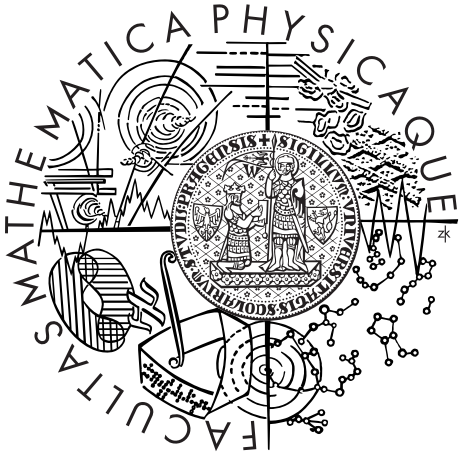
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Adéla Haníková

Eliptické křivky a testování prvočíselnosti

Katedra algebry

Vedoucí diplomové práce:	prof. RNDr. Aleš Drápal, CSc., DSc.
Studijní program:	Matematika
Studijní obor:	Matematické metody informační bezpečnosti

Praha 2015

Chtěla bych poděkovat svému vedoucímu prof. RNDr. Aleši Drápalovi, CSc., DSc. za odborné vedení, za pomoc a rady při zpracování této práce. Dále bych ráda poděkovala Mgr. Robertovi El Bashirovi, Dr. a Mgr. Janu Jeronýmovi Zvánovcovi za pomoc při programování algoritmu.

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Eliptické křivky a testování prvočíselnosti

Autor: Adéla Haníková

Katedra: Katedra algebry

Vedoucí diplomové práce: prof. RNDr. Aleš Drápal, CSc., DSc., Katedra algebry

Abstrakt: Cílem této práce je popsat a implementovat metodu faktorizace pomocí eliptických křivek s využitím křivek v Edwardsově tvaru. Práce se dá pomyslně rozdělit na dvě části, přičemž první část práce se zabývá teorií Edwardsových křivek, zejména vlastnostmi příslušných eliptických funkčních těles. Druhá část pak popisuje využití ve faktorizačním algoritmu a to čistě teoreticky i prakticky tak, jak je algoritmus skutečně implementován. Přínosem této práce je přiložená implementace faktorizace pomocí eliptických křivek využívající grafickou kartu, která je díky paralelizaci rychlejší než obecně nejpoužívanější implementace GMP-ECM.

Klíčová slova: ECM, eliptické křivky, Edwardsovy křivky, faktorizace

Title: Elliptic curves and primality testing

Author: Adéla Haníková

Department: Department of Algebra

Supervisor: prof. RNDr. Aleš Drápal, CSc., DSc., Department of Algebra

Abstract: The aim of the thesis is to desribe and implement the elliptic curve factorization method using curves in Edwards form. The thesis can be notionally divided into two parts. The first part deals with the theory of Edwards curves especially with properties of elliptic function fields. The second part deals with the factorization algorithm using Edwards form both formally and practically in the way the algorithm is really implemented. The contribution of this thesis is the enclosed implementation of the elliptic curve factorisation algorithm which can be run on a graphic card and which is faster than the state-of-the-art implementation GMP-ECM.

Keywords: ECM, elliptic curves, Edwards curves, factorization

Obsah

1 Základní teorie **5**
- 1.1 Algebraická funkční tělesa . . . 5
- 1.2 Eliptická funkční tělesa . . . 8
- 1.3 Rozšíření algebraických funkčních těles . . . 11
- 1.4 Racionální zobrazení . . . 12

2 Polynom stupně 4 **13**
- 2.1 Nečtvercový vedoucí koeficient . . . 17

3 Edwardsovy křivky **18**
- 3.1 Přehled . . . 18
- 3.2 Biracionální ekvivalence . . . 19
- 3.3 Algebraický pohled . . . 20
  - 3.3.1 Pokud d není čtverec . . . 24
  - 3.3.2 Prvek řádu 4 . . . 24
- 3.4 Sčítání . . . 26
  - 3.4.1 d je druhou mocninou . . . 26
- 3.5 Geometrický pohled . . . 27
  - 3.5.1 Divisory . . . 27
  - 3.5.2 Sčítání v $\mathbb{R}$ . . . 29

4 Faktorizace metodou ECM **32**
- 4.1 Pollardova $p-1$ metoda . . . 32
- 4.2 Základní popis . . . 33
  - 4.2.1 Shrnutí . . . 36
- 4.3 ECM a Edwardsovy křivky . . . 36
  - 4.3.1 Počítání modulo N . . . 36
  - 4.3.2 Algoritmus formálně . . . 37
  - 4.3.3 Volba paramterů . . . 39
- 4.4 Souřadnice a vzorce pro implementaci . . . 40
  - 4.4.1 Vzorce . . . 40
  - 4.4.2 Souřadnice . . . 40
  - 4.4.3 Kombinace . . . 41
  - 4.4.4 Algoritmus neformálně . . . 43
- 4.5 Druhá fáze . . . 45
- 4.6 Parametry . . . 45

5 Implementace **46**
5.1 Programování grafických karet 46
5.1.1 OpenCL 47
5.1.2 Jazyky 47
5.2 Předvýpočty 47
5.3 Počítání na křivce 48
5.3.1 Výpočty 49
5.4 Aritmetika velkých čísel 50
5.4.1 Montgomeryho reprezentace 51
5.4.2 Algoritmy 51
5.5 Měření 53

Literatura **57**

Seznam obrázků **59**

Seznam tabulek **60**

Přílohy **61**

Úvod

Prvočísla, tedy čísla, která jsou dělitelná pouze sama sebou a jedničkou, jsou známa již z dob před začátkem našeho letopočtu. Jejich objev je přisuzován pythagorejcům, skupině založené Pythagorem, která byla činná okolo roku 500 př.n.l. Navzdory tomu se problémem rozkladu čísla na prvočísla, případně určení, zda je dané číslo prvočíslo, až do nedávné doby nikdo příliš nezabýval. Vědělo se, že nalézt rozklad je možné pro každé číslo, dokonce se vědělo jak, ovšem počítat rozklady ručně se pravděpodobně nikomu příliš nechtělo. Zvláště pak za situace, kdy by tato aktivita neměla žádný větší význam.

Změna nastala s vývojem výpočetní techniky, díky které bylo najednou možné faktorizovat větší čísla a mechanická práce byla přenechána stroji. Ještě více podnítila zájem o tuto oblast kryptologie, ve které se objevily šifry, jejichž bezpečnost stojí právě na faktu, že nejsme schopni v rozumném čase faktorizovat velká čísla, jako např. RSA. Kromě vývoje „superpočítačů“ se pochopitelně začaly vyvíjet i složitější algoritmy pro faktorizaci. Aktuálním rekordmanem je číselné síto, pomocí kterého se podařilo v roce 2009 podařilo faktorizovat číslo RSA-768, číslo dlouhé 768 bitů (232 decimálních míst), které je součinem dvou přibližně stejně velkých prvočísel. Při použití číselného síta je ovšem potřeba faktorizovat další, pomocná čísla a zde se často jako pomocný algoritmus používá faktorizace pomocí eliptických křivek, „Elliptic Curve Method“ nebo-li ECM.

Využít grupu bodů eliptické křivky pro faktorizaci navrhl poprvé H. W. Lenstra roku 1987 [14] a nápad se rychle ujal. Od té doby je snaha co nejvíce vylepšit reálnou časovou náročnost algoritmu, díky čemuž můžeme nacházet ve stejném čase větší dělitele. Délka běhu ECM záleží z velké části na počtu aritmetických operací, které potřebujeme k sečtení dvou bodů křivky, a to záleží na volbě reprezentace křivky. Donedávna se používaly křivky v Montgomeryho tvaru, než H. M. Edwards roku 2007 uvedl ve svém článku [10] nový tvar eliptických křivek, pro který D. J. Bernstein a T. Lange [6] vzápětí uvedli „rychlejší“ vzorce.

Možnost dalšího zrychlení je dána využitím grafických karet, které umožňují vysokou míru paralelizace a stále více se využívají nejen pro vykreslování grafiky, ale také pro obecné výpočetní úkony. ECM je sice vcelku komplexní algoritmus, pro který nejsou grafické karty dimenzované, ovšem na druhé straně může velmi efektivně využít paralelizace, je proto přirozené se alespoň pokusit využít grafické karty pro ECM.

Jelikož ECM patří k jednoduchým algoritmům, které stojí na složité teorii, jsou první tři kapitoly věnovány právě teorii eliptických křivek. V kapitole 1 jsou shrnuty základy z algebraických křivek, které jsou potřeba pro další práci v kapitolách 2 a 3. V první ze zmíněných kapitol zkoumáme algebraické vlastnosti křivek daných předpisem $y^2 = f(x)$ pro $f(x)$ stupně čtyři. Na ní navazuje kapitola zabývající se z algebraického hlediska Edwardsovými křivkami jako takovými.

V kapitole 4 je pak popsáno využití Edwardsových křivek ve faktorizačním algoritmu, nejprve čistě formálně a následně matematicky nepřesně tak, jak se skutečně implementuje. Závěrečná kapitola 5 obsahuje stručný popis přiložené implementace, která může být spuštěna na grafické kartě a jsou zde prezentovány výsledky několika testovacích měření.

Kapitola 1

Základní teorie

Zde shrneme teorii algebraických a eliptických křivek, na které staví zbytek této práce. Definujeme základní pojmy, se kterými budeme pracovat a bez důkazu také uvedeme některá důležitá tvrzení, o která se opírají pozdější důkazy. Látka zde vyložená je obsahem kurzů Křivky a funkční tělesa a Eliptické křivky a kryptografie a je také možné ji dohledat například v [17] nebo v [16].

V této kapitole bude K značit libovolné, pevně zvolené těleso a $f \in K[x_1,x_2]$ ireducibilní polynom.

1.1 Algebraická funkční tělesa

Křivkou rozumíme algebraickou množinu $C = V_f = \{(x,y) \in \bar{K}^2 | f(x,y) = 0\}$, kde $\bar{K}$ značí algebraický uzávěr tělesa K, $C(K) = \{(x,y) \in K^2 | f(x,y) = 0\}$ jsou K-racionální prvky křivky. *Souřadnicový okruh* $K[C] = K[x_1,x_2]/(f)$. *Funkční těleso* $K(C)$ je podílové těleso $K[C]$, jeho prvky jsou zlomky ve tvaru $(a + (f))/(b + (f))$, $a,b \in K[x_1,x_2]$, $b \notin (f)$ a nazývají se *racionální funkce na* C. Také se na $K(C)$ můžeme dívat jako na $K(\bar{x_1},\bar{x_2}), \bar{x_i} = x_i + (f), i \in \{1,2\}$ a prvky vnímat jako $a(\bar{x_1},\bar{x_2})/b(\bar{x_1},\bar{x_2})$, $a,b \in K[x_1,x_2]$.

Uvažujme nyní ekvivalenci $\sim$ na množině dvojic (a,b), $a,b \in K[x_1,x_2], b \notin (f)$ takovou, že $(a,b) \sim (c,d)$, pokud $a(\alpha)/b(\alpha) = c(\alpha)/d(\alpha)\ \forall \alpha \in C \setminus (V_b \cup V_d)$. Platí, že $(a,b) \sim (c,d)$ právě tehdy, když $(a + (f))/(b + (f)) = (c + (f))/(d + (f))$ a to je právě tehdy, když existuje nekonečně mnoho $\alpha \in C$ takových že $b(\alpha) \neq 0$ a $d(\alpha) \neq 0$ a $a(\alpha)/b(\alpha) = c(\alpha)/d(\alpha)$.

Pro prvek $\rho \in K(C)$ nazveme *reprezentantem* libovolnou racionální funkci $a/b, a,b \in K[x_1,x_2], b \notin (f)$ takovou, že $\rho = (a + (f))/(b + (f))$. Z předchozího odstavce plyne, že pokud chceme spočítat $\rho(\alpha), \alpha \in C$, nezáleží na volbě reprezentanta, pokud je pro něj hodnota v α definovaná. Můžeme tedy určit $\rho(\alpha)$ pro každé α, pro které existuje reprezentant a/b, že $b(\alpha) \neq 0$. Množinu všech těchto α budeme značit $\mathrm{Dom}(\rho)$.

Definujme $O_\alpha\ \forall \alpha \in C$ jako množinu všech $\rho \in K(C)$, pro které $\alpha \in \mathrm{Dom}(\rho)$ a P_α jako všechny prvky O_α, které dávají v α nulu. Zapsáno množinově $O_\alpha = \{\rho \in K(C) | \alpha \in \mathrm{Dom}(\rho)\}$, $P_\alpha = \{\rho \in O_\alpha | \rho(\alpha) = 0\}$ a platí, že O_α je lokální okruh (tedy okruh s jediným, netriviálním maximálním ideálem) a P_α je onen maximální ideál. Právem se proto O_α nazývá *lokální okruh v* α.

Algebraické funkční těleso

Rozšíření těles F/K nazveme *algebraické funkční těleso*, zkráceně a.f.t., pokud existuje $x \in F$ transcendentní nad K takové, že $[F : K(x)] < \infty$. Řekneme o něm, že je dáno rovnicí $f(x,y) = 0$, pokud existují $x,y \in F$ taková, že $K(x,y) = F$ a $f \in K[x_1,x_2]$ je ireducibilní polynom.

Pro algebraické funkční těleso F/K budeme pak $\tilde{K}$ značit algebraický uzávěr K v F. Libovolný polynom $f \in K[x_1,x_2]$ nazveme *absolutně ireducibilní*, pokud je ireducibilní v $\bar{K}$.

Zmiňme nyní vlastnosti týkající se právě definovaných pojmů. Pro algebraické funkční těleso F/K zadané rovnicí $f(x,y) = 0$ platí, že existuje právě jeden K-izomorfismus $K(C) \cong F$, který zobrazuje $x_1+(f)$ na x a $x_2+(f)$ na y. V takovém případě je x transcendentní nad K právě tehdy, když nejvyšší mocnina proměnné x_2 v f je větší než nula, označme ji d, pak $d = [F : K(x)]$. Navíc platí $\tilde{K} = K$, pokud f je absolutně ireducibilní polynom.

Naopak, pokud K je perfektní těleso, tak každé algebraické funkční těleso je zadáno nějakou rovnicí $f(x,y) = 0$. Za určitých podmínek si tedy algebraická funkční tělesa a tělesa racionálních funkcí křivek odpovídají.

Pro $g \in K[x_1,x_2]$ a $\alpha \in V_g$ označme $(c_1,c_2) = ((\partial g/\partial x_1)(\alpha),(\partial g/\partial x_2)(\alpha))$. Polynom g je *hladký* nebo *nesingulární* v α, pokud $(c_1,c_2) \neq (0,0)$ a *singulární* v opačném případě. Pokud je g hladký ve všech $\alpha \in V_g$, nazýváme ho *hladkým polynomem*.

Diskrétní valuace

Nechť F je těleso. Okruh $R \subsetneq F$ se nazývá *valuační okruh* F, pokud $\forall a \in F^*$ je $a \in R$ nebo $a^{-1} \in R$. Obecně se okruh R nazývá valuační, pokud je to valuační okruh svého podílového tělesa. Dále pro obor integrity R a jeho podílové těleso F definujme *lomený ideál*. To je takový R-modul $A \subset F$, že existuje $c \in R, c \neq 0$, že cA je konečně generovaným ideálem R.

Tvrzení 1.1. *Nechť R je lokální noetherovský obor integrity s maximálním ideálem M. Ať $M = tR$, $t \in R$ a F je podílové těleso R. Pak každé $x \in F^*$ může být vyjádřeno jednoznačně jako $t^i r$, $r \in R^*$, $i \in \mathbb{Z}$. Definujme zobrazení $\nu : F \to \mathbb{Z} \cup \{\infty\}$ následovně: $\nu(x) = i$ a $\nu(0) = \infty$. Platí následující tvrzení:*

- $\nu(xy) = \nu(x) + \nu(y)\ \forall x,y \in F$;
- $\nu(x + y) \geq \min\{\nu(x), \nu(y)\}\ \forall x,y \in F$;
- $\nu(x) = \infty \iff x = 0$;
- *pokud $x \in F$, pak $M = xR \iff \nu(x) = 1$.*

Pro každý vlastní lomený ideál A existuje jednoznačně určené $i \in \mathbb{Z}$, že $A = t^i R = \{x \in F;\ \nu(x) \geq i\}$.

Nyní řekněme, co je to *diskrétní valuace*. Nazýváme tak každé zobrazení $\nu : F \to \mathbb{Z} \cup \{\infty\}$, které splňuje první tři podmínky z tvrzení 1.1. Pokud $\nu(x) = 0\ \forall x \in F^*$, mluvíme i triviální diskrétní valuaci a pokud ν splňuje i čtvrtou podmínku, nazýváme ji *normalizovanou*. Není složité ověřit, že platí také $\nu(1) = 0$

a pro $x \in F^*$ je $\nu(x^{-1}) = -\nu(x)$. Označíme-li $R = \{x \in F|\ \nu(x) \geq 0\}$, R je podokruh F a $x \in R^* \iff \nu(x) = 0$.

Okruh $R \subset F$ se nazývá *diskrétní valuační okruh tělesa* F, pokud existuje netriviální diskrétní valuace ν taková, že $R = \{x \in F|\nu(x) \geq 0\}$. Okruh R se nazývá diskrétní valuační okruh (zkráceně DVO), pokud je to diskrétní valuační okruh svého podílového tělesa.

Platí, že každý DVO je valuační okruh a že R je DVO právě tehdy, když je to noetherovský lokální okruh s hlavním maximálním ideálem. Libovolný prvek $t \in R$ takový, že maximální ideál je roven tR, nazýváme *uniformizující*.

Valuačním okruhem algebraického funkčního tělesa F/K pak nazýváme každý valuační okruh R tělesa F takový, že $R \supseteq K$. Podobně diskrétní valuace ν F/K je každá diskrétní valuace F taková, že $\nu(a) = 0\ \forall a \in K^*$. Každý valuační okruh R algebraického funkčního tělesa F/K je zároveň diskrétní valuační okruh F.

Místa a divisory

Mějme a.f.t. F/K. Množina $M \subset F$ se nazývá *místo*, pokud existuje R valuační okruh F/K, že M je maximální ideál R.

Poznámka. Místo určuje valuační okruh R jednoznačně, neboť $R^* = \{ab^{-1},\ a, b \in M \setminus M^2\}$.

Množinu všech míst F/K značíme $\mathbb{P}_{F/K}$ a platí, že je nekonečná. Pro každé místo $P \in \mathbb{P}_{F/K}$ existuje právě jeden valuační okruh, který ho obsahuje, tento valuační okruh budeme značit O_P. Dále zavedeme pojem *stupeň místa* $\deg(P)$ jako $[O_P/P : (K + P)/P]$. Tato hodnota se rovná $\dim_K(O_P/P)$ a je vždy konečná. Každý valuační okruh R algebraického funkčního tělesa F/K je DVO F a obsahuje $\tilde{K}$. Pro každé $P \in \mathbb{P}_{F/K}$ tedy existuje jednoznačně určená normalizovaná valuace F, kterou budeme značit ν_P.

Formální součet tvaru $\Sigma_{P\in\mathbb{P}_{F/K}} a_P P$, ve kterém $a_P \in \mathbb{Z}$ a $a_P \neq 0$ jen v konečně mnoha případech, nazýváme *divisorem* F/K, přičemž množina P, pro která jsou a_P nenulová, se nazývá *nosič divisoru*. Množina všech divisorů tvoří abelovskou grupu a značíme ji $\mathrm{Div}(F/K)$. Pro dva divisory $A = \Sigma_{P\in\mathbb{P}_{F/K}} a_P P$ a $B = \Sigma_{P\in\mathbb{P}_{F/K}} b_P P$ definujeme $\max\{A,B\} = \Sigma_{P\in\mathbb{P}_{F/K}} \max\{a_P, b_P\}P$ a $\min\{A,B\} = \Sigma_{P\in\mathbb{P}_{F/K}} \min\{a_P, b_P\}P$ a $A \geq B$, pokud $a_P \geq b_P\ \forall P \in \mathbb{P}_{F/K}$. Dále definujeme kladnou část divisoru $A_+ = \Sigma a'_P P$ tak, že $a'_P = a_P$, pokud $a_P \geq 0$ a $a'_P = 0$ ve zbylých případech. Záporná část divisoru A_- se definuje jako $(-A)_+$.

Ačkoli míst je nekonečně mnoho, pro každé $x \in F^*$ existuje jen konečně mnoho míst takových, že $x \in P$, a tudíž $\nu_P(x) > 0$, a také jen konečně mnoho míst, že $x \notin O_P$, a proto $\nu_P < 0$. Můžeme proto pro každé $x \in F^*$ definovat *hlavní divisor* $\mathrm{div}_{F/K}(x)$ nebo zkráceně pouze (x) jako $\Sigma\nu_P(x)P$. Množina všech hlavních divisorů se značí $\mathrm{Princ}(F/K)$ a tvoří podgrupu $\mathrm{Div}(F/K)$.

Tvrzení 1.2. *Nechť F/K je algebraické funkční těleso, $\tilde{K} = K$. Pak pro $x \in F \setminus K$ platí* $\deg(x)_- = \deg(x)_+ = [F : K(x)]$.

Rod

Pro libovolný divisor A označme

$$\mathcal{L}(A) = \{x \in F^*|\ (x) + A\ \geq 0\} \cup \{0\}.$$

Do této množiny náleží právě ta x, pro která $\nu_P(x) \geq -a_P \ \forall P \in \mathbb{P}_{F/K}$. Z vlastností diskrétní valuace pak pro $x,y \in \mathcal{L}(A)$ $\nu_P(x+y) \geq \min\{\nu_P(x),\nu_P(y)\} \geq -a_P$ a pro $\lambda \in K^*$ $(\lambda x) = (x)$. $\mathcal{L}(A)$ je tudíž vektorový prostor nad K a nazývá se *Riemann-Rochův prostor*. Jeho dimenze nad K je konečná a značí se $\ell(A)$.

Riemann-Rochův prostor pak figuruje v jedné z dalších důležitých vět teorie algebraických křivek:

Věta 1.3 (Riemannova)**.** *Ať F/K je algebraické funkční těleso, $\tilde{K} = K$. Pak existuje $\gamma \in \mathbb{N} \cup \{0\}$, že $\deg(A) - \ell(A) < \gamma$ pro každý $A \in \mathrm{Div}(F/K)$.*

Zjevně můžeme ze všech takových γ zvolit to nejmenší, to se značí g a nazývá se rod F/K.

Na Riemannovu větu navazuje Riemann-Rochova věta, která explicitně vyjadřuje rozdíl $\deg(A) - \ell(A)$ a g, pro její formulaci bychom potřebovali zavést další netriviální pojmy. Větu samotnou ale nebudeme potřebovat, pouze jeden z jejích důsledků.

Tvrzení 1.4 (Hlavní důsledek Riemann-Rochovy věty)**.** *Nechť F/K je a.f.t. rodu g. Pokud $A \in Div(F/K)$ takový, že $\deg(A) \geq 2g-1$, pak $\ell(A) = \deg(A) + 1 - g$.*

Zmiňme ještě užitečné tvrzení, které říká, že $K(x)/K$ je a.f.t. rodu nula. A naopak, máme-li F/K rodu nula, ve kterém existuje místo stupně jedna, tak $F = K(x)$ pro nějaké $x \in F$.

1.2 Eliptická funkční tělesa

Nyní již známe všechny pojmy potřebné k tomu, abychom mohli vyslovit definici eliptického funkčního tělesa: Algebraické funkční těleso F/K se nazývá *eliptické funkční těleso*, zkráceně e.f.t., pokud je rodu jedna a obsahuje místo stupně jedna. Ve zbytku kapitoly budeme e.f.t. vždy značit E/K, zatímco pokud použijeme F/K, budeme myslet obecné algebraické funkční těleso.

Důvodů, proč se nám líbí zrovna e.f.t., je hned několik. Jednak místa stupně jedna tvoří grupu. Jednak pro každé e.f.t. E/K existuje f ve tvaru $a - b$, kde $a(x_1,x_2) = b(x_1,x_2)$ je tzv. Weierstrassova rovnice, že E/K zadáno $f(x,y) = 0$. A nakonec místa stupně jedna a prvky $V_f(K)$ si (skoro) odpovídají. Tím pádem (až na drobnosti) tvoří i $V_f(K)$ (konečnou) grupu. Nyní ale popořadě a trochu podrobněji.

Picardova grupa

V libovolném F/K je $\mathrm{Ker}(\deg)/\mathrm{Princ}(F/K)$ správně definovaná abelovská grupa, neboť deg je homomorfismus abelovských grup a jeho jádro $\deg(0)^{-1}$ obsahuje $\mathrm{Princ}(F/K)$. Nazýváme ji *Picardova grupa* a značíme $\mathrm{Pic}(F/K)$.

Množinu všech míst stupně jedna budeme značit $\mathbb{P}^{(1)}_{E/K}$. Ukažme nyní, že tato množina tvoří ve skutečnosti grupu, pomocí izomorfismu s Picardovou grupou.

