks use malicious implementations of a cryptographic scheme to leak secret information through deliberately generated outputs that look like legitimate ones. Young and Yung introduced SETUP attacks [12] and later used the term kleptography and gave an attack on classical Diffie–Hellman [13]. Later work studied when algorithm-substitution attacks are undetectable [1, 2], and how state resets can help detect them [11]. For CSIDH, the study of compromised implementations has so far focused mainly on side-channel and fault attacks [3, 7]. This note asks whether CSIDH admits a kleptographic attack in which ordinary-looking public curves leak secret information. We show that it does.

Alice’s subverted device stores a CSIDH public key $E_{\mathbf{r}} = [\mathbf{r}] \star E_0$ chosen by the kleptographer, together with a random string r (we will show later that this random string can be derived during Alice’s use of the device, so there is no need to store it in the device beforehand). To leak a target bit b , the device repeatedly samples secret keys $\mathbf{a}$ from the honest CSIDH secret-key distribution until the inner product $\langle \text{bin}(A_{[\mathbf{a}] \star E_{\mathbf{r}}}), r \rangle \bmod 2 = b$, and outputs the curve $E_{\mathbf{a}} = [\mathbf{a}] \star E_0$ as Alice’s public key, where $\text{bin}(A_E)$ is the binary representation of the canonical Montgomery coefficient A_E of E . After observing $E_{\mathbf{a}}$, the kleptographer computes $[\mathbf{r}] \star E_{\mathbf{a}} = [\mathbf{a}] \star E_{\mathbf{r}}$ and recovers b using the same r . Assuming the hardness of group-action Diffie–Hellman and using the Goldreich–Levin theorem [10], we show that the bit $\langle \text{bin}(A_{[\mathbf{a}] \star E_{\mathbf{r}}}), r \rangle \bmod 2$ is a hard-core bit given $(E_{\mathbf{a}}, E_{\mathbf{r}}, r)$. We further show that the rejection-sampled public key $E_{\mathbf{a}}$ is computationally indistinguishable from an honestly generated CSIDH public key, even when $(E_{\mathbf{r}}, r)$ is known. Repeating the procedure over later sessions using different random strings r leaks a CSIDH shared curve or any other secret one bit at a time.

This attack applies naturally to other non-interactive key-exchange schemes, such as SWOOSH [9], and to key encapsulation mechanisms built from non-interactive key-exchange schemes. In particular, the two parties must be able to derive the same shared value from a secret key and the other party's public key; this value must be computationally hard to recover from the public keys alone; and it must admit a canonical binary encoding from which the shared secret can be recovered. Under these conditions, the same Goldreich–Levin and rejection-sampling mechanism can be used to leak any polynomial-length secret. A notable feature of the attack is that it relies on no hardness assumption beyond those already underlying the security of the original scheme.

2. THE ATTACK

Let $p = 4 \prod_{i=1}^n \ell_i - 1$ be a prime with $\ell_1, \dots, \ell_n$ distinct small odd primes. CSIDH [4] uses Montgomery curves $E_A : y^2 = x^3 + Ax^2 + x$ over $\mathbb{F}_p$ with base curve $E_0 : y^2 = x^3 + x$. Let π be the p -power Frobenius and $\mathcal{O} = \text{End}_{\mathbb{F}_p}(E_0) = \mathbb{Z}[\pi] \cong \mathbb{Z}[\sqrt{-p}]$ the associated quadratic order. Each ℓ_i splits as $(\ell_i) = \mathfrak{l}_i \bar{\mathfrak{l}}_i$ with $\mathfrak{l}_i = (\ell_i, \pi - 1)$ and $\bar{\mathfrak{l}}_i = (\ell_i, \pi + 1)$. For a nonzero invertible integral ideal $\mathfrak{a} \subseteq \mathcal{O}$ of norm coprime to p , write $E[\mathfrak{a}] = \bigcap_{\alpha \in \mathfrak{a}} \ker(\alpha)$ and let $[\mathfrak{a}] \star E$ be the codomain of the separable isogeny with kernel $E[\mathfrak{a}]$. The class group $\text{cl}(\mathcal{O})$ acts freely and transitively on the orbit $\mathcal{E} = \{[\mathfrak{a}] \star E_0 : [\mathfrak{a}] \in \text{cl}(\mathcal{O})\}$. For $e = (e_1, \dots, e_n) \in \mathbb{Z}^n$ write $[e] = \prod_i [\mathfrak{l}_i]^{e_i}$.

In CSIDH, Alice and Bob sample secret vectors $e \in \{-m, \dots, m\}^n$ and $f \in \{-m, \dots, m\}^n$ respectively, exchange the public curves $[e] \star E_0$ and $[f] \star E_0$, and obtain the shared curve $([e][f]) \star E_0 = [e + f] \star E_0$. The group-action inverse problem (GAIP) asks to recover an efficiently actionable representation of $[e]$ given $(E_0, [e] \star E_0)$; the group-action computational Diffie–Hellman problem (GACDH) asks to compute $([e][f]) \star E_0$ given $(E_0, [e] \star E_0, [f] \star E_0)$.

