Indice

Introduzione

La nascita della crittografia, l'arte di nascondere i messaggi risale ad epoca antica. Oggi riveste un'importanza assoluta in molte situazioni, per garantire la sicurezza dei dati personali, dell'invio di messaggi riservati ecc... L'introduzione della crittografia basata sulle curve ellittiche è relativamente recente, ma si è imposta, come valida alternativa ai sistemi crittografici a chiave pubblica ampiamente utilizzati, come l'RSA.

La teoria delle curve ellittiche è uno dei crocevia fondamentali della matematica in quanto vi si incontrano analisi, geometria, algebra e negli ultimi anni anche l'informatica. Inizialmente il loro studio era puramente astratto, in quanto riguardava il campo della teoria dei numeri o quello della geometria algebrica, fino ad arrivare alla dimostrazione del matematico Andrew Wiles dell'ultimo teorema di Fermat nel 1995.

Solo negli ultimi anni, le curve ellittiche sono state studiate anche per risolvere molti problemi nel campo della crittografia, ultima nata tra le applicazioni pratiche della teoria dei numeri. Nonostante tracce di sistemi crittografici si ritrovino nell'antichità a partire da 2000 anni a.C., la nascita della crittografia moderna si può collocare durante la seconda guerra mondiale, quando gli sforzi dei crittoanalisti alleati ebbero la meglio sul sistema tedesco basato sulla macchina Enigma. L'avvento dei calcolatori e dell'informatica, a partire dai lavori di Alan Turing, fece sì che si potessero sviluppare algoritmi basati su proprietà matematiche, in precedenza intrattabili. Gli elementi fondamentali della crittografia al giorno d'oggi sono i gruppi finiti con un numero di elementi dell'ordine di 10^{200} ed i cosiddetti problemi 'NP' (non-deterministic polynomial), ovvero algoritmi estremamente difficili da calcolare a meno di avere un'informazione aggiuntiva. Il problema NP utilizzato nel caso delle curve ellittiche riguarda i gruppi: data una di queste strutture possiamo, a partire da un elemento a dato, calcolare l'elemento $b = a + a + ... + a$, n volte. Tuttavia, affinché il calcolo sia effettivamente impraticabile anche sui calcolatori attuali, è necessario che n sia molto grande e soprattutto che tutte le somme parziali $a + a$, $a + a + a$, e così via fino a

n, diano sempre elementi diversi: se così non fosse si creerebbe un ciclo ed esisterebbe inevitabilmente un $m < n$ tale che b sia m volte a. Questo può essere evitato prendendo un gruppo che sia allo stesso tempo molto 'grande' (ovvero che contenga molti elementi) e ciclico, ovvero in cui esiste almeno un elemento che sommato tante volte (fino al numero di elementi del gruppo) dia sempre risultati diversi; nonostante esistano gruppi di questo tipo 'facili' da realizzare (come Z/pZ, con p primo), essi sono anche 'facili' da risolvere dai calcolatori se p è troppo piccolo.

L'algoritmo probabilmente più noto tra quelli utilizzati comunemente è il cosiddetto RSA, un sistema di cifratura a chiave pubblica. Per mantenere un livello adeguato di sicurezza, l'RSA necessita oggi di una chiave lunga almeno 1024 bit. L'utilizzo delle curve ellittiche può dare un netto vantaggio proprio sotto questo punto di vista. I punti di queste curve formano un gruppo in cui, a parità di numero di elementi, la ricerca di n risulta sensibilmente più difficile, tanto che per avere un livello di sicurezza equivalente a quello fornito dall'RSA con chiave di 1024 bit (circa 338 cifre decimali) è sufficiente utilizzare curve 'buone' con chiavi lunghe solo 160 bit (circa 53 cifre decimali).

Per approfondire si può consultare il sito di Alfred Menezes:
`http://www.math.uwaterloo.ca/~ajmeneze` per lunghe discussioni e lavori (spesso in collaborazione con N. Koblitz) sul concetto di sicurezza garantita.

La dimostrazione di Wiles dell'ultimo teorema di Fermat (1995)

Nel 1637 Fermat lesse l'*aritmetica* di Diofanto, un monumentale libro del terzo secolo, e annotò sul margine la seguente osservazione:

'Dividere un cubo in due cubi, o in generale una potenza n-esima in due potenze n-esime, è impossibile se n è maggiore di 2: ho trovato una dimostrazione veramente notevole di ciò, ma il margine è troppo ristretto per contenerla.'

Questa osservazione era stata anticipata per i cubi nel 1070 da Omar Khayyam, matematico e poeta. Nella sua forma generale divenne nota come l'*ultimo teorema di Fermat*, ed è stata per 350 anni uno dei problemi più famosi della matematica.

Fermat richiedeva che n fosse maggiore di 2 perché già i Babilonesi, e poi i Pitagorici, sapevano che ci sono quadrati che si possono scrivere come somma di due quadrati, per esempio

$$3^2 + 4^2 = 5^2$$

Si è trovata, nella corrispondenza di Fermat, una dimostrazione del teorema per $n = 4$: essa usa un ingegnoso metodo detto *discesa infinita*, che consiste nel supporre per assurdo che ci sia una soluzione, e far vedere che allora ce ne deve essere un'altra i cui numeri non sono più grandi della precedente, e almeno uno non è strettamente più piccolo, il che porta a un impossibile regresso infinito.

Nel corso degli anni i migliori matematici si impegnarono nel problema, e confermarono il teorema in vari casi: $n = 3$ Eulero nel 1753, $n = 5$ Dirichelet e Legendre nel 1825, $n = 7$ Lamé nel 1839, ogni n minore di 100 Kummer fra il 1847 ed il 1857. Benché nel 1980 la verifica fosse arrivata ad ogni n minore di 125000, mancava però una dimostrazione generale del teorema.

Il primo vero risultato generale fu ottenuto in maniera piuttosto indiretta. Il punto di partenza è l'osservazione che il teorema di Fermat richiede soluzioni *intere* di equazioni del tipo

$$a^n + b^n = c^n$$

Poiché allora

$$(a/b)^n + (b/c)^n = 1$$

si tratta dunque di trovare soluzioni *razionali* di equazioni del tipo

$$x^n + y^n = 1$$

Queste equazioni definiscono una curva se considerate sui numeri reali, e una superficie se considerate sui numeri complessi: queste superfici si possono poi classificare in base al numero di buchi che hanno. Per esempio, per $n = 2$ non ci sono buchi, poiché l'equazione precedente definisce un cerchio come curva e una sfera come superficie; e ci sono infinite soluzione razionali, che già Diofanto sapeva come descrivere completamente. Nel caso di n maggiore di 2 ci sono invece buchi: uno per $n = 3$, tre per $n = 4$, sei per $n = 5$, e così via. Naturalmente col crescere dei buchi cresce la complessità della superficie e diminuisce la possibilità di trovare soluzioni razionali.

Oltre alle soluzioni precedenti, un altro tipo era nel frattempo risultato particolarmente interessante: le cosiddette *curve ellittiche*. In questo caso il numero dei buchi della corrispondente superficie è uno, e anche qui è possibile avere infinite soluzioni razionali. Nel 1922 Leo Mordell propose la *congettura di Mordell*: gli unici tipi di equazioni che ammettono infinite soluzioni razionali sono quelli che definiscono superfici o senza buchi, o con un buco solo.

Il che significa che, se vale la congettura di Mordell, il teorema di Fermat è *quasi* vero, perché per tutti gli n maggiori di 3 (e il caso $n = 3$ era già stato risolto da Eulero) l'equazione definisce una superficie con più di un buco, e può dunque avere al massimo un numero *finito* di soluzioni razionali.

Nel 1962 Igor Shafarevich propose, a sua volta, la *congettura di Shafarevich*: in certe condizioni, si possono trovare le soluzioni intere di un'equazione smontando dapprima l'equazione stessa, considerandone cioè i vari analoghi ottenuti limitando gli interi al di sotto dei vari numeri primi, risolvendo questi analoghi finiti, e rimontando poi le soluzioni per ottenere una soluzione dell'equazione di partenza. In altre parole, si cerca di ricostruire le soluzioni sulla base della conoscenza dei loro resti rispetto alla divisione per vari numeri primi.

Un legame tra le due congetture fu trovato nel 1968 da Parshin, il quale provò che dalla congettura di Shafarevich discende la congettura di Mordell. E la congettura di Shafarevich venne dimostrata nel 1983 da Gerd Faltings, che ottenne per questo la medaglia Fields nel 1986. La dimostrazione utilizza in maniera essenziale la soluzione di Deligne dell'ulteriore congettura di Weil.

La dimostrazione della congettura di Mordell è un risultato talmente interessante da essere stato propagandato come il *teorema del secolo*, ma sembra non aiutare molto per quanto riguarda il teorema di Fermat: anche una sola soluzione razionale dell'equazione

$$x^n + y^n = 1$$

produrrebbe infatti una soluzione intera dell'equazione

$$a^n + b^n = c^n$$

e quindi infinite soluzioni (ottenute moltiplicando la precedente per una costante). In realtà, nel 1985 Andrew Granville e Roger Heath-Brown riuscirono a derivare dal teorema di Faltings la validità del teorema di Fermat per infiniti esponenti primi. Anzi, per *quasi tutti* gli esponenti, dal punto di vista di teoria della misura. Alla dimostrazione del teorema di Fermat per *tutti* gli esponenti maggiori di 2 si arrivò ancora una volta per una strada molto indiretta, attraverso la cosiddetta *congettura di Taniyama*. Il punto di partenza è ora l'osservazione che l'equazione

$$x^2 + y^2 = 1$$

si può parametrizzare mediante le funzioni trigonometriche, che soddisfano appunto l'equazione fondamentale:

$$\sin^2 \alpha + \cos^2 \alpha = 1$$

Risolvere l'equazione di Fermat per $n = 2$ significa dunque trovare un angolo α i cui seno e coseno siano razionali. In maniera analoga, le cosiddette funzioni trigonometriche iperboliche parametrizzano l'equazione

$$x^2 - y^2 = 1$$

Passando alle equazioni quadratiche che definiscono le coniche alle cubiche, Taniyama congetturò nel 1955 che certe *funzioni modulari*, più generali di quelle trigonometriche, parametrizzano in maniera analoga qualunque curva ellittica.