Lemma 1.5. *Ať E/K je eliptické funkční těleso. Pokud pro $P_1, P_2 \in \mathbb{P}^{(1)}_{E/K}$ platí $P_1 - P_2 \in \mathrm{Princ}(E/K)$, pak $P_1 = P_2$. Navíc pro každý divisor $A \in Div(E/K)$ stupně 1 existuje právě jedno $P \in \mathbb{P}^{(1)}_{E/K}$, že $P - A \ \in \mathrm{Princ}(E/K)$.*

Můžeme tedy pevně vybrat libovolné místo $Q \in \mathbb{P}^{(1)}_{E/K}$ a definovat bijekci $\psi : \mathbb{P}^{(1)}_{E/K} \to \mathrm{Pic}(E/K)$ tak, že místu P přiřadíme rozkladovou třídu $[P-Q]$. Pomocí této bijekce pak můžeme přenést grupovou operaci z Picardovy grupy na $\mathbb{P}^{(1)}_{E/K}$. Shrňme předchozí řádky v lemmatu.

Lemma 1.6. *Ať E/K je e.f.t., $Q \in \mathbb{P}^{(1)}_{E/K}$. Pak $P \to [P-Q]$ je bijekce $\mathbb{P}^{(1)}_{E/K} \to \mathrm{Pic}(E/K)$. Definujme na $\mathbb{P}^{(1)}_{E/K}$ operaci $\oplus$ tak, aby tato bijekce byla izomorfismem grup. Platí $P_1 \oplus P_2 = P_3 \iff [P_1+P_2] = [P_3+Q]$ pro libovolná $P_i \in \mathbb{P}^{(1)}_{E/K}$, $i \in \{1,2,3\}$. Neutrální prvek v $\mathbb{P}^{(1)}_{E/K}$ je Q a pro $k>0$ a libovolná $P_1,\dots,P_k \in \mathbb{P}^{(1)}_{E/K}$ platí $P_1 \oplus \cdots \oplus P_k = R \iff P_1+\cdots+P_k - R - (k-1)Q \in \mathrm{Princ}(E/K)$.*

Body křivky

Už tedy víme, že místa stupně jedna v eliptickém funkčním tělese lze interpretovat jako abelovskou grupu, nyní si řekneme poslední důležitou věc, která propojí body křivky a místa stupně jedna.

Mějme $f \in K[x_1,x_2]$ ireducibilní polynom, $C = V_f$, který je hladký v $\alpha = (\alpha_1,\alpha_2) \in C(K)$. Pak P_α (maximální ideál lokálního okruhu O_α) je místo stupně jedna, $P_\alpha \cap K[C] = (\bar{x}_1 - \alpha_1, \bar{x}_2 - \alpha_2)$, $K[C] \subseteq O_\alpha$ a P_α je jediné místo obsahující $\bar{x}_i - \alpha_i, i \in \{1,2\}$.

Zároveň $\forall P \in \mathbb{P}^{(1)}_{K(C)/K}$ taková, že $K[C] \subseteq O_P$, existuje $\alpha \in C(K) : P \supseteq P_\alpha$. Rovnost nastává, pokud je f v α hladký.

Máme-li hladký polynom, jehož těleso racionálních funkcí je eliptické funkční těleso, pak si K-racionální body křivky a všechna místa $P \in \mathbb{P}^{(1)}_{K(C)/K}, K[C] \subseteq O_P$ odpovídají. Stačí tedy k $C(K)$ přidat prvky, které budou odpovídat zbylým místům stupně jedna a máme grupu.

Místa P taková, že $K[C] \nsubseteq O_P$, budeme nazývat *místa v nekonečnu*, ne vždy to musí být místa stupně 1.

Weierstrassovy křivky

Pokud budeme dohledávat informace o eliptických křivkách, první, na co narazíme, je Weierstrassova rovnice. To je obecně rovnice s koeficienty $a_i \in K, i \in \{1,2,3,4,6\}$ tvaru

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

Označíme levou stranu jako polynom $a(x,y)$ a pravou $b(x,y)$, často se označuje $\omega(x_1,x_2) = 0$ jako Weierstrassova rovnice, kde ω je polynom $a(x_1,x_2) - b(x_1,x_2)$

Nad tělesem charakteristiky různé od dvou a tří můžeme pomocí změny souřadnic každou Weierstrassovu rovnici napsat v tzv. krátkém tvaru

$$x_2^2 = x_1^3 + a_4x_1 + a_6. \tag{1.1}$$

Taková rovnice dává algebraické funkční těleso rodu jedna, pokud $x_1^3 + a_4x_1 + a_6$ nemá vícenásobné kořeny. A naopak v každém e.f.t. E/K najdeme prvky x, y, že E/K bude dáno $y^2 = x^3 + a_4x + a_6$ pro nějaká $a_4, a_6 \in K$ (pokud $\mathrm{char}K \notin \{2,3\}$).

Tvrzení 1.7. *Mějme $C = V_\omega$, kde $\omega(x_1, x_2)$ je Weierstrassova rovnice ve tvaru 1.1 a označme $C(K) = V_\omega \cup \{\mathcal{O}\}$. Na $C(K)$ pak umíme definovat inverz (-) a sčítání (+) tak, že tvoří abelovskou grupu s neutrálním prvkem $\mathcal{O}$. Operace jsou pro body $\alpha = (\alpha_1, \alpha_2)$ a $\beta = (\beta_1, \beta_2)$, $\alpha, \beta \in V_\omega(K)$ definovány následovně:*

$$\alpha + \mathcal{O} = \alpha.$$

Opačný prvek k α je

$$-\alpha = (\alpha_1, -\alpha_2).$$

Pokud $\alpha = -\beta$, pak $\alpha + \beta = \mathcal{O}$, a pokud $\alpha \neq -\beta$, tak

$$\alpha + \beta = (\gamma_1, \gamma_2) = (-\alpha_1 - \beta_1 + \lambda^2, \lambda(\alpha_1 - \gamma_1) - \alpha_2), \tag{1.2}$$

kde $\lambda = \frac{\beta_2 - \alpha_2}{\beta_1 - \alpha_1}$, pokud $\alpha_1 \neq \beta_1$ a $\lambda = \frac{3\alpha_1^2 + a_4}{2\alpha_2}$, pokud $\alpha_1 = \beta_1$.

V prezentacích vysvětlujících, co jsou eliptické křivky a jaké je jejich použití, nebo v některých stručných, přehledových článcích o eliptických křivkách (jako např. na Wikipedii) typicky najdeme definici eliptické křivky jako křivky danou rovnicí 1.1. Jak je ovšem vidět ze stručné teorie, kterou jsme zde vyložili, odpovídající grupovou operaci můžeme zavést na libovolné křivce, jejíž funkční těleso je eliptické funkční těleso.

Montgomeryho křivky

Jedním z používaných typů křivek, které také určují eliptické funkční těleso jsou Montgomeryho křivky.

Definice 1. *Montgomeryho křivka nad tělesem K,* char$K \notin \{2,3\}$ *je definována rovnicí*

$$E_{M,A,B} : By^2 = x^3 + Ax^2 + x,$$

pro $A, B \in K$ takové, že $B(A^2 - 4) \neq 0$.

Přičemž Weierstrassovu křivku $y^2 = x^3 + a_4x + a_6$ lze převést na Montgomeryho právě tehdy, když ji lze ekvivalentně vyjádřit jako $y^2 = x^3 + ux^2 + v^2x$ pro nějaké u,$v \in K$, a to lze právě tehdy, když $\exists \alpha \in K$ kořen polynomu $f(x) = x^3 + a_4x + a_6$ tak, že $3\alpha^2 + a_4$ je v K čtverec.

Vlastnosti grupy

Známe-li předpis eliptické křivky, není jednoduché popsat strukturu grupy tvořené jejími body, přesto není úplně neznámá. Označme pro $n \in \mathbb{N}$ jako $E[n]$ množinu $\{P \in E(\bar{K}) | nP = \infty\}$. Nechť je E eliptická křivka nad K, $n \in \mathbb{N}$ a char K je nula nebo nedělí n. Pak $E[n] \simeq \mathbb{Z}_n \times \mathbb{Z}_n$.

Předně pro každou grupu eliptické křivky nad $\mathbb{F}_q$ platí, že je buď cyklická, nebo izomorfní $\mathbb{Z}_m \times \mathbb{Z}_n$, kde $m, n \in \mathbb{N}$, $m \mid n$ a $m \mid q - 1$. Kromě toho také víme přibližnou hodnotu řádu grupy, což je jedna z klíčových vlastností pro využití ve faktorizaci.

Věta 1.8 (Hasseho)**.** *Mějme eliptickou křivku nad tělesem s q prvky $\mathbb{F}_q$ a označme k počet bodů na této křivce. Platí*

$$|k - (q+1)| \leq 2\sqrt{q}.$$

Kromě omezení řádu je podstatná také „opačná implikace“. Tedy fakt, že pro každé číslo z intervalu $(q+1-\sqrt{q}, q+1+\sqrt{q}$ nalezneme křivku nad $\mathbb{F}_q$, která bude mít daný řád.

1.3 Rozšíření algebraických funkčních těles

Mějme v této podkapitole dvě algebraická funkční tělesa F'/K' a F/K. Řekneme, že F'/K' je rozšířením F/K, pokud $F' \supseteq F$ a $K' \supseteq K$. O rozšíření a.f.t. říkáme, že má nějakou vlastnost (je algebraické, Galoisovo apod.), pokud má tuto vlastnost rozšíření F'/F.

Máme tedy dvě algebraická funkční tělesa, která jsou do sebe vnořená. Pro místa a valuační okruhy platí:

$$O'_P \supseteq O_P \iff P' \supseteq P \iff O_P = O'_P \cap F \iff P = P' \cap F.$$

Navíc pro každé místo P z F/K existuje alespoň jedno místo P' z F'/K', které ho obsahuje, říkáme že P' je nad P a píšeme $P'|P$. Pro každou takovou dvojici míst existuje *ramifikační index*, což je právě jedno přirozené číslo e, že $\nu_{P'}(x) = e\nu_P(x)\ \forall x \in F$. Ramifikační index pro P', P se značí $e(P'|P)$. Kromě něj je ještě definován *relativní stupeň* $P'|P$, který se značí $f(P'|P)$, a je dán jako hodnota $[O_{P'}/P' : (O_P + P')/P']$.

Lemma 1.9. *Ať F'/K' je algebraické rozšíření, $P'|P$, $P' \in \mathbb{P}_{F'/K'}$, $P \in \mathbb{P}_{F/K}$. Pak $f(P'|P) < \infty$ právě když jde o konečné rozšíření. Je-li tomu tak, platí*

$$\deg(P')[K' : K] = f(P'|P)\deg(P).$$

Tvrzení 1.10 (Fundamentální rovnost)**.** *Ať $F'/K' \supseteq F/K$ je konečné. Pak pro každé $P \in \mathbb{P}_{F/K}$ existuje jen konečně mnoho míst P', že $P'|P$. Označme je $P_1, \ldots, P_m$, pak platí*

$$[F' : F] = \Sigma_{i=1}^{m} e(P_i|P)f(P_i|P).$$

Pro „hezká“ rozšíření pak platí další vlastnosti.

Tvrzení 1.11. *Ať $F'/K' \supseteq F/K$ je Galoisovo rozšíření a.f.t. Ať $P'|P$, $P' \in \mathbb{P}_{F'/K'}, P \in \mathbb{P}_{F/K}$. Pak $\forall P'' \in \mathbb{P}_{F'/K'}$ platí $P''|P \iff \exists \sigma \in \mathrm{Gal}(F'/F)$, že $\sigma(P') = P''$.*

Ať P_i, $i \in \{1, \ldots, m\}$ jsou právě všechna místa nad P. Pak existují $e > 0, f > 0$, že $e(P_i|P) = e$, $f(P_i|P) = f\ \forall i$.

Předpokládejme nyní, že K je perfektní těleso, F/K a.f.t., $K \subseteq K'$ algebraické rozšíření a F' nejmenší těleso takové, že obsahuje K' a F. Pak F'/K' má stejný rod jako F/K.

1.4 Racionální zobrazení

Ať $C_1 = V_{f_1}$, $C_2 = V_{f_2}$, f_1,$f_2 \in K[x_1,x_2]$ ireducibilní pro tuto sekci.

Definice 2. $\rho\colon C_1 \to C_2$ *nazvu* racionální zobrazení, *pokud existují* $\rho_1, \rho_2 \in K(C_1)$, *že* $\rho(\alpha) = (\rho_1(\alpha), \rho_2(\alpha))$ $\forall \alpha \in \operatorname{Dom}\rho_1 \cap \operatorname{Dom}\rho_2$.

Ačkoli je v definici řečeno „pro všechna α“, racionální zobrazení určují i taková $\rho_1, \rho_2 \in K(C_1)$, že $(\rho_1(\alpha), \rho_2(\alpha)) \in C_2$ pro nekonečně mnoho $\alpha \in \operatorname{Dom}\rho_1 \cap \operatorname{Dom}\rho_2$.

Tvrzení 1.12. *Je-li* $\rho : C_1 \to C_2$ *nekonstantní racionální zobrazení,* $\rho = (\rho_1, \rho_2)$, *tak* $\rho^* : \sigma \to \sigma \circ \rho$ *je* K*-homomorfismus* $K(C_2) \to K(C_1)$.

Dvě křivky nazveme nazveme *biracionálně ekvivalentní*, pokud jsou izomorfní jejich tělesa racionálních funkcí, tedy pokud ρ^* z 1.12 je K-izomorfismus.

Kapitola 2

Polynom stupně 4

Ve svém článku [10] H. M. Edwards definuje eliptické křivky jako křivky, které mohou být zapsány ve tvaru $z^2 = f(x)$, $\deg(f) \in \{3,4\}$ a f nemá vícenásobné kořeny. Explicitně pak vyjádřil jednoduché racionální zobrazení - izomorfismus - mezi tělesem racionálních funkcí Edwardsovy křivky a křivky $z^2 = f(x)$, $\deg f = 4$.

Toto zobrazení dokonce převádí x z tělesa racionálních funkcí jedné křivky na x z tělesa racionálních funkcí druhé křivky. Podíváme se tedy nejprve, co můžeme říci o a.f.t. zadaném $z^2 = f(x)$, $\deg(f) = 4$. K bude v této kapitole značit těleso charakteristiky různé od dvou.

Lemma 2.1. *Nechť $f \in K[x]$ je polynom stupně 4. Pak polynom $g \in K[x_1, x_2]$, $g = x_2^2 - f(x_1)$ je absolutně ireducibilní $\iff$ je ireducibilní $\iff$ f není druhou mocninou.*

Důkaz. Předpokládejme, že g není ireducibilní, pak tedy $x_2^2 - f(x_1) = uv$, $u, v \in K[x_1, x_2]$ nekonstatní polynomy. Podívejme se na u,v jako na prvky $K(x_1)[x_2]$ a označme $a(x_1)$, $b(x_1)$ jejich vedoucí koeficienty. Platí, že $a(x_1)b(x_1) = 1$, můžeme proto předpokládat, že oba vedoucí koeficienty jsou rovny 1.

Neboť g je monický polynom stupně 2 v proměnné x_2, musí být $u, v \in K(x_1)[x_2]$ polynomy stupně 1. Označme $u = x_2 + s(x_1)$ a $v = x_2 + t(x_1)$, $uv = x_2^2 + x_2(s(x_1) + t(x_1)) + s(x_1)t(x_1)$. Musí tedy platit $s(x_1) = -t(x_1)$ a $f(x_1) = (s(x_1))^2$.

□

Mějme po zbytek kapitoly polynom $f \in K[x_1]$ stupně 4, který není čtverec a označme F/K algebraické funkční těleso zadané $y^2 = f(x)$. Díky lemmatu 2.1 víme, že $\tilde{K} = K$. Mimo to také platí $[F : K(x)] = 2$ a $[F : K(y)] = 4$, tudíž neplatí triviální rovnost $F = K(x)$ nebo $F = K(y)$, kdy bychom rovnou dostali a.f.t. rodu 0.

Lemma 2.2. *Pro každé $P \in \mathbb{P}_{F/K}$ platí, že buď $\nu_P(x) \geq 0$ a $\nu_P(y) \geq 0$, nebo $\nu_P(x) = 2\nu_P(y) < 0$.*

Důkaz. Vyplývá ze základních vlastností diskrétní valuace a rovnosti $y^2 = f(x)$.

Předpokládejme, že $\nu_P(x) < 0$ a $\nu_P(y) \geq 0$ pro nějaké $P \in \mathbb{P}_{F/K}$. $\nu_P(y^2) = 2\nu_P(y) \geq 0$ a $\nu_P(f(x)) = \nu_P(x^4) = 4\nu_P(x) < 0 \Rightarrow \nu_P(y^2) > \nu_P(f(x))$, což je spor, neboť $y^2 = f(x)$. Obdobně dojdeme ke sporu, pokud předpokládáme $\nu_P(y) < 0$ a $\nu_P(x) \geq 0$. Obě valuace musí mít tedy stejné znaménko.

Ať P je takové, že valuace x i y je záporná. Pak stejnými úvahami dostáváme $2\nu_P(y) = 4\nu_P(x)$.

□

Následující dvě lemmata nám pomohou téměř určit rod F/K.

Lemma 2.3. *Označme $D = (x)_-$. Pak $\deg(D) = 2$ a $\forall k > 1$ platí $x^k \in \mathcal{L}(kD)$ a $x^{k-2}y \in \mathcal{L}(kD)$.*

Důkaz. Podle tvrzení 1.2 je $\deg(D) = [F : K(x)] = 2$.

Zjevně $x \in \mathcal{L}(D)$ a tedy $x^k \in \mathcal{L}(kD)$. Z lemmatu 2.2 plyne, že $(y)_- = 2(x)_- \Rightarrow (x^{k-1}y)_- = (k-1)(x)_- + (y)_- = (k-1)(x)_- + 2(x)_- = (k+1)(x)_-$ a z definice Riemann-Rochova prostoru $x^{k-1}y \in \mathcal{L}((k+1)D)$.

□

Lemma 2.4. *Pro každé $k > 1$ jsou prvky $1, x, y, x^2, yx, x^3, \ldots, yx^{k-2}, x^k$ lineárně nezávislé.*

Důkaz. Pokud by pro $k > 1$ neplatilo, znamenalo by to, že existují $a_i \in K$, $i \in \{0, \ldots, k\}$, $b_i \in K$, $i \in \{2, \ldots, k\}$, kde alespoň jedno a_i nebo b_i je nenulové, že platí: $a_0 + a_1x + \Sigma_{i=2}^{k}(b_iyx^{i-2} + a_ix^i) = 0$. Neboť nejvyšší mocnina x_2 v f je rovna 2, je x transcendentní nad K, a musí proto být nenulový jeden z koeficientů b_i. Pak ale $y \in K(x)$, což je spor.

□

Pro $k = 1$ v $\mathcal{L}(D)$ leží $1, x$ a jsou lineárně nezávislé, dostáváme tedy následující důsledky.

Důsledek. $\forall k \geq 1$ $\ell(kD) \geq 2k = \deg(kD)$.

Důsledek. Rod g je roven 0 nebo 1.

Důkaz. Existuje k, že $2k > g$, pak podle důsledku 1.4 platí $\ell(kD) = \deg(kD) + 1 - g$, ale zároveň $\ell(kD) \geq \deg(kD)$.

□

Než se podíváme na to, kdy je rod F/K roven nule a kdy jedné, formulujeme tvrzení, které popisuje zápornou část (x).

Tvrzení 2.5. *Nechť F/K je algebraické funkční těleso dané $y^2 = f(x)$, kde f je monický polynom stupně 4 a označme $D = (x)_-$. Pak platí, že $D = P_1 + P_2$, kde $P_1, P_2 \in \mathbb{P}_{F/K}$, $\deg(P_1) = \deg(P_2) = 1$, $\nu_{P_1}(x) = \nu_{P_2}(x) = -1$ a $\nu_{P_1}(y) = \nu_{P_2}(y) = -2$.*

Důkaz. Protože D je stupně 2, může to být buď místo stupně 2, nebo součet dvou (ne nutně různých) míst stupně 1.

Předpokládejme, že $D \in \mathbb{P}_{F/K}$, $\deg(D) = 2$. Pak $\nu_D(x) = -1$, $\nu_D(y) = -2$, $\nu_D(y \pm x^2) \geq -2$ a zároveň

$$\nu_D(y + x^2) + \nu_D(y - x^2) = \nu_D(f(x) - x^4) \geq -3,$$

z čehož plyne, že valuace alespoň jednoho z prvků $y \pm x^2$ je ≥ -1, nechť je to prvek $y - x^2$. Pak $1, x, y - x^2 \in \mathcal{L}(D) \Rightarrow \ell(D) = 3 = 2 + 1 - g \Rightarrow g = 0$. Pak ale $\ell(2D) = 5$ a prvky $1, x, y - x^2, x^2, (y - x^2)^2, x(y - x^2) \in \mathcal{L}(2D)$ jsou lineárně

závislé. Ze vztahu $(y-x^2)^2 = f(x) + x^4 - 2yx^2$ následně plyne $y \in K(x)$, což je spor.

D je tedy součtem dvou míst, předpokládejme, že $D = 2P \Rightarrow \nu_P(x) = -2$, $\nu_P(y) = -4$.

$$\nu_P(x^2 - y) + \nu_P(x^2 + y) = \nu_P(x^4 - f(x)) = \nu_P(a_1x^3 + a_2x^2 + a_3x + a_4) \geq -6$$

a zároveň $\nu_P(x^2 \pm y) \geq -4$. Ať je valuace alespoň jednoho prvku větší rovna 2, nechť je to $x^2 - y$. Pak opět $1, x, y - x^2 \in \mathcal{L}(D)$ a dojdeme ke sporu. Jediná zbývající možnost tedy je, že valuace obou prvků je rovna -3, pak ale $-(x^2-y)+(x^2+y) = 2y \in \mathcal{L}(3P)$, což je spor.

Zbývá jediná možnost, $D = P_1 + P_2$.

□

V důkazu jsme využívali prvky $y-x^2$ a $y+x^2$, které se zdají být docela důležité, podívejme se na ně podrobněji. Zkraťme značení ν_{P_i} na ν_i pro $i \in \{1,2\}$, $D = P_1 + P_2$, jako v předchozím tvrzení.

Tvrzení 2.6. *Divisor D lze vyjádřit jako $P_1 + P_2$ tak, že platí $\nu_1(y+x^2) = -2 = \nu_2(y-x^2)$, $\nu_1(y-x^2) \geq -1$ a $\nu_2(y+x^2) \geq -1$.*

Důkaz. Vytvořme tabulku.

	$x^2 - y$	$x^2 + y$
ν_1	?	?
ν_2	?	?

Víme, že $\nu_i(x^2 \pm y) \geq -2$. Naším cílem je ukázat, že v každém sloupci i řádku je právě jedna hodnota -2.

Předpokládejme, že v jednom řádku jsem obě hodnoty větší nebo rovny -1, pak $\nu_i(x^2 + y + x^2 - y) = \nu_i(x^2) \geq -1$, což je spor.

Pokud by hodnoty $\nu_1(x^2 \pm y)$ i $\nu_2(x^2 \pm y)$ byly větší než -1, pak $x^2 \pm y \in \mathcal{L}(P_1 + P_2)$, to je opět spor.

V každém řádku i sloupci tedy musí být alespoň jedna hodnota -2. Protože $\nu_i(x^2+y)+\nu_i(x^2-y) = \nu_i(x^4-f(x)) \geq -3$, nemůžou být dvě -2 v jednom řádku. Pokud by byly obě -2 v jednom sloupci, nebude v druhém žádná a to je spor.

□

Označme místo P_i, pro které platí $\nu_i(y + x^2) = -2$ jako P_+ a to druhé jako P_-.

Věta 2.7. *F/K zadané $y^2 = f(x)$, $f(x) = x^4 + a_2x^2 + a_1x + a_0$, je rodu 0 právě tehdy, když $f(x)$ má vícenásobné kořeny.*

Důkaz. „⇒“: Předpokládejme $g = 0$, pak divisor A hlavní $\iff \deg(A) = 0$. Tudíž existuje $t \in F$, že $P_+ - P_- = (t)$, $\nu_+(t) = 1$ a $\nu_-(t) = -1$.

Označme $D = (x)_-$, v $\mathcal{L}(D)$ leží prvky 1,t,t^{-1},x a platí, že $\ell(D) = 3$, což implikuje $x = a_2t + a_1 + a_0t^{-1} = (a_2t^2 + a_1t + a_0)/t = a(t)/t$. V $\mathcal{L}(2D)$ leží 1,t,t^{-1}, t^2, t^{-2}, y a $\ell(D) = 5$, což implikuje $y = b_4t^2 + b_3t + b_2 + b_1t^{-1} + b_0t^{-2} = b(t)/t^2$.

Vyjádřeme nyní $f(x)$ jako součin $(x-\alpha_i)$, kde α_i značí kořeny f, $i \in \{1,\ldots,4\}$ a dosaďme za x, y do $y^2 = \prod_{i=1,\ldots,4}(x-\alpha_i)$ vyjádření pomocí t:

$$\frac{(b(t))^2}{t^4} = \prod_{i=1}^{4}(\frac{a(t)}{t} - \alpha_i).$$

Vytkneme t, zkrátíme:

$$(b(t))^2 = \prod_{i=1}^{4}(a(t) - t\alpha_i)$$

a podíváme se na rovnici jako na rovnost dvou polynomů. Pokud by platila, musí mít oba polynomy shodné kořeny. Na levé straně máme polynom stupně 4 umocněný na druhou - má tedy celkem 8 kořenů, ale maximálně čtyři různé.