The attack adapts the Young–Yung kleptographic attack for Diffie–Hellman [13] to CSIDH. It replaces the cross-session key-dependent method of Young–Yung with a cross-session key-independent method by using the Goldreich–Levin theorem [10] together with rejection sampling. It works as follows.

Let Eve be the kleptographer, possibly a device vendor. Before selling the device to Alice, Eve generates a CSIDH key pair (x, E_x) with $E_x = [x] \star E_0$ and embeds the public key E_x into the device. She also embeds a sequence of polynomially many independently and uniformly sampled strings $r_1, r_2, \dots \in \{0, 1\}^L$ into the device, where $L = \lceil \log_2 p \rceil = \Theta(\lambda)$ with λ the security parameter. Alice runs the device as a black box and the device only outputs CSIDH public keys $[a] \star E_0$ and CSIDH shared secrets $([a][b]) \star E_0$ to Alice. For each candidate secret key a sampled by the device, let $E_a = [a] \star E_0$ and define $Z_a = [a] \star E_x$. By commutativity of the class-group action, we have $Z_a = [a] \star E_x = [x] \star E_a$. The device can compute Z_a from (a, E_x) , while Eve can compute Z_a from her secret key x and Alice's public key E_a , without interacting with the device. Write $Z_a = E_{A_{Z_a}}$ in canonical Montgomery form. For any of the random strings r shared between the device and Eve, define the Goldreich–Levin predicate $h_r(Z_a) = \langle \text{bin}(A_{Z_a}), r \rangle \pmod{2}$. We show that this is a hard-core predicate given (E_a, E_x, r) .

Lemma 1 (Hard-core predicate). *Let x and a be independently sampled from the CSIDH secret-key distribution, and let $r \leftarrow \{0, 1\}^L$ be uniform. If GACDH is hard, then no polynomial-time algorithm can predict $h_r(Z_a)$ from (E_a, E_x, r) with non-negligible advantage over $1/2$.*

Proof. Suppose there exists a polynomial-time predictor P with non-negligible advantage in predicting $h_r(Z_a)$ from (E_a, E_x, r) . Averaging over the honestly sampled secrets a and x , there is a non-negligible fraction of “bad” pairs (E_a, E_x) for which $P(E_a, E_x, \cdot)$ predicts $h_r(Z_a)$ with non-negligible advantage over uniform r . Fix any such bad pair; this also fixes the curve Z_a . Applying the Goldreich–Levin reconstruction lemma [10, Lemma 1] to $P(E_a, E_x, \cdot)$ gives a polynomial-size list containing $\text{bin}(A_{Z_a})$ with non-negligible probability in polynomial time. Thus, given a random GACDH instance (E_0, E_a, E_x) , one may run the Goldreich–Levin reconstruction algorithm and select a uniformly random element of the resulting list. With non-negligible probability the pair (E_a, E_x) is bad, and conditioned on that event the random choice from the final list equals $\text{bin}(A_{Z_a})$ with non-negligible probability. This means that GACDH can be solved in polynomial time with non-negligible probability, a contradiction. $\square$

Now, suppose Alice has already completed an honest CSIDH exchange with Bob and obtained a shared curve E_A . Write $b_1, b_2, \dots$ for the bits of $\text{bin}(A)$. To leak b_i in a later session, the device samples at most L independent candidate secret keys $a \leftarrow \{-m, \dots, m\}^n$ and uses the first one that satisfies $h_{r_i}([a] \star E_x) = b_i$. If no candidate succeeds, it samples an honest secret a instead. It then sets $a_i := a$ and outputs $E_{a_i} = [a_i] \star E_0$. After observing E_{a_i} , Eve computes $Z_{a_i} = [x] \star E_{a_i}$ and recovers $b_i = h_{r_i}(Z_{a_i})$ if $h_{r_i}([a_i] \star E_x) = b_i$.

Note that the pair (E_x, r) is fixed before sampling a . There is a chance that for some bad pairs (E_x, r) , an a such that $h_{r_i}([a] \star E_x) = b_i$ occurs with very low probability or does not exist at all. Namely, most or all a give $h_{r_i}([a] \star E_x) = 1 - b_i$. For these pairs (E_x, r) , rejection sampling and hence the attack may fail under the sampling-time bound L . We show that for overwhelmingly many pairs (E_x, r) , the predicate $h_r([a] \star E_x)$ is nearly balanced over the choice of a . We then use this balancing property to show that there are only negligibly many bad pairs that would fail most a .

Lemma 2 (Balanced predicate). *Fix $b \in \{0, 1\}$. For each pair (E_x, r) , define the probability $p_b(E_x, r) := \Pr_a[h_r([a] \star E_x) = b]$ over a which is sampled from the CSIDH secret-key distribution. If GACDH is hard, then for honestly and independently sampled (E_x, r) and every inverse-polynomial $\varepsilon > 0$, we have $\Pr_{E_x, r} [|p_b(E_x, r) - \frac{1}{2}| \geq \varepsilon] = \text{negl}(\lambda)$. Moreover, $\mathbb{E}_{E_x, r} [|p_b(E_x, r) - \frac{1}{2}|] = \text{negl}(\lambda)$.*