Il legame tra congettura e il teorema di Fermat fu notato nel 1985 da Gerhard Frey, il quale propose di associare all'equazione di Fermat

$$a^n + b^n = c^n$$

la curva ellittica

$$y^2 = x(x + a^n)(x - b^n)$$

Frey notò che la sua curva ellittica ha proprietà troppo belle per essere vere: per esempio, il discriminante che determina l'esistenza di radici del polinomio

$$(x + a^n)(x - b^n) = x^2 + x(a^n - b^n) - a^n b^n$$

e cioè,

$$\Delta = \sqrt{(a^n - b^n)^2 + 4a^n b^n} = a^n + b^n = c^n$$

è una potenza n-esima perfetta. Nel 1986 Ken Ribet dimostrò che la curva di Frey non può essere parametrizzata da funzioni modulari: il che, detto altrimenti, significa che dalla congettura di Taniyama discende il teorema di Fermat.

Rimaneva soltanto da dimostrare la congettura. Nel 1995 Andrew Wiles riuscì a provarne una parte, per una classe di curve ellittiche dette semistabili, a cui appartiene la curva di Frey, risolvendo così uno dei più famosi problemi aperti della matematica moderna. Wiles ottenne per questo storico risultato il *premio Wolf* nel 1995-96, ma non poté aggiudicarsi una medaglia Fields nel 1998 perché aveva da poco i quarant'anni. Nel 1999 Brian Conrad, Richard

Taylor, Christophe Breuil e Fred Diamond hanno completato il lavoro di Wiles, dimostrando che la congettura di Taniyama vale anche per le curve ellittiche non semistabili.

Il settore 'Elliptic curves' di Wikipedia (sezione inglese) è molto accurato ed aggiornatissimo (anche quello di Number Theory). Naturalmente non ci si impara la Matematica (non ci sono dimostrazioni e non si distingue il livello di difficoltà dei vari argomenti), ma è utilissimo per links e citazioni ad articoli recenti su arXiv, riviste SIAM ed AMS e sembra essere estremamente imparziale (riporta tutte le varie campane) in un settore così delicato ed economicamente importante come la crittografia. Ad esempio ha immediatamente segnalato l'attacco temporale alle chiavi che riportiamo in sezione [3.5], ma senza enfatizzarlo.

Circa un anno fa l'annuncio di un importante risultato di Richard Taylor sul sito dell'AMS recava come suggerimento per più dettagli il link alla corrispondente voce su Wikipedia (English). Nel sito di preprints [6] c'è anche una sezione di discussione sui singoli preprints, ma (30 giugno 2011) nulla sul preprint [10] in questione. Sul sito web di N. Koblitz al momento non c'è alcun articolo sul problema affrontato in [10]: aspettiamo con ansia!

Su questo settore la versione italiana di Wikipedia è quasi muta, mentre è ottima nella parte storica di matematica collegata con matematici italiani (la parte sui geometri sembra essere stata suggerita da precedenti gruppi di lavoro e di divulgazione matematica, ad esempio quello di Trieste). Altri varianti linguistiche di Wikipedia sono erratiche per la parte di Matematica: miserrima quella in spagnolo, catastrofica quella in portoghese, ottimo il sito in tedesco (N. Schappacher è uno degli esperti di teoria dei numeri e uno dei massimi storici del settore, plenary speaker a ICM2010 in India), utile quella francese, se uno si interessa a matematici francesi; la Societé Mathematique de France ha raccolto on line molto materiale di archivio e penso anche altre società matematiche lo abbiano fatto (sicuramente quella italiana e quella tedesca).

Obiettivi e organizzazione della tesi

Verrà introdotta la crittografia basata sulle curve ellittiche, analizzando sia gli aspetti teorici che i protocolli utilizzati e soprattutto confrontando questa teoria con le moderne soluzioni a chiave pubblica.

Il primo capitolo si apre con una breve esposizione di alcuni momenti significativi della storia della crittografia e di alcuni concetti di base per affrontare un completo studio su tale argomento. Vengono presentati poi i principali sistemi asimmetrici.

Il secondo capitolo descrive le curve ellittiche e la loro aritmetica, comprese le formule relative alle regole di addizione.

Il terzo capito descrive i sistemi crittografici basati sulle curve ellittiche (ECC).

Il quarto capitolo definisce le curve iperellittiche allo scopo di introdurre la crittografia iperellittica (HCC).

Nel quinto capitolo viene fatto un confronto tra l'ECC e l'HCC e si tratta del problema dell'efficienza al variare del genere e della dimensione del campo.

Capitolo 1

Introduzione alla crittografia

Il termine **crittografia** deriva dalla lingua greca (*kryptòs*, nascosto e *gràphein*, scrivere) ed è l'arte di scrivere messaggi nascosti. Lo scopo della crittografia è quello di inventare codici oscuri che possano nascondere ad eventuali 'intrusi' un messaggio riservato. Per rendere incomprensibile un messaggio lo si altera per mezzo di un procedimento concordato dal mittente e dal destinatario. Questi può quindi invertire il procedimento, e ricavarne il messaggio iniziale. Il vantaggio della crittografia è che se un intruso intercetta il messaggio esso risulta incomprensibile. Infatti l'estraneo, non conoscendo il metodo di alterazione, troverà estremamente complicato, se non impossibile, ricostruire il messaggio originale. In questo contesto un ruolo molto importante è ricoperto dalla *chiave* del codice.

Agli inizi della storia della crittografia venivano usati metodi in cui la chiave non giocava un ruolo essenziale; oggi, invece, ogni codice deve poter avere moltissime chiavi. La matematica è apparsa in modo esplicito in crittografia solo a partire dagli anni '40 in poi; essa serve sia per costruire un codice, sia per romperlo. Precisamente la crittografia studia le tecniche matematiche che consentono di modificare un certo messaggio in modo da renderlo incomprensibile ad un intruso malintenzionato, ma leggibile soltanto al destinatario. La modifica del testo in chiaro viene detta *cifratura* e produce il cosiddetto testo cifrato. Il procedimento inverso che dal messaggio cifrato ricostruisce il messaggio in chiaro è chiamato *decifratura*. Come detto precedentemente, il mittente ed il legittimo destinatario devono condividere a priori una conoscenza che consenta la cifratura del messaggio in chiaro e la successiva decifratura.

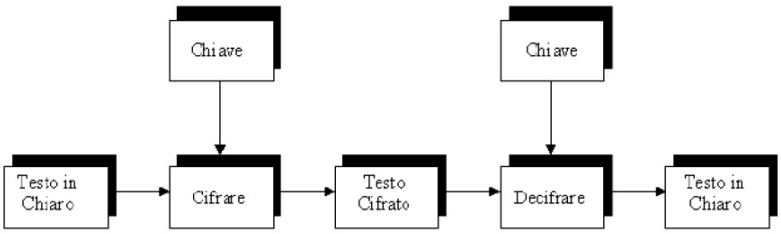


Figura 1.1: generico schema crittografico

Tale conoscenza però non è il processo di modifica stesso ma la cosiddetta chiave ossia una stringa alfanumerica che costituisce un parametro della funzione di cifratura e della funzione di decifratura.

Il metodo di alterazione pertanto è noto a chiunque (chiave pubblica), e quindi anche all'intruso, ma ogni volta viene parametrizzato da una chiave segreta (privata) condivisa solamente tra mittente e destinatario.

Se la crittografia si occupa dei possibili metodi di cifratura, la **crittoanalisi** studia le tecniche che consentono di decifrare, ovvero di rendere vani, tali metodi ricavando il testo in chiaro dal messaggio cifrato senza però conoscere la chiave segreta ma sfruttando le debolezze dell'algoritmo impiegato. L'obiettivo primario della crittoanalisi è quindi quello di scoprire il valore della chiave privata utilizzata.

Crittografia e crittoanalisi formano ciò che viene chiamata **crittologia**.

I *codici a trasposizione* ed i *codici a sostituzione* costituiscono due metodi crittografici usati nell'antichità che sviluppano i prototipi di due tipi di cifrari. Un esempio del primo tipo è la *scitala lacedemonica*, usata a Sparta circa 2500 anni fa. La scitala era un cilindro usato per trasmettere un messaggio in modo segreto. Il metodo consisteva nell'avvolgere intorno alla scitala un nastro, sul quale veniva scritto il messaggio in righe longitudinali. Finita l'operazione si svolgeva il nastro, che veniva mandato al destinatario. E' chiaro che sul nastro le lettere che componevano le parole del messaggio risultavano permutate. Il destinatario era in possesso di un cilindro identico a quello usato dal mittente; riavvolgendo il nastro su di esso, il messaggio si ricomponeva, così da poter essere letto. La chiave di questo codice è data

dal diametro del cilindro. Un codice basato su questo principio si dice *codice a trasposizione*.

Figura 1.2: scitale lacedemonica

Codici basati su un diverso principio sono quelli a sostituzione. In questo tipo di codici ogni lettera del testo in chiaro viene trasformata in un'altra lettera. Il prototipo storico di questi codici è il *codice di Cesare*, basato sull'uso di un alfabeto in chiaro ed un alfabeto segreto. Una qualsiasi lettera del testo in chiaro viene cifrata nella lettera dell'alfabeto segreto. L'alfabeto segreto usato da Cesare si ottiene trasformando le lettere dell'alfabeto in chiaro in quella posizionata tre posizioni a sinistra. Si può notare facilmente che il codice di Cesare è soggetto ad una facile crittoanalisi. Infatti, poiché ogni lettera viene trasformata sempre nello stesso modo, il cifrato conserva la frequenza con cui compare ogni data lettera. Schemi di questo tipo conservano ormai un mero valore enigmistico.

Gli algoritmi crittografici possono essere divisi in due classi principali: gli algoritmi a chiave privata (o simmetrica) e quelli a chiave pubblica (o asimmetrica).