Na pravé straně pak máme 4 polynomy stupně 2, které se liší koeficienty u lineárního členu a mají společný nenulový konstantní člen. Pokud by byly α_i různé, měli bychom 4 polynomy, jejichž kořeny by byly různé a maximálně dva z nich by měly dvojnásobný kořen. To je celkem 6 různých hodnot, což je spor. Musí tedy existovat i, j, že $\alpha_i = \alpha_j$.

„$\Leftarrow$“: Předpokládejme nyní, že $f(x)$ má vícenásobné kořeny. Zároveň pořád platí předpoklad, že $f(x)$ není čtverec, může tedy mít jeden dvojnásobný kořen (a dva jednoduché), nebo jeden trojnásobný kořen (a k němu jeden jednoduchý).

Uvažujme první možnost trojnásobného kořene, tedy že $f(x) = (x-\alpha)^3(x-\beta)$. Označme $x-\alpha$ jako s, $x-\beta$ jako $s-\delta$ a také $t = y/s^2$. Tedy:

$$t^2 = \frac{s-\delta}{s},$$
$$t^2 = 1 - \frac{\delta}{s},$$

z čehož plyne, že $s \in K(t) \Rightarrow x \in K(t)$ a $y = ts^2 \Rightarrow y \in K(t)$.

Nechť $f(x) = (x-\alpha)^2(x-\beta)(x-\gamma)$. Opět zjednodušme zápis: $x-\alpha = s$, $x-\beta = s\ -\delta$, $x-\gamma = s\ -\epsilon$ a $t = \frac{y}{s(s-\delta)}$.

$$t^2 = \frac{s-\epsilon}{s-\delta},$$

z čehož plyne $s \in K(t) \Rightarrow x \in K(t)$ a $y = ts(s-\delta) \Rightarrow y \in K(t)$.

Poslední možnost je, že $f(x) = (x-\alpha)^2 g(x)$, kde $g(x)$ bude ireducibilní polynom stupně 2. Pak pomocí předchozího víme, že $K'(C)/K'$ je rodu 0, kde K' je kořenové rozšíření tělesa K, ve kterém se $g(x)$ rozkládá. A tedy i $K(C)/K$ je rodu 0.

□

Víme nyní, jak vypadá $(x)_-$ a jaký je rod a.f.t. F/K, pokud je zadáno rovnicí $y^2 = f(x)$, f monický. Stejně tak se postupuje i u „klasické“ teorie pro $y^2 = f(x)$, kde $\deg(f) = 3$, tam ale každou rovnici můžeme jednoduchou substitucí převést do tohoto tvaru. Ukážeme, že u polynomu stupně čtyři tomu tak není, a zjistíme, jak se změní dané algebraické funkční těleso. Upozorněme ještě na skutečnost, že předpoklad o monickém polynomu jsme potřebovali pouze v posledních tvrzeních 2.5, 2.6 a 2.7, všechna předchozí platí pro libovolný nečtvercový polynom.

Rozepišme se

$$y^2 = ax^4 + bx^3 + cx^2 + dx + e$$
$$ay^2 = a^2x^4 + abx^3 + acx^2 + adx + ae.$$

Položíme-li $\bar{y} = \sqrt{a}y$ a $\bar{x} = \sqrt{a}x$, dostaneme $\bar{y}^2 = f(\bar{x})$, f monický. Podmínkou ovšem je, že v tělese K leží $\sqrt{a}$. Co se stane v případě, že v K není odmocnina z vedoucího koeficientu f?

2.1 Nečtvercový vedoucí koeficient

Mějme $f(x_1) \in K[x_1]$, jeho vedoucí koeficient označme d a nechť $\sqrt{d} \notin K$. Označme $K' = K(\sqrt{d})$ a vezměme rozšíření algebraických funkčních těles $F'/K' = K'(C)/K' \supseteq K(C)/K = F/K$. Toto rozšíření je Galoisovo, neboť $K'(C) = K(\sqrt{d})(C)$ je rozkladové nadtěleso polynomu $x_1^2 - d \in K(C)[x_1]$.

V K' můžeme $y^2 = f(x)$ převést do tvaru $\bar{y}^2 = \bar{f}(\bar{x})$, kdy $\bar{f}$ bude monický. Platí proto všechna předchozí tvrzení a $(\bar{x})_- = \bar{P}_+ + \bar{P}_-$. Tato místa dávají stejnou valuaci na $\bar{x}$ i $\bar{y}$ a rozeznáme je podle valuací prvků $\bar{x}^2 - \bar{y}$ a $\bar{x}^2 + \bar{y}$. Pokud neprovedeme substituci na monický polynom, bude $(x)_-$ také součtem dvou míst, ty se ale budou lišit valuacemi na prvcích $dx^2 - \sqrt{d}y$ a $dx^2 + \sqrt{d}y$.

$\mathrm{Gal}(F'/F)$ obsahuje dva homomorfismy: identitu a σ, pro který platí $\sigma(\sqrt{d}) = -\sqrt{d}$. $\sigma(P_+)$ bude tedy nějaké místo P, pro které bude také platit $\nu_P(x) < 0$, ale $\nu_P(dx^2 - \sqrt{d}y) = -2$. Tudíž $\sigma(P_+) = P_-$. Podle věty 1.11 tedy obě místa leží nad stejným $P \in \mathbb{P}_{F/K}$. V F/K je proto $(x)_- = P$, $\deg(P) = 2$.

Znamená to, že v tomto případě tvoří grupu samotné K-racionální body křivky, bez nutnosti počítat s bodem v nekonečnu, což může výpočty zjednodušit.

Kapitola 3

Edwardsovy křivky

Na začátku této kapitoly definujeme Edwardsovy křivky a připojíme hrubý přehled, který typicky nalezneme v článcích, které Edwardsovy křivky zmiňují. V celé kapitole bude K značit těleso charakteristiky různé od dvou.

3.1 Přehled

V roce 2007 představil Edwards ve svém článku [10] novou normální formu eliptických křivek $x^2 + y^2 = a^2 + a^2x^2y^2$ a dokázal, že nad algebraicky uzavřeným tělesem lze každou eliptickou křivku vyjádřit v tomto tvaru. Pro obecné těleso to však neplatí a právě proto, aby se v obecném případě pokrylo více křivek, byla v článku [6] zobecněna tato forma na $x^2 + y^2 = a^2(1 + dx^2y^2)$.

Definice 3 (Edwardsova křivka)**.** *Mějme těleso K,* char$K \neq 2$, $d \in K \setminus \{0,1\}$. *Edwardsova křivka je dána rovnicí*

$$E_d : \; x^2 + y^2 = 1 + dx^2y^2. \tag{3.1}$$

Poznámka. Rovnice 3.1 je K-ekvivalentní s každou tvaru $\bar{x}^2 + \bar{y}^2 = a^2(1 + \bar{d}\bar{x}^2\bar{y}^2)$ takovou, že platí $d = \bar{d}a^4$. Jednoduše definujeme $\bar{x} = ax$ a $\bar{y} = ay$.

Poznámka. Ještě obecnější formou jsou twisted Edwardsovy křivky, které jsou nad K, char $K \neq 2$, dány předpisem $E_{a,d} : ax^2 + y^2 = 1 + dx^2y^2$, kde a,d jsou navzájem různé nenulové prvky K.

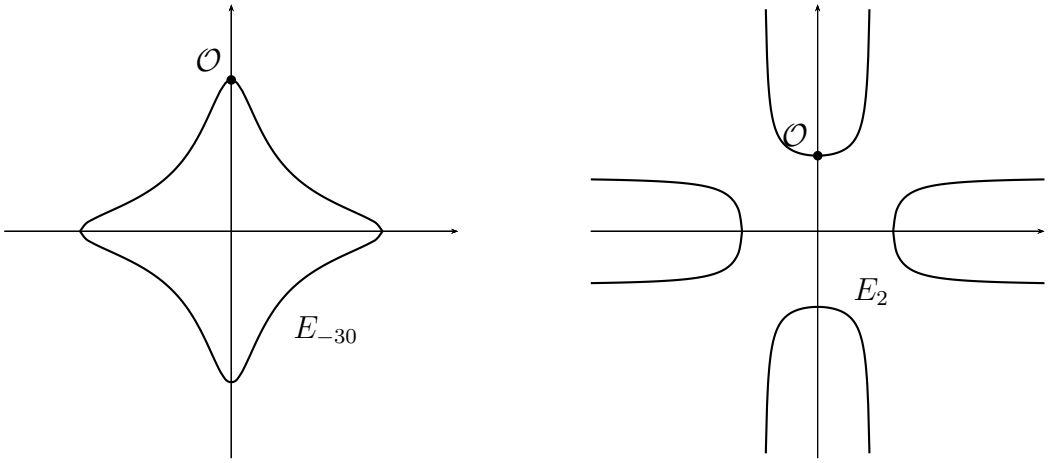
Často se dočteme, že sčítání na Edwardsově křivce je dáno následujícím vzorcem:

$$(x_1, y_1) + (x_2, y_2) = \left(\frac{x_1y_2 + y_1x_2}{1 + dx_1x_2y_1y_2}, \frac{y_1y_2 - x_1x_2}{1 - dx_1x_2y_1y_2} \right).$$

Jak uvidíme, není situace tak jednoduchá, neboť můžou existovat body, pro které není výsledek definován. Pokud ale definován je, tedy pokud $1 \pm dx_1x_2y_1y_2 \neq 0$, platí, jak je dokázáno v článku [6], že výsledný bod je opět bod křivky a že operace odpovídá klasickému sčítání na izomorfní Weierstrassově křivce.

Pokud přijmeme tuto definici sčítání, není těžké ověřit, že neutrálním prvkem sčítání je bod (0,1). Kromě něj leží na každé Edwardsově křivce také bod $(0, -1)$ řádu 2 a body$(1,0), (-1, 0)$ řádu 4. Podrobnější informace jsou v podkapitole 3.4.

Jak uvidíme, liší se vlastnosti Edwardsovy křivky v závislosti na tom, zda d je nebo není čtverec v tělese K. Pro čtenářovu představu jsou zde znázorněny Edwardsovy křivky nad $\mathbb{R}$ na obrázku 3.1a pro d záporné, tedy v $\mathbb{R}$ nečtvercové a na obrázku 3.1b pro d kladné.



(a) Křivka $x^2 + y^2 = 1 - 30x^2y^2$ (b) Křivka $x^2 + y^2 = 1 + 2x^2y^2$

Obrázek 3.1: Rozdíl mezi Edwardsovými křivkami

3.2 Biracionální ekvivalence

Lemma 3.1. *Polynom $x_1^2 + x_2^2 - 1 - dx_1^2x_2^2 \in K[x_1,x_2]$ je absolutné ireducibilní pro libovolné $d \in K \setminus \{0,1\}$.*

Důkaz. Označme tento polynom f a podívejme se na něj jako na prvek $\bar{K}(x_2)[x_1]$. Pak f je tvaru $x_1^2(1-dx_2^2)-(1-x_2^2)$. Můžeme ho vynásobit prvkem $(1-dx_2^2)^{-1} \in K(x_2)$, abychom získali monický polynom ve tvaru $x_1^2 - a$. Takový polynom se může rozkládat jen na součin $(x-\sqrt{a})(x+\sqrt{a})$, kde a je tvaru $\frac{1-x_2^2}{1-dx_2^2}$. Z $1-x_2^2 = (1-x_2)(1+x_2)$ a $1-dx_2^2 = (1-\sqrt{d}x_2)(1+\sqrt{d}x_2)$, plyne, že a není čtverec a tedy f je ireducibilní.

□

Tvrzení 3.2. *Označme $K(C)$ těleso racionálních funkcí Edwardsovy křivky dané $f = x_1^2 + x_2^2 - 1 - dx_1^2x_2^2$ a $K(\bar{C})$ těleso racionálních funkcí křivky $g = x_2^2 - (1-x_1^2)(1-dx_1^2)$. Platí, že $K(C) = K(x,y)$, kde $x = x_1 + (f)$, $y = x_2 + (f)$ a $K(\bar{C}) = K(\bar{x},\bar{y})$, kde $\bar{x} = x_1 + (g)$, $\bar{y} = x_2 + (g)$ a existuje izomorfismus $\sigma\colon K(\bar{C}) \cong K(C)$ takový, že $\sigma(\bar{x}) = x$ a $\sigma(\bar{y}) = y(1-dx^2)$.*

Důkaz. Označme ρ zobrazení z C do $\bar{C}$, $(\alpha_1,\alpha_2) \to (\alpha_1,\alpha_2(1-d\alpha_1))$. ρ_1,ρ_2 z definice racionálního zobrazení 2 jsou popořadě tvaru x a $y(1-dx^2)$ a jejich definiční obor je celé C. Každé $\alpha = (\alpha_1,\alpha_2) \in C$ zobrazíme na bod $(\alpha_1,\alpha_2(1-d\alpha_1^2))$, který leží v $\bar{C}$, pokud je splněna rovnost

$$\alpha_2^2(1-d\alpha_1^2)^2 = (1-\alpha_1^2)(1-d\alpha_1^2).$$

Postupnými úpravami dostáváme:

$$\begin{aligned}\alpha_2^2(1-d\alpha_1^2)^2 - (1-\alpha_1^2)(1-d\alpha_1^2) &= 0\\ (1-d\alpha_1^2)(\alpha_2^2(1-d\alpha_1^2) - (1-\alpha_1^2)) &= 0\\ (1-d\alpha_1^2)(\alpha_2^2+\alpha_1^2-d\alpha_1^2\alpha_2^2-1) &= 0,\end{aligned}$$

přičemž poslední rovnost platí, neboť $\alpha \in C$ a tedy $\alpha_2^2 + \alpha_1^2 = d\alpha_1^2\alpha_2^2 + 1$. ρ je racionální zobrazení a dle 1.12 je $\rho^* = \sigma$ K-homorfismus z $K(\bar{C})$ do $K(C)$ a tedy

prostý. Zároveň je σ na, neboť pro každý prvek $K(C)$ najdeme jeho vzor v $K(\bar{C})$, díky vztahu $y = \bar{y}/(1 - d\bar{x}^2)$. Tím pádem je σ izomorfismus.

□

Máme dvě biracionálně ekvivalentní křivky, a tedy dvě izomorfní funkční tělesa, z nichž jedno je dáno $x_2^2 = (1 - x_1^2)(1 - dx_1^2)$, po roznásobení $x_2^2 = dx_1^4 - x_1^2(1 + d) + 1$. Protože $d \neq 1$ z definice, určuje každá Edwardsova křivka eliptické funkční těleso, neboť $(1-x_1^2)(1-dx_1^2)$ není čtverec ani nemá vícenásobné kořeny (viz lemma 2.7).

Nabízí se říci, že pokud d není čtverec, tvoří i u Edwardsovy křivky grupu pouze K-racionální body. To je pravda, leč ne úplně zjevně a podrobnosti následují v podkapitole 3.3. Připočítáme-li k tomu fakt, že máme jeden jediný vzorec, je sčítání na Edwardsově křivce daleko jednodušší než na Weierstrassově. Horší je situace v případě, kdy d je čtverec.

Stejně jako v předchozí kapitole přibudou místa v nekonečnu. Protože ale σ^{-1} nezobrazuje prvky $K[C]$ do $K[\bar{C}]$, nezobrazuje místa v nekonečnu jedné křivky na místa v nekonečnu té druhé. Vyplývá to i z tvaru předpisu křivek, zatímco pro $y^2 = f(x)$, $\deg(f) = 4$ platí $(y)_- = 2(x)_-$, u Edwardsovy křivky je role x a y symetrická a takový vztah by u ní intuitivně platit neměl.

Biracionální ekvivalence z 3.2 je ta „nejhezčí“, x zobrazí na $\bar{x}$ a předpis pro y není složitý, a poprvé ji zmínil už Edwards ve svém prvním článku [10]. Nicméně neboť Edwardsova křivka určuje eliptické funkční těleso, budou v $K(C)/K$ dané Edwardsovou křivkou i prvky u, v takové, že $K(C)/K$ je zadáno $v^2 = u^3 + au + b$ a nejen to. Uveďme dvě důležitá tvrzení z článku [2].

Věta 3.3. *Každá twisted Edwardsova křivka $E_{a,d}$ je biracionálně ekvivalentní s Montgomeryho křivkou $E_{M,A,B}$ pro $A = 2\frac{a+d}{a-d}$ a $B = \frac{4}{a-d}$. Izomorfismus je dán $(x,y) \to \left(\frac{1+y}{1-y}, \frac{1+y}{x(1-y)}\right)$ a opačným směrem $(u,v) \to \left(\frac{u}{v}, \frac{u-1}{u+1}\right)$.*

A naopak, každá Montgomeryho křivka je biracionálně ekvivalentní twisted Edwardsově křivce $E_{a,d}$ pro $a = \frac{A+2}{B}$ a $d = \frac{A-2}{B}$.

Poznámka. Důkaz 3.3 je proveden pomocí matematického software Sage. Ten ověří že $u = (1 + y)/(1 - y)$ a $v = (1 + y)/(x(1 - y))$ splňují rovnost $Bv^2 = u^3 + Au^2 + u$ ve funkčním tělese Edwardsovy křivky $E_{a,d}$.

Poznámka. Zúžíme-li větu pouze na Edwardsovy křivky, vidíme, že si odpovídají s Montgomeryho křivkami, pro které platí $B = A + 2$.

Věta 3.4. *Nechť K je těleso charakteristiky různé od 2 a E eliptická křivka nad K daná předpisem $v^2 = u^3 + au^2 + bu$. Pak grupa $E(K)$ obsahuje bod řádu 4 právě tehdy, když E je biracionálně ekvivalentní Edwardsově křivce nad K.*

Poznámka. Označme bod řádu 4 jako (u_4, v_4), izomorfismus mezi křivkami je dán předpisem $(u,v) \to \left(\frac{v_4 u}{u_4 v}, \frac{u-u_4}{u+u_4}\right)$ a zobrazí bod (u_4, v_4) na bod (1,0). Parametr Edwardsovy křivky je $d = 1 - \frac{4u_4^3}{v_4^2}$. Důkaz věty je opět proveden pomocí ověření vztahů v softwaru Sage.

3.3 Algebraický pohled

Ponechejme v této podkapitole značení použité v 3.2 a mimo to označme $K(C)$ jako F a $K(\bar{C})$ jako $\bar{F}$.

Biracionální ekvivalence mezi $K(C)$ a $K(\bar{C})$ je velmi jednoduchá a pomůže nám při zkoumání $K(C)$. Protože ale, jak již bylo řečeno, nezobrazuje na sebe $K[C]$ a $K[\bar{C}]$, některé věci se změní. Popíšeme nyní $K(C)$ podobně jako $K(\bar{C})$ v kapitole 2.

Mohli bychom obdobným postupem dokázat, že F/K je rodu 1, tato vlastnost je ale z biracionální ekvivalence zjevná, a proto ji dokazovat nebudeme. Podívejme se místo toho rovnou na místa v nekonečnu.

Tvrzení 3.5. *Pro každé $P \in \mathbb{P}_{F/K}$ platí $\nu_P(x) < 0 \Rightarrow \nu_P(y) = 0$, $\nu_P(y) < 0 \Rightarrow \nu_P(x) = 0$ a $\nu_P(1 - dx^2) = -\nu_P(1 + dy^2)$.*

Důkaz. Postupně z rovnosti $x^2 + y^2 = 1 + dx^2y^2$ dostáváme

$$\begin{aligned} y^2 &= \frac{x^2 - 1}{dx^2 - 1} \\ y^2 &= d^{-1}\left(\frac{x^2 - d^{-1}}{x^2 - d^{-1}} + \frac{d^{-1} - 1}{x^2 - d^{-1}}\right) \\ y^2 - d^{-1} &= d^{-1}\frac{d^{-1} - 1}{x^2 - d^{-1}} \\ (y^2 - d^{-1})(x^2 - d^{-1}) &= d^{-1}(d^{-1} - 1). \end{aligned}$$

Z poslední rovnosti plyne, že $\forall P \in \mathbb{P}_{F/K}$ platí $\nu_P(y^2 - d^{-1}) = -\nu_P(x^2 - d^{-1})$ a tedy i $\nu_P(1 - dx^2) = -\nu_P(1 + dy^2)$. Připomeňme, že $\nu_P(a+b) \geq \min\{\nu_P(a), \nu_P(b)\}$, přičemž rovnost nastává vždy, když jsou $\nu_P(a)$,$\nu_P(b)$ rozdílné.

Pokud tedy máme P takové, že $\nu_P(y) < 0$, pak $\nu_P(y^2 - d^{-1}) = 2\nu_P(y) < 0 \Rightarrow \nu_P(x^2 - d^{-1}) > 0 \Rightarrow \nu_P(x) = 0$. Pro P takové, že $\nu_P(x) < 0$ provedeme totožnou úvahu.

□

Důsledek. Z tvrzení bezprostředně vyplývá, že nosiče záporných částí hlavních divisorů (x) a (y) budou disjunktní.

Tvrzení 3.6. *Nechť F/K je zadáno $x^2 + y^2 = 1 + dx^2y^2$ a d je v K čtverec. Pak $(x)_- = A_1 + A_2$ a $(y)_- = B_1 + B_2$, kde $A_1, A_2, B_1, B_2 \in \mathbb{P}^{(1)}_{F/K}$ jsou navzájem různá.*

Důkaz. Provedeme důkaz téměř stejný jako v tvrzení 2.5. Víme, že $\deg((x)_-) = 2$ a máme tedy jen tři možnosti, jak může vypadat.

Předpokládejme, že $(x)_- = P$, $\deg(P) = 2$. Pak by z vlastností valuace a tvrzení 3.5 platily následující vztahy:

$$\begin{aligned} \nu_P(x) = -1, \quad \nu_P(1 - dy^2) = 2, \\ \nu_P(1 - dx^2) = -2, \qquad \nu_P(y) = 0. \end{aligned} \tag{3.2}$$

Z rovnosti

$$\begin{aligned} &(\sqrt{d}x^2 - y(1 - dx^2))(\sqrt{d}x^2 + y(1 - dx^2)) = dx^4 - y^2(1 - dx^2)^2 = \\ &= dx^4 - (1 - dx^2)^2\frac{1 - x^2}{1 - dx^2} = x^2(d + 1) - 1 \end{aligned}$$

vyplývá, že $\nu_P(\sqrt{d}x^2 - y(1-dx^2)) + \nu_P(\sqrt{d}x^2 + y(1-dx^2)) \geq -2$, tím pádem alespoň jeden z prvků musí mít valuaci ≥ -1. Nechť je to první z nich, označme ho θ. Platí, že $\theta \in \mathcal{L}(P)$? Jinými slovy, platí, že $\forall Q \neq P\ \nu_Q(\theta) \geq 0$? Neboť $\theta \in K[C]$, záleží to jen na $\nu_Q(\theta)$ pro $Q \in \mathbb{P}_{F/K}$ takové, že $\nu_Q(y) < 0$.

Pro takové Q je $\nu_Q(\sqrt{d}x^2) = 0$ a

$$\nu_Q(y(1-dx^2)) = \nu_Q(y)+\nu_Q(1-dx^2) = \nu_Q(y)-\nu_Q(1-dy^2) = \nu_Q(y)-2\nu_Q(y) > 0.$$

Tudíž $\nu_Q(\theta) \geq 0$ a $\theta \in \mathcal{L}(P)$ společně s prvky $1, x$. Pokud by byly tyto prvky lineárně závislé, znamenalo by to $y \in K(x)$. Musí tedy být lineárně nezávislé, $\ell(P) = 3$ a zároveň $\deg(P) = 2 > 2g-1 = 1$, $\ell(P) = \deg(P)+1-g = 2+1-1 = 2$ a to je spor.

Nechť tedy $(x)_- = 2P$, $\deg(P) = 1$. Pak platí:

$$\begin{aligned} \nu_P(x) &= -2, & \nu_P(1-dy^2) &= 4, \\ \nu_P(1-dx^2) &= -4, & \nu_P(y) &= 0 \end{aligned}$$

a $\nu_P(\sqrt{d}x^2 - y(1-dx^2)) + \nu_P(\sqrt{d}x^2 + y(1-dx^2)) \geq -4$. Proto valuace alespoň jednoho prvku musí být ≥ -2, ať je to opět první z nich θ. Znovu nás zajímá, zda $\theta \in \mathcal{L}(2P)$, tedy zda $\nu_Q(\theta) \geq 0\ \forall Q \neq P$, což, jak jsme již ukázali, platí a leží zde společně s prvky $1, x$. Buďto tedy jsou $1, x, \theta$ lineárně závislé a $y \in K(x)$, což je spor, nebo $\ell(2P) = 3$ a zároveň $\deg(2P) = 2 > 2g-1 = 1$, $\ell(2P) = \deg(2P)+1-g = 2+1-1 = 2$ a to je také spor.

$(x)_-$ tedy musí být součet dvou různých míst stupně 1, označme je A_1 a A_2.