Proof. Suppose for some inverse-polynomial $\varepsilon > 0$ the probability $\Pr_{E_x, r} [|p_b(E_x, r) - \frac{1}{2}| \geq \varepsilon]$ is non-negligible. We construct a predictor for the hard-core bit in Lemma 1. Given a challenge (E_a, E_x, r) , sample an independent honest secret a' and output $h_r([a'] \star E_x)$ as the prediction for $h_r([a] \star E_x)$. For a fixed pair (E_x, r) , the challenge bit $h_r([a] \star E_x)$ and the predicted bit $h_r([a'] \star E_x)$ are independent samples from the same Bernoulli distribution. Hence the predictor is correct with probability $p_b(E_x, r)^2 + (1 - p_b(E_x, r))^2 = \frac{1}{2} + 2(p_b(E_x, r) - \frac{1}{2})^2$. Whenever the pair (E_x, r) satisfies $|p_b(E_x, r) - 1/2| \geq \varepsilon$, the advantage of the predictor is at least $2\varepsilon^2$. If such pairs occurred with non-negligible probability, then given (E_a, E_x, r) , the overall advantage of predicting $h_r([a] \star E_x)$ would be non-negligible, contradicting Lemma 1. This proves the first claim.

For the second claim, let $X = |p_b(E_x, r) - \frac{1}{2}|$. For every inverse-polynomial $\varepsilon > 0$, the first claim gives $\Pr[X \geq \varepsilon] = \text{negl}(\lambda)$. Since $0 \leq X \leq 1/2$, we have $\mathbb{E}_{E_x, r}[X] \leq \varepsilon + \frac{1}{2} \Pr[X \geq \varepsilon]$. Now take $\varepsilon = 1/(2q(\lambda))$ for any polynomial q . For sufficiently large λ , the second term of

$\mathbb{E}_{E_x, r}[X]$ is negligible and hence smaller than $1/(2q(\lambda))$, so $\mathbb{E}_{E_x, r}[X] < 1/q(\lambda)$. Since q is any polynomial, we have that the expected bias $\mathbb{E}_{E_x, r}[X]$ is negligible. $\square$

Now we show that the probability of sampling a bad pair (E_x, r) which fails many a is negligible. This conclusion is also used in the proof of Lemma 4. It allows us to neglect the effect of bad pairs when calculating certain probabilities.

Lemma 3 (Bad-pair probability). *For $b \in \{0, 1\}$, define $\text{Bad}_b := \{(E_x, r) : p_b(E_x, r) \leq \frac{1}{4}\}$. If GACDH is hard, then $\Pr_{E_x, r}[(E_x, r) \in \text{Bad}_b] = \text{negl}(\lambda)$. Consequently, except with negligible probability over (E_x, r) , both target bits have per-trial success probability greater than $1/4$.*

Proof. If $p_b(E_x, r) \leq 1/4$, then $|p_b(E_x, r) - \frac{1}{2}| \geq \frac{1}{4}$. Therefore, $\{(E_x, r) : p_b(E_x, r) \leq 1/4\} \subseteq \{(E_x, r) : |p_b(E_x, r) - \frac{1}{2}| \geq \frac{1}{4}\}$. Since the event on the left is precisely Bad_b , it follows that $\Pr_{E_x, r}[(E_x, r) \in \text{Bad}_b] \leq \Pr_{E_x, r}[|p_b(E_x, r) - \frac{1}{2}| \geq \frac{1}{4}]$, which is negligible by Lemma 2. The same argument applies to either value of b . $\square$

We are now ready to prove indistinguishability or undetectability results. Our final goal is to show that the output distribution of L -bounded rejection sampling is indistinguishable from the output distribution of honest CSIDH key generation. For this, we first prove that the output distributions of unbounded rejection-sampling key generation and honest key generation are indistinguishable, and then show that the output distributions of bounded and unbounded rejection-sampling key generation are indistinguishable.

Lemma 4 (Indistinguishability between unbounded and honest). *Fix $b \in \{0, 1\}$. Define unbounded distribution R_b to be the joint output distribution obtained by sampling (E_x, r) honestly and then repeatedly sampling honest secrets a until $h_r([a] \star E_x) = b$, and outputting (E_x, r, E_a) , where $E_a = [a] \star E_0$. If $p_b(E_x, r) = 0$, instead sample an independent honest secret a and output (E_x, r, E_a) . Then, under GACDH, R_b is computationally indistinguishable from the honest CSIDH joint distribution obtained by sampling E_x, r , and a independently and outputting (E_x, r, E_a) .*