1.1 Crittografia a chiave privata

Gli algoritmi di crittografia a chiave privata o simmetrica sono chiamati così perché utilizzano la stessa chiave sia per la cifratura che per la decifratura. Questo tipo di algoritmi storicamente sono nati per primi e fanno uso di tecniche di trasposizione e di sostituzione.

Una delle caratteristiche più significative dei sistemi a chiave simmetrica è la velocità dell'implementazione. L'aspetto negativo riguarda il cosiddetto

problema della distribuzione delle chiavi. Se si hanno N entità che devono comunicare tra loro in modo sicuro attraverso un sistema crittografico a chiave simmetrica si deve fare in modo che ogni entità possegga $N-1$ chiavi differenti, una per ciascuna delle altre entità. Ciò significa che si deve essere in grado di generare, trasmettere e conservare $\frac{N(N-1)}{2}$ chiavi diverse. Naturalmente, per valori molto grandi di N la gestione del sistema diventa troppo complessa. Altro problema è poi la trasmissione della chiave simmetrica perché è molto difficile realizzare un canale sicuro.

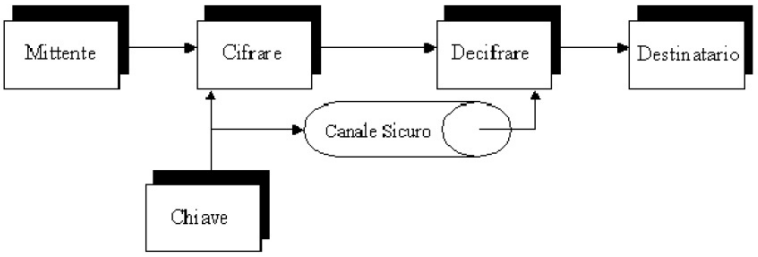


Figura 1.3: schema crittografico a chiave simmetrica

Un esempio di crittografia a chiave simmetrica è il *codice Vernam*. In tale codice ogni dato è una stringa m definita sull'alfabeto $Z_2 = 0, 1$. I due interlocutori posseggono una stessa chiave k, una stringa definita sullo stesso alfabeto e di lunghezza pari a quella del messaggio che deve essere trasmesso. Il mittente trasmette $f(x) = x + k$, e il destinatario, ricevuto il messaggio criptato, dovrà semplicemente calcolare $f(x) + k = x$. Se un estraneo intercetta il messaggio $f(x)$, ma non conosce k, non può risalire al messaggio x, perché $\forall y\ \exists k$ tale che $f(x) = x + k = y$ quindi tale codice è inattaccabile.

1.2 Crittografia a chiave pubblica

Viste le difficoltà dovute alla distribuzione delle chiavi dei cifrari simmetrici, caratterizzati da un'unica chiave per cifrare e decifrare, che deve essere protetta e allo stesso tempo distribuita a tutti gli utenti del sistema, nel 1976 i ricercatori Diffie ed Hellman proposero il concetto di crittografia a chiave pubblica.

Un sistema a chiave pubblica prevede che la chiave di decifratura non sia facilmente derivabile da quella di cifratura. Nella proposta di Diffie-Hellman l'algoritmo di cifratura, $C(.)$ e quello di decifratura $D(.)$ devono soddisfare due condizioni:

1. $D(C(M)) = M$
2. deve essere difficile ricavare $D(.)$ da $C(.)$

Questo metodo funziona nel seguente modo. Alice che vuole ricevere messaggi riservati da Bob, trova due funzioni C e D, che soddisfino le due precedenti condizioni, parametrizzati dalla chiave per cifrare e da quella per decifrare. La chiave di cifratura viene resa pubblica mentre quella di decifratura viene mantenuta segreta da Alice.

Se Bob vuole mandare un messaggio M ad Alice, si procura la sua chiave pubblica, calcola $S = C(M)$, che è il messaggio cifrato, e glielo invia. A questo punto Alice, utilizzando la sua chiave privata, calcola $M = D(S)$.

La crittografia a chiave pubblica, quindi, richiede che l'utente possegga due chiavi: una *pubblica* usata per cifrare i messaggi da chiunque voglia comunicare con lui ed una *privata* per decifrare i messaggi ricevuti. Queste due chiavi devono essere tra loro correlate, ma deve essere difficile poter calcolare la chiave privata da quella pubblica, affinché il sistema rimanga sicuro. Questa difficoltà si basa sulla intrattabilità di alcuni problemi matematici, come:

1. **Problema della fattorizzazione degli interi IFP**: dato un numero intero $n = pq$, con p e q primi molto grandi, trovare p e q. Su tale problema si basa il sistema RSA;

2. **Problema del logaritmo discreto DLP**: calcolare il logaritmo x di un elemento $b = a^x$ su un campo finito. Su tale problema si basa l'algoritmo di firma digitale DSA ed il sistema di ElGamal;

3. **Problema del logaritmo discreto su curve ellittiche**: è una forma generalizzata del DLP sui punti di una curva ellittica. Su tale problema si basa l'analogo su curva ellittica di ElGamal.

La caratteristica dei sistemi crittografici a chiave pubblica è che sono molto più lenti di quelli a chiave simmetrica, ma allo stesso tempo, risolvono il problema della distribuzione delle chiavi. Proprio per questo si preferisce l'utilizzo di un sistema che accomuni entrambi: gli schemi a chiave pubblica vengono utilizzati per lo scambio delle chiavi simmetriche (in genere molto più brevi del messaggio) e la crittografia simmetrica viene utilizzata per cifrare i messaggi.

1.2.1 RSA e problema della fattorizzazione intera

Il sistema RSA deve il suo nome ai suoi ideatori R. Rivest, A. Shamir e L. Adleman. Venne introdotto nel 1977 ed è uno dei sistemi crittografici a chiave pubblica maggiormente utilizzati per cifrare e per fornire la firma digitale.

Definizione 1.2.1. Il problema della fattorizzazione intera (IFP) di un numero intero positivo n consiste nel trovare i numeri primi p_1, p_2,...,p_k tali che:

$$n = p_1^{e_1} p_2^{e_2} ... p_k^{e_k} \tag{1.1}$$

con $e_i \geq 1$.

L'algoritmo RSA si basa sul seguente problema:

dato un intero positivo n, prodotto di due distinti numeri primi p e q, un intero positivo e (esponente di cifratura) tale che $MCD(e, \phi) = 1$ (dove $\phi = (p-1)(q-1)$), ed un intero c, si trovi quell'unico intero m tale che $m^e \equiv c (mod\ n)$.

Il primo passaggio dell'RSA è la generazione della coppia di chiavi, che avviene secondo il seguente algoritmo. La chiave pubblica consiste nella coppia di interi (n,e), mentre quella privata è l'intero d, chiamato esponente di decifratura, tale che $ed \equiv 1 (mod\ \phi)$.

Algoritmo di generazione delle chiavi RSA

1. Scegliere due numeri primi grandi, distinti e casuali p e q;
2. calcolare $n = pq$ e $\phi = (p-1)(q-1)$;
3. scegliere un intero casuale e, $1 \leq e \leq \phi$, tale che $MCD(e, \phi) = 1$;
4. usando l'algoritmo di Euclide, calcolare l'unico intero d, $1 \leq d \leq \phi$ tale che $ed \equiv 1 (mod\ \phi)$;
5. la chiave pubblica è (n, e) e quella privata d.

Supponiamo che Bob voglia mandare un messaggio cifrato ad Alice. L'algoritmo per cifrare un messaggio è il seguente:

Algoritmo di cifratura RSA

1. Bob ottiene la chiave pubblica di Alice (n, e);
2. rappresenta il messaggio come un intero m nell'intervallo $[0, n-1]$;
3. calcola $c \equiv m^e (mod\ n)$;
4. invia c ad Alice.

Decifratura RSA

Alice può calcolare m utilizzando la propria chiave privata d, calcolando $m \equiv c^d (mod\ n)$.

Ora vediamo come utilizzare gli stessi parametri $(n = pq, e)$ e d dell'algoritmo RSA per *firmare* un messaggio. Supponiamo che Alice voglia firmare un messaggio m da inviare a Bob. Viene utilizzata una funzione $H(.)$ detta *funzione hash*, come chiave pubblica (n, e) e come chiave privata d. La funzione H è pubblica e serve soltanto a ridurre le dimensioni del messaggio, nel caso in cui queste fossero troppo grandi.

Firma e verifica RSA

Alice genera la firma s del messaggio m:

- calcola $h = H(m)$;
- calcola $s = h^d (mod\ n)$;
- s è la firma di Alice del messaggio m, che verrà trasmessa assieme al messaggio a Bob.

Bob riceve la coppia (s, m) e verifica la firma di Alice:

- calcola $h = H(m)$;
- calcola $h' = s^e (mod\ n)$;
- se $h = h'$ la firma è accettata, altrimenti è rifiutata.

La verifica della firma quindi si basa anch'essa sul fatto che $h^{ed} \equiv h (mod\, n)$.

Osservazione 1.2.2. La sicurezza dell'RSA è tutta basata sulla difficoltà di trovare p e q partendo da n, dal momento che, se un estraneo viene a conoscenza di p e q, egli può, allo stesso modo di Bob calcolare ϕ e quindi d.

1.3 Sistemi basati sul logaritmo discreto

La sicurezza di molti sistemi crittografici si basa sull'intrattabilità del problema del logaritmo discreto. Particolari esempi di sistemi di questo tipo sono il protocollo di Diffie-Hellman per la condivisione di una chiave segreta, il crittosistema ElGamal e l'algoritmo di firma digitale DSA.

Definizione 1.3.1. Sia G un gruppo ciclico finito di ordine n. Sia α un generatore di G e $\beta \in G$. Il logaritmo discreto (DL) di β in base α, indicato con $\log_\alpha \beta$, è l'unico intero x, $0 \leq x \leq n-1$, tale che $\alpha^x = \beta$.