Stejně, jen s „prohozenými“ x a y dokážeme to samé pro y. Neboť $(y)_-$ má stupeň dva, máme tři možnosti, jak může vypadat. Pokud by $(y)_- = P$, $\deg P = 2$, tak

$$\begin{aligned} \nu_P(y) &= -1, & \nu_P(1-dx^2) &= 2, \\ \nu_P(1-dy^2) &= -2, & \nu_P(x) &= 0. \end{aligned}$$

Z rovnosti

$$\begin{aligned} &(\sqrt{d}y^2 - x(1-dy^2))(\sqrt{d}y^2 + x(1-dy^2)) = dy^4 - x^2(1-dy^2)^2 = \\ &= dy^4 - (1-dy^2)^2\frac{1-y^2}{1-dy^2} = y^2(d+1) - 1 \end{aligned}$$

vyplývá, že $\nu_P(\sqrt{d}y^2 - x(1-dy^2)) + \nu_P(\sqrt{d}y^2 + x(1-dy^2)) \geq -2$, tím pádem alespoň jeden z prvků musí mít valuaci ≥ -1. Nechť je to $\sqrt{d}y^2 - x(1-dy^2)$ a označme ho μ. Platí, že $\mu \in \mathcal{L}(P)$, podobně jako na začátku důkazu, neboť $\nu_Q(\theta) \geq 0\ \forall Q \neq P$. To záleží to jen na $\nu_Q(\theta)$ pro $Q \in \mathbb{P}_{F/K}$ takové, že $\nu_Q(x) < 0$.

Pro takové Q je $\nu_Q(\sqrt{d}y^2) = 0$ a

$$\nu_Q(x(1-dy^2)) = \nu_Q(x) + \nu_Q(1-dy^2) = \nu_Q(x) - 2\nu_Q(x) > 0.$$

Tudíž $\mu \in \mathcal{L}(P)$ společně s prvky $1, y$. Pokud by byly tyto prvky lineárně závislé, znamenalo by to $x \in K(y)$. Musí tedy být lineárně nezávislé, $\ell(P) = 3$ a zároveň $\deg(P) = 2 > 2g-1 = 1$, $\ell(P) = \deg(P)+1-g = 2+1-1 = 2$ a to je spor.

Další možnost je, že $(y)_- = 2P$, $\deg P = 1$, pak

$$\begin{aligned} \nu_P(y) &= -2, & \nu_P(1-dx^2) &= 4, \\ \nu_P(1-dy^2) &= -4, & \nu_P(x) &= 0 \end{aligned}$$

a $\nu_P(\sqrt{d}y^2 - x(1-dy^2)) + \nu_P(\sqrt{d}y^2 + x(1-dy^2)) \geq -4$. Tentokrát musí být valuace alespoň jednoho z prvků ≥ -2, ať je to opět první z nich označený μ. Tentokrát $\mu \in \mathcal{L}(2P)$ společně s prvky $1, y$ a buďto jsou prvky lineárně závislé, cože je ve sporu s tím, že $x \notin K(y)$, nebo jsou lineárně nezávislé, což je ve sporu s tím, že $\ell(2P) = 2$.

Musí tedy platit, že $(y)_-$ je součet dvou různých míst stupně 1, označme je B_1, B_2. To, že $\{A_1, A_2\}$ a $\{B_1, B_2\}$ jsou disjunktní je důsledkem tvrzení 3.5. □

Znamená to, že pokud d je čtverec, neplatí, že by pouze K-racionální body křivky tvořily grupu. Je k nim ještě potřeba přidat 4 „body v nekonečnu“. V takovém případě se sčítání na křivce, alespoň v afinních souřadnicích, značně komplikuje; jaké existuje řešení je popsáno v 3.4.

Pro místa A_1, A_2 platí 3.2 a po prohození x a y platí tyto vztahy pro B_1, B_2. Stejně jako v kapitole 2, máme i zde „speciální“ prvky, které byly v důkazu důležité, jen tentokrát nejsou v tak „pěkném“ tvaru. Označme je takto:

$$\begin{aligned} \rho_x^+ &= \sqrt{d}x^2 + y(1-dx^2) \\ \rho_x^- &= \sqrt{d}x^2 - y(1-dx^2) \\ \rho_y^+ &= \sqrt{d}y^2 + x(1-dy^2) \\ \rho_y^- &= \sqrt{d}y^2 - x(1-dy^2). \end{aligned}$$

Lemma 3.7. *$(x)_-$ se dá zapsat ve tvaru $A_+ + A_-$ a $(y)_-$ se dá zapsat ve tvaru $B_+ + B_-$ tak, že platí:*

	ρ_x^+	ρ_x^-	ρ_y^+	ρ_y^-
ν_{A_+}	-2	≥ 0	0	0
ν_{A_-}	≥ 0	-2	0	0
ν_{B_+}	0	0	-2	≥ 0
ν_{B_-}	0	0	≥ 0	-2

Důkaz. Ukažme nejprve nulové hodnoty:

$$\nu_{B_\pm}(\rho_x^\pm) = \nu_{B_\pm}(\underbrace{\sqrt{d}x^2}_{0} \pm \underbrace{y(1-dx^2))}_{-1+2} = 0,$$

$$\nu_{A_\pm}(\rho_y^\pm) = \nu_{A_\pm}(\underbrace{\sqrt{d}y^2}_{0} \pm \underbrace{x(1-dy^2))}_{-1+2} = 0.$$

Pro obě dvě podtabulky 2×2 obsahující -2 pak použijeme argumentaci stejnou jako v tvrzení 2.6. Pokud by byly v jednom řádku pro $A_\pm$ obě hodnoty ≥ -1, bude také $-1 \leq \nu_{A_\pm}(\rho_x^+ + \rho_x^-) = \nu_{A_\pm}(2\sqrt{d}x^2) = -2$, což je spor.

Pokud by byly ve sloupci obě hodnoty ≥ -1, bude $\rho_x^+ \in \mathcal{L}(A_+ + A_-)$ (případně ρ_x^-). $\mathcal{L}(A_+ + A_-)$ obsahuje prvky $1, x$ a $\ell(A_+ + A_-) = 2$, tím pádem by existovaly koeficienty $a_1, a_2, a_3 \in \mathbb{Z}$ takové, že $a_1 + a_2 x + a_3\sqrt{d}x^2 + a_3 y(1-dx^2)$. Jelikož x je transcendentní nad K, a_3 musí být nenulové a platilo by $y \in K(x)$, což je spor.

V každém řádku i sloupci tedy musí být alespoň jedna hodnota -2. Z

$$\nu_{A_\pm}(\rho_x^+) + \nu_{A_\pm}(\rho_x^-) = \nu_{A_\pm}(\rho_x^+ \cdot \rho_x^-) = \nu_{A_\pm}(x^2(d+1) - 1) \geq -2$$

pak vyplývá, že pokud je valuace jednoho z prvků rovna -2, valuace druhého musí být ≥ 0.

□

3.3.1 Pokud d není čtverec

Stejne jako v kapitole 2 jsme i zde nejprve předpokládali, že d je v K čtverec a našli 4 místa v nekonečnu. K rozeznání A_+ od A_- případně B_+ od B_- potřebujeme určit jejich valuace na prvcích, ve kterých se vyskytuje $\sqrt{d}$.

Pro $d \neq a^2,\ a \in K$ se opět podívejme na $K'(C)/K'$, $K' = K(\sqrt{d})$, rozšíření $K(C)/(K)$. Toto rozšíření je Galoisovo stupně 2. V $\mathrm{Gal}(K'(C)/K(C))$ leží identita a σ, které zobrazuje $\sqrt{d}$ na $-\sqrt{d}$. Podle věty 1.11 tedy A_+ a A_- leží nad stejným $A \in \mathbb{P}_{K(C)/K}$. Toto A je stupně 2 a platí $(x)_- = A$. Podobně leží v $K(C)$ místo B stupně 2 a $(y)_- = B$. Pro nečtvercové d tedy tvoří grupu bodů Edwardsovy křivky pouze K-racionální body.

Z fundamentální rovnosti 1.10 $2 = mef = 2ef$ pro toto Galoisovo rozšíření plyne $e = 1 = f$, a tedy pro $t \in K(C)$ je $\nu_A(t) = \nu_{A_+}(t) = \nu_{A_-}(t)$. Dále z rovnosti $\deg(P')[K' : K] = f(P'|P)\deg(P)$ pro každé $P \in K(C)/K$ stupně 1 dostáváme $\deg(P') = 1$ a $f(P'|P) = 2 \Rightarrow e = 1, m = 1$.

3.3.2 Prvek řádu 4

Předpokládejme opět, že d je čtverec a vraťme se k vyjádření $(y^2 - d^{-1})(x^2 - d^{-1}) = d^{-1} - 1$, přičemž označme $a(x) = (x^2 - d^{-1}) = a_1(x)a_2(x)$ a $b(y) = (y^2 - d^{-1}) = b_1(y)b_2(y)$. Polynomy a,b se díky předpokladu, že d je čtverec, rozkládají na lineární členy.

Protože

$$(y^2 - d^{-1}) = \frac{d^{-1} - 1}{(x^2 - d^{-1})},$$

tak

$$(a(x))_- = 2(A_+ + A_-) = (b(y))_+$$
$$(b(y))_- = 2(B_+ + B_-) = (a(x))_+.$$

Platí $(a_1)_- = A_+ + A_- = (a_2)_-$, z čehož plyne $(a_1)_+ + (a_2)_+ = 2(B_- + B_+)$ a jelikož a_1 není lineární násobek a_2, nemůže nastat $(a_1)_+ = (a_2)_+ = B_- + B_+$, tudíž $(a_1)_+ = 2B_-$ a $(a_2)_+ = 2B_+$. Podobnou argumentaci můžeme použít pro $b(y)$ a máme $(b_1) = 2A_- - B_+ - B_-$ a $(b_2) = 2A_+ - B_+ - B_-$. Shrňme, co jsme zjistili, s obecnějším značením pro divisory:

$$\begin{aligned} (a_1) &= 2B - A - C \\ (a_2) &= 2D - A - C \\ (b_1) &= 2A - B - D \\ (b_2) &= 2C - B - D. \end{aligned} \tag{3.3}$$

V případě, ze kterého jsme vztahy odvodily, je $A = A_+$, $B = B_+$, $C = A_-$, $D = B_-$. Jak ukážeme v následujících lemmatech, existence takových hlavních divisorů v E/K je ekvivalentní s tím, že v $\mathrm{Pic}(E/K)$ existuje prvek řádu 4.

Lemma 3.8. *Nechť E/K eliptické funkční těleso a* $\mathrm{Princ}(E/K)$ *obsahuje hlavní divisory ve tvaru 3.3 a $[A+C] \neq [B+D]$. Pak $[A-B]$ je v* $\mathrm{Pic}(E/K)$ *řádu* 4 *a* $2[A-B] = [A-C] = [C-A] = [B-D] = [D-B]$.

Důkaz. $[2A-B-D] = 0 \Rightarrow [2A] = [B+D]$ a $[2B-A-C] = 0 \Rightarrow [2B] = [A+C]$. Z rovností

$$2[A-B] = [2A-2B] = [B+D-A-C],$$
$$2[B+D-A-C] = [2B-A-C] + [2D-A-C] = 0$$

vyplývá, že $4[A-B] = 0$, tedy $[A-B]$ je řádu 4.

$2[B-D] = 0$ neboť $2B-2D = 2B-A-C-(2D-A-C)$ a to je hlavní divisor a_1/a_2. Stejně tak $2[A-C] = 0$, a tedy $[B-D] = [D-B]$ a $[A-C] = [C-A]$.

Zbývající rovnosti pak vyplývají ze vztahu $2[A-B] = [B+D-A-C] = [(2B-A-C)-B+D] = [D-B]$ a $[B+D-A-C] = [-(2A-B-D)+A-C] = [A-C]$.

□

Lemma 3.9. *Ať $[A-B] \in \mathrm{Pic}(E/K)$ je řádu* 4. *Pak existují jednoznačně určená, navzájem různá místa C, D, že $2[A-B] = [A-C]$ a $2[A-B] = [D-B]$ a v* $\mathrm{Princ}(E/K)$ *platí vztahy 3.3 pro vhodná $a_i, b_i \in F$, $i \in \{1,2\}$.*

Důkaz. Existence a jednoznačnost míst C, D plyne lemmatu 1.5. Pokud by $C = D$, pak

$$[2A-2B] = [A-C] \qquad [2A-2B] = [C-B]$$
$$[2B-A] = [C] \qquad [2A-B] = [C]$$

$$[2A-B] = [2B-A]$$
$$3[A-B] = 0$$

a to je spor, neboť řád $[A-B]$ je roven 4.

Z $2[A-B] = [A-C] = [D-B]$ triviálně dostáváme $[2A-B-D] = 0$ a $[A+C-2B] = 0 \Rightarrow [2B-A-C] = 0$.

Neboť $[A-B]$ je řádu 4, tak $2[A-C] = 0$ a tedy $[2A] = [2C]$ a podobně $[2B] = [2D]$. Proto i $[2C-B-D] = [2A-B-D] = 0$ a $[2D-A-C] = [2B-A-C] = 0$ a všechny divisory tvaru 3.3 jsou hlavní.

□

Důsledek. Každé eliptické funkční těleso dané Edwardsovou křivkou obsahuje místa stupně jedna, pro která platí vztahy 3.3, přičemž můžeme položit $A = P_{(1,0)}$, $B = P_{(0,1)}$, $C = P_{(-1,0)}$ a $D = P_{(0,-1)}$.

Důkaz. Víme, že $[P_{(1,0)} - P_{(0,1)}] \in \mathrm{Pic}(E/K)$ je řádu 4, proto můžeme označit $A = P_{(1,0)}$, $B = P_{(0,1)}$. Ze vztahů $2[A-B] = [D-B]$ a $2(1,0) = (0,-1)$ snadno odvodíme $D = P_{(0,-1)}$, zbývá tedy určit C. Pro to použijeme vztah $[C+A] = [B+B]$ plynoucí z $2[A-B] = [A-C] \Rightarrow [2A-2B] = [A-C] \Rightarrow [C+A] = [2B]$, vlastnosti $P_1 + P_2 = P_3 \Leftrightarrow [P_1 + P_2] = [P_3 + P_{(0,1)}]$ a toho, že víme, že opačný

prvek k $P_{(1,0)}$ je $P_{(-1,0)}$ a tím pádem $C = P_{(-1,0)}$.

□

Označme pro E/K zadané Edwardsovou křivkou $E[4]$ podgrupu prvků $\alpha \in \mathrm{Pic}(E/K)$, pro které platí $4\alpha = 0$. Z obecné struktury grupy bodů eliptické křivky víme, že $E[4]$ je izomorfní $\mathbb{Z}_4$ nebo $\mathbb{Z}_4 \times \mathbb{Z}_4$, neboť, jak jsme ukázali, pro Edwardsovy křivky body řádu 4 existují vždy. Navíc platí, že pokud $d \neq a^2\ \forall a \in K$, je $E[4]$ izomorfní $\mathbb{Z}_4$, v opačném případě je izomorfní $\mathbb{Z}_4 \times \mathbb{Z}_4$.

3.4 Sčítání

V krátkém přehledu na začátku kapitoly jsme psali, že sčítání na Edwardsově křivce $E_d : x^2 + y^2 = 1 + dx^2y^2$ je definováno jako

$$(x_1, y_1) + (x_2, y_2) = (x_3,y_3) = \left(\frac{x_1y_2 + y_1x_2}{1 + dx_1x_2y_1y_2}, \frac{y_1y_2 - x_1x_2}{1 - dx_1x_2y_1y_2} \right). \tag{3.4}$$

V článku [6] lze najít ověření, že pokud je výsledný bod (x_3, y_3) definován, leží opět na křivce E_d a vzorec odpovídá standardnímu sčítání na Weiersrassově křivce. Pokud d není čtverec, je (x_3, y_3) definován vždy, neboli $1 \pm dx_1x_2y_1y_2$ není nikdy rovno nule, což se dá jednoduše výpočetně ověřit, jak je v [6](Theorem 3.3). Vyplývá to ale také z podkapitoly 3.3.1, neboť pokud by jmenovatel byl roven nule, znamenalo by to, že výsledkem sčítání je nevlastní bod, který ale na křivce s nečtvercovým parametrem d žádný není.

Porovnáme-li vzorce s těmi pro Weierstrassovy křivky 1.7, vidíme, že zde při počítání $P + Q$ nemusíme rozlišovat případy $P = Q$, $P = -Q$ a $P \neq \pm Q$. Takové vzorce označujeme jako *uniformní* a jejich využití může být přínosné převážně v kryptografii jako ochrana proti útokům postranními kanály. V případě, že d není čtverec, je navíc (x_3, y_3) definován pro všechny body (jmenovatel je vždy nenulový) a takovým vzorcům říkáme *úplné*.

3.4.1 d je druhou mocninou

Případem, kdy d je čtverec a na křivce tedy leží 4 body v nekonečnu se poměrně dlouhou dobu nikdo nezabýval. Čistě formálně je možné to vyřešit pomocí izomorfismu s křivkou v jiném tvaru, například Weierstrassově.

Jak se izomorfismu vyhnout a počítat přímo na Edwardsově křivce je popsáno včetně důkazů v článku [4]. Kvůli využití ve formálním popisu algoritmu zde hlavní větu z článku uvedeme, nejprve potřebujeme ale zavést nové souřadnice.

Úplné souřadnice

Množina úplných bodů Edwardsovy křivky je množina tvaru

$$\{((X : Z),(Y : T)) \in \mathbb{P}^1 \times \mathbb{P}^1 | X^2T^2 + Y^2Z^2 = Z^2T^2 + dX^2Y^2\}.$$

Afinní body jsou vnořeny do úplných pomocí $(x_1,y_1) \to ((x_1 : 1),(y_1 : 1))$. Navíc, pokud d je čtverec, přibudou k afinním 4 další body: $((1 : \pm\sqrt{d}),(1 : 0))$ a $((1 : 0),(\pm 1/\sqrt{d} : 1))$.

Poznámka. Vidíme, že body v úplných souřadnicích mají stejnou vlastnosti, jako prvky Picardovy grupy - pokud je d čtverec, přibudou čtyři body „v nekonečnu“.

Věta 3.10. *Nechť K je těleso,* char$K \neq 2$, $d \in K \setminus \{0,1\}$ *a* $P_1 = ((X_1 : Z_1),(Y_1 : T_1))$, $P_2 = ((X_2 : Z_2),(Y_2 : T_2))$ *dva body v úplných souřadnicích křivky E_d. Označme*

$$\begin{aligned} X_3 &= X_1Y_2Z_2T_1 + X_2Y_1Z_1T_2, \\ Z_3 &= Z_1Z_2T_1T_2 + dX_1X_2Y_1Y_2, \\ Y_3 &= Y_1Y_2Z_1Z_2 - X_1X_2T_1T_2, \\ T_3 &= Z_1Z_1T_1T_2 - dX_1X_2Y_1Y_1; \end{aligned} \tag{3.5}$$

a

$$\begin{aligned} X'_3 &= X_1Y_1Z_2T_2 + X_2Y_2Z_1T_1, \\ Z'_3 &= X_1X_2T_1T_2 + Y_1Y_2Z_1Z_2, \\ Y'_3 &= X_1Y_1Z_2T_2 - X_2Y_2T_1Z_1, \\ T'_3 &= X_1Y_2Z_2T_1 - X_2Y_1Z_1T_2. \end{aligned} \tag{3.6}$$

Platí $X_3Z'_3 = X'_3Z_3$, $Y_3T'_3 = Y'_3T_3$ a nastává alespoň jedna z následujících možností:

- $(X_3,Z_3) \neq (0,0)$ *a* $(Y_3,T_3) \neq (0,0)$. *Pak* $P_1 + P_2 = ((X_3 : Z_3),(Y_3 : T_3))$.
- $(X'_3,Z'_3) \neq (0,0)$ *a* $(Y'_3,T'_3) \neq (0,0)$. *Pak* $P_1 + P_2 = ((X'_3 : Z'_3),(Y'_3 : T'_3))$.

Pokud $P_1 = P_2$ vždy nastává první případ.

Poznámka. Vzorce 3.5 jsou odvozeny z 3.4 dosazením X/Z za x a Y/T za y. Vzorce 3.6 jsou odvozeny z alternativního vzorce pro sčítání 4.3 představeného v kapitole 4.4.1.

3.5 Geometrický pohled

Sčítání na Weierstrassově křivce nad $\mathbb{R}$ se dá velmi pěkně vysvětlit geometricky. Pokusíme se nyní o to samé na Edwardsově křivce, přičemž při odvozování budeme od Weierstrassova tvaru „opisovat“.

Sčítání na Edwardsově křivce nad $\mathbb{R}$ bývá často přirovnáváno ke sčítání bodů na jednotkové kružnici pomocí úhlů. Definujeme-li $P_i = (x_i, y_i)$, $x_i = \sin\alpha_i$, $y_i = \cos\alpha_i$ a součet dvou bodů jako součet úhlů α_i, platí

$$\begin{aligned} x_3 &= \sin(\alpha_1 + \alpha_2) = \sin\alpha_1\cos\alpha_2 + \sin\alpha_2\cos\alpha_1 = x_1y_2 + x_2y_1 \\ y_3 &= \cos(\alpha_1 + \alpha_2) = \cos\alpha_1\cos\alpha_2 - \sin\alpha_2\sin\alpha_1 = y_1y_2 - x_1x_2. \end{aligned}$$

Vidíme, že to přesně odpovídá čitatelům v 3.4. Na pohled sčítání opravdu připomíná sčítání na jednotkové kružnici, ovšem přesné hodnoty úhlů tomu neodpovídají. Jak se dá sčítání doopravdy geometricky reprezentovat přiblížíme ve zbytku kapitoly.

3.5.1 Divisory

Uvažujme libovolnou Edwardsovu křivku E_d a v závislosti na d označme A a B místa v nekonečnu, jako v kapitole 3.3.1, pokud d není čtverec, případně označme $A = A_+ + A_-$ a $B = B_+ + B_-$, pokud je d čtverec, kde $A_\pm, B_\pm$ jako v lemma 3.7.

Inverzní prvek

Tvrzení 3.11. *Pro Edwardsovu křivku E_d nad K a každé $y_1 \in K \setminus \{\pm 1\}$ takové, že existuje $x_1 \in K$, pro které $(x_1,y_1) \in C(K)$, platí, že hlavní divisor prvku $\frac{y-y_1}{1-y} \in K(C)$ je roven*

$$P_{(x_1,y_1)} + P_{(-x_1,y_1)} - 2P_{(0,1)}.$$

Důkaz. Spočtěme zvlášť hlavní divisory čitatele a jmenovatele a poté využijme vztahu $(xy^{-1}) = (x) + (y^{-1}) = (x) - (y)$.

Záporná část $(y - y_1)_- = B$ a má stupeň dva, protože valuace y_1 je vždy nulová a valuace y je záporná pouze pro místo B (resp. místa B_+, B_-). Zároveň protože $y_1 \neq \pm 1$, leží na $y - y_1$ právě dva body $C(K)$: (x_1,y_1) a $(-x_1,y_1)$ a proto $(y - y_1) = P_{(x_1,y_1)} + P_{(-x_1,y_1)} - B$.

Stejně tak záporná část $(1 - y)_- = B$ a má stupeň 2. Na $1 - y$ leží z bodů křivky pouze (0,1). Vyjádřeme $y - 1$ z předpisu křivky jako $x^2\frac{(dy^2-1)}{y+1}$, valuace zlomku v bodě (0,1) je 0, x je uniformizující prvek a tedy valuace $y - 1$ je 2, z čehož plyne $(1 - y) = 2P_{(0,1)} - B$.

Využitím vztahu ze začátku důkazu dostaneme $\left(\frac{y-y_1}{1-y}\right) = P_{(x_1,y_1)} + P_{(-x_1,y_1)} - B - 2P_{(0,1)} + B = P_{(x_1,y_1)} + P_{(-x_1,y_1)} - 2P_{(0,1)}$.

□

Důsledek. Z lemmatu lemma 1.6 plyne, že prvek $P_{(-x_1,y_1)}$ je v Picardově grupě inverzním prvkem k $P_{(x_1,y_1)}$. Přeneseme-li operaci na body křivky, platí $-(x_1,y_1) = (-x_1,y_1)$.

Poznámka. Pro $y = 1$ dostaneme nulový divisor, což odpovídá tomu, že (0,1) je neutrálním prvkem, pro $y = -1$ pak divisor $2P_{(0,-1)} - 2P_{(0,1)}$, platí tedy, že $(0,-1)$ je sám sobě inverzním prvkem.

Sčítání

Tvrzení 3.12. *Nechť C je Edwardsova křivka nad K daná předpisem $x^2 + y^2 = 1 + dx^2y^2$, $d \neq a^2$ pro libovolné $a \in K$, $\alpha = (\alpha_1, \alpha_2), \beta = (\beta_1, \beta_2) \in C(K) \setminus \{(0,1),(0,-1)\}$ takové, že $\alpha \neq \pm\beta$ a $f(x,y) = (y + 1) + pxy + qx \in K(C)$, kde $p,q \in K$ jsou taková, že $f(\alpha) = 0$ a $f(\beta) = 0$. Poté platí, že hlavní divisor $\frac{(y+1)+pxy+qx}{x(1-y)}$ je roven*

$$P_\alpha + P_\beta + P_\gamma - 3P_{(0,1)}$$

pro nějaké $\gamma \in C(K)$.