Proof. Let H be the honest distribution. Define R_0 and R_1 as above, and let $M := \frac{1}{2}R_0 + \frac{1}{2}R_1$. We first compare M with H . Fix (E_x, r) and write $q = p_0(E_x, r)$. If $0 < q < 1$, then under honest key generation the event $h_r([a] \star E_x) = 0$ occurs with probability q and the event $h_r([a] \star E_x) = 1$ occurs with probability $1 - q$. Therefore, conditioned on this fixed pair (E_x, r) , the honest output distribution is the mixture $H \mid (E_x, r) = q(R_0 \mid (E_x, r)) + (1 - q)(R_1 \mid (E_x, r))$. By contrast, M mixes the same two conditional distributions with equal weights, namely $M \mid (E_x, r) = \frac{1}{2}(R_0 \mid (E_x, r)) + \frac{1}{2}(R_1 \mid (E_x, r))$. Since $R_0 \mid (E_x, r)$ assigns probability only to outputs satisfying $h_r([a] \star E_x) = 0$, while $R_1 \mid (E_x, r)$ assigns probability only to outputs satisfying $h_r([a] \star E_x) = 1$, their supports are disjoint. By the definition of statistical distance, the statistical distance between the two mixtures is determined by the difference in their mixture weights, namely $\Delta(H \mid (E_x, r), M \mid (E_x, r)) = \frac{1}{2}(|q - \frac{1}{2}| + |(1 - q) - \frac{1}{2}|) = |q - \frac{1}{2}|$. If $q = 0$ or $q = 1$, one predicate value never occurs. By definition, the sampler for the impossible value outputs an independently generated honest public key, while conditioning on the value that occurs with probability one also gives the honest distribution. Hence both components of M are honest for this fixed pair, so the conditional statistical distance is zero. Thus, in all cases, $\Delta(H \mid (E_x, r), M \mid (E_x, r)) \leq |p_0(E_x, r) - \frac{1}{2}|$.

Since $M = \frac{1}{2}R_0 + \frac{1}{2}R_1$ and both R_0 and R_1 sample (E_x, r) according to the same honest distribution as H , for every fixed (e_x, ρ) we have $\Pr_{R_0}[(E_x, r) = (e_x, \rho)] = \Pr_{R_1}[(E_x, r) =$

$(e_x, \rho)] = \Pr_H[(E_x, r) = (e_x, \rho)]$. Therefore, $\Pr_M[(E_x, r) = (e_x, \rho)] = \frac{1}{2} \Pr_{R_0}[(E_x, r) = (e_x, \rho)] + \frac{1}{2} \Pr_{R_1}[(E_x, r) = (e_x, \rho)] = \Pr_H[(E_x, r) = (e_x, \rho)]$. Hence H and M have the same marginal distribution on (E_x, r) . Therefore, the statistical distance between their joint distributions is the average of the corresponding conditional statistical distances, namely $\Delta(H, M) = \mathbb{E}_{E_x, r} [\Delta(H | (E_x, r), M | (E_x, r))]$. Plug in the bound above, namely $\Delta(H | (E_x, r), M | (E_x, r)) \leq |p_0(E_x, r) - \frac{1}{2}|$, together with Lemma 2, we have $\Delta(H, M) \leq \mathbb{E}_{E_x, r} [|p_0(E_x, r) - \frac{1}{2}|] = \text{negl}(\lambda)$.

We now show that R_b is computationally indistinguishable from H . Suppose for contradiction that there exists a polynomial-time distinguisher $\mathcal{D}$ that distinguishes R_b from H with non-negligible advantage. Since H and M have negligible statistical distance, $\mathcal{D}$ also distinguishes R_b from M with non-negligible advantage. Let q_b and q_{1-b} be the probabilities that $\mathcal{D}$ outputs 1 on samples from R_b and R_{1-b} , respectively. Since $M = \frac{1}{2}R_b + \frac{1}{2}R_{1-b}$, we have $\Pr[\mathcal{D}(M) = 1] = \frac{q_b + q_{1-b}}{2}$. Then the non-negligible distinguishing advantage between R_b and M is $|q_b - \frac{q_b + q_{1-b}}{2}| = \frac{1}{2}|q_b - q_{1-b}|$. Hence $|q_b - q_{1-b}|$ is non-negligible.

Now consider the following two polynomial-time predictors for an honest hard-core challenge (E_a, E_x, r) . The first runs $\mathcal{D}(E_x, r, E_a)$ and predicts b if $\mathcal{D}$ outputs 1, and predicts $1 - b$ otherwise. The second uses the opposite prediction rule. Consider first a sample from M . Outside $\text{Bad}_0 \cup \text{Bad}_1$, we have $p_b(E_x, r) > 0$ and $p_{1-b}(E_x, r) > 0$, so neither R_b nor R_{1-b} falls back to honest sampling of a . Hence a sample from R_b satisfies $h_r([a] \star E_x) = b$, while a sample from R_{1-b} satisfies $h_r([a] \star E_x) = 1 - b$. By Lemma 3, the exceptional event $\text{Bad}_0 \cup \text{Bad}_1$ has negligible probability. Hence, on $M = \frac{1}{2}R_b + \frac{1}{2}R_{1-b}$, the first predictor succeeds with probability $\frac{1}{2}q_b + \frac{1}{2}(1 - q_{1-b}) \pm \text{negl}(\lambda) = \frac{1}{2} + \frac{1}{2}(q_b - q_{1-b}) \pm \text{negl}(\lambda)$, while the second predictor succeeds with probability $\frac{1}{2}(1 - q_b) + \frac{1}{2}q_{1-b} \pm \text{negl}(\lambda) = \frac{1}{2} - \frac{1}{2}(q_b - q_{1-b}) \pm \text{negl}(\lambda)$. Since $|q_b - q_{1-b}|$ is non-negligible, at least one of these two predictors predicts the hard-core bit on M with non-negligible advantage over $1/2$.