Gruppi di particolare interesse in crittografia per il problema del logaritmo discreto sono i gruppi moltiplicativi Z_p^* degli interi modulo p, con p primo. Quindi:

Definizione 1.3.2. Il problema del logaritmo discreto (DLP) è il seguente: dato un numero primo p, un generatore α di Z_p^*, trovare l'intero x, $0 \leq x \leq p-2$, tale che $\alpha^x \equiv \beta (mod\ p)$.

L'utilizzo del DLP in crittografia deriva dalla difficoltà di risolvere questo problema. Uno dei più importanti algoritmi che si basano su questo problema è il *protocollo Diffie-Hellman*.

1.3.1 Protocollo Diffie-Hellman

Il protocollo Diffie-Hellman venne presentato nello stesso articolo che introdusse per la prima volta la crittografia a chiave pubblica, da Diffie ed Hellman nel 1976. Si tratta di un protocollo che consente ad Alice ed a Bob di condividere una chiave segreta K, senza la necessità di un canale sicuro.

1. Vengono fissati un numero primo p ed un generatore α di Z_p^*;
2. Alice sceglie un intero casuale compreso tra 1 e $p-1$ e calcola $X = \alpha^x (mod\ p)$;
3. Alice invia X a Bob;
4. Bob sceglie un intero casuale y compreso tra 1 e $p-1$ e calcola $Y = \alpha^y (mod\ p)$;

5. Bob invia Y ad Alice;
6. Alice riceve Y e calcola $K_X = Y^x (mod\ p) = (\alpha^y)^x (mod\ p)$;
7. Bob riceve X e calcola $K_Y = X^y (mod\ p) = (\alpha^x)^y (mod\ p)$;
8. Alice e Bob condividono la chiave $K = K_X = K_Y$.

Si noti che per determinare la chiave segreta K, il nemico non deve necessariamente risolvere il DLP ma il seguente problema:

Definizione 1.3.3. Il problema di Diffie-Hellman (DHP) è il seguente: dato un numero primo p, un generatore α di Z_p^* e gli elementi $\alpha^a (mod\ p)$ e $\alpha^b (mod\ p)$ trovare $\alpha^{ab} (mod\ p)$.

Nonostante sia facile dimostrare che la risoluzione del DLP in Z_p^* implichi la risoluzione del DHP, non è ancora stato provato che valga il contrario.

1.3.2 Crittosistema di ElGamal

La sicurezza di questo sistema si basa sull'intrattabilità del problema del logaritmo discreto. Questo algoritmo costituisce il primo crittosistema basato su DLP. Gli algoritmi che seguono, descrivono rispettivamente la generazione delle chiavi e gli schemi di cifratura e decifratura del sistema ElGamal.

Generazione delle chiavi per ElGamal

Ogni entità crea una chiave pubblica ed una corrispondente chiave privata.

Alice svolge le seguenti azioni:

1. sceglie un numero primo grande p e un generatore α del gruppo moltiplicativo Z_p^* degli interi modulo p;
2. sceglie un intero casuale a compreso tra 1 e $p-2$ e calcola $\alpha^a (mod\ p)$;
3. la chiave pubblica di Alice è (p, α, α^a), mentre quella privata è a.

Cifratura per il sistema ElGamal

Bob per cifrare un messaggio m da inviare ad Alice svolge le seguenti azioni:

1. ottiene la chiave pubblica di Alice (p, α, α^a);
2. rappresenta il messaggio come un intero m compreso tra 0 e $p-1$;
3. sceglie un numero casuale k compreso tra 1 e $p-2$;
4. calcola $\gamma = \alpha^k (mod\ p)$ e $\delta = m(\alpha^a)^k (mod\ p)$;
5. invia $c = (\gamma, \delta)$ ad Alice.

Decifratura per il sistema ElGamal

Alice per ricavare m da c svolge le seguenti azioni:

1. usa la chiave privata a per calcolare $\gamma^{p-1-a} (mod\ p)$,
   (osserviamo che $\gamma^{p-1-a} = \gamma^{-a} = \alpha^{-ak}$);
2. ottiene m calcolando $(\gamma^{-a})\delta (mod\ p)$.

La decifratura di tale algoritmo si basa sull'osservazione che:

$$\gamma^{-a}\delta \equiv \alpha - akm\alpha^{ak} \equiv m (mod\ p) \tag{1.2}$$

Il processo di cifratura è non-deterministico in quanto il messaggio cifrato c dipende dal messaggio in chiaro m e dal valore casuale k, scelto da Bob. Questo significa che uno stesso messaggio può essere cifrato in vari modi a seconda del valore di k.

Firma e verifica ElGamal

L'algoritmo che segue descrive il processo con il quale Alice usa la propria chiave privata a per firmare il messaggio m e il modo in cui Bob verifica tale firma utilizzando la chiave pubblica (p, α, α^a) di Alice.

Come nel caso dello schema RSA, entrambi possono usare una funzione hash $H(.)$ pubblica nel caso in cui il messaggio sia eccessivamente lungo.

Alice genera la firma (r, s) del messaggio m svolgendo le seguenti azioni:

1. sceglie un numero casuale compreso tra 1 e $p-2$ con $MCD(k, p-1) = 1$;
2. calcola $r = \alpha^k (mod\ p)$;
3. calcola $s = k^{-1}[H(m) - ar](mod\ p - 1)$;

La coppia (r, s) è la firma di Alice del messaggio m.

Bob verifica la firma di Alice svolgendo le seguenti azioni:

1. verifica che r sia compreso tra 1 e $p - 1$ altrimenti rifiuta la firma;
2. calcola $v_1 = \alpha^{ar} r^s \mod p$;
3. calcola $v_2 = \alpha^{H(m)} \mod p$;
4. accetta la firma se e solo se $v_1 = v_2$.

La verifica della firma si basa sulle seguenti considerazioni:

se s è stato veramente generato da Alice allora:

$$ks \equiv (H(m) - ar) \mod (p - 1)$$

che equivale a scrivere

$$H(m) \equiv ar + ks \mod (p - 1)$$

Ciò significa che

$$\alpha^{H(m)} \equiv \alpha ar + ks \equiv (\alpha^a)^r r^s \mod p$$

e quindi $v_1 = v_2$.

1.3.3 L'algoritmo DSA

Il DSA (Digital Signature Algorithm) è uno schema di firma digitale che rappresenta una variante dello schema di firma ElGamal. Gli algoritmi seguenti indicano come fare la generazione delle chiavi e gli schemi di firma e verifica DSA.

Generazione delle chiavi DSA

1. Si sceglie un primo grande q;
2. si sceglie un primo p tale che q divida $p-1$
3. si sceglie un elemento $\alpha \in Z_p^*$ di ordine q;
4. si sceglie un intero compreso tra 1 e $q-1$;
5. si calcola $y = \alpha^x \mod p$;
6. la chiave pubblica è $p, q, \alpha, y)$ e quella privata x.

Firma DSA

Alice firma con la propria chiave privata un messaggio m secondo i seguenti passaggi:

1. sceglie un intero k compreso tra 1 e $q-1$;
2. calcola $r = (\alpha^k \mod p) \mod q$;
3. calcola $k^{-*1} \mod q$;
4. calcola $s = k^{-1}[H(m) + xr] \mod q$;
5. la coppia (r, s) è la firma di Alice del messaggio m.

Verifica DSA

Bob verifica la firma di Alice sul messaggio m facendo i seguenti passaggi:

1. verifica che r sia compreso tra 1 e $q-1$ altrimenti rifiuta la firma;
2. calcola $w = s^{-1} \mod q$;
3. calcola $u_1 = wH(m) \mod q$ e $u_2 = rw \mod q$;
4. calcola $v = (\alpha^{u_1} y^{u_2} \mod p)(\mod q)$;
5. accetta la firma se e solo se $v = r$.

I passi di questo algoritmo ci consentono di dimostrare come il meccanismo firma/verifica DSA funzioni correttamente: se (r, s) è la firma legittima di Alice sul messaggio m allora vale $H(m) \equiv -xr + ks \mod q$ e quindi, moltiplicando per w, $wH(m) + xrw \equiv k \mod q$. L'ultima congruenza può essere riscritta come $u_1 + xu_2 \equiv k \mod q$, pertanto se si eleva ad entrambi i membri si ottiene

$$\alpha^{u_1} y^{u_2} \mod p(\mod q) = \alpha^k \mod p \tag{1.3}$$

cioè $v = r$.

Capitolo 2

Curve ellittiche

2.1 Definizioni

Nonostante la crittografia basata sulle curve ellittiche (ECC) sia stata introdotta solo vent'anni fa, lo studio delle curve ellittiche, in particolare nel campo della teoria dei numeri, dell'algebra e della geometria, risale alla metà del XIX secolo. Questi oggetti matematici sono stati impiegati per risolvere problemi di varia natura, come la fattorizzazione di numeri interi ed il problema dei numeri congruenti e test di primalità.

All'inizio degli anni Novanta il matematico Andrew Wiles ha dimostrato l'ultimo teorema di Fermat utilizzando una teoria avanzata delle curve ellittiche.

L'obiettivo di questo capitolo è quello di dare una descrizione matematica delle curve ellittiche con particolare attenzione all'aritmetica basata su curve ellittiche (legge di gruppo, moltiplicazione scalare,...) e verranno presentati i concetti di ordine di curva ed ordine di punto. Infine si studieranno famiglie di curve ellittiche considerate crittograficamente insicure, quindi da evitare nella pratica.

Consideriamo un campo qualsiasi K e la sua chiusura algebrica $\overline{K}$ e sia $K \leq F \leq \overline{K}$.

Definizione 2.1.1. Si definisce piano affine $A^2(K)$ ogni insieme, i cui elementi chiameremo punti, che sia in corrispondenza biunivoca con lo spazio vettoriale K^2. Identificando A con K^2 per mezzo di tale corrispondenza scriveremo direttamente:

$$A^2(K) = \{(x, y) : \ x, y \in K\}$$

Una coppia ordinata di punti (P, Q) è detta *vettore affine* ed è indicata con $\overrightarrow{(P, Q)}$

Definizione 2.1.2. Una curva ellittica E, su un campo F è una curva data dall'equazione nella forma:

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{2.1}$$

con $a_i \in F$.
Questa equazione viene chiamata equazione di Weierstrass.