Důkaz. Spočtěme opět zvlášť hlavní divisory čitatele a jmenovatele a poté využijme vztahu $(xy^{-1}) = (x) + (y^{-1}) = (x) - (y)$.

Pro jmenovatel platí $(x(1-y)) = (x)+(1-y) = P_{(0,1)}+P_{(0,-1)}-A+2P_{(0,1)}-B$, jak plyne z předchozího textu a důkazu tvrzení 3.11.

Nyní vyjádřeme hlavní divisor čitatele. Víme, že pro čitatele platí $f(\alpha) = f(\beta) = f((0,-1)) = 0$, z čehož plyne $\nu_{P_\alpha}(f(x,y)) > 0$, $\nu_{P_\beta}(f(x,y)) > 0$ a $\nu_{P_{(0,-1)}}(f(x,y)) > 0$ a tedy

$$\deg((y + 1) + pxy + qx)_+ \geq 3 \Rightarrow \deg((y + 1) + pxy + qx)_- \geq 3.$$

Valuace může být záporná jen pro místa A, B, která mají stupeň dva. Pro B vyjádřeme $(y+1)+pxy+qx$ jako $y(1+px)+qx+1$, pak

$$\nu_B(y(1+px)+qx+1) \geq \min\{\nu_B(y(1+px)),k\},$$

kde $k = \nu_B(qx+1)$ je hodnota ≥ 0 a $\nu_B(y(1+px)) = \nu_B(y)+\nu_B(1+px) \geq -1$, neboť $\nu_B(1+px) \geq 0$ stejně jako $\nu_B(qx+1)$.

Podobně $\nu_A(x(py+q)+y+1) \geq \min\{\nu_A(x(py+q),k\}$, kde $k = \nu_A(y+1)$ je hodnota ≥ 0 a $\nu_A(x(py+q)) = \nu_A(x)+\nu_A(py+q) \geq -1$.

Nejmenší hodnota $\nu_A(f(x,y))$ i $\nu_B(f(x,y))$ může být v obou případech -1, protože ale pro žádná jiná místa neplatí, že by daly pro prvek $K[C]$ zápornou valuaci a $\deg((y+1)+pxy+qx)_- \geq 3$, musí platit v obou případech rovnost, tedy $\nu_A(f(x,y)) = -1$ i $\nu_B(f(x,y)) = -1$, $((y+1)+pxy+qx)_- = A\ +B$ a $\deg((y+1)+pxy+qx)_- = 4$.

Z předchozího plyne $\deg((y+1)+pxy+qx)_+ = 4$ a nutně tedy musí existovat ještě jedno místo $P \in K(C)$ stupně 1 takové, že $((y+1)+pxy+qx)_+ = P_\alpha + P_\beta + P_{(0,-1)} + P$. Jelikož $d \neq a^2$, odpovídají všechna místa stupně 1 místům P_γ pro $\gamma \in C(K)$ a proto i $P = P_\gamma$ pro nějaké $\gamma \in C(K)$.

Potom $\left(\frac{(y+1)+pxy+qx}{x(1-y)}\right) = P_\alpha + P_\beta + P_\gamma + P_{(0,-1)} - A\ -B - 3P_{(0,1)} - P_{(0,-1)} + A\ +B = P_\alpha + P_\beta + P_\gamma - 3P_{(0,1)}$.

□

Poznámka. Při předpokladech pro α, β, p, q platí vztah $\left(\frac{(y+1)+pxy+qx}{x(1-y)}\right) = P_\alpha + P_\beta + P - 3P_{(0,1)}$ pro P místo stupně 1 i na křivce nad K, kde $d = a^2$, tam ale není zaručeno, že P bude odpovídat nějakému P_γ, neboť to může být jedno ze čtyř míst v nekonečnu.

Důsledek. Z lemmatu 1.6 plyne, že prvek $-P_\gamma$ je v Picardově grupě roven $P_\alpha + P_\beta$.

3.5.2 Sčítání v $\mathbb{R}$

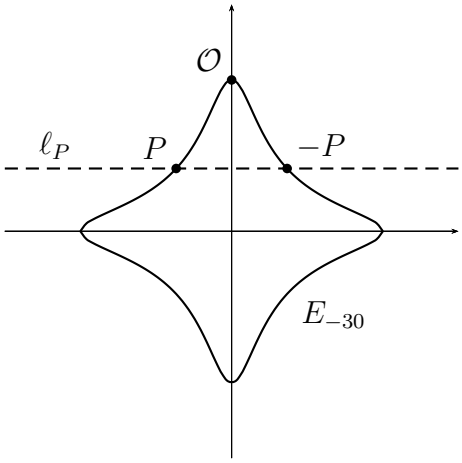
Vezměme nyní křivku nad $\mathbb{R}$ $E_{-30} : x^2 + y^2 = 1 - 30x^2y^2$. Parametr $d = -30$ tedy není čtvercem v $\mathbb{R}$ a ukažme sčítání na několika příkladech, souřadnice bodů v příkladech jsou zaokrouhlena na dvě desetinná místa.

Pokud chceme najít bod $-(x_1, y_1) = -P$, hledáme průsečíky E_{-30} s křivkou $(y-y_1)/(y-1) = 0$, což je jednoduše přímka rovnoběžná s osou x procházející bodem $(0,y_1)$. Na obrázku 3.2 je hledání inverzu znázorněno pro bod křivky $P = (-0.37, 0.42)$, přímka označená ℓ_P tedy prochází bodem $(0, 0.42)$ a inverzní bod $-P = (0.37, 0.42)$.

Pro sečtení bodů α, β hledáme průsečíky s $\frac{(y+1)+kxy+lx}{x(1-y)} = 0$. Nehleďme nyní na body, ve kterých není křivka definována a upravme ji:

$$\begin{aligned}
&(y+1)+kxy+lx = 0,\\
&y(1+kx) = -1-lx,\\
&y = -\frac{1+lx}{1+kx}.
\end{aligned}$$

Je to tedy lineární lomená funkce určená pouze dvěma parametry, které ze dvou různých bodů jednoduše dopočítáme. Na obrázku 3.3 je ukázán příklad s body

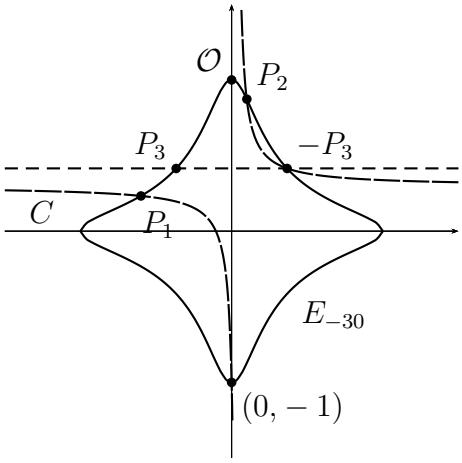


Obrázek 3.2: Inverzní prvek

$P_1 = (-0.6,0.23)$ a $P_2 = (0.1, 0.87)$. Pro zjištění parametrů k a l řešíme soustavu rovnic

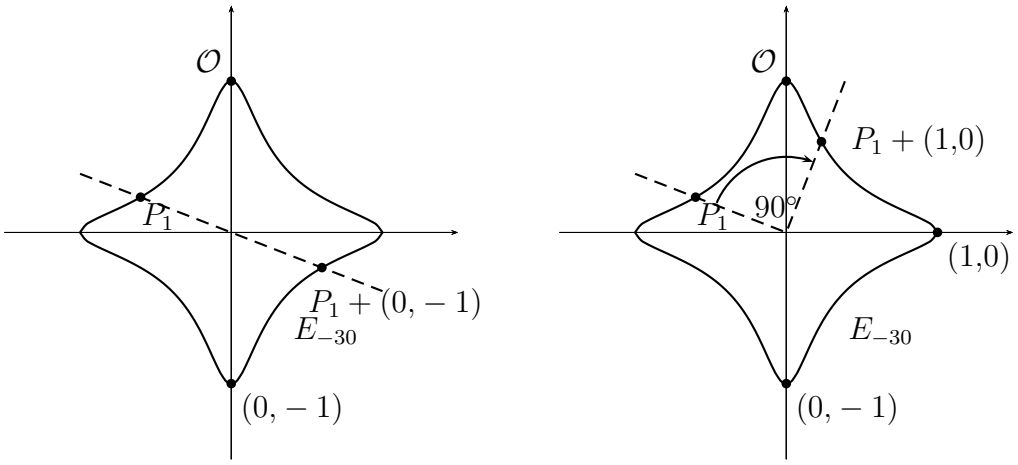
$$0.23 = -\frac{1 + l(-0.6)}{1 + k(-0.6)},$$
$$0.87 = -\frac{1 + l0.1}{1 + k0.1},$$

jejíž řešení je $k = -32.48$, $l = 9.62$. Narýsujeme tedy křivku $y = -\frac{1+9.62x}{1-32.48x}$, která protne Edwardsovu křivku v bodech $(0, -1)$, P_1, P_2 a posledním, čtvrtém bodě, který odpovídá $-P_3$, kde $P_3 = P_1 + P_2$. Souřadnice bodu $-P_3$ jsou $(0.37, 0.42)$ a výsledný bod $P_3 = (-0.37, 0.42)$ dostaneme překlopením podle osy y, jako na obrázku 3.2.



Obrázek 3.3: Součet dvou různých bodů

Výjimečné vzhledem ke sčítání jsou body $(\pm1,0)$ a $(0,\pm1)$. Ty jediné se chovají, jako kdybychom sčítali úhly (měřené od osy y po směru hodinových ručiček) na jednotkové kružnici.



(a) Chování bodu $(0,-1)$ (b) Chování bodu (1,0)

Obrázek 3.4: „Speciální“ body Edwardsovy křivky

Kapitola 4

Faktorizace metodou ECM

V této kapitole popíšeme samotný algoritmus pro faktorizaci. Pro snazší pochopení je jako první popsána v podkapitole 4.1 Pollardova $p-1$ metoda, ze které ECM vzešlo, následně je pak v podkapitole 4.2 ECM popsáno tak, jak jej uvedl ve svém článku [14] Lenstra. V podkapitole 4.3.2 je velmi podobně popsán ECM s využitím Edwardsových křivek.

Oba tyto popisy jsou však pouze formální a mohou posloužit k matematické analýze algoritmu, ale nejsou vhodné k implementaci. V kapitole 4.4.4 je stručně shrnut zásadní rozdíl mezi formálním popisem a implementací a následně jsou v podkapitole 4.4 popsány potřebné informace pro kapitolu 5.

V následujícím textu bude N značit přirozené číslo takové, že NSD(N,6) = 1 a N je dělitelné alespoň dvěma různými prvočísly, a p jednoho z jeho prvočíselných dělitelů.

Definice 4. *Číslo $m \in \mathbb{N}$ nazveme B-hladké pro $B \in \mathbb{N}$, pokud pro všechna prvočísla q dělící m platí $q < B$. m nazveme B-mocné, pokud pro každou mocninu prvočísla q^i takovou, že $q^i \mid m$ platí $q^i < B$.*

4.1 Pollardova $p-1$ metoda

Pro snazší pochopení nejprve stručně popíšeme Pollardovu $p-1$ metodu, ze které ECM vychází. Její popis je s malými úpravami převzat z knihy [9].

Metoda je postavena na malé Fermatově větě, která říká, že pro každé a nesoudělné s p platí:

$$a^{p-1} \equiv 1 \pmod{p}.$$

Stejně tak $a^M \equiv 1 \pmod{p}$, pokud $p-1|M$. Za daných předpokladů tedy platí, že $a^M - 1 = kp$ pro nějaké celé $k > 0$ a tím pádem $p \mid \text{NSD}(a^{p-1}-1, N)$. Idea Pollardovy faktorizační metody je vybrat pro výpočet $a^M - 1$ číslo M s co nejvíce děliteli tvaru $p-1$ (p prvočíslo), neboť pro každý takový dělitel bude p dělit výsledné $a^M - 1$ a při výpočtu NSD($a^{p-1}-1, N$) tak zkusíme najednou co nejvíce potenciálních dělitelů p čísla N .

Na M pak závisí úspěšnost, ale také délka běhu algoritmu, typicky je to nejmenší společný násobek všech čísel menších než volená mez B_1. Čím menší dělitele má $p-1$, tím menší mez B_1 je potřeba.

Algoritmus pak může selhat dvěma způsoby, buďto, v případě, že $p-1 \nmid M$ pro žádné p, vyjde NSD($a^{p-1}-1, N$) = 1, nebo naopak bude největší společný

dělitel roven N. Při druhé variantě může být řešením zvolit jiné a nebo zmenšit mez B_1, každopádně tato možnost nenastává příliš často.

Mnohem častější případ je, že NSD je rovno 1, pak můžeme zvolit větší mez B_1 a algoritmus opakovat nebo pokračovat takzvanou *druhou fází*. Pro tu se volí další mez, $B_2 > B_1$ a postupně se počítá $\mathrm{NSD}(a^{q_i M} - 1, N)$ pro q_i prvočísla z intervalu (B_1,B_2). Pomocí druhé fáze odkryjeme takové dělitele $p|N$, pro které $p - 1 = qu$ kde q je prvočíslo z intervalu (B_1,B_2) a $u|M$.

V praxi se ukázalo, že druhá část je často nápomocná a zároveň se dá implementovat dostatečně rychle.

Input: N, B_1, B_2
Output: netriviální dělitel nebo neúspěch

```
1  k := ∏_{prvočíslo q<B1} q^⌊log_q B1⌋;
2  zvol a ∈ (0,N) náhodně ;
3  if N > NSD(a,N) > 1 then
4  |  return NSD(a,N);
5  b = a^k − 1  mod N ; // Neboť potřebujeme hodnotu jen pro výpočet
   NSD s N, můžeme počítat modulo N
6  g := NSD(b, N) ;
7  if 1 < g < N then
8  |  return g;
9  else // Druhá fáze
10 |  for p_i prvočíslo z (B1,B2) do
11 |  |  g := NSD(b^{p_i}, N);
12 |  |  if 1 < g < N then
13 |  |  |  return g
14 |  end
15 return Neúspěch
```

Algoritmus 4.1: Pollardova $(p-1)$-metoda

Nevýhoda Pollardovy metody je v tom, že záleží přímo na prvočíselných dělitelích $p - 1$, které nemůžeme nijak ovlivnit. Pokud má $p - 1$ příliš velké prvočíselné dělitele, rozklad N touto metodou v rozumném čase nezískáme. Tento nedostatek řeší modifikace algoritmu využívající eliptické křivky - ECM.

4.2 Základní popis

V této kapitole uvedeme algoritmus tak, jak jej popsal poprvé ve svém článku [14] Lenstra, neboť je to jeden z mála matematicky korektních popisů.

Definice 5. *Nechť K je těleso, $a, b \in K$, že $4a^3 + 27b^2 \neq 0$. Pak označme projektivní eliptickou křivku danou $y^2 = x^3 + ax + b$ následovně*

$$E_{a,b}(K) = \{(x : y : z) \in \mathbb{P}^2(K) | y^2 z = x^3 + axz^2 + bz^3\}.$$

Pokud jsou a, b nebo těleso z kontextu zjevné, můžeme je vynechávat.

Definice 6. *Označme pro* $N \in \mathbb{N}$

$$V_N = \{(x : y : 1) \mid x,y \in \mathbb{Z}_N\} \cup \{\mathcal{O}\},$$

kde $\mathcal{O}$ *značí bod* $(0 : 1 : 0)$. *Pro* $P \in V_N$ *a prvočíslo* p *dělící* N *označme* P_p *bod z* $\mathbb{P}^2(\mathbb{Z}_p)$, *který obdržíme zredukováním souřadnic* P *modulo* p.

Lemma 4.1. $P_p = \mathcal{O}_p \iff P = \mathcal{O}$.

Následně Lenstra popsal algoritmus, který pro dva body P,$Q \in V_N$ buď nalezne netriviálního dělitele N, nebo vrátí bod $R \in V_N$ s následující vlastností: pokud p je dělitel N takový, že $6(4\bar{a}^3 + 27\bar{b}^2) \neq 0$ pro $\bar{a} = a \mod p$, $\bar{b} = b \mod p$ a $P_p \in E_{\bar{a},\bar{b}}(\mathbb{Z}_p)$, $Q_p \in E_{\bar{a},\bar{b}}(\mathbb{Z}_p)$, pak $R_p = Q_p + P_p$ v grupě $E_{\bar{a},\bar{b}}$.

Tento algoritmus je založen na vzorcích pro sčítání na weierstrassově křivce z tvrzení 1.2, musí ale brát v potaz, že ne vždy je v $\mathbb{Z}_N$ možné najít inverz a že pro dva body $P, Q \in V_N$ a p může platit $P \neq Q \wedge P_p = Q_p$.

Mějme dva body P,$Q \in V_N$ a definujme algoritmus „pseudosčítání“ následovně:

- pokud $P = \mathcal{O}, R := Q$; pokud $Q = \mathcal{O}, R := P$;
- pokud $P \neq \mathcal{O}, Q \neq \mathcal{O}$, označme $P = (x_1 : y_1 : 1)$ a $Q = (x_2 : y_2 : 1)$.
- Spočti $g := \mathrm{NSD}(x_1 - x_2, N)$
  - pokud $1 < g < N$ vrať g,
  - pokud $g = 1$, spočti $(x_1 - x_2)^{-1} \in \mathbb{Z}_N$ a dosaď do 1.2 s $\lambda = (y_1 - y_2)/(x_1 - x_2)^{-1}$,
  - pokud $g = N \Rightarrow x_1 = x_2$, pokračuj dalším bodem.
- Spočti $f := \mathrm{NSD}(y_1 + y_2, N)$
  - pokud $1 < f < N$ vrať f,
  - pokud $f = N$ (a tedy platí $x_1 = x_2$ a $y_2 = -y_1$) $R := \mathcal{O}$
  - pokud $f = 1$ spočti $(y_1 + y_2)^{-1} \in \mathbb{Z}_N$ a dosaď do 1.2 s $\lambda = (3x_1^2 + a)/(y_1 + y_2)^{-1}$,

Označme pro každé celé číslo $r \geq 2$ a dané v největší číslo m takové, že platí $r^m \leq v + 2\sqrt{v} + 1$, jako $e(r)$. Nyní můžeme formulovat faktorizační algoritmus 4.2.

V tvrzení 4.2 zformulujme postačující podmínky pro úspěch algoritmu.

Tvrzení 4.2. *Mějme* N, v, w *přirozená čísla větší než* 1 *a* $a, x, y \in \mathbb{Z}_N$. *Definujme* $b = y^2 - x^3 - ax \in \mathbb{Z}_N$ *a* $P = (x : y : 1) \in V_N$. *Předpokládejme, že* N *má dělitele* p *a* q *splňující následující podmínky.*

1. $p \leq v$;
2. $6(4\bar{a}^3 + 27\bar{b}^2) \neq 0$ *v* $\mathbb{Z}_p$, *kde* $\bar{a} = a \mod p$ *a* $\bar{b} = b \mod p$;
3. $|E_{\bar{a},\bar{b}}(\mathbb{Z}_p)|$ *je* w*-hladký;*
4. $6(4\widehat{a}^3 + 27\widehat{b}^2) \neq 0$ *v* $\mathbb{Z}_q$, *kde* $\widehat{a} = a \mod q$ *a* $\widehat{b} = b \mod q$;

Input: $N, v, w \in \mathbb{N} \setminus \{1\}$, $a, x, y \in \mathbb{Z}_N$
Output: netriviální dělitel nebo neúspěch

```
1 b = y^2 − x^3 − ax ∈ Z_N;
2 P = (x : y : 1);
3 k = ∏_{r=2}^{w} r^{e(r)};
4 Pokus se spočítat kP pomocí „pseudosčítání“ na V_N;
5 if předchozí krok selhal then
6 |   return dělitel ;
7 else
8 |   return neúspěch
9 end
```

Algoritmus 4.2: Lenstrova faktorizace na jedné křivce

5. $|E_{\widehat{a},\widehat{b}}(\mathbb{Z}_q)|$ *není dělitelné největším prvočíslem, které dělí řád bodu* P *na křivce* $E_{\bar{a},\bar{b}}(\mathbb{Z}_p)$

Pak algoritmus 4.2 nalezne netriviálního dělitele N.

Důkaz. Z podmínky $p \leq v$ a Hasseho věty 1.8 plyne, že $|E_{\bar{a},\bar{b}}(\mathbb{F}_p)| \leq v + 2\sqrt{v} + 1$ a tedy každý prvočíselný dělitel r řádu grupy $E_{\bar{a},\bar{b}}(\mathbb{F}_p)$ dělí řád nejvýše v $e(r)$-té mocnině. To samé musí platit i pro r dělitele řádu bodu P_p, který označme ω. Označme l onoho největšího prvočíselného dělitele ω a m jeho exponent v ω, platí $1 \leq m \leq e(l)$. Označme

$$k_0 = \left(\prod_{r=2}^{l-1} r^{e(r)} \right) l^{m-1}.$$

Platí $k_0 \not\equiv 0 \mod \omega$, ale $k_0 l \equiv 0 \mod \omega$ a tudíž $k_0 P_p \neq \mathcal{O}_p$ a $k_0 l P_p = \mathcal{O}$ v grupě $E_{\bar{a},\bar{b}}(\mathbb{Z}_p)$.

Ze třetího bodu víme, že $l < \omega$ a proto k_0 i $k_0 l$ dělí číslo k z algoritmu 4.2, tím pádem, pokud je úspěšně spočítán násobek kP, spočítal se při tom i $k_0 P$ a $k_0 l P$. Stačí nám proto dokázat, že $k_0 P$ a $k_0 l P$ nemohou být definovány naráz. Využijeme k tomu faktu, že $P_p = \mathcal{O}_p \iff 0 = \mathcal{O}$.

Pokud $k_0 l P \in V_N$ existuje, potom $(k_0 l P)_p = k_0 l P_p = \mathcal{O}_p$ v $E_{\bar{a},\bar{b}}(\mathbb{Z}_p)$ a proto $k_0 l P = \mathcal{O}$ ve V_N, ale potom také $(k_0 l P)_q = k_0 l P_q = \mathcal{O}_q$ v $E_{\bar{a},\bar{b}}(\mathbb{Z}_q)$ a z bodu 5 pak také $k_0 P_q = \mathcal{O}_q$. Proto, pokud $k_0 P \in E_{a,b}(\mathbb{Z}_N)$ je také definováno, muselo by platit $k_0 P = \mathcal{O} \Rightarrow k_0 P_p = \mathcal{O}_p$, což je spor.

□

Poznámka. Mez v může být na základě tvrzeni 4.2 chápána jako horní hranice pro dělitele čísla N. Číslo w ovlivňuje čas, který jsme ochotni strávit na jedné křivce a zároveň pravděpodobnost úspěchu, obojí (čas i pravděpodobnost) roste společně s w.

Výhodou proti Pollardovu $p-1$ algoritmu je, že můžeme při neúspěchu zvolit nová a, b a pro stejná p, q získáme nové grupy $E_{\bar{a},\bar{b}}(\mathbb{Z}_p)$, $E_{\widehat{a},\widehat{b}}(\mathbb{Z}_q)$ s novými řády. Obecně uznávanou heuristikou pro eliptické křivky je, že pro náhodně volené křivky jsou jejich řády náhodná čísla rozložena rovnoměrně v intervalu z Hasseho věty.

4.2.1 Shrnutí

Oprostíme-li se od matematické přesnosti, dá se zjednodušeně říci (a často je tak algoritmus také popsán), že pro číslo N zvolíme $a, x, y \in \mathbb{Z}_N$, dopočteme $b = y^2 - x^3 - ax \in \mathbb{Z}_N$ a počítáme pro $P = (x : y : 1)$ jeho násobek kP stejně, jako by N bylo prvočíslo a zvolené parametry určovaly křivku. Přitom jediná operace, která může v $\mathbb{Z}_N$ selhat, je hledání inverzu, čímž odhalíme netriviálního dělitele a algoritmus uspěje.

Takto popsaný algoritmus staví na specifikách sčítání pomocí vzorců 1.7. Při počítání na Edwardsově křivce máme ale, v ideálním případě, kdy d není čtverec, jen jeden vzorec a jmenovatelé v 3.4 jsou v $\mathbb{Z}_p$ vždy nenulové. Nikdy by tedy nedošlo k situaci, kdy bychom chtěli najít inverz prvku, který je nenulový modulo N, ale nulový modulo p. Při výpočtech budeme muset proto vždy kontrolovat, zda jsme již nenarazili na neutrální prvek (0,1).

4.3 ECM a Edwardsovy křivky

Použijme pro Edwardsovy křivky podobné značení jako v kapitole 4.2.

Definice 7. *Označme*

$$U_N = \{((X : Z),(Y : T)) \in \mathbb{P}^1(\mathbb{Z}_N) \times \mathbb{P}^1(\mathbb{Z}_N)\}$$

a pro $P \in U_N$ a prvočíslo p dělící N označme P_p čtveřici $((x : z),(y : t))$, kterou obdržíme zredukováním souřadnic P modulo p.

Poznámka. V tomto případě nemusí platit, že $P_p \in \mathbb{P}^1(\mathbb{Z}_p) \times \mathbb{P}^1(\mathbb{Z}_p)$, neboť se může stát, že souřadnice X i Z nebo Y i T bodu P budou dělitelné p.