Finally, since M and H have negligible statistical distance, and statistical distance bounds the difference in the probability of any event, for the event that a predictor correctly predicts $h_r([a] \star E_x)$ we have $|\Pr_M[\text{succ}] - \Pr_H[\text{succ}]| \leq \Delta(M, H) = \text{negl}(\lambda)$. Hence changing the input distribution from M to H changes the predictor's success probability by at most a negligible amount. Thus one of the two polynomial-time predictors predicts $h_r([a] \star E_x)$ on an honest challenge with non-negligible advantage, contradicting Lemma 1. Hence R_b is computationally indistinguishable from H . $\square$

Lemma 5 (Indistinguishability between bounded and unbounded). *Fix $b \in \{0, 1\}$. Let $R_b^{(L)}$ be the joint output distribution obtained by sampling (E_x, r) honestly, then sampling at most L independent candidate secrets $a_1, a_2, \dots$ and using the first a_j satisfying $h_r([a_j] \star E_x) = b$. If no such candidate exists, sample a fresh honest secret a instead. In either case, output (E_x, r, E_a) , where $E_a = [a] \star E_0$. Then $\Delta(R_b^{(L)}, R_b) = \text{negl}(\lambda)$.*

Proof. Fix (E_x, r) and write $q = p_b(E_x, r)$. We first compare the conditional distributions of $R_b^{(L)}$ and R_b for this fixed pair.

If $q = 0$, then by definition both samplers output an independently generated honest public key, so $\Delta(R_b^{(L)} | (E_x, r), R_b | (E_x, r)) = 0$.

Now suppose that $q > 0$. For the fixed pair (E_x, r) , the bounded procedure independently samples at most L honest secrets $a_1, a_2, \dots$, where each a_j satisfies $h_r([a_j] \star E_x) = b$ with probability q . The unbounded procedure keeps sampling honest secrets until the first one

satisfying this condition appears and outputs its public key. The bounded procedure does the same if one of $a_1, \dots, a_L$ satisfies the condition, and otherwise outputs a fresh honest public key. Therefore, the bounded and unbounded procedures differ only when all L sampled secrets fail, which occurs with probability $(1 - q)^L$. Hence $\Delta(R_b^{(L)} \mid (E_x, r), R_b \mid (E_x, r)) \leq (1 - q)^L$.

Since $R_b^{(L)}$ and R_b have the same marginal distribution on (E_x, r) , we have $\Delta(R_b^{(L)}, R_b) = \mathbb{E}_{E_x, r} [\Delta(R_b^{(L)} \mid (E_x, r), R_b \mid (E_x, r))]$. We evaluate this expectation by separating the pairs (E_x, r) according to whether $q \leq \frac{1}{4}$ or $q > \frac{1}{4}$. The contribution from the event $q \leq \frac{1}{4}$ is $\Pr[q \leq \frac{1}{4}] \mathbb{E} [\Delta(R_b^{(L)} \mid (E_x, r), R_b \mid (E_x, r)) \mid q \leq \frac{1}{4}]$. Since statistical distance is at most 1, this contribution is at most $\Pr[q \leq \frac{1}{4}]$. The contribution from the event $q > \frac{1}{4}$ is $\Pr[q > \frac{1}{4}] \mathbb{E} [\Delta(R_b^{(L)} \mid (E_x, r), R_b \mid (E_x, r)) \mid q > \frac{1}{4}]$. Since $q > \frac{1}{4}$ implies $\Delta(R_b^{(L)} \mid (E_x, r), R_b \mid (E_x, r)) \leq (1 - q)^L \leq (3/4)^L$, this contribution is at most $(3/4)^L$. Hence $\Delta(R_b^{(L)}, R_b) \leq \Pr[q \leq \frac{1}{4}] + (3/4)^L$. By Lemma 3, the first term is negligible. Since $L = \lceil \log_2 p \rceil = \Theta(\lambda)$, the second term is also negligible. Therefore $\Delta(R_b^{(L)}, R_b) = \text{negl}(\lambda)$. $\square$

Theorem 1 (Rejection-sampling indistinguishability). *Fix $b \in \{0, 1\}$. Let $R_b^{(L)}$ be the joint output distribution of the bounded rejection-sampling procedure defined above, which honestly samples (E_x, r) and outputs (E_x, r, E_a) . If GACDH is hard, then $R_b^{(L)}$ is computationally indistinguishable from the honest joint distribution obtained by sampling E_x, r , and a independently and outputting (E_x, r, E_a) . In particular, the kleptographically generated public key E_a is computationally indistinguishable from an honestly generated public key even when (E_x, r) is known.*