Denotiamo con $E(F)$ l'insieme dei punti $(x, y) \in F^2$ che soddisfa questa equazione, aggiungendo il cosiddetto punto all'infinito, denotato con O.

Se F non ha caratteristica 2, allora senza perdita di generalità possiamo supporre che $a_1 = a_3 = 0$.

Nel caso che la caratteristica sia 2, abbiamo il caso supersingolare con $y^2 + a_3y = x^3 + a_2x^2 + a_4x + a_6$, il caso non-supersingolare con $y^2 + a_1xy = x^3 + a_2x^2 + a_4x + a_6$. Negli altri casi senza perdita di generalità possiamo supporre che $a_1 = 1$.

In caratteristica 2 possiamo anche supporre che $a_2 = 0$ nel caso supersingolare e che $a_4 = 0$ nel caso non-supersingolare.

Se la caratteristica di F è diversa da 2 e da 3, allora, facendo un cambio di variabile $(x \to x - \frac{1}{3}a_2)$, possiamo anche rimuovere il termine x^2 ottenendo quindi un'equazione nella forma:

$$y^2 = x^3 + ax + b \tag{2.2}$$

con a, $b \in F$, e char $(F) \neq 2, 3$.

Richiedere che la curva sia regolare equivale a richiedere che x^3+ax+b non abbia radici multiple. Quindi segue che il discriminante di $x^3 + ax + b$, che è $\Delta = -(4a^3+27b^2)$ sia diverso da 0. Con il termine regolare si vuole specificare che in $E(F)$ non ci siano punti *singolari*, ossia punti in cui le derivate parziali di (2.1) si annullino contemporaneamente. Per questo motivo le curve regolari sono anche chiamate non-singolari.

Ricordiamo che il discriminante di un polinomio monico di grado d con radici $r_1, ..., r_d$ è

$$\prod_{i \neq j}(r_i - r_j) = (-1)^{\frac{d(d-1)}{2}} \prod_{i<j}(r_i - r_j)^2 \tag{2.3}$$

Per ogni estensione K di F, l'insieme $E(K)$ forma un gruppo abeliano il cui elemento identità è O.

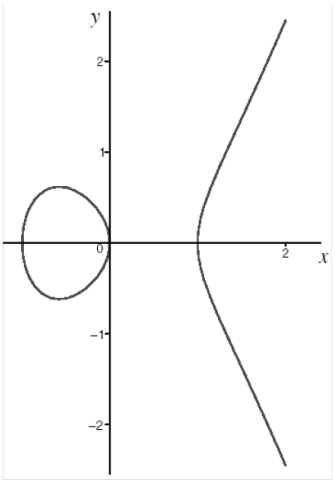


Figura 2.1: grafico della curva $y^2 = x^3 - x$

I due grafici che seguono sono due esempi di curve singolari.

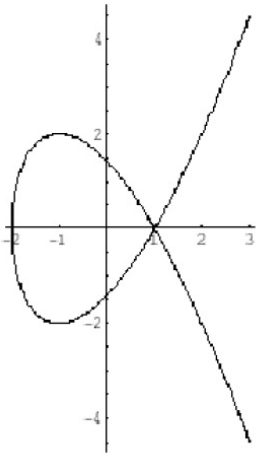


Figura 2.2: curva $y^2 = x^3 - 3x + 2 = (x-1)^2(x+2)$ con nodo nel punto $(1, 0)$

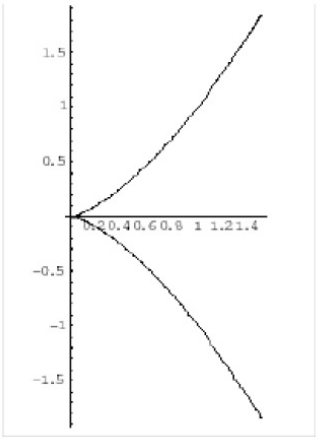


Figura 2.3: curva $y^2 = x^3$ con punto di cuspide in $(0, 0)$

2.2 La legge di gruppo

Definizione 2.2.1. Sia E una curva ellittica sull'insieme dei numeri reali data da un'equazione $y^2 = x^3 + ax + b$, e siano P e Q due punti su E. Definiamo il negativo di P e la somma $P + Q$ secondo le seguenti regole:

1. Se P è il punto all'infinito O, allora definiamo $-P = O$. Per ogni punto Q definiamo $O + Q = Q$; cioè, O è l'elemento neutro della somma.

2. In quello che segue supponiamo che né P né Q siano punti all'infinito. $-P$ è il punto con la stessa coordinata x di P ma con coordinata y opposta; cioè, $-(x, y) = (x, -y)$. Segue dall'equazione $y^2 = x^3 + ax + b$ che $(x, -y)$ appartiene alla curva se anche (x, y) le appartiene.

3. Se P e Q hanno coordinate x diverse, allora consideriamo la retta l passante per i due punti P e Q; questa retta intersecherà la curva in esattamente un altro punto Z della curva. Quindi definiamo $P + Q = -Z = R$. Se invece la retta l fosse verticale segue che $P + Q = O$

4. Se $P = Q$. Allora considero la retta l tangente alla curva. Sia Z l'unico altro punto di intersezione di l con la curva, e definiamo $2P = -Z = R$.

Denotiamo con (x_1, y_1), (x_2, y_2) e (x_3, y_3) le coordinate di P, Q e $P + Q$, rispettivamente. Vogliamo esprimere x_3 e y_3 in termini di x_1, y_1, x_2, y_2. Supponiamo che siamo nel caso in cui i punti P e Q siano distinti punti della curva ellittica e sia $y = \alpha x + \beta$, l'equazione della retta l passante per P e Q, che non sia la retta verticale. Allora segue che:

$$x_3 = \alpha^2 - x_1 - x_2$$

$$y_3 = -y_1 + \alpha(x_1 - x_3)$$

con $\alpha = \frac{y_2 - y_1}{x_2 - x_1}$ coefficiente angolare della retta l.

Il caso in cui $P = Q$ è simile, eccetto che α è la derivata $\frac{dy}{dx}$ in P. La derivazione implicita di $y^2 = x^3 + ax + b$ porta alla formula $\alpha = \frac{3x_1^2 + a}{2y_1}$, e quindi si ottengono le seguenti formule per calcolare le coordinate del punto $2P$:

$$x_3 = \alpha^2 - 2x_1$$

$$y_3 = -y_1 + \alpha(x_1 - x_3)$$

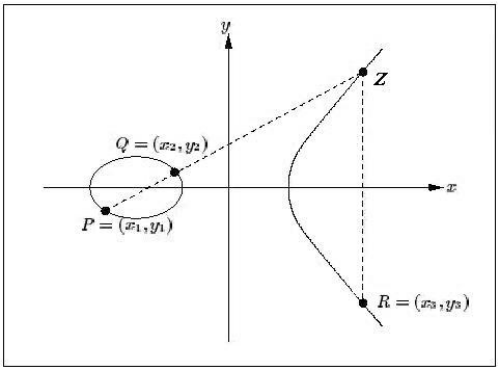


Figura 2.4: P+Q

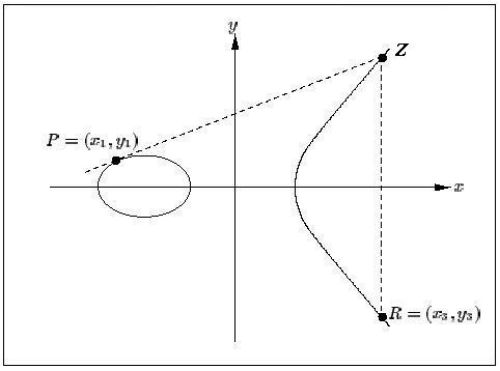


Figura 2.5: 2P

2.2.1 Legge di gruppo per curve su campi di caratteristica 2 o 3

Se la caratteristica del campo F è 2, allora la curva ellittica non può essere messa nella forma $y^2 = x^3 + ax + b$; infatti, la curva non può essere regolare in caratteristica 2. Nel caso di caratteristica 3, non è possibile eliminare il termine a_2x^2 se non è già 0. Così non è possibile usare le formule per sommare i punti visti precedentemente.

Comunque, è possibile trovare delle formule analoghe applicate a curve ellittiche scritte in forma generale, che possono essere usate per ogni caratteristica.

1. Quando $a_1 = a_3 = 0$ ma a_2 non necessariamente 0, così che possiamo lavorare in caratteristica 3:

$$x_3 = \alpha^2 - a_2 - x_1 - x_2$$

$$y_3 = -y_1 + \alpha(x_1 - x_3)$$

quando si sommano punti distinti $P(x_1, y_1)$ e $Q(x_2, y_2)$; e:

$$x_3 = (\frac{3x_1^2 + 2a_2x_1 + a_4}{2y_1})^2 - a_2 - 2x_1$$

$$y_3 = -y_1 + (\frac{3x_1^2 + 2a_2x_1 + a_4}{2y_1})(x_1 - x_3)$$

quando si raddoppia un punto P. (Si noti che in caratteristica 3 il coefficiente angolare α si semplifica a $\frac{a_2x_1 - a_4}{y_1}$);

2. quando $a_3 = a_4 = 0$, ma a_1 è diverso da 0 e può essere assunto uguale a 1, e la caratteristica di F è 2 (caso non-supersingolare):

$$x_3 = (\frac{y_1 + y_2}{x_1 + x_2})^2 + \frac{y_1 + y_2}{x_1 + x_2} + x_1 + x_2 + a_2$$

$$y_3 = \frac{y_1 + y_2}{x_1 + x_2}(x_1 + x_3) + x_3 + y_1$$

quando si sommando due punti distinti; e:

$$x_3 = x_1^2 + \frac{a_6}{x_1^2}$$

$$y_3 = x_1^2 + (x_1 + \frac{y_1}{x_1})x_3 + x_3$$

quando si raddoppia un punto.