Definice 8. *Nechť p je prvočíslo dělící N a $d \in \mathbb{Z}_N$, označme $d' = d \mod p$ a*

$$E_d^u(\mathbb{Z}_p) = \{((X : Z),(Y : T)) \in \mathbb{P}^1(\mathbb{Z}_p) \times \mathbb{P}^1(\mathbb{Z}_p) | X^2T^2 + Y^2Z^2 = Z^2T^2 + d'X^2Y^2\}.$$

Pokud $d' \notin \{0, 1\}$, nazýváme množinu $E_d^u(\mathbb{Z}_p)$ podkřivkou U_N.

4.3.1 Počítání modulo N

Neboť se při počítání s Edwardsovými křivkami vyhneme počítání inverzu, víme, že veškeré výpočty v $\mathbb{Z}_N$ budou definované a pomocí homomorfismu je můžeme převést do tělesa $\mathbb{Z}_p$ pro libovolného prvočíselného dělitele p.

Pro číslo $d \in \mathbb{Z}_N$ pak U_N obsahuje všechny $E_d^u(\mathbb{Z}_p)$, kde p je prvočíselný dělitel N, navíc pro ta p, pro která $(d \mod p) \notin \{0,1\}$ tvoří podmnožina $E_d^u(\mathbb{Z}_p)$ grupu, proto ji nazýváme podkřivkou.

Pokud pro $d \in \mathbb{Z}_N$ vybereme z U_N bod $P = ((X : Z),(Y : T))$ takový, že splňuje $X^2T^2 + Y^2Z^2 = Z^2T^2 + dX^2Y^2$ v $\mathbb{Z}_N$ a budeme počítat s využitím vzorců pro Edwardsovy křivky, můžeme teoreticky počítat souběžně na všech podkřivkách $E_d^u(\mathbb{Z}_p)$, pro které $P_p \in E_d^u(\mathbb{Z}_p)$. Problémem je, že úplné uniformní vzorce existují jen pro určitý typ Edwardsových křivek (a nikde není zaručeno, že všechny podkřivky budou v tomto tvaru, naopak by člověk přirozeně předpokládal, že tomu tak nebude) a pro obecné úplné vzorce z věty 3.10 se může stát, že bude na každé podkřivce potřeba použít jiný vzorec. Druhá možnost ovšem vede k nalezení netriviálního dělitele N, čehož v algoritmu (alespoň formálně) využijeme.

4.3.2 Algoritmus formálně

Popišme nyní ECM s využitím Edwardsových křivek formálně, včetně tvrzení o postačujících podmínkách velmi podobně, jako popsal algoritmus poprvé Lenstra. Základní princip algoritmu - výpočet kP pro nějaké $k \in \mathbb{N}$ a bod křivky P, jehož výsledkem je při úspěchu neutrální prvek dané křivky - zůstává stejný, liší se způsob počítání na křivce.

Ponechejme význam hodnot v, w a definici $e(r)$ jako v kapitole 4.2

Tvrzení 4.3 (Algoritmus „pseudosčítání“)**.** *Mějme $N \in \mathbb{N}$, které je součinem alespoň dvou různých prvočísel a $d \in \mathbb{Z}_N$. Pro dva body P,$Q \in U_N$ označme $P = ((X_1 : Z_1),(Y_1 : T_1))$, $Q = ((X_2 : Z_2),(Y_2 : T_2))$ a použijme vzorce z věty 3.10, pro určení hodnot X_3,Y_3,Z_3,T_3 a X_3',Y_3',Z_3',T_3'. Definujme algoritmus „pseudosčítání“ následovně:*

- *spočti X_3,Y_3,Z_3,T_3 a X_3',Y_3',Z_3',T_3',*
- *spočti $g_1 = \mathrm{NSD}(X_3,N)$ a $g_1' = \mathrm{NSD}(X_3',N)$,*
- *pokud $1 < g_1 < N$ nebo $1 < g_1' < N$, vrať dělitele a algoritmus ukonči.*
- *Pokud $(X_3 : Z_3) = (0,0)$ nebo $(Y_3 : T_3) = (0,0)$, polož $P + Q = ((X_3' : Z_3'),(Y_3' : T_3'))$ a algoritmus ukonči.*
- *Je-li $(X_3 : Z_3) \neq (0,0)$ a $(Y_3 : T_3) \neq (0,0)$ spočti $g_2 = \mathrm{NSD}(Z_3,N), g_3 = \mathrm{NSD}(Y_3,N), g_4 = \mathrm{NSD}(T_3,N)$,*
- *pokud $1 < g_i < N$ pro nějaké $i \in \{2,3,4\}$, vrať dělitele a algoritmus ukonči,*
- *a nakonec pokud $g_1 = g_2 = g_3 = g_4 = 1$, polož $P_1 + P_2 = ((X_3 : Z_3),(Y_3 : T_3))$.*

Pokud existuje p prvočíselný dělitel N takový, že $(d \mod p) \notin \{0,1\}$ a $P_p, Q_p \in E_d^u(\mathbb{Z}_p)$, tak algoritmus odhalí netriviálního dělitele čísla N, nebo vrátí bod $R \in U_N$ takový, že pro každou podkřivku $E_d^u(\mathbb{Z}_q)$ takovou, že $P_q, Q_q \in E_d^u(\mathbb{Z}_q)$ platí $R_q = P_q + Q_q$ v grupě tvořené body $E_d^u(\mathbb{Z}_q)$.

Důkaz. Potřebujeme dokázat, že pokud nebude nalezen dělitel, bude $R = ((X : Z),(Y : Z)) \in U_N$, tedy $(X : Z) \neq (0,0)$ a $(Y : T) \neq (0,0)$. Neboť dle předpokladů existuje podkřivka modulo p a P_p,$Q_p \in E_d^u(\mathbb{Z}_p)$, platí dle věty 3.10, že $(X_3 : Z_3) \neq (0,0)$ a $(Y_3 : T_3) \neq (0,0)$ v $\mathbb{Z}/p\mathbb{Z}$ nebo $(X_3' : Z_3') \neq (0,0)$ a $(Y_3' : T_3') \neq (0,0)$ v $\mathbb{Z}/p\mathbb{Z}$, tím pádem i v $\mathbb{Z}_N$.

Stejně tak přímo z věty 3.10 plyne, že výsledný bod R splňuje $R_q = P_q + Q_q$ pro každou podkřivku $E_d^u(\mathbb{Z}_q)$.

□

Poznámka. Ukončení algoritmu při nalezení neutrálního prvku na alespoň jedné, ale ne všech podkřivkách, tedy bodu (0,1) (v úplných souřadnicích $((0 : 1),(1 : 1))$) je zajištěno výpočtem g_1 a g_1'.

Formulujme tedy algoritmus pro faktorizaci pomocí eliptických křivek v Edwardsově tvaru 4.3, který se liší pouze tvarem křivky (tedy vstupními hodnotami) a algoritmem pro sčítání bodů.

Input: $N, v, w \in \mathbb{N}$, $d, x, y \in \mathbb{Z}_N \setminus \{0,1\}$ takové, že $x^2 + y^2 = 1 + dx^2y^2$
Output: netriviální dělitel nebo neúspěch

```
1  g = NSD(x, N);
2  if 1 < g < N then
3  |  return g;
4  P = ((x : 1),(y : 1));
5  k = ∏_{r=2}^{w} r^{e(r)};
6  Pokus se spočítat kP pomocí „pseudosčítání“ na U_N pro Edwardsovy
   křivky;
7  if předchozí krok selhal then
8  |  return dělitel ;
9  else
10 |  return neúspěch ;                  // Pokud výpočet úspěšně doběhl
11 end
```

Algoritmus 4.3: ECM pro Edwardsovy křivky s jediným náhodným d

Tvrzení 4.4 (Postačující podmínky)**.** *Mějme dány $N, v, w \in \mathbb{N}$ a čísla $d, x, y \in \mathbb{Z}_N \setminus \{0,1\}$ taková, že v $\mathbb{Z}_N$ platí $x^2 + y^2 = 1 + dx^2y^2$ a označme $P = ((x : 1),(y : 1))$. Předpokládejme, že N má prvočíselné dělitele p a q, které splňují následující podmínky*

1. *$p \leq v$;*
2. *$d_p \notin \{0,1\}$ pro $d_p = d \mod p$;*
3. *$|E_d^u(\mathbb{Z}_p)|$ je w-mocné ;*
4. *$d_q \notin \{0,1\}$ pro $d_q = d \mod q$;*
5. *řád $E_d^u(\mathbb{Z}_q)$ není dělitelný největším prvočíslem dělící řád P_p.*

Pak algoritmus 4.3 nalezne netriviálního dělitele N.

Důkaz. Body 2, 4 implikují, že $E_d^u(\mathbb{Z}_p)$ a $E_d^u(\mathbb{Z}_p)$ jsou podkřivky. Zároveň podmínky pro x,y implikují, že $P_p \in E_d^u(\mathbb{Z}_p)$ a již na začátku algoritmu jsou splněny podmínky pro P z tvrzení 4.3.

Stejně jako v důkazu 4.2 z $p \leq v$ a Hasseho věty plyně, že $|E_d(\mathbb{Z}_p)| \leq v + 2\sqrt{v} + 1$ a tedy že pro každé prvočíslo r bude jeho mocnina dělící $|E_d(\mathbb{Z}_p)|$ nejvýše $e(r)$ a to samé bude platit i pro řád ω bodu P na $E_d(\mathbb{Z}_p)$.

Označme opět l onoho největšího prvočíselného dělitele ω, m jeho exponent v ω a

$$k_0 = \left(\prod_{r=2}^{l-1} r^{e(r)} \right) l^{m-1}.$$

K dokončení důkazu stačí ukázat, že výpočet $k_0 lP$ nebude dokončen. Připomeňme že na Edwardsově křivce je neutrální prvek roven (0,1), budeme tedy počítat NSD x-ové souřadnice a N, drobnou komplikací je, že existuje ještě jeden bod, konkrétně $(0, -1)$ s nulovou x-ovou souřadnicí.

Po výpočtu k_0P získáme bod $P_0 = ((x_0 : z_0),(y_0 : t_0))$, který může být roven $((0 : 1),(1 : 1))$ na $E_d(\mathbb{Z}_q)$, ale vzhledem k bodu 3 není roven $((0 : 1),(1 : 1))$ na $E_d(\mathbb{Z}_q)$. Pokud je P_0 neutrální prvkem $E_d(\mathbb{Z}_q)$, bude platit $q \mid x_0$.

Pokud by P_0 byl neutrálním prvkem $E_d(\mathbb{Z}_q)$ a $(k_0P)_p = ((0 : 1),(-1,1))$, neboť by pak $p \mid x_0$. V takovém případě by však $l = 2$ a řád bodu P_q by byl buďto dělitelný dvěma, což je spor s předpoklady a nebo by samotný P_q byl neutrální prvek. V druhém případě bychom ale nalezli netriviálního dělitele při výpočtu g. Proto $p \nmid x_0$ a $1 < \mathrm{NSD}(x_0,N) < N$.

Pokud $(k_0P)_q$ není neutrálním prvkem $E_d(\mathbb{Z}_q)$, nemůže jím být vzhledem k 5 ani $(k_0lP)_q$, stejně tak nemůže být $(k_0lP)_q$ roven $((0 : 1),(-1 : 1))$, neboť pak by řád P_q byl roven $2k_0l$, což je spor s bodem 5. Zároveň je $(k_0lP)_p$ neutrálním prvek $E_d(\mathbb{Z}_p)$

Označme $k_0lP = ((x_1 : z_1),(y_1 : t_1))$, z předchozího plyne, že $p \mid x_1$, $q \nmid x_1$ a proto $1 < \mathrm{NSD}(x_0,N) < N$.

□

Z věty 4.4, stejně jako 4.2, vyplývá výhoda, kterou má metoda ECM oproti Pollardovu $p - 1$ algoritmu a to fakt, že úspěšnost algoritmu závisí na řádech podkřivek $E_d^u(\mathbb{Z}_q)$ a $E_d^u(\mathbb{Z}_p)$, přičemž ty se pro různá d mění.

4.3.3 Volba paramterů

Oproti původnímu algoritmu je u Edwardsových křivek složitější volba parametrů „křivek“. Opět platí, že je jednodušší zvolit náhodný bod (x,y) a k němu dopočítat parametr d, ne vždy se to ale povede.

V požadavcích algoritmu 4.3 stojí, že vstupem jsou $d, x, y \in \mathbb{Z}_N \setminus \{0,1\}$ takové, že $x^2 + y^2 = 1 + dx^2y^2$. Abychom je získali, zvolíme nejprve nenulové náhodné $x,y \in \mathbb{Z}_N$ a zkontrolujeme, zda $\mathrm{NSD}(x,N) = 1$ a $\mathrm{NSD}(y,N) = 1$, pokud tomu tak není, nalezli jsme netriviálního dělitele N a nemusíme pokračovat v algoritmu. Zjevně tato možnost nenastává nijak často, jinak by bylo výhodnější faktorizovat N pomocí náhodných voleb a počítání největšího společného dělitele.

Pokud jsou oba dělitele rovni jedné, položíme $d = (x^2 + y^2 - 1)/(x^2y^2)$ v $\mathbb{Z}_N$, v případě, že vyjde $d \in \{0,1\}$, hodnoty zahodíme a generujeme nové x a y. Pokud by $d \in \{0,1\}$ nastávalo příliš často, byl by tento postup nepoužitelný, nebo krajně nevýhodný. To, že k tomu ve skutečnosti nedochází téměř nikdy, jsem ověřila pouze prakticky v implementaci.

V algoritmu z kapitoly 4.3.2 platí, že buď najdeme dělitele, nebo počítáme souběžně na všech podkřivkách. To samé platí i pro Lenstrův algoritmus popsaný v kapitole 4.2. Skutečnost je ale taková, že ani jeden z algoritmů se tak neimplementuje.

Zásadním rozdílem u reálné implementace algoritmu 4.3 je, že se nedbá na správnost výpočtů, ale pouze na jejich rychlost a nevolí se dva parametry v, w. Vzorce a souřadnice i jejich kombinace pro nejrychlejší implementaci jsou popsány v kapitole 4.4, na závěr v kapitole 4.4.4 je ve stručnosti popsán (bez implementačních detailů) algoritmus tak, jak se reálně používá.

4.4 Souřadnice a vzorce pro implementaci

4.4.1 Vzorce

Neboť v některých případech není uniformita vzorců nijak významnou výhodou a také z toho důvodu, že pokud je d čtverec, nejsou vzorce 4.1 úplné, objevují se v článcích modifikace původního vzorce 3.4.

Není složité si všimnout, že pro určení Edwardsovy křivky stačí jediný bod, pokud má obě souřadnice nenulové. Pomocí jeho souřadnic si pak můžeme vyjádřit parametr d a dosadit ho do vzorce pro sčítání. Přesně to udělali autoři v článku [12] v jeho druhé části.

Poznámka. Pro twisted Edwardsovu křivku $ax^2 + y^2 = 1 + dx^2y^2$ potřebujeme body s nenulovými souřadnicemi dva, abychom vyjádřili parametr d i a. V [12] ve skutečnosti uvedli vzorce právě pro twisted Edwardsovy křivky, vzorce uvedené zde jsou převzaté z tohoto článku a upravené pro Edwardsovy křivky, tzn. za a je dosazena 1.

Připomeňme naposledy, že vzorec pro sčítání vypadá následovně

$$(x_1, y_1) + (x_2, y_2) = (x_3, y_3) = \left(\frac{x_1y_2 + y_1x_2}{1 + dx_1x_2y_1y_2}, \frac{y_1y_2 - x_1x_2}{1 - dx_1x_2y_1y_2} \right). \tag{4.1}$$

Upravit vzorec pro zdvojení je jednoduché. Dosadíme-li za $d = (x_1^2 + y_1^2 - 1)/(x_1^2y_1^2)$ do 4.1 v případě, že $(x_1, y_1) = (x_2, y_2)$, dostáváme

$$\frac{2x_1y_1}{1 + dx_1^2y_1^2} = \frac{2x_1y_1}{1 + \frac{x_1^2+y_1^2-1}{x_1^2y_1^2}x_1^2y_1^2} = \frac{2x_1y_1}{1 + x_1^2 + y_1^2 - 1}$$

a

$$\frac{y_1^2 - x_1^2}{1 - dx_1^2y_1^2} = \frac{y_1^2 - x_1^2}{1 - \frac{x_1^2+y_1^2-1}{x_1^2y_1^2}x_1^2y_1^2} = \frac{y_1^2 - x_1^2}{1 - x_1^2 - y_1^2 + 1}.$$

Upravený vzorec tedy vypadá následovně

$$2(x_1,y_1) = \left(\frac{2x_1y_1}{y_1^2 + x_1^2}, \frac{y_1^2 - x_1^2}{2 - y_1^2 - x_1^2} \right). \tag{4.2}$$

Upravování vzorce pro sčítání je daleko složitější, zkrácený postup lze najít v článku [12], každopádně i ten lze upravit do „hezkého“ tvaru:

$$(x_1, y_1) + (x_2, y_2) = (x_3,y_3) = \left(\frac{x_1y_1 + y_2x_2}{y_1y_2 + x_1x_2}, \frac{x_1y_1 - y_2x_2}{x_1y_2 - y_1x_2} \right). \tag{4.3}$$

Poznámka. Obecně můžeme říci, že všechny tři vzorce dávají stejné výsledky, liší se ale body, pro které nejsou definované.

Poznámka. Vzorec 3.6 ve větě 3.10 je rozepsán ze vzorce 4.3.

4.4.2 Souřadnice

Kromě různých vzorců se pro urychlení výpočtů používají také různé reprezentace bodů. Nejtypičtější jsou projektivní souřadnice, nejsou to ale jen ty. V této kapitole uvedeme pouze ty souřadnice, které se v implementaci reálně použijí. Více reprezentací je uvedeno v článku [3], ze kterého tato kapitola čerpá.

Projektivní souřadnice

Množina projektivních bodů E_d je množina

$$\{(X : Y : Z) \in \mathbb{P}^2 | X^2Z^2 + Y^2Z^2 = Z^4 + dX^2Y^2\}.$$

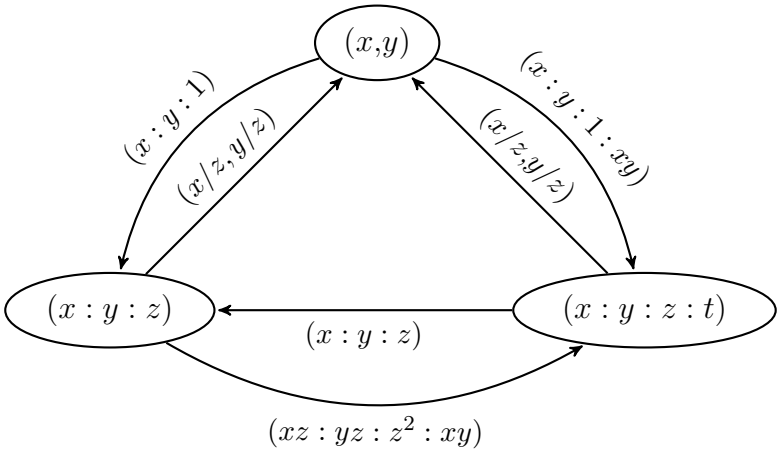
Tvoří je afinní body vnořené do $\mathbb{P}^2$ jako $(x,y) \to (x : y : 1)$ spolu se 2 body v nekonečnu $(0 : 1 : 0)$ a $(1 : 0 : 0)$. Projektivní reprezentace se u Edwardsových využívá k eliminaci zlomků ve vzorečcích.

Rozšířené souřadnice

Množina rozšířených bodů E_d je množina tvaru

$$\{(X : Y : Z : T) \in \mathbb{P}^3 | X^2 + Y^2 = Z^2 + dT^2 \wedge XY = ZT\}.$$

Jsou to afinní body (x_1,y_1) vnořené do $\mathbb{P}^3$ pomocí zobrazení $(x_1,y_1) \to (x_1 : y_1 : 1 : x_1y_1)$ a k tomu 4 body navíc, pokud d je čtverec. Přidané body jsou tvaru $(0 : \pm\sqrt{d} : 0 : 1)$ a $(\pm\sqrt{d} : 0 : 0 : 1)$.



Obrázek 4.1: Převod mezi souřadnicemi

4.4.3 Kombinace

V článku [12] navrhli kombinaci vzorců a souřadnic pro co nejrychlejší implementaci. Jejich návrh popíšeme zde, postup implementace pak v kapitole 5.3.

Označme ϵ systém projektivních souřadnic a ϵ^e systém rozšířených souřadnic, výpočet skalárního násobku na křivce, který zahrnuje počítání zdvojení a součtů, můžeme urychlit následující kombinací souřadnic:

- Pokud po zdvojení následuje opět zdvojení, počítej v projektivních souřadnicích, $\epsilon \leftarrow 2\epsilon$.
- Pokud po zdvojení následuje sčítání dvou různých bodů, postupuj takto:
  - $\epsilon^e \leftarrow 2\epsilon$ pro zdvojení následováno
  - $\epsilon \leftarrow \epsilon^e + \epsilon^e$ pro sečtení dvou různých bodů.

Okomentujme nyní jednotlivé body.

$\epsilon \leftarrow 2\epsilon$

Pro tento výpočet je použit vzorec 4.2 převedený do projektivních souřadnic, jehož odvození uvádíme zde.

Dosaďme do vzorce 4.2 X_1/Z_1 za x_1 a Y_1/Z_1 za y_1, pak

$$X_3 = \frac{2\frac{X_1Y_1}{Z_1Z_1}}{\frac{Y_1}{Z_1^2}\frac{X_1}{Z_1^2}} = \frac{2X_1Y_1}{Z_1^2}\frac{Z_1^2}{X_1^2+Y_1^2} = \frac{2X_1Y_1}{X_1^2+Y_1^2}$$
$$Y_3 = \frac{\frac{Y_1^2}{Z_1^2} - \frac{X_1^2}{Z_1^2}}{2 - \frac{Y_1^2}{Z_1^2} - \frac{X_1^2}{Z_1^2}} = \frac{Y_1^2-X_1^2}{Z_1^2}\frac{Z_1^2}{2Z_1^2-Y_1^2-X_1^2} = \frac{Y_1^2-X_1^2}{2Z_1^2-Y_1^2-X_1^2}$$

Využijme toho, že $(X : Y : Z) = (\lambda X : \lambda Y : \lambda Z)$ a vynásobme nyní vyjádřené X_3, Y_3 a Z_3, které je rovno jedné, hodnotou $\lambda = (X_1^2+Y_1^2)(2Z_1^2-Y_1^2-X_1^2)$, čímž se zbavíme zlomků. Dostaneme pak pro $2(X_1 : Y_1 : Z_1) = (X_3 : Y_3 : Z_3)$ vyjádření

$$\begin{aligned} X_3 &= 2X_1Y_1(2Z_1^2 - Y_1^2 - X_1^2), \\ Y_3 &= (Y_1^2 - X_1^2)(X_1^2 + Y_1^2), \\ Z_3 &= (X_1^2 + Y_1^2)(2Z_1^2 - Y_1^2 - X_1^2), \end{aligned} \tag{4.4}$$

$\epsilon^e \leftarrow 2\epsilon$

Pro realizaci této operace bychom mohli spočítat dvojnásobek bodu v projektivních souřadnicích ($\epsilon \leftarrow 2\epsilon$) a poté bod zobrazit do rozšířených tak, jak je na obrázku 4.1. To by moc výhodné nebylo, ale platí, že daný výpočet lze provést i bez zobrazení bodů.

Bod v rozšířených souřadnicích $(X : Y : Z : T)$, kde $Z \neq 0$ zobrazíme do afinních stejně, jako bod projektivní, tedy na $(X/Z,Y/Z)$, proto můžeme upravit vzoreček pro zdvojení stejným způsobem. Při počítání $2(X_1 : Y_1 : Z_1 : T_1) = (X_3 : Y_3 : Z_3 : T_3)$ musíme akorát ze vztahu $X_3Y_3 = Z_3T_3$ vyjádřit T_3 a dostaneme

$$\begin{aligned} X_3 &= 2X_1Y_1(2Z_1^2 - Y_1^2 - X_1^2), \\ Y_3 &= (Y_1^2 - X_1^2)(X_1^2 + Y_1^2), \\ Z_3 &= (X_1^2 + Y_1^2)(2Z_1^2 - Y_1^2 - X_1^2), \\ T_3 &= (Y_1^2 - X_1^2)2X_1Y_1. \end{aligned} \tag{4.5}$$

Všimněme si, že pro výpočet není potřeba souřadnice T_1 a můžeme proto rovnou spočítat i $2(X_1 : Y_1 : Z_1) = (X_3 : Y_3 : Z_3 : T_3)$. To je důvod, proč je navrhovaná kombinace souřadnic výhodná.