Proof. Let H be the honest joint distribution obtained by independently and honestly sampling E_x, r , and a and outputting (E_x, r, E_a) . By Lemma 4, $R_b \approx_c H$. Suppose for contradiction that $R_b^{(L)} \not\approx_c H$. Then there exists a polynomial-time distinguisher $\mathcal{D}$ such that $|\Pr[\mathcal{D}(R_b^{(L)}) = 1] - \Pr[\mathcal{D}(H) = 1]|$ is non-negligible. By Lemma 5, $\Delta(R_b^{(L)}, R_b) = \text{negl}(\lambda)$. Let $q_L := \Pr[\mathcal{D}(R_b^{(L)}) = 1]$ and $q := \Pr[\mathcal{D}(R_b) = 1]$. Since statistical distance bounds the difference in the probability of any event, we have $|q_L - q| \leq \Delta(R_b^{(L)}, R_b) = \text{negl}(\lambda)$. Therefore, by the reverse triangle inequality, $|\Pr[\mathcal{D}(R_b) = 1] - \Pr[\mathcal{D}(H) = 1]| \geq |\Pr[\mathcal{D}(R_b^{(L)}) = 1] - \Pr[\mathcal{D}(H) = 1]| - |\Pr[\mathcal{D}(R_b^{(L)}) = 1] - \Pr[\mathcal{D}(R_b) = 1]|$. The first term on the right is non-negligible by the contradiction assumption and the second is negligible by Lemma 5, so the left-hand side is non-negligible. Thus $\mathcal{D}$ distinguishes R_b from H with non-negligible advantage, contradicting Lemma 4. Hence $R_b^{(L)} \approx_c H$. $\square$

Theorem 1 can be interpreted in terms of the commutative CSIDH action. In leakage session i , the four curves form the following commutative square:

$$\begin{array}{ccc} E_0 & \xrightarrow{[x]} & E_x \\ [a_i] \downarrow & & \downarrow [a_i] \\ E_{a_i} & \xrightarrow{[x]} & Z_i. \end{array}$$

6

The attack can be viewed as sampling a CSIDH public key E_{a_i} according to one bit of information about the parallelization curve Z_i . Theorem 1 says that this conditioning cannot be detected from (E_x, E_{a_i}, r_i) under GACDH. In particular, knowing E_x and r_i does not help determine how a_i was selected.

Theorem 1 concerns a single rejection-sampled public key. The following corollary extends the claim to the joint distribution of polynomially many rejection-sampled public keys.

Corollary 1 (Multi-session indistinguishability). *Let $t = t(\lambda)$ be polynomial, let E_x be honestly generated, and let $r_1, \dots, r_t$ be independent uniform strings. Fix any target bits $b_1, \dots, b_t \in \{0, 1\}$. For each i , generate E_{a_i} using the bounded rejection-sampling procedure above with target bit b_i , string r_i , and fresh independent randomness. Then the joint distribution $(E_x, r_1, \dots, r_t, E_{a_1}, \dots, E_{a_t})$ is computationally indistinguishable from the honest joint distribution obtained by sampling E_x and $r_1, \dots, r_t$ in the same way and sampling $E_{a_1}, \dots, E_{a_t}$ as independent honest CSIDH public keys.*

Proof. We prove the claim by a hybrid argument using the single-session indistinguishability of Theorem 1. For $j = 0, \dots, t$, define a hybrid distribution H_j in which the first j public keys $E_{a_1}, \dots, E_{a_j}$ are generated honestly, while the remaining $t - j$ public keys are generated by the bounded rejection-sampling procedure with target bits $b_{j+1}, \dots, b_t$. In every hybrid, the same honestly generated E_x and independently sampled strings $r_1, \dots, r_t$ are used. Thus H_0 is the multi-session rejection-sampling distribution of the corollary, while H_t is the honest multi-session distribution.

Suppose for contradiction that there exists a polynomial-time distinguisher $\mathcal{D}$ that distinguishes H_0 from H_t with non-negligible advantage $\varepsilon(\lambda)$. By the triangle inequality, the total distinguishing advantage between H_0 and H_t is at most the sum of the distinguishing advantages between the adjacent hybrids H_{j-1} and H_j . Therefore, for some $j \in \{1, \dots, t\}$, $\mathcal{D}$ distinguishes H_{j-1} from H_j with advantage at least $\varepsilon(\lambda)/t$. Since t is polynomial, $\varepsilon(\lambda)/t$ is still non-negligible.

The distributions H_{j-1} and H_j differ only in session j : in H_{j-1} , E_{a_j} is generated by the bounded rejection-sampling procedure with target bit b_j and string r_j , whereas in H_j , E_{a_j} is generated honestly.

We now use $\mathcal{D}$ to distinguish the two distributions in Theorem 1 for the fixed target bit b_j . Given a challenge (E_x, r, E) from that theorem, set $r_j = r$ and $E_{a_j} = E$. Sample all other strings r_i independently and uniformly. For each $i < j$, generate E_{a_i} honestly. For each $i > j$, generate E_{a_i} using the bounded rejection-sampling procedure with the same E_x , string r_i , target bit b_i , and fresh independent randomness.

If E is generated by the bounded rejection-sampling procedure, then the resulting joint distribution is exactly H_{j-1} . If E is honestly generated, then the resulting joint distribution is exactly H_j . Therefore $\mathcal{D}$ distinguishes the two challenge distributions of Theorem 1 with non-negligible advantage, contradicting that theorem. Hence H_0 and H_t are computationally indistinguishable. $\square$

Therefore, using successive strings r_i across sessions leaks the target bits b_i one at a time, with Alice's sequence of kleptographically generated public keys remaining computationally indistinguishable from a sequence of honestly generated CSIDH public keys.