3. quando $a_1 = a_2 = 0$ ma $a_3 \neq 0$, e la caratteristica di F è 2 (caso supersingolare):

$$x_3 = (\frac{y_1 + y_2}{x_1 + x_2})^2 + x_1 + x_2$$

$$y_3 = (\frac{y_1 + y_2}{x_1 + x_2})(x_1 + x_3) + y_1 + a_3$$

quando si sommano due punti distinti; e:

$$x_3 = \frac{x_1^4 + a_4^2}{a_3^2}$$

$$y_3 = (\frac{x_1^2 + a_4}{a_3})(x_1 + x_3) + y_1 + a_3$$

quando si raddoppia un punto.

In tutti i casi le regole per sommare un punto $P(x_1, y_1)$ a $Q(x_2, y_2)$ ed ottenere $R = P + Q$ possono essere scritte in forma di espressioni razionali per calcolare x_3 e y_3 in termini di x_1, y_1, x_2, y_2 ed i coefficienti a_i.

Notare che per una curva ellittica scritta in forma generale il negativo di un punto $P(x, y)$ è il punto $-P = (x, -a_1x - a_3 - y)$.

Ottenute queste espressioni possiamo enunciare il seguente:

Teorema 2.2.2. *La somma di punti su una curva ellittica E soddisfa le seguenti proprietà:*

1. *(Commutativa) $P_1 + P_2 = P_2 + P_1$, per ogni $P_1, P_2 \in E$.*
2. *(Esistenza dell'elemento neutro) $P + O = P$, per ogni $P \in E$.*
3. *(Esistenza dell'elemento inverso) Dato $P \in E$, esiste $P' = -P \in E$ con $P + P' = O$.*
4. *(Associativa) $P_1 + (P_2 + P_3) = (P_1 + P_2) + P_3$, per ogni $P_1, P_2, P_3 \in E$.*

I punti di E formano quindi un gruppo abeliano addittivo, dove O è l'elemento neutro.

Dimostrazione.
La proprietà commutativa è ovvia e deriva dal fatto che la retta che unisce P_1 e P_2 è la stessa che unisce P_2 e P_1. Che l'elemento neutro sia O deriva dalla definizione. Per l'inverso, sia $P' = -P$. Allora $P + P' = O$ perché la retta che unisce i due punti è verticale. Per quanto riguarda la proprietà associativa può essere verificata applicando direttamente le formule di addizione.

□

2.2.2 Multipli

Se k è positivo, si indica con kP la somma

$$Q = kP = P + P + ... + P$$

k volte.

Se k è negativo, kP rappresenta $-k(-P)$.

Se invece $k = 0$, si pone $0P = O$.

Esempio:

Per calcolare $17P$ conviene calcolare:

$$2P = P + P$$

$$4P = 2P + 2P$$

$$8P = 4P + 4P$$

$$16P = 8P + 8P$$

infine

$$17P = 16P + P$$

Chiaramente se k è molto grande diventa difficile calcolare velocemente kP. Esistono comunque degli algoritmi per la determinazione dei multipli che sfruttano le proprietà delle curve ellittiche. I protocolli dell'ECC si basano proprio sulla determinazione dei multipli e sulla difficoltà di ricavare k anche se sono noti P e kP. Questo è, infatti, il problema del logaritmo discreto su curva ellittica (ECDLP) che verrà presentato in seguito.

2.3 Ordine di una curva ellittica

Un parametro molto importante e difficile da determinare è il numero di punti razionali su un dato campo di una certa curva ellittica definita su un campo finito. Tale grandezza verrà indicata con N ed è detta *ordine della curva E sul campo finito F*.

Esempio:

Consideriamo la curva E data dall'equazione $y^2 = x^3 + x + 1$ sul campo finito F_5. Il numero di punti si ottiene in questo modo: si elencano tutti i possibili valori che può assumere x, si calcolano i valori $x^3 + x + 1 \mod 5$ e successivamente si determina, se esiste, la radice quadrata.

x	$x^3 + x + 1 \ (\text{mod}5)$	y	Punti
0	1	± 1	$(0,1), (0,4)$
1	3		
2	1	± 1	$(2,1), (2,4)$
3	1	± 1	$(3,1), (3,4)$
4	4	± 2	$(4,2), (4,3)$

Figura 2.6: punti di $E(F)$

Contando anche il punto O, si ottiene che $N = 9$.

Naturalmente, se consideriamo un campo finito F_q di $q = p^f$ elementi, con q grande, questo metodo non è conveniente. E' chiaro che vi sono al più $2q + 1$ punti. Infatti si possono assegnare ad x tutti i valori del campo F_q e per ognuno di essi vi sono al più due y corrispondenti. Se poi si aggiunge il punto all'infinito O, si ottiene proprio $2q + 1$.

Il seguente teorema fornisce dei limiti più precisi per il calcolo di N.

Teorema 2.3.1 (Hasse)**.** *Il numero N dei punti di F_q appartenenti a una curva ellittica definita su F_q sta nell'intervallo*

$$q + 1 - 2\sqrt{q} \leq N \leq q + 1 + 2\sqrt{q}$$

L'intervallo $[q + 1 - 2\sqrt{q}, q + 1 + 2\sqrt{q}]$ è chiamato intervallo di Hasse.

Capitolo 3

La crittografia ellittica

3.1 Storia

La proposta di usare le curve ellittiche per creare dei sistemi crittografici risale al 1985 da parte di Victor Miller e Neal Koblitz. I vantaggi sono la grande flessibilità nella scelta del gruppo e l'assenza di algoritmi con tempo sub-esponenziale in grado di rompere il sistema, a condizione che E venga scelta in modo opportuno.

Inizialmente la crittografia basata sulle curve ellittiche sembrava non poter essere applicata, se non in un distante futuro, ma tuttavia, come spesso accade in crittografia, il futuro è arrivato velocemente, infatti oggi molte persone hanno sviluppato implementazioni utili.

Pochi anni dopo l'invenzione di sistemi crittografici basati su curve ellittiche, Menezes, Okamoto e Vanstone nel 1993 trovarono un nuovo modo di affrontare il problema del logaritmo discreto su cui si basa la sicurezza di questi sistemi crittografici.

Vale a dire, data una curva ellittica E definita su F_q, usarono la Weil pairing per incastrare E in un gruppo moltiplicativo di un'estensione F_{q^k}. Questo riduce il problema al problema del logaritmo discreto in ${F_{q^k}}^*$. Tuttavia, in modo che questo sia di qualche utilità l'estensione di grado k deve essere piccola. In sostanza solo se le curve ellittiche per i quali k è piccolo sono quelle supersingolari.

Così la domanda di fondo in crittografia ellittica è se si può trovare un algoritmo di tempo sub-esponenziale per il problema del logaritmo discreto su alcune classi di curve ellittiche non-supersingolari. Al momento nessuno sembra avere la risposta. Nel frattempo, a causa del progresso in calcolo di logaritmi discreti in campi finiti e in fattorizzazione di interi, la dimensione

delle chiavi necessarie a garantire la sicurezza per i più popolari sistemi a chiave pubblica, è in crescita sostanziale.

3.2 Scambio di chiavi e trasmissione di messaggi

Una dei più importanti usi di un sistema a chiave pubblica è per lo scambio di chiavi. La chiave può essere un intero più o meno 'casuale' che i due utenti Alice e Bob concordano ma nessuno altra persona conosce. L'unica caratteristica della crittografia a chiave pubblica per lo scambio di chiavi è che Alice e Bob possono arrivare alla condivisione delle chiavi usando solo la comunicazione in chiaro.

Il primo sistema a chiave pubblica fu lo scambio di chiavi Diffie-Hellman. Può essere adattato per le curve ellittiche come segue.

Notare prima che un punto casuale su una curva ellittica E può servire come chiave, dato che Alice e Bob sono d'accordo preliminarmente sul metodo per convertirlo in un intero. (Per esempio, potrebbero prendere l'immagine della coordinata x con una determinata mappa da F_q in N).

Supponiamo che E è una curva ellittica su un campo F_q, e Q è un punto concordato (e noto pubblicamente) sulla curva. Alice sceglie segretamente un intero casuale k_A e calcola il punto k_AQ, che manda a Bob.

Allo stesso modo, Bob sceglie segretamente un intero casuale k_B, calcola k_BQ e lo manda ad Alice. La chiave comune è $P = k_Ak_BQ$. Alice calcola P moltiplicando il punto che ha ricevuto da Bob per la sua chiave segreta k_A; Bob calcola P moltiplicando il punto che ha ricevuto per la sua chiave segreta k_B.

Un eventuale intruso che vuole spiare Alice e Bob deve determinare $P = k_Ak_BQ$ conoscendo Q, k_AQ e k_BQ, ma non conoscendo k_A e k_B. Questo compito è chiamato il *problema Diffie-Hellman per le curve ellittiche*.

Non è difficile modificare il protocollo Diffie-Hellman per trasmettere messaggi, usando l'idea di ElGamal. Supponiamo che l'insieme messaggi unità sia stata inserita in una curva ellittica E in qualche modo concordato, e che Bob voglia inviare un messaggio $M \in E$ ad Alice.

Alice e Bob hanno già condiviso k_AQ e k_BQ con la procedura vista precedentemente. Bob ora sceglie un altro intero casuale segreto l, ed invia ad Alice la coppia di punti $(lQ, M + l(k_AQ))$. Per decifrare il messaggio e ricavare M,

Alice moltiplica il primo punto della coppia per il suo intero segreto k_A e poi sottrae il risultato dal secondo punto della coppia.

I sistemi Diffie-Hellman ed ElGamal possono essere rotti se si può risolvere il *problema del logaritmo discreto* nel gruppo E.

Definizione 3.2.1. Il *problema del logaritmo discreto* nel gruppo G con la base $g \in G$ è il problema, dato $y \in G$, di trovare un intero x tale che $g^x = y$ ($xg = y$ quando l'operazione di gruppo in G è scritta additivamente), provando che un tale intero esiste. Così nel caso $G = E$, il *problema del logaritmo discreto su curva ellittica* alla base $Q \in E$ è il problema, dato $P \in E$, di trovare un intero x tale che $P = xQ$ se questo x esiste.