$\epsilon = \epsilon^e + \epsilon^e$

Stejně tak můžeme dosadit X/Z a Y/Z do vzorců pro sčítání 4.1 nebo 4.3. Dosadíme-li X_1/Z_1, Y_1/Z_1, X_2/Z_2, Y_2/Z_2 do 4.1 a využijeme vztahů $X_1Y_1 = Z_1T_1$ a $X_2Y_2 = Z_2T_2$, pak pro $(X_3 : Y_3 : Z_3 : T_3) = (X_1 : Y_1 : Z_1 : T_1) + (X_2 : Y_2 : Z_2 :$

T_2) dostáváme

$$\begin{aligned}
X_3 &= \frac{\frac{X_1Y_2}{Z_1Z_2} + \frac{Y_1X_2}{Z_1Z_2}}{1 + d\frac{X_1X_2Y_1Y_2}{Z_1^2Z_2^2}} = \frac{X_1Y_2 + Y_1X_2}{Z_1Z_2} \cdot \frac{Z_1^2Z_2^2}{Z_1^2Z_2^2 + dX_1X_2Y_1Y_2} = \\
&= \frac{X_1Y_2 + Y_1X_2}{1} \cdot \frac{Z_1Z_2}{Z_1^2Z_2^2 + dT_1Z_1T_2Z_2} = \frac{X_1Y_2 + Y_1X_2}{Z_1Z_2 + dT_1T_2}, \\
Y_3 &= \frac{\frac{Y_1Y_2}{Z_1Z_2} - \frac{X_1X_2}{Z_1Z_2}}{1 - d\frac{X_1X_2Y_1Y_2}{Z_1^2Z_2^2}} = \frac{Y_1Y_2 - X_1X_2}{Z_1Z_2} \cdot \frac{Z_1^2Z_2^2}{Z_1^2Z_2^2 - dX_1X_2Y_1Y_2} = \\
&= \frac{Y_1Y_2 - X_1X_2}{1} \cdot \frac{Z_1Z_2}{Z_1^2Z_2^2 - dT_1Z_1T_2Z_2} = \frac{Y_1Y_2 - X_1X_2}{Z_1Z_2 - dT_1T_2},
\end{aligned}$$

z čehož po zbavení se zlomků máme hodnoty

$$\begin{aligned}
X_3 &= (X_1Y_2 + Y_1X_2)(Z_1Z_2 - dT_1T_2), \\
Y_3 &= (Y_1Y_2 - X_1X_2)(Z_1Z_1 + dT_1T_2), \\
Z_3 &= (Z_1Z_1 - dT_1T_2)(Z_1Z_2 + dT_1T_2), \\
T_3 &= (Y_1Y_2 - X_1X_2)(X_1Y_2 + Y_1X_2).
\end{aligned} \tag{4.6}$$

V druhém případě, dosadíme-li X_1/Z_1, Y_1/Z_1, X_2/Z_2, Y_2/Z_2 do 4.3 a využijeme vztahů $X_1Y_1 = Z_1T_1$ a $X_2Y_2 = Z_2T_2$, pak pro $(X_3 : Y_3 : Z_3 : T_3) = (X_1 : Y_1 : Z_1 : T_1) + (X_2 : Y_2 : Z_2 : T_2)$ dostáváme

$$\begin{aligned}
X_3 &= \frac{\frac{X_1Y_1}{Z_1Z_1} + \frac{Y_2X_2}{Z_2Z_2}}{\frac{Y_1Y_2}{Z_1Z_2} + \frac{X_1X_2}{Z_1Z_2}} = \frac{X_1Y_1Z_2^2 + Y_2X_2Z_1^2}{Z_1^2Z_2^2} \cdot \frac{Z_1Z_2}{Y_1Y_2 + X_1X_2} = \\
&= \frac{Z_1T_1Z_2^2 + Z_2T_2Z_1^2}{Z_1Z_2} \cdot \frac{1}{Y_1Y_2 + X_1X_2} = \frac{T_1Z_2 + Z_1T_2}{Y_1Y_1 + X_1X_2}, \\
Y_3 &= \frac{\frac{X_1Y_1}{Z_1^2} - \frac{X_2Y_2}{Z_2^2}}{\frac{X_1Y_2}{Z_1Z_1} - \frac{Y_1X_2}{Y_1Y_2}} = \frac{X_1Y_1Z_2^2 - X_2Y_2Z_1^2}{Z_1^2Z_2^2} \cdot \frac{Z_1Z_2}{X_1Y_2 - Y_1X_2} = \\
&= \frac{T_1Z_2 + Z_1T_2}{X_1Y_2 - Y_1X_2},
\end{aligned}$$

z čehož po zbavení se zlomků máme hodnoty

$$\begin{aligned}
X_3 &= (T_1Z_2 + Z_1T_2)(X_1Y_2 - Y_1X_2), \\
Y_3 &= (T_1Z_2 + Z_1T_2)(Y_1Y_2 + X_1X_2), \\
Z_3 &= (X_1Y_2 - Y_1X_2)(Y_1Y_2 + X_1X_2), \\
T_3 &= (T_1Z_2 + Z_1T_2)(T_1Z_2 + Z_1T_2).
\end{aligned} \tag{4.7}$$

Tyto vzorce jsou nezávislé na d a můžou být výhodnější, pokud je d velké.

Jak konkrétně, pomocí kolika proměnných spočítat hodnoty a jak využít navrhované kombinace projektivních a rozšířených souřadnic je popsáno v kapitole 5.3.

4.4.4 Algoritmus neformálně

Shrňme zásadní rozdíly algoritmu pro implementaci ECM s Edwardsovými křivkami od formálního popisu:

- namísto dvou parametrů v, w se volí jeden parametr, který se značí B_1 a k je voleno jako nejmenší společný násobek všech čísel z množiny $\{2, \ldots, B_1\}$. V k se tedy vyskytují všechna prvočísla $p_i < B_1$, každé z nich umocněno na $\lfloor \log B_1 / \log p_i \rfloor$.
- Pro výpočet $kP = (x_k : y_k : z_k)$ se používá některý z algoritmů pro binární umocňování s využitím kombinace vzorců popsaných v 4.4.3.
- kP se spočte úplně (během výpočtu se nekontroluje, zda používáme správné vzorce) a dělitel čísla N se hledá na závěr výpočtem NSD(x_k,N).

Input: $N, B_1 \in \mathbb{N}$, $d, x, y \in \mathbb{Z}_N \setminus \{0,1\}$ takové, že $x^2 + y^2 = 1 + dx^2y^2$
Output: netriviální dělitel nebo neúspěch
1 $k := \prod_{\text{prvočísla } q<B} q^{\lfloor \log_q B \rfloor}$;
2 $P = (x : y : 1)$;
3 Spočti $kP = (x_k : y_k : z_k)$ pomocí binárního umocňování a kombinace vzorců z kapitoly 4.4.3;
4 $g = \text{NSD}(x_k, N)$;
5 **if** $1 < g < N$ **then**
6 | **return** g ;
7 **else**
8 | **return** neúspěch ;
9 **end**

Algoritmus 4.4: ECM neformálně s využitím jednoho parametru Edwardsovy křivky

Poznámka. V algoritmu 4.4 je vynechán výpočet NSD(x,N), ten bychom mohli klidně ponechat, jak ale plyne z kapitoly 4.3.3, kontroluje se ve skutečnosti již při generování náhodných parametrů.

Algoritmus 4.4 uspěje, pokud se pro výpočet $P+Q$ použijí vzorce 4.7 a N má dělitele p,q pro které platí:

- $(d \mod p) \notin \{0,1\}$,
- $(d \mod q) \notin \{0,1\}$,
- v grupě generované bodem P_p neleží body v nekonečnu,
- v grupě generované bodem P_q neleží body v nekonečnu,
- řád P_p je B_1-mocný,
- řád P_q ani $2P_q$ není B_1-mocný.

Dané podmínky vlastně říkají, že pro dané d určuje $x^2 + y^2 = 1 + dx^2y^2$ eliptickou křivku nad $\mathbb{Z}_p$ i nad $\mathbb{Z}_q$, během výpočtů si vystačím s afinními body, tudíž budou použité vzorce správně definované a výsledný bod bude neutrálním prvkem modulo p, ale nebude neutrálním prvkem, ani bodem $(0, -1)$ modulo q $\Rightarrow p \mid x_k \wedge q \nmid x_k$ a $1 <$ NSD(x_k,N) $< N$.

Jako všechny zmíněné podmínky úspěšnosti jsou i tyto postačující, ale zdaleka ne nutné.

4.5 Druhá fáze

Pro zvýšení pravděpodobnosti nalezení dělitele se často přidává druhá fáze algoritmu, která je principiálně stejná, jako u Pollardova $p-1$ algoritmu: předpokládá se, že B_1-hladkost řádu bodu P_0 nám kazí jen jeden prvočíselný dělitel, který je větší než B_1, a zvolí se druhá mez, která se typicky značí B_2. Postupně pak pro každé prvočíslo q z intervalu (B_1, B_2) spočítáme $Q' = qQ$ a zkontrolujeme, zda neodhalíme dělitele N pomocí Q'.

Vidíme, že druhá fáze se od první příliš neliší. Důvod, proč se v algoritmu používá je ten, že se dá implementovat výrazně rychleji, než fáze první. Praxe navíc potvrzuje, že přidání druhé fáze často pomůže. Velmi pěkně jsou vylepšení druhé fáze popsány například v článku [3]. Bohužel druhá fáze není v přiložením softwaru implementována a navíc by její podrobnější popis byl jen přepsáním již známých faktů, proto se jí v této práci podrobněji zabývat nebudeme.

Zmíníme jen jeden „do očí bijící“ důvod, proč je implementace výrazně rychlejší. Ačkoli se v matematickém popisu říká, že pro každé prvočíslo $q \in (B_1, B_2)$ spočteme $Q' = qQ$, bylo by takové počítání velmi neefektivní. Vezměme $q_1 < q_2$, q_1,$q_2 \in (B_1, B_2)$ a spočtěme q_1Q. Poté nám ke zjištění hodnoty q_2Q stačí spočítat $q_1Q + (q_2 - q_1)Q$. q_1Q již známe a $q_2 - q_1$ je typicky mnohem menší, než q_2, tím pádem spočteme $(q_2 - q_1)Q$ mnohem rychleji než q_2Q.

4.6 Parametry

Jako vhodné parametry (mez B_1 a počet křivek) pro faktorizaci jsou obecně považovány hodnoty, které jsou uvedené v README souboru „state-of-the-art“ implementace GMP-ECM, zde uvedené v tabulce 4.1. Při použití parametrů pro dělitele s D ciframi by měla být pravděpodobnost, že takového dělitele nenalezneme přibližně $1/e \approx 37\%$.

Počet cifer p	B_1	Křivek
15	2000	?
20	11000	86
25	50000	214
30	250000	430
35	1000000	910

Tabulka 4.1: Doporučené parametry pro ECM

Kapitola 5

Implementace

Princip metody ECM vybízí k tomu, naprogramovat ji paralelně tak, aby proběhly výpočty na všech křivkách najednou. Jak je patrné z kapitoly 4.6, potřebovali bychom spouštět desítky až stovky vláken, přičemž procesory v dnešní době nemají takovou kapacitu (typicky zvládnout 4 až 8 vláken).

Cílem této kapitoly je popsat algoritmy přiložené implementace, která využívá pro paralelizaci grafickou kartu a OpenCL.

Délka běhu algoritmu závisí na dvou věcech: za prvé musí být efektně implementována v pozadí ležící aritmetika velkých čísel, kde je největším problémem násobení modulo N. Za druhé je potřeba zvolit správnou reprezentaci bodů křivky a vzorce pro sčítání bodů, abychom minimalizovali počet aritmetických operací, přičemž tyto dva požadavky jsou na sobě nezávislé.

Podkapitola 5.1.1 se stručně věnuje programování na grafických kartách, vybranému jazyku a pomocnému kódu, který je prováděn na procesoru. Implementaci aritmetiky se pak věnuje podkapitola 5.4, zatímco volba souřadnic a další aspekty „vyšší úrovně“ jsou v podkapitole 5.3.

Názvy funkcí z kódu jsou psané `tímto fontem` a všechny funkce zmiňované v kapitolách 5.3 a 5.4 jsou ze souboru ECM_kernels.cl.

5.1 Programování grafických karet

Ruku v ruce s rozvojem počítačových her se zlepšovala také výkonnost grafických karet, které v dnešní době často převyšují svým výkonem vícejádrové procesory. Toto tvrzení bychom ale neměli nechat bez komentáře, neboť neplatí univerzálně. Procesor je univerzálním nástrojem pro všechny výpočetní úlohy, zatímco grafické karty jsou specializované na paralelizaci jednodušších operací. Ideální jsou proto pro vykonání stejného kusu kódu pro větší množství dat, což je náš případ, neboť naším cílem je počítat na mnoha eliptických křivkách zároveň.

Snaha využít grafické pro výpočetní účely se začala vyvíjet začátkem tohoto století. Jedním z významných kroků bylo roku 2006 uvedení prostředí CUDA firmou NVIDIA, jehož prostřednictvím se dá programovat pro grafické karty NVIDIA v jazyce C. Do dnešní doby je to jedna z nejvyužívanějších možností pro programování grafických karet, následovaly prostředí od firem AMD či Intel, všechny mají ale společnou jednu zásadní nevýhodu: jsou závislé na hardwaru konkrétního výrobce.

5.1.1 OpenCL

Řešením je využití OpenCL, aplikačního rozhraní, jehož cílem je umožnit psaní kódu (nejen) pro grafickou kartu bez nutnosti výběru konkrétního výrobce. To jsem si vybrala pro implementaci právě pro svoji univerzálnost a potenciál do budoucna. Přenositelnost kódu je ale zároveň nevýhodou OpenCL.

Při programování grafických karet se kód dělí na dvě části, první z nich se spouští na hostitelském systému, typicky na procesoru, a druhá část kódu, která se nazývá kernel, se spouští na grafické kartě. V případě OpenCL se kernel dá spustit na jakémkoli zařízení daného systému, které podporuje OpenCL, tedy klidně na samotném procesoru.

Při použití OpenCL je nejprve potřeba na hostitelském systému nalézt dostupné zařízení, která podporují OpenCL, vybrat z nich to(ta), na kterém se bude kernel spouštět a kernel pro toto zařízení přeložit. Veškerý kód specifický pro OpenCL (volba tzv. platformy a zařízení) je v souboru `ECM_setup.cpp` ve funkci `setupCL`.

Dále se pak připraví data, viz kapitola 5.2, přesunou se na zařízení (grafickou kartu) a zadá se příkaz ke spuštění kernelu. Program čeká, až kernel doběhne a následně zkopíruje výsledky zpět na hostitelský systém a vyhodnotí je. Zkopírování dat, spuštění kódu na kartě i překopírování dat zpátky se děje ve funkci `runKernels`.

5.1.2 Jazyky

Kód v hostitelském systému je psán v jazyce C++, zatímco kernel v jazyce OpenCL C verze 1.1. OpenCL C vychází ze standardu C99, proti C99 zahrnuje různá rozšíření i omezení, podrobnější informace o OpenCL včetně standardů a výpisu všech verzí jsou k nalezení na webové stránce Khronos group [13]. V březnu tohoto roku (2015) uvedli verzi OpenCL 2.1, ve které představili jazyk OpenCL C++ postavený na novém C++14, díky čemuž by mohla oblíbenost OpenCL dále vzrůstat.

5.2 Předvýpočty

Příprava dat pro grafickou kartu zahrnuje vygenerování prvočísel menších než zadaná mez B_1 ve funkci `getPrimes` pomocí Eratosthenova síta. V poli `primes`, které je předáno grafické kartě, jsou rovnou ukládány hodnoty p^k, kde p je prvočíslo a k největší celé číslo takové, že $p^k < B_1$.

Dále vygenerování náhodných paramterů (d,x,y) ve funkci `naivni`, přičemž pro práci s velkými čísly na procesoru je využita knihovna GMP. Při generování se postupuje tak, jak je popsáno v kapitole 4.3.3: nejprve se vygenerují náhodná $x, y < N$ pomocí náhodného generátoru GMP inicializovaného aktuálním časem, pokud $x \in \{0,1\}$ nebo $y \in \{0,1\}$, vygenerovaná x, y se zahodí a vygenerují se nová. Pokud se stane, že x nebo y je soudělné s N, je rovnou vrácen dělitel a program ukončen. A pokud nenastane ani jedna z těchto možností, je možné dopočítat parametr $d = (x^2 + y^2 - 1)/(x^2y^2)$.

Po vygenerování každé trojice parametrů (d,x,y) se tyto parametry převedou do Montgomeryho tvaru (viz kapitola 5.4.1) a následně do vlastní reprezentace

velkých čísel (viz kapitola 5.4).

5.3 Počítání na křivce

Celý princip ECM spočívá v tom, spočítat pro nějaký bod P na křivce E hodnotu $(\prod_{\text{prvočísla } q<B_1} q^{\lfloor \log_q B_1 \rfloor})P$. Jak je v psáno článku [3] část 4.3, můžeme si vybrat, zda spočítat celý násobek a poté vynásobit P jediným velkým číslem, nebo zvolit postupné násobení. Podle [3] snad veškeré implementace používají druhou možnost, kterou jsem, tak jak je popsána v algoritmu 5.1, použila i já.

Input: P bod křivky, prvočísla p_i a celá čísla k_i taková, že $p_i^{k_i} < B$
Output: $(\prod_i p_i^{k_i})P$
1 **for** i **do**
2 | $P := p_i^{k_i} P$; `// Pomocí algoritmu 5.2`
3 **end**
4 **return** P

Algoritmus 5.1: Výpočet $(\prod_i p_i^{k_i})P$, `factorize`

Nespornou výhodou postupného násobení při implementaci na grafické kartě je ten, že jednotlivé $p_i^{k_i}$ se vejdou do rozsahu nativního typu `int`.

V každém průchodu for cyklem potřebujeme spočítat pro nějaké přirozené číslo a hodnotu aP, nebo-li $\overbrace{P + \cdots + P}^{a\text{-krát}}$. Použít postupné přičítání by bylo příliš pomalé, pro vypočtení skalárního násobku v aditivní grupě lze použít libovolný z algoritmů pro rychlý výpočet mocniny v multiplikativní grupě. Vzhledem k navrhované kombinaci souřadnic z článku [12] popsané v této práci v kapitole 4.4.3, jsem zvolila algoritmus binárního umocňování od nejvíce významného bitu.

Označme pro libovolný bod P na křivce jeho projektivní reprezentaci jako P_p a jako P_r jeho rozšířenou reprezentaci, výpočet skalárního násobku je popsán v algoritmu 5.2.

Input: E_d křivka v Edwardsově tvaru, P bod křivky, $a \in \mathbb{N}$, $a = \Sigma_{i=0}^{j} a_i 2^i$
Output: aP
1 $Q = \mathcal{O}$;
2 **for** $i = j, \ldots, 0$ **do**
3 | **if** $a_j = 1$ **then**
4 | | $Q_r := 2Q_p$; `// Pomocí 5.2`
5 | | $Q_p := Q_r + P_r$; `// Pomocí 5.3`
6 | **else**
7 | | $Q_p := 2Q_p$; `// Pomocí 5.1`
8 **end**
9 **return** Q;

Algoritmus 5.2: Výpočet aP, `mulPoint`

Poznámka. Jak je vidět, při sčítání je výsledný bod v projektivních souřadnicích, proto je při použití vzorců 4.6 úplně vynechán výpočet T_3.

Označme **M** modulární násobení, **S** modulární umocňování na druhou, **A** modulární sčítání (nebo odčítání) a **D** modulární násobení parametrem křivky d. V tabulace 5.1 uvádíme porovnání výpočetní náročnosti kombinací jednotlivých vzorců a souřadnic popsaných v kapitole 4.4, údaje v tabulce jsou převzaty z webové stránky [5].

	4.1	4.3	4.2
Projektivní souřadnice	$10\mathbf{M}+\mathbf{S}+\mathbf{D}+7\mathbf{A}$	N/A	$3\mathbf{M}+4\mathbf{S}+6\mathbf{A}$
Rozšířené souřadnice	$9\mathbf{M}+\mathbf{D}+7\mathbf{A}$	$9\mathbf{M}+7\mathbf{A}$	$4\mathbf{M}+4\mathbf{S}+6\mathbf{A}$

Tabulka 5.1: Počet aritmetických operací

V tabulce 5.1 jasně vidíme, že zatímco výpočet $2P$ je výhodnější v projektivních souřadnicích, součet dvou různých bodů se vyplatí v rozšířených souřadnicích.

5.3.1 Výpočty

V kapitole 4.4 jsme sepsali vzorce v jednotlivých souřadnicích, v této kapitole ukážeme přesný postup, jak vzorce implementovat za pomocí počtu aritmetických operací z tabulky 5.1. Způsob výpočtu je převzat z webové stránky [5].

Zdvojení v projektivních souřadnicích

Zdvojení v projektivních souřadnicích 4.4 se dá spočítat pomocí $3\mathbf{M}+4\mathbf{S}+6\mathbf{A}$ následovně

$$\begin{aligned} &B=(X_1+Y_1)^2;\ C=X_1^2;\ D=Y_1^2;\ F=C+D;\ H=Z_1^2;\ J=F-2H;\\ &X_3=(B-C-D)J;\ Y_3=F(C-D);\ Z_3=FJ. \end{aligned} \tag{5.1}$$

Tato posloupnost vzorců je implementována ve funkci `projectiveDoubling`, která pro projektivní bod P vrátí $2P$ v projektivních souřadnicích.

Ověření:

$$\begin{aligned} X_3 &= (B-C-D)J = ((X_1+Y_1)^2-X_1^2-Y_1^2)(F-2H) = \\ &= (2X_1Y_1)(C+D-2Z_1^2) = 2X_1Y_1(X_1^2+Y_1^2-2Z_1^2),\\ Y_3 &= F(C-D) = (C+D)(C-D) = (X_1^2+Y_1^2)(X_1^2-Y_1^2),\\ Z_3 &= FJ = (C+D)(F-2H) = (X_1^2+Y_1^2)(X_1^2+Y_1^2-2Z_1^2). \end{aligned}$$

Zdvojení v rozšířených souřadnicích

Zdvojení v rozšířených souřadnicích 4.5 se dá spočítat pomocí $4\mathbf{M}+4\mathbf{S}+6\mathbf{A}$ takto:

$$\begin{aligned} &A=X_1^2;\ B=Y_1^2;\ C=2Z_1^2;\ E=(X_1+Y_1)^2-A-B;\\ &G=A+B;\ F=G-C;\ H=A-B;\\ &X_3=EF;\ Y_3=GH;\ T_3=EH;\ Z_3=FG. \end{aligned} \tag{5.2}$$

Toto je implementováno ve funkci `extendedDoubling`.

Ověření:

$$\begin{aligned}
X_3 = EF &= ((X_1+Y_1)^2 - A\ - B)(G-C) = 2X_1Y_1(X_1^2+Y_1^2-2Z_1^2),\\
Y_3 = GH &= (A+B)(A-B) = (X_1^2+Y_1^2)(X_1^2-Y_1^2),\\
Z_3 = FG &= (G-C)(A+B) = (A\ + B - 2Z_1^2)(X_1^2+Y_1^2) =\\
&= (X_1^2+Y_1^2-2Z_1^2)(X_1^2+Y_1^2),\\
T_3 = EH &= ((X_1+Y_1)^2 - A - B)(A-B) = 2X_1Y_1(X_1^2-Y_1^2).
\end{aligned}$$

Sčítání v rozšířených souřadnicích

A nakonec sčítání v rozšířených souřadnicích 4.6 se dá spočítat pomocí $9\mathbf{M}+7\mathbf{A}$ následovně

$$\begin{aligned}
&A\ = X_1X_2;\ B = Y_1Y_2;\ C = Z_1T_2;\ D = T_1Z_2;\ E = D+C;\\
&F = (X_1-Y_1)(X_2+Y_2)+B-A;\ G = B+A;\ H = D-C;\\
&X_3 = EF;\ Y_3 = GH;\ Z_3 = FG;\ T_3 = EH.
\end{aligned} \tag{5.3}$$

Tyto vzorce jsou implementovány v `extendedAddition`.

Ověření:

$$\begin{aligned}
X_3 = EF &= (D+C)((X_1-Y_1)(X_2+Y_2)+B-A) =\\
&= (T_1Z_2+Z_1T_2)(X_1X_2+X_1Y_2-Y_1X_2-Y_1Y_2+Y_1Y_2-X_1X_2) =\\
&= (T_1Z_2+Z_1T_2)(X_1Y_2-Y_1X_2),\\
Y_3 = GH &= (B+A)(D-C) = (X_1X_2+Y_1Y_2)(T_1Z_2-Z_1T_2,)\\
Z_3 = FG &= ((X_1-Y_1)(X_2+Y_2)+B-A)(B+A) = (X_1Y_2-Y_1X_2)\\
&(X_1X_2+Y_1Y_2),\\
T_3 = EH &= (D+C)(D-C) = (T_1Z_2+Z_1T_2)(T_1Z_2-Z_1T_2).
\end{aligned}$$

5.4 Aritmetika velkých čísel

Neboť potřebujeme počítat s velkými čísly i v kernelu a nenašla jsem žádnou dostupnou implementaci v OpenCL C, bylo potřeba si napsat vlastní.

Čísla jsou reprezentována v bázi 2^{32}, koeficienty jsou uloženy v poli délky `DELKA` tvořeném 32-bitovými integery a jsou chápána jako kladná:

```
typedef struct{ unsigned int coeff[DELKA]; } mpz;
```

Konstanta `DELKA` je definována v `ECM.h` i v `ECM_kernels.cl` a v obou souborech musí být pro správné přenesení dat definována stejně. Hodnota konstanty omezuje maximální délku faktorizovaného čísla a to tak, že N může mít maximálně $\lfloor$`DELKA`$/2\rfloor * 32$ bitů, protože se do pole musí vejít i násobek dvou čísel menších než N. Při testování byla použita hodnota 12.