There is one more thing we want to address to optimize the attack scenario, which is about how the strings r_i can be obtained. Although the attack above pre-embeds polynomially

many independent uniform strings in Alice's device, they need not be fixed at setup, and their number need not be fixed in advance. In fact, fresh public random bits can be derived from honestly generated CSIDH public keys E_a as the device is used. In the following, we first show that the parity of the canonical Montgomery coefficient of an honest CSIDH public key E_a is nearly balanced, and then show how to use von Neumann debiasing to generate independent uniform bits.

Lemma 6 (Parity of honest public keys). *Let a be sampled from the honest CSIDH secret-key distribution, let $E_a = [a] \star E_0 = E_A$ be an honest public curve, and let $P(E_a) = A \bmod 2$ be the parity of the canonical Montgomery coefficient $A \in \{0, \dots, p-1\}$. Under GACDH, we have $\Pr[P(E_a) = 0] = \frac{1}{2} + \text{negl}(\lambda)$ and $\Pr[P(E_a) = 1] = \frac{1}{2} - \text{negl}(\lambda)$.*

Proof. Note that the honest secret-key distribution is invariant under $a \mapsto -a$. Moreover, if $[a] \star E_0 = E_A$, then $[-a] \star E_0 = E_{-A}$, since inversion of the CSIDH class-group action on E_0 corresponds to quadratic twisting [4]. Thus every nonzero coefficient A is paired with $-A$ with the same probability. Since p is odd, A and $-A$ have opposite parity. This means that, apart from the fixed point $A = 0$ of the map $A \mapsto -A$, the two parities occur with the same probability.

To handle the fixed point $A = 0$, let $\delta = \Pr_a[[a] \star E_0 = E_0]$. We can see that under GACDH, δ is negligible, since otherwise if δ were non-negligible, then given a random GACDH instance (E_0, E_a, E_x) , with non-negligible probability we will have $E_a = E_0$, and one could simply output E_x as the answer of GACDH, and this would solve GACDH with non-negligible probability. Hence $\delta = \text{negl}(\lambda)$. Since $A = 0$ has parity 0, we have $\Pr[P(E_a) = 0] = \frac{1-\delta}{2} + \delta = \frac{1}{2} + \frac{\delta}{2}$ and $\Pr[P(E_a) = 1] = \frac{1-\delta}{2} = \frac{1}{2} - \frac{\delta}{2}$. This completes the proof. $\square$

This lemma gives a simple way for the device and Eve to derive public uniform bits from honestly generated CSIDH public keys E_a using the von Neumann debiasing procedure. Consider the parity bits of two successive independently generated honest public keys. If the pair is 01, output 0; if it is 10, output 1; and if it is 00 or 11, discard the pair and try again. We show that this generates one uniform bit using about four public keys E_a on average.

Let $q = \Pr[P(E_a) = 1]$. A pair is accepted with probability $\Pr[01] + \Pr[10] = 2q(1-q)$. By Lemma 6, we have $q = \frac{1}{2} - \text{negl}(\lambda)$, and hence the acceptance probability is $2q(1-q) = \frac{1}{2} - \text{negl}(\lambda)$. Moreover, $\Pr[01] = (1-q)q = \Pr[10]$, so conditioned on acceptance, the resulting bit is exactly uniform. Since each trial uses two honest public keys, the expected number of honest public keys E_a required to obtain one uniform bit is $\frac{1}{q(1-q)} = 4 + \text{negl}(\lambda)$. Applying this procedure to fresh independent honest public keys gives independent uniform bits. Collecting L such bits gives a fresh uniform string r_i .

Notice also that Alice will learn these strings r_i as well. This is not a problem as r_i were expected to be public for using the Goldreich–Levin theorem. The Goldreich–Levin argument only requires r_i to be uniform and independent of the current leakage-session secret key a_i and the independently generated embedded key E_x . Thus the device and Eve can derive each random string r_i from earlier honestly generated public keys before using it in a leakage session.

We remark that this attack can leak any secret of Alice instead of just leaking a shared curve E_A . For example, the device may leak the ephemeral CSIDH secret e of Alice, and Eve can later use e together with the public key $[f] \star E_0$ of Bob to compute the shared curve $E_A = ([e][f]) \star E_0$. Unlike E_A , which is known to Alice, the ephemeral secret e can be erased

after the leakage is complete, so Alice may never recover it even if she later reverse-engineers the device.

We have now proved all the main results of this note. They provide the foundation for studying how this attack satisfies the definitions of several notions in the kleptography and algorithm-substitution literature, including regular, weak, strong, and polynomially bounded SETUPS and SETUP bandwidth [12–14], undetectability and strong undetectability of algorithm-substitution attacks [1, 2], and state-reset detection and asymmetric ASAs [11]. Here we do not go into details.