E' facile vedere che il problema Diffie-Hellman può essere risolto se il problema del logaritmo discreto può essere risolto. Un intruso, che conosce Q e k_AQ, trova il segreto k_A e quindi rompe il cifrario. Il contrario, l'affermazione che il problema Diffie-Hellman è *equivalente* al problema del logaritmo discreto è una congettura che non è stata ancora provata anche se gli ultimi risultati trovati sono a sostegno della congettura dell'equivalenza dei due problemi.

3.3 L'algoritmo del logaritmo discreto in gruppi di ordine regolare

Definizione 3.3.1. Sia B un numero reale positivo. Un intero è detto B-regolare se non è divisibile per ogni primo più grande di B.

Se l'ordine del nostro gruppo G è B-regolare per un ragionevole valore B piccolo, allora il logaritmo discreto in G può essere efficientemente calcolato con il metodo di Silver-Pohlig-Hellman.

Sia G un gruppo di ordine $\#G = \prod p_i^{s_i}$. Scriviamo la legge addittiva del gruppo, e denotiamo con O l'elemento neutro. Supponiamo che $\#G$ sia B-regolare; cioè che $p_i \leq B$ per ogni i. Se il limite B è abbastanza piccolo, allora è possibile usare il seguente algoritmo per trovare il logaritmo discreto di $y \in G$ con la base g.

Per prima cosa troviamo l'esatto ordine di g. Può essere fatto calcolando $(\#G/p_i)g$ per i differenti p_i, e poi $(\#G/{p_i}^2)g$ ogni volta che $(\#G/p_i)g = O$, e così via fino a trovare il più piccolo $N = \prod p_i^r$ tale che $Ng = O$.

Il nostro compito è quello di trovare un intero positivo $x < N$ tale che $xg = y$. Se nessun x esiste, allora l'algoritmo che segue si romperà prima, e in tal caso non esisterà soluzione.

Il nostro metodo è trovare $x \mod p_r$, dove p_r è uno delle prime potenze nella fattorizzazione di N, e poi usiamo il teorema del resto cinese per trovare $x \mod N$ con $N = \prod p_i^r$. Così supponiamo che p è un fissato primo divisore di N, e sia $x \equiv x_0 + x_1 p + ... + x_{r-1} p^{r-1} \mod p^r$, con $0 \leq x_i \leq p - 1$ per $i = 0, 1, ..., r - 1$.

Per trovare il valore incognito di x_0, moltiplichiamo entrambi le parti dell'uguaglianza $xg = y$ per $N' = N/p$, ottenendo $x_0(N'g) = N'y$. Potremo provare le p differenti possibilità per x_0 fino a trovare quella per cui l'ultima equazione vale. Se nessuna di tali $x_0 \in 0, 1, ..., p - 1$ esiste, ciò significa che y non è nel sottogruppo generato da g. Questa procedura prevede $O(p)$ passi.

3.4 ECDSA

Ora verrà descritto l'algoritmo di firma digitale del Governo Usa, utilizzando le curve ellittiche (ECDSA). Questo metodo è stato studiato dalle commissioni di diverse organizzazione professionali, e potrebbe presto essere adottato come firma digitale standard per poter essere usato come alternativa efficace al DSA.

3.4.1 Generazione delle chiavi

Per semplicità, useremo curve ellittiche definite su F_p, comunque la costruzione può essere facilmente adattata ad altri campi finiti. Sia E una curva ellittica definita su F_p, e sia P un punto di ordine q in $E(F_p)$. Questi sono i parametri segreti del sistema.

(Si noti che come nel DSA, q non denota una potenza di p, ma un altro numero primo. A differenza che in DSA, dove q è molto più piccolo di p, nel ECDSA q è simile a p).

Alice sceglie un numero intero x casuale nell'intervallo $1 < x < q - 1$ e calcola $Q = xP$. La chiave pubblica di Alice è Q; la sua chiave privata è x.

3.4.2 Firma

Per firmare un messaggio m, Alice fa i seguenti passaggi:

1. sceglie un intero casuale k nell'intervallo $1 < k < q - 1$;
2. calcola $kP = (x_1, y_1)$ e $r = x_1 \mod q$ (cioè, x_1 è considerato come un intero compreso tra 0 e $p - 1$, e r è preso come il minimo residue non negativo modulo q). Se $r = 0$, si ritorna al passo 1.
3. calcola $k^{-1} \mod q$;
4. calcola $s = k^{-1}(H(m) + xr) mod q$, dove $H(m)$ è il valore della funzione Hash del messaggio. Se $s = 0$, si ritorna al passo 1. Si noti che k è scelto casualmente, quindi la probabilità che si abbia $r = 0$ o $s = 0$ è trascurabile.
5. La firma del messaggio m è la coppia di interi (r, s).

3.4.3 Verifica

Bob per verificare la firma di Alice (r, s) del messaggio m, deve:

1. ottenere una copia autentica della chiave pubblica di Alice Q;
2. verificare che r ed s siano interi nell'intervallo $[1, q - 1]$;
3. calcolare $w = s^{-1} \mod q$ e $H(m)$;
4. calcolare $u_1 = H(m)w \mod q$ e $u_2 = rw \mod q$;
5. calcolare $u_1P + u_2Q = (x_0, y_0)$ e $v = x_0 \mod q$;
6. accettare la firma se e solo se $v = r$.

La differenza sostanziale tra ECDSA e DSA è nella generazione di r. Il DSA fa che prendendo una potenza casuale ($\alpha^k \mod p$) e riducendola modulo q, si ottenga un intero nell'intervallo $[1, q - 1]$. (Ricordiamo che in DSA q è un 160-bit primo divisore di $p - 1$, e che α è un elemento di ordine q in F_q^*). L'ECDSA genera l'intero r nell'intervallo $[1, q - 1]$ prendendo la coordinata x del multiplo kP e riducendolo modulo g.

Per ottenere un livello di sicurezza simile a quello del DSA, il parametro q dev'essere almeno di 160 bits. Se questo è il caso, le firme di DSA e ECDSA hanno la stessa lunghezza (320 bits).

3.5 Attacco ai lati e scambio di chiavi

In [10] Billy Bob Brumley e Nicola Tuveri hanno proposto un attacco temporale per lo scambio di chiavi nella crittografia con curve ellittiche (e questo attacco, se non verrà neutralizzato, è potenzialmente devastante). Si ricorda che un attacco ai lati (a side-channel attack) in crittografia è un attacco che usa informazioni che vengono carpite durante l'esecuzione del protocollo. Un attacco ai lati si dice 'temporale' se usa il fatto che in molti protocolli dati diversi usano tempi molto diversi per essere analizzati. Quindi l'attaccante misurando i tempi che usano i vari protocolli ottiene informazioni su quali dati i protocolli stavano usando. Per famosi attacchi temporali a Diffie-Hellman ed RSA vedi [12], [8], [9]. Il caso di ECDSA è particolarmente preoccupante per i seguenti motivi. Di solito ECDSA usa chiavi corte. Infatti, poiché per parametri appropriati non sono noti algoritmi subesponenziali per risolvere il problema del logaritmo discreto, si possono usare chiavi brevi. Inoltre, nell'algoritmo di Montgomery per ECC ci sono tanti vantaggi e tra questi, oltre alla velocità, anche il fatto che 'in ogni iterazione del loop principale sono usate le stesse operazione, potenzialmente aumentando la sua resistenza ad attacchi temporali' ([11], p. 103, citato in [10]). Invece proprio questo è stato usato in [10]. Qui non si descrive [10] e non si sono verificate le sue simulazioni.

Capitolo 4

Curve iperellittiche

In questo capitolo vedremo le principali definizione e le proprietà delle curve iperellittiche e i loro jacobiani.

Sia F un campo, e sia $\overline{F}$ la sua chiusura algebrica.

4.1 Definizioni e proprietà

Definizione 4.1.1. Una *curva iperellittica* C di genere g su F ($g \geq 1$) è un'equazione nella forma

$$C: \quad v^2 + h(u)v = f(u) \tag{4.1}$$

in $F[u, v]$, dove $h(u) \in F[u]$ è un polinomio di grado al massimo g e $f(u) \in F[u]$ è un polinomio monico di grado $2g + 1$. Questa curva deve essere regolare in tutti in punti $(x, y) \in \overline{F} \times \overline{F}$ che soddisfino l'equazione $y^2 + h(x)y = f(x)$ (cioè, nessun punto soddisfi simultaneamente le equazioni $2y + h(x) = 0$ e $h'(x)y - f'(x) = 0$).

Un *punto singolare* di C è una soluzione $(u, v) \in \overline{F} \times \overline{F}$ che simultaneamente soddisfi l'equazione (4.1 e le equazioni delle derivate parziali ($2v + h(u) = 0$ e $h'(u)v - f'(u) = 0$). Si può quindi dire che una curva iperellittica non ha alcun punto singolare.

Lemma 4.1.2. *Sia C una curva iperellittica su F definita dall'equazione (4.1).*

1. *Se $h(u) = 0$, allora $charF \neq 2$.*
2. *Se $char(F) \neq 2$, allora il cambio di variabili $u \to u$, $v \to (v - h(u)/2)$ trasforma C nella forma $v^2 = f(u)$ dove $\deg_u f = 2g + 1$.*

3. *Sia C un'equazione nella forma (4.1) con $h(u) = 0$ e $char(F) \neq 2$. Allora C è una curva iperellittica se e solo se $f(u)$ non ha radici ripetute in $\overline{F}$.*

Dimostrazione.

1. Supponiamo che $h(u) = 0$ e $char(F) = 2$. Allora l'equazione della derivata parziale si riduce a $f'(u) = 0$. Si noti che $\deg_u f'(u) = 2g$. Sia $x \in \overline{F}$ una radice dell'equazione $f'(u) = 0$, e sia $y \in \overline{F}$ una radice dell'equazione $v^2 = f(u)$. Allora il punto (x, y) è un punto singolare di C. Segue quindi (1).