Jak je patrné z kapitoly 5.3, je pro počítání na křivce potřeba sčítání, odčítání a násobení, to vše modulo N. Vzhledem k délce čísel a využití grafické karty jsou pro sčítání, odčítání a násobení nejvhodnější klasické školské algoritmy, důležité je správně zvolit algoritmus pro počítání modulo. Stejně jako v implementaci z článku [7], jsem i já k tomu využila tzv. Montgomeryho reprezentaci popsanou v článku [15].

5.4.1 Montgomeryho reprezentace

Zvolme celé číslo $r > N$, takové, že NSD(N,r) = 1 a spočtěme (například pomocí rozšířeného Euklidova algoritmu) $0 < N' < r$ a $0 < r' < N$ takové, že $rr' - NN' = 1$. Pro číslo $a \in \mathbb{Z}_N$ budeme jeho Montgomeryho reprezentaci značit $\overline{a}$ a ta je definována jako $\overline{a} = ar \mod N$. Jelikož jsou r a N nesoudělné, je zobrazení $a \to ar \mod N$ bijekce.

Sčítání a odčítání v Montgomeryho reprezentaci se nijak neliší od klasického sčítání modulo, neboť

$$\overline{a} \pm \overline{b} = ar \pm br = (a \pm b)r = \overline{a \pm b}.$$

Pro násobení to ovšem neplatí:

$$\overline{a}\overline{b} = arbr = \overline{abr} \mod N.$$

Pro správný výsledek bychom potřebovali spočítat $\overline{ab}r^{-1} \mod N$, čehož se dá docílit bez nutnosti invertovat r, ale hlavně bez nutnosti počítání samotné modulo N, jak je popsáno v algoritmu 5.3.

Input: $\overline{a}, \overline{b} \in \mathbb{Z}_N$ v Montgomeryho reprezentaci, r, N, N' jako v 5.4.1
Output: $\overline{ab} \in \mathbb{Z}_N$

```
1 T := ā b̄ ;                                   // Pomocí algoritmu 5.8
2 m := ((T mod r)N') mod r;
3 t := (T + mN)/r;
4 if t ≥ N then
5 |  return t − N ;                             // Pomocí algoritmu 5.5
6 else
7 |  return t;
```

Algoritmus 5.3: Montgomeryho modulární násobení, `mpz_Monmul`

Okomentujme algoritmus 5.3 stejně jako v [15]. Platí, že $T + mN = T + TN'N = T - T \mod r$ a tedy $T + mN$ je dělitelné r a t je celé číslo. Dále $tr = T \mod N$, z čehož plyne $t = Tr^{-1} \mod N$, tedy $t = \overline{ab}r^{-1}$. A nakonec $0 \leq T + mN < rN + rN$ a po vydělení r platí $0 \leq t < 2N$.

Výpočet modulo N jsme převedli na počítání modulo a dělení číslem r, které si můžeme při dodržení jistých kritérií volit. Typicky se r volí jako mocnina dvojky, čímž se z dělení stane bitový posun a z modulení bitový AND. V implementaci se r volí jako 2^{32s} pro s nejmenší takové, že $2^{32s} > N$. Spočítat $\mod r$ pak v dané reprezentaci znamená vynulovat prvky pole s indexem s a vyšším, zatímco dělení r znamená „zapomenout“ prvky pole s indexy 0 až $s - 1$.

5.4.2 Algoritmy

Jak již bylo zmíněno, pro sčítání je použit klasický školský algoritmus popsaný v pseudokódu 5.4. Pokud do něj vstupují dvě celá čísla $a, b \in [0,N-1]$, výsledek $c = a + b$ je z intervalu $[0, 2N - 1]$, po každém sčítání je tedy potřeba spočítat modulo N. Protože ale víme, že $c < 2N$, stačí zkontrolovat, zda $c > N$ a pokud ano, jednou od c odečíst N. Přesně to dělá funkce `mpz_onemod` 5.7.

Input: $a = \Sigma_{i=0}^{n-1} a_i B^i$, $b = \Sigma_{i=0}^{n-1} b_i B^i$
Output: $a + b$

```
1  p := 0;
2  for i = 0, ..., n − 1 do
3  |   c_i = a_i + b_i + p;
4  |   if c_i > B then
5  |   |   c_i := c_i − B;
6  |   |   p := 1;
7  |   else
8  |   |   p := 0;
9  end
10 c_n := p;
11 return Σ_{i=0}^{n} c_i B^i
```

Algoritmus 5.4: Školské sčítání, `mpz_add`

Input: $a = \Sigma_{i=0}^{n-1} a_i B^i$, $b = \Sigma_{i=0}^{n-1} b_i B^i$, $a > b$
Output: $a - b$

```
1  p := 0;
2  for i = 0, ..., n − 1 do
3  |   c_i = a_i − b_i − p;
4  |   if c_i < 0 then
5  |   |   c_i := c_i + B;
6  |   |   p := 1;
7  |   else
8  |   |   p := 0;
9  end
10 return Σ_{i=0}^{n−1} c_i B^i
```

Algoritmus 5.5: Školské odčítání, `mpz_sub_only`

Pro algoritmus 5.7 ovšem potřebujeme umět odčítat. Algoritmus odčítání je v principu stejný, jako algoritmus sčítání, problém však nastává při odčítání většího čísla od menšího ve zvolené reprezentaci, která je bez-znaménková. Proto jsou v kódu funkce pro odčítání dvě. Funkce `mpz_sub_only` 5.5 počítá s tím, že menšenec je větší, než menšitel a používá se právě v případě, kdy chceme modulit číslo $< 2N$. Funkce `mpz_sub` 5.6 pak předpokládá, že vstupem jsou a,$b \in \mathbb{Z}_N$ a vrátí hodnotu $(a - b) \in \mathbb{Z}_N$.

Poslední a nejsložitější je násobení, pro které je opět použit tradiční školský algoritmus 5.8.

Poznámka. Kromě násobení se běžně implementuje zvlášť umocnění na druhou, tedy násobek čísla se sebou samým, neboť se využívá vztahu $a_i b_j = a_j b_i$. Pokoušela jsem se implementovat umocňování jako samostatnou funkci také, ale na grafické kartě to vedlo ke zhoršení času programu.

Právě implementace aritmetiky je, podle mého názoru, největší slabinou programu. Jak se píše v článku [19], zatímco využití chytrých algoritmů může urychlit výpočet o 10% až 20%, za cenu i několika měsíců, až dvojnásobného zrychlení lze docílit během pár dní přepsáním některých rutin do assembleru. V OpenCL

```
   Input: N; a,b ∈ {[0, ..., N − 1}
   Output: a − b  mod N
1  if a < b then
2  |  a := a + N;
3  return a − b ;                                  // Využij algoritmus 5.5
```

Algoritmus 5.6: Funkce `mpz_sub`

```
   Input: N; a ∈ {[0, ..., 2N − 1}
   Output: a  mod N
1  if a > N then
2  |  return a − N ;                                // Využij algoritmus 5.5
3  else
4  |  return a;
```

Algoritmus 5.7: Funkce `one_mod`

ovšem možnost využití assembleru chybí, právě pro jeho přenositelnost.

Kromě toho i implementace algoritmů není pravděpodobně ideální, neboť ne dokonale rozumím programování na grafických kartách a některé postupy (způsob využití pointerů apod.), které kód na procesoru zrychlují, vedly na grafické kartě naopak ke zpomalení.

5.5 Měření

S využitím přiložené implementace jsem provedla několik testů. K testování jsem použila svůj notebook s grafickou kartou NVIDIA GeForce GT 740M, na které jsou k dispozici 2 tzv. SMX bloky po 194 CUDA jádrech (teoreticky je tedy možné spustit paralelně až 388 vláken, v praxi je to poněkud složitější) a podpora OpenCL 1.1 CUDA a procesorem Intel(R) Core(TM) i5-4200M s podporou OpenCL 1.2, který má 2 jádra na kterých umí spustit celkem 4 vlákna. Kromě porovnání času programu na kartě a procesoru jsem také porovnávala výsledky s časem dosud nejuznávanější, avšak čistě sériové, implementace GMP-ECM 6.4.4, která je ke stažení na stránce [11]. GMP-ECM, jak z názvu vyplývá, využívá pro aritmetiku velkých čísel známou a velmi dobře optimalizovanou knihovnu GMP.

K testování jsem vždy zadala k rozložení 20 stejných čísel, jejichž decimální délka je 48 až 50 cifer a které jsou součinem dvou přibližně stejně velkých prvočísel, konkrétně jsou čísla vypsány v příloze v tabulce 1 s rozklady v tabulce 2. V přiložené implementaci byla konstanta `DELKA` = 12 a vždy použity ty samé křivky (náhodný generátor knovny `gmp` byl vždy inicializován 0), program GMP-ECM byl spuštěn bez druhé fáze. Čas implementace je měřen přímo v kódu pomocí funkce OpenCL (běh kernelu a kopírování dat), případně pomocí `time.h`, čas běhu GMP-ECM vypisuje sám program.

Časy pro jednotlivá čísla jsem pak zprůměrovala a výsledky pro parametr $B_1 = 5000$ jsou v tabulce 5.2, zatímco výsledky pro $B_1 = 20000$ jsou v tabulce 5.3, názornější reprezentace dat v grafu je v příloze v grafu 3 a 4.

Z dat lze vyčíst hned několik věcí. Zaprvé, že ačkoli by tomu tak teoreticky být nemělo, i u grafické karty svým způsobem platí, že čím více vláken, tím více

Input: $a = \Sigma_{i=0}^{n-1} a_i B^i$, $b = \Sigma_{i=0}^{n-1} b_i B^i$
Output: ab

```
1  for i = 0, ..., n − 1 do
2      carry = 0;
3      for j = 0, ..., n − 1 do
4          c_{i+j−1} := c_{i+j−1} + carry + a_j b_i;
5          carry := c_{i+j−1} ÷ B;
6          c_{i+j−1} := c_{i+j−1} mod B;
7      end
8      c_{i+n} := c_{i+n} + carry;
9  end
10 return Σ_{i=0}^{2n−1} c_i B^i
```

Algoritmus 5.8: Školské násobení, `mpz_mul`

	128	256	384	512	214
CPU	1.51098	3.01152	4.44690	6.26486	2.82811
GPU	1.57824	1.97588	2.07762	2.12961	1.90169
GMP-ECM	1.48305	3.01855	4.45985	6.3890	2.98995

Tabulka 5.2: Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru GMP-ECM pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 5000$ a konstantu `DELKA` rovnou 12

potřebujeme času. Drobný časový nárůst může být způsoben větším množstvím kolizí při přístupu do sdílené paměti nebo větší režií spojenou se synchronizací vláken, pravý důvod bohužel neznám. Z tabulek je také vidět, že grafická karta na jedné křivce výrazně zaostává, nicméně i tak se již u 256 křivek projeví síla paralelizace. Při optimalizaci aritmetiky velkých čísel by se hranice posunula pravděpodobně ještě níže.

V tabulce 5.3 si pak můžeme všimnout, že přiložená implementace je na CPU rychlejší, až na případ 512 křivek, zde se projevuje jistá výhoda sériového zpracování křivek. Pro $B_1 = 20000$ se některá čísla rozloží a GMP-ECM v tom momentě skončí, pokud jíž známe dělitele, není třeba zkoušet zbylé křivky, ve skutečnosti tedy nezkusila při každém běhu všech 512 křivek.

Kromě kratšího času má využití grafické karty ještě jednu významnou výhodu: nezatěžuje procesor, který může být využit k jiným účelům. Při měření pro tabulku 5.2 program při spuštění na procesoru zabíral všechen dostupný výkon (okolo 90%), pokud by ve skutečnosti mohl zabírat například jen 2 vlákna, čas by se ještě zdvojnásobil. GMP-ECM pak vzhledem ke své implementaci plně zaměstnávalo pouze jedno vlákno procesoru.

	128	256	384	512	214
CPU	5.97019	11.93622	17.87136	24.35554	10.04580
GPU	6.14315	7.65652	8.03347	8.19564	7.38011
GMP-ECM	6.3944	13.0104	19.1306	21.81795	10.3122

Tabulka 5.3: Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru GMP-ECM pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 20000$ a konstantu `DELKA` rovnou 12

Závěr

Obsahově se celá tato práce dá rozdělit více méně do dvou celků. V prvním z nich jsme se zabývali Edwardsovými křivkami z hlediska teorie algebraických funkčních těles, která pomáhá objasnit některá fakta, předložená v článcích bez větších komentářů. Vzhledem k tomu, že Edwardsův tvar je vcelku nový a jeho přínos se hojně zkoumá spíše z implementačního hlediska, je možné tuto teorii považovat za původní. Stejně tak je možné považovat za vlastní přínos geometrickou interpretaci sčítání na Edwardsově křivce, ačkoli jsme posléze popis téhož tématu nalezli v článku [1].

V druhé části jsme se zabývali samotnou faktorizací pomocí Edwardsových křivek. Ta, ač je tématem mnohem diskutovanějším, nebyla dosud nikde formálně popsána tak, jako zde v kapitole 4.3. Zajímavé by bylo navázat a odhadnout pravděpodobnost úspěchu algoritmu, případně zjistit, zda je vyšší, než u původního popisu Lenstry [14]. Potenciál ke zlepšení je také u přiložené implementace, ve které by se dala pravděpodobně vylepšit práce s pamětí či implementace aritmetiky velkých čísel v kernelu nebo generování parametrů křivek. Vzhledem k nedávným novinkám v OpenCL se nabízí také možnost přepsat kernel do nového jazyka OpenCL C++.

ECM je v dnešní době považován za jeden z nejlepších algoritmů pro faktorizaci, jehož časová složitost závisí převážně na velikosti dělitele a z menší části na velikosti dělitele samotného. Jeho význam podle mého názoru ještě vzroste, vzhledem k jednoduché možnosti paralelizace a současnému rozmachu využívání grafických karet i pro vědecké účely.

Literatura

[1] Arene, Christophe and Lange, Tanja and Naehrig, Michael and Ritzenthaler, Christophe. Faster Computation of the Tate Pairing. In: *IACR Cryptology ePrint Archive*. 2009. S. 155. [online, cit. 24.7.2015]. Dostupné z: http://eprint.iacr.org/2009/155

[2] Bernstein, Daniel J. and Birkner, Peter and Jove, Marc and and Lange, Tanja and Peters, Christiane. Twisted Edwards Curves, In: *IACR Cryptology ePrint Archive*, 2008, s. 13 [online, cit. 24.7.2014]. Dostupné z: `https://eprint.iacr.org/2008/013`

[3] Bernstein, Daniel J. and Birkner, Peter and Lange, Tanja and Peters, Christiane. *ECM using Edwards curves.* Cryptology ePrint Archive, Report 2008/016 [online, cit. 24.7.2014]. Dostupné z: `http://eprint.iacr.org/2008/016`

[4] Bernstein, Daniel J. and Lange, Tanja. A complete set of addition laws for incomplete Edwards curves. In: *IACR Cryptology ePrint Archive*, Report 2009/580 [online, cit. 24.7.2015]. Dostupné z: `http://eprint.iacr.org/2009/580`

[5] Bernstein, Daniel J. and Lange, Tanja. *Explicit formula database*[online, cit. 24.7.2015]. Dostupné z: `http://hyperelliptic.org/EFD`

[6] Bernstein, Daniel J. and Lange, Tanja. Faster addition and doubling on elliptic curves. In: *Advances in Cryptology – ASIACRYPT 2007.* Lecture Notes in Computer Science Volume 4833, 2007, pp 29-50. [online, cit. 24.7.2014]. Dostupné z: `http://link.springer.com/chapter/10.1007%2F978-3-540-76900-2_3`

[7] Bernstein, Daniel J. and Chen, Tien-Ren and Cheng, Chen-Mou and Lange, Tanja and Yang, Bo-Yin. ECM on Graphic Cards. In: *IACR Cryptology ePrint Archive*, 2008, s. 480.[online, cit. 24.7.2015]. Dostupné z: `https://eprint.iacr.org/2008/480`

[8] Brent, Richard P. Some integer factorization algorithms using elliptic curves. In: *Australian Coomputer Sciende Communications 8*, 1986, retyped and postscript added 1998, 149-163 [online, cit. 24.7.2014]. Dostupné z: `http://maths-people.anu.edu.au/ brent/pub/pub102.html`

[9] Crandall, Richard a Pomerance, Carl. *Prime Numbers: A Computational Perspective.* New York: Springer Publishing Company, Incorporated, 2001. ISBN: 0-387-94777-9

[10] Edwards, Harold M. A normal form for elliptic curves, *Bulletin of the American Mathematical Society*, Volume 44, Number 3, July 2007, pages 393-422 [online, cit. 24.7.2014]. Dostupné z: http://www.ams.org/bull/2007-44-03/S0273-0979-07-01153-6/-home.html

[11] GMP-ECM [online, cit. 24.7.2015]. Dostupné z: http://ecm.gforge.inria.fr/

[12] Hisil, Huseyin and Wong, Kenneth Koon-Ho and Carter, Gary and Dawson, Ed. Twisted Edwards curves revisited. In *IACR Cryptology ePrint Archive*, 2008, s. 522 [online, cit. 24.7.2015]. Dostupné z: http://eprint.iacr.org/2008/522.

[13] Khronos Group. *OpenCL - The open standard for parallel programming of heterogeneous systems*[online, cit. 24.7.2015]. Dostupné z: https://www.khronos.org/opencl/

[14] Lenstra, Hendrik W. Jr. Factoring integers with elliptic curves. In:*Annals of Mathematics*, **126**, 1987, 649-673[online, cit. 24.7.2014]. Dostupné z: https://openaccess.leidenuniv.nl/bitstream/1887/3826/1/346_086-.pdf

[15] Montgomery, Peter L. Modular multiplication without trial division. In:*Mathematics of Computation*, **44**, 1985, s. 519-521 [online, cit. 24.7.2015]. Dostupné z: http://http://www.ams.org/journals/mcom/1985-44-170/S0025-5718-1985-0777282-X/.

[16] Silverman, Joseph H. *The Arithmetic of Elliptic Curves*. Graduate Texts in Mathematics, 106. Printed in USA: Springer-Verlag New York Inc. ISBN: 0-387-96203.

[17] Stichtenoth, Henning. *Algebraic Function Fields and Codes*. Second edition. New York: Springer Publishing Company, Incorporated, 2008. Graduate Texts in Mathematics 111. ISBN: 9783540768777

[18] Washington, Lawrence C., *Elliptic Curves: Number Theory and Cryptography*. Second edition. CRC Press, Taylor & Francis Group, 2008. ISBN: 9781420071467

[19] Zimmermann, Paul and Dodson, Bruce. 20 Years of ECM. In: *Algorithmic Number Theory*, Springer, Lecture Notes in Computer Science Volume 4076, 2006, 525-542.

Seznam obrázků

3.1 Rozdíl mezi Edwardsovými křivkami 19
3.2 Inverzní prvek 30
3.3 Součet dvou různých bodů 30
3.4 „Speciální“ body Edwardsovy křivky 31

4.1 Převod mezi souřadnicemi 41

3 Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru gmp-ecm pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 5000$ a konstantu `DELKA` rovnou 12 62
4 Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru gmp-ecm pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 20000$ a konstantu `DELKA` rovnou 12 63

Seznam tabulek

4.1 Doporučené parametry pro ECM 45

5.1 Počet aritmetických operací 49
5.2 Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru GMP-ECM pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 5000$ a konstantu `DELKA` rovnou 12 54
5.3 Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru GMP-ECM pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 20000$ a konstantu `DELKA` rovnou 12 55
1 Čísla použitá pro testování času přiložené implementace a GMP-ECM. 61
2 Rozklad čísel z tabulky 1 62

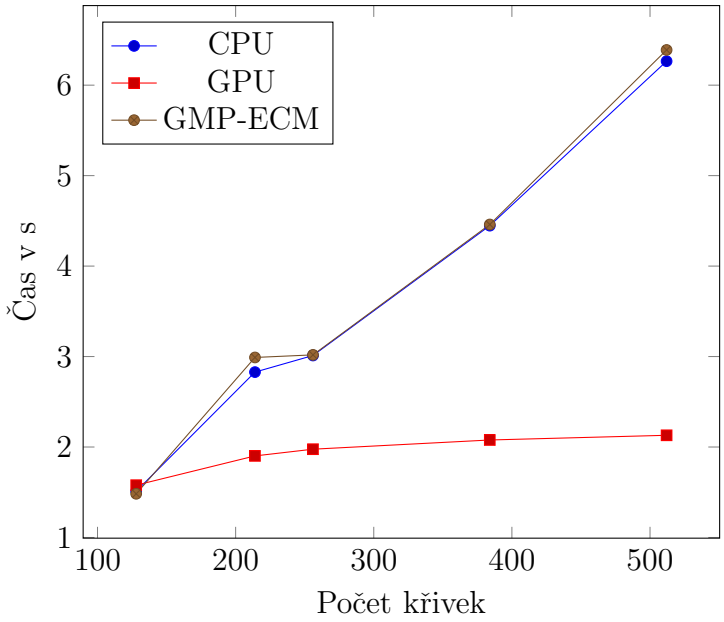
Přílohy

45231803683118295073086341937464869888104487123 7
53184743153014841213483716208064335950735151876189
72072814827719855657539928035157124673898974905019
96081301683822541789318962712104671474126489845 1
15178031594793585186242968400397168366772935383711
29224029841387651533567451206539030330286462704731
80040956326960473263320044145134334413015142018 21
30284336243318998954333657771854343054096611207073
61741085745721566735660970119580512137837785251869
13561716485736085305702032772929672069861251961017
23128977396319891051727260907588358322114113458503
29313287531889914077715993890927817515478880924167
23577666026576291543575665609771536089215220243711
14268963995458261503624912923893058289031409867711
48669904753316426645851859157874943187839504608679
15356451996252465911217676329661801357463164676623
56215801785690708593035499894791852080441137497 93
16183566467521436482726804978369930497920780585237
30669454767536631830065241159960454532862346727081
49755384920370025349374757517865379632525912622549

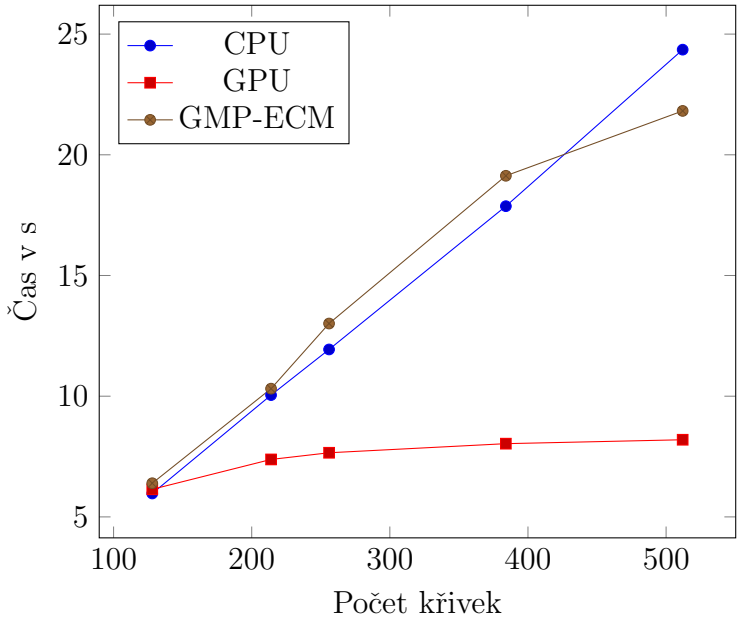
Tabulka 1: Čísla použitá pro testování času přiložené implementace a GMP-ECM.

213379889007584782100623	211977819903685906870781 9
679931611204644063146769 7	782207243737156299905623 7
807862168367777996888137 7	892142467487211490443394 7
367433349643926166519409	261493143659750558519353 9
275266830625000157577242 3	551393408364222034117725 7
541243471210202995379869	539942399231969683537339 9
191626438126172190242571 1	417692658224223088292901 1
477184418160291096642040 1	634646377601252322739787 3
760115686709142180602921 9	812259065630186853348435 1
182516498608492270046776 3	743040579297255660820105 9
324079093956952232240396 9	713683104760597038280288 7
372087052516369416028406 3	787807243859965259748520 9
285921388348616497643629 3	824620577101726219801662 7
357799623901948789343822 3	398797624207914776614305 7
552886620534535757474873 9	880287258647386420496046 1
208715391150846307928345 9	735760401356974766885299 7
116442000318818093583717 7	482779423505023042971760 9
277602765181727303256655 7	582975693953450946718524 1
339626829808804444053812 3	903033920635841375650114 7
566076446393880643093228 9	878951689958670699760834 1

Tabulka 2: Rozklad čísel z tabulky 1



Obrázek 3: Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru gmp-ecm pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 5000$ a konstantu `DELKA` rovnou 12



Obrázek 4: Porovnání času v sekundách přiložené reprezentace pro grafickou kartu (GPU), procesoru (CPU) a softwaru gmp-ecm pro N součin dvou prvočísel délky až 83 bitů, $B_1 = 20000$ a konstantu `DELKA` rovnou 12