We end this note by explaining why a natural character-based detection method does not detect our attack. In the classical Diffie–Hellman SETUP [13], Young and Yung considered the setting in which the device is not reverse-engineered but the owner knows her own ephemeral secret keys. In their Diffie–Hellman SETUP, the algebraic relation between successive secret exponents can be detected through quadratic residuity, which motivates the additional blinding term in their attack. A natural analogue for CSIDH would be a non-trivial character $\chi : \text{cl}(\mathcal{O}) \rightarrow \{\pm 1\}$. To see how such a detector could work, suppose the hidden relation were encoded by the condition $\chi([x][a_i]) = (-1)^{b_i}$. Since χ is multiplicative, this gives $(-1)^{b_i} \chi([a_i]) = \chi([x])$. An owner who knows a_i and b_i can therefore compute $s_i := (-1)^{b_i} \chi([a_i])$ in every session. Although $\chi([x])$ is unknown, the SETUP relation forces all the values s_i to be equal. Provided that $s_i = 1$ and $s_i = -1$ each occur with non-negligible probability under honest key generation, repeating the test over several sessions would reveal the existence of a subversion attack, without learning either the secret class $[x]$ or the public curve E_x embedded in the device. However, this detection does not apply to our attack for at least two reasons. First, there is no fixed multiplicative character relation across sessions: each a_i is freshly and independently rejection-sampled using the condition $h_{r_i}([a_i] \star E_x) = b_i$ for an independent string r_i . These conditions are not evaluations of a fixed multiplicative character on $[x][a_i]$, so the cancellation producing the statistic s_i above does not apply. Second, for the standard CSIDH setting $p \equiv 3 \pmod{8}$ and $\mathcal{O} = \mathbb{Z}[\sqrt{-p}]$, genus theory gives odd class number $h(-p)$ for the maximal order, while the conductor-two class-number formula gives $h(-4p) = 3h(-p)$ [8]. Thus $\mathcal{O}$ also has odd class number. Hence $\text{cl}(\mathcal{O})$ has no quotient of order 2, so $\text{Hom}(\text{cl}(\mathcal{O}), \{\pm 1\})$ is trivial. In particular, the assigned genus characters [5, 6, 8] provide no non-trivial quadratic character analogous to the Legendre symbol. Thus the specific character-based detection that motivates blinding in the Young–Yung Diffie–Hellman attack requires no special treatment here.

REFERENCES

- [1] M. Bellare, J. Jaeger and D. Kane, *Mass-surveillance without the State: Strongly Undetectable Algorithm-Substitution Attacks*, ACM CCS 2015, ACM, 2015, pp. 1431–1440.
- [2] M. Bellare, K. G. Paterson and P. Rogaway, *Security of Symmetric Encryption against Mass Surveillance*, CRYPTO 2014, Part I, LNCS 8616, Springer, 2014, pp. 1–19.
- [3] G. Banegas, J. Krämer, T. Lange, M. Meyer, L. Panny, K. Reijnders, J. Sotáková and M. Trimoska, *Disorientation Faults in CSIDH*, EUROCRYPT 2023, Part V, LNCS 14008, Springer, 2023, pp. 310–342.
- [4] W. Castryck, T. Lange, C. Martindale, L. Panny and J. Renes, *CSIDH: An Efficient Post-Quantum Commutative Group Action*, ASIACRYPT 2018, Part III, LNCS 11274, Springer, 2018, pp. 395–427.
- [5] W. Castryck, J. Sotáková and F. Vercauteren, *Breaking the Decisional Diffie–Hellman Problem for Class Group Actions Using Genus Theory*, CRYPTO 2020, Part II, LNCS 12171, Springer, 2020, pp. 92–120.
- [6] W. Castryck, J. Sotáková and F. Vercauteren, *Breaking the Decisional Diffie–Hellman Problem for Class Group Actions Using Genus Theory: Extended Version*, J. Cryptology **35** (2022), no. 4, Article 24, 30 pp.

- [7] D. Cervantes-Vázquez, M. Chenu, J.-J. Chi-Domínguez, L. De Feo, F. Rodríguez-Henríquez and B. Smith, *Stronger and Faster Side-Channel Protections for CSIDH*, LATINCRYPT 2019, LNCS 11774, Springer, 2019, pp. 173–193.
- [8] D. A. Cox, *Primes of the Form $x^2 + ny^2$: Fermat, Class Field Theory, and Complex Multiplication*, Wiley, 1989.
- [9] P. Gajland, B. de Kock, M. Quaresma, G. Malavolta and P. Schwabe, *SWOOSH: Efficient Lattice-Based Non-Interactive Key Exchange*, 33rd USENIX Security Symposium (USENIX Security 24), USENIX Association, 2024, pp. 487–504.
- [10] O. Goldreich and L. A. Levin, *A Hard-Core Predicate for All One-Way Functions*, STOC 1989, ACM, 1989, pp. 25–32.
- [11] P. Hodges and D. Stebila, *Algorithm Substitution Attacks: State Reset Detection and Asymmetric Modifications*, IACR Trans. Symmetric Cryptol. **2021**(2), 2021, pp. 389–422.
- [12] A. Young and M. Yung, *The Dark Side of “Black-Box” Cryptography, or: Should We Trust Capstone?*, CRYPTO 1996, LNCS 1109, Springer, 1996, pp. 89–103.
- [13] A. Young and M. Yung, *Kleptography: Using Cryptography Against Cryptography*, EUROCRYPT 1997, LNCS 1233, Springer, 1997, pp. 62–74.
- [14] A. Young and M. Yung, *Bandwidth-Optimal Kleptographic Attacks*, CHES 2001, LNCS 2162, Springer, 2001, pp. 235–250.