2. Sotto questo cambio di variabile, l'equazione (4.1) si trasforma in:

$$(v - f(u)/2)^2 + h(u)(v - h(u)/2) = f(u)$$

che si semplifica a $v^2 = f(u) + h(u)^2/4$. Quindi si nota che:

$$\deg_u(f + h^2/4) = 2g + 1$$

3. Un punto singolare $(x, y) \in C$ deve soddisfare $y^2 = f(x)$, $2y = 0$ e $f'(x) = 0$. Quindi $y = 0$ e x è una radice ripetuta del polinomio $f(u)$.

□

Definizione 4.1.3. Sia K un'estensione di campo di F. L'*insieme di K-punti razionali su C*, denotato con $C(K)$, è l'insieme di tutti i punti $P = (x, y) \in K \times K$ che soddisfi l'equazione (4.1 della curva C, assieme a un punto speciale chiamato *punto all'infinito* denotato con ∞. L'insieme dei punti $C(\overline{F})$ sarà semplicemente denotato da C. I punti in C, eccetto ∞ sono chiamati *punti finiti*.

Definizione 4.1.4. Sia $P = (x, y)$ un punto finito su una curva iperellittica C. L'*opposto* di P è il punto $\widetilde{P} = (x, -y - h(x))$. Si noti che anche $\widetilde{P}$ appartiene alla curva C. Si può anche definire l'opposto di ∞ come $\widetilde{\infty} = \infty$ anch'esso. Se un punto finito P soddisfa $P = \widetilde{P}$, allora il punto è detto *speciale*; altrimenti, il punto è chiamato *ordinario*.

Definizione 4.1.5. Se $P = (x, y)$ è un K-punto della curva iperellittica, si definisce il suo *opposto* $\widetilde{P}$ come il punto con la stessa coordinata x che soddisfa l'equazione della curva: $\widetilde{P} = (x, -y - h(x))$. Se $P = \infty$, si prende $\widetilde{P} = \infty$.

Definizione 4.1.6. Un *divisore* su C è un somma formale finita di $\overline{F}$-punti $D = \sum m_i P_i$. Il suo *grado* è la somma dei suoi coefficienti $\sum m_i$. Se K è un'estensione algebrica di F, diciamo che D è definito su K se per ogni automorfismo ρ di $\overline{F}$ che fissa K uno ha $\sum m_i P_i^\rho = D$, dove P^ρ denota il punto ottenuto applicando ρ alle coordinate di P (e $\infty^\rho = \infty$). Si denoti con D il gruppo addittivo di divisori definiti su K (dove K è fissato), e denotiamo con D^0 il sottogruppo consistente nei divisori di grado 0.

Definizione 4.1.7. Il *massimo comune divisore* di $D = \sum m_i P_i \in D^0$ e $D' = \sum m'_i P_i \in D^0$ è definito come $(\sum \min(m_i, m'_i) P_i) - (*)\infty$, dove il coefficiente $*$ è scelto tale che il massimo comune divisore abbia grado 0.

Definizione 4.1.8. Dato un polinomio $G(u, v) \in \overline{F}[u, v]$, possiamo considerare $G(u, v)$ come una funzione sulla curva (o in modo equivalente come un elemento dell'anello quoziente $\overline{F}[u, v]/(v^2+h(u)v-f(u))$. Questo significa che si abbassa la potenza di v in $G(u, v)$ per mezzo dell'equazione della curva finché non si avrà l'espressione nella forma $G(u, v) = a(u)-b(u)v$. Denotiamo con $G(u, v) = (\sum m_i P_i) - (*)\infty \in D^0$ il *divisore della funzione polinomiale* $G(u, v)$, dove il coefficiente m_i è l'ordine di eliminazione di $G(u, v)$ nel punto P_i.

Definizione 4.1.9. Un divisore nella forma $(G(u, v)) - (H(u, v))$, cioè, il *divisore della funzione razionale* $G(u, v)/H(u, v)$, è chiamato il *divisore principale*. Denotiamo con J (più precisamente $J(K)$, dove K è un campo contenente F), il quoziente del gruppo D^0 dei divisori di grado zero definiti su K con il sottogruppo P di divisori principali che provengono da G, $H \in K[u, v]$. $J = D^0/P$ è chiamato il *jacobiano* della curva.

4.2 Funzioni polinomiali e razionali

Questa sezione introduce proprietà di base su funzioni polinomiali e razionali che si mettono in evidenza quando vengono viste come funzioni di curve iperellittiche.

Definizione 4.2.1. L'anello coordinato di C su F, denotato con $F[C]$, è l'anello quoziente

$$F[C] = F[u, v]/(v^2 + h(u)v - f(u)) \tag{4.2}$$

dove $(v^2 + h(u)v - f(u))$ denota l'ideale in $F[u, v]$ generato dal polinomio $v^2 + h(u)v - f(u)$. Allo stesso modo, l'anello coordinato di C su $\overline{F}$ è definito come

$$\overline{F}[C] = \overline{F}[u,v]/(v^2 + h(u)v - f(u)) \tag{4.3}$$

Lemma 4.2.2. *Il polinomio $r(u,v) = v^2 + h(u)v - f(u)$ è irriducibile su $\overline{F}$, e quindi $\overline{F}[C]$ è un dominio integrale.*

Dimostrazione.
Se $r(u,v)$ fosse riducibile su $\overline{F}$, si fattorizzerebbe come $(v-a(u))(v-b(u))$ per qualche $a, b \in \overline{F}[u]$. Ma allora $\deg_u(ab) = \deg_u f = 2g+1$ e $\deg_u(a+b) = \deg_u h \leq g$, che è impossibile. □

Si osservi che per ogni funzione polinomiale $G(u,v) \in \overline{F}[C]$, possiamo ripetere ogni occorrenza di v^2 con $f(u) - h(u)v$, così da ottenere eventualmente una rappresentazione del tipo

$$G(u,v) = a(u) - b(u)v$$

dove $a(u), b(u) \in \overline{F}[u]$. E' facile vedere che la rappresentazione di $G(u,v)$ in questa forma è unica.

Definizione 4.2.3. Sia $G(u,v) = a(u) - b(u)v$ una funzione polinomiale in $\overline{F}[C]$. Il coniugato di $G(u,v)$ è definito come la funzione polinomiale $\overline{G}(u,v) = a(u) + b(u)(h(u) + v)$.

Definizione 4.2.4. Sia $G(u,v) = a(u) - b(u)v$ una funzione polinomiale in $\overline{F}[C]$. La norma di G è la funzione polinomiale $NG = G\overline{G}$.

La funzione norma sarà usata nelle trasformazioni da domande riguardanti funzioni polinomiali in due variabili in più semplici domande relative e funzioni in una singola variabile.

Lemma 4.2.5. *Siano G, $H \in \overline{F}[C]$ due funzioni polinomiali.*

1. *$N(G)$ è un polinomio in $\overline{F}[u]$.*
2. *$N(\overline{G}) = N(G)$.*
3. *$N(GH) = N(G)N(H)$.*

Dimostrazione.
Sia $G = a - bv$ e $H = c - dv$, dove a, b, c, $d \in \overline{F}[u]$.

1. Ora, $\overline{G} = a + b(h + v)$ e

$$N(G) = G\overline{G} = (a - bv)(a + b(h + v)) = a^2 + abh - b^2 f \in \overline{F}[u]$$

2. Il coniugato di $\overline{G}$ è

$$\overline{\overline{G}} = (a + bh) + (-b)(h + v) = a - bv = G$$

Quindi $N(\overline{G}) = \overline{G}\,\overline{\overline{G}} = \overline{G}G = N(G)$.

3. $GH = (ac + bdf) - (bc + ad + bdh)v$, e il suo coniugato è

$$\begin{aligned}
\overline{GH} &= (ac + bdf) + (bc + ad + bdh)(h + v) \\
&= ac + bdf + bch + adh + bdh^2 + bcv + adv + bdhv \\
&= ac + bc(h + v) + ad(h + v) + bd(h^2 + hv + f) \\
&= ac + bc(h + v) + ad(h + v) + bd(h^2 + 2hv + v^2) \\
&= (a + b(h + v))(c + d(h + v)) = \overline{GH}
\end{aligned}$$

Quindi $N(GH) = GH\overline{GH} = GH\overline{G}\,\overline{H} = G\overline{G}H\overline{H} = N(G)N(H)$.

□

Definizione 4.2.6. Il campo della funzione $F(C)$ di C su F è il campo delle frazioni di $F[C]$. Allo stesso modo, il campo della funzione $\overline{F}(C)$ di C su $\overline{F}$ è il campo delle frazioni di $\overline{F}[C]$. Gli elementi di $\overline{F}(C)$ sono chiamati *funzioni razionali* su C.

Si noti che $\overline{F}[C]$ è un sottoanello di $\overline{F}(C)$ per esempio ogni funzione polinomiale è anche una funzione razionale.

Definizione 4.2.7. Siano $R \in \overline{F}(C)$ e $P \in C$ con $P \neq \infty$. Allora R è detto essere *definito su* P se esistono due funzioni polinomiali $G, H \in \overline{F}[C]$ tali che $R = G/H$ e $H(P) \neq 0$; se non esistono tali G e H, allora R non è definito su P. Se R è definita su P, il valore di R su P è definito come $R(P) = G(P)/H(P)$.

E' facile vedere che il valore $R(P)$ è ben definito, per esempio, se non dipende dalla scelta di G e H. La seguente definizione introduce la nozione del grado di una funzione polinomiale.

Definizione 4.2.8. Sia $G(u, v) = a(u) - b(u)v$ una funzione polinomiale non zero in $\overline{F}[C]$. Il *grado* di G è definito come:

$$\deg(G) = \max\{2\deg_u(a), 2g + 1 + 2\deg_u(b)\}$$

Lemma 4.2.9. *Siano $G, H \in \overline{F}[C]$.*

1. $\deg(G) = \deg_u(N(G))$.
2. $\deg(GH) = \deg(G) + \deg(H)$.
3. $\deg(G) = \deg(\overline{G})$.

Dimostrazione.

1. Sia $G = a(u) - b(u)v$. La norma di G è $N(G) = a^2 + abh - b^2 f$. Siano $d_1 = \deg_u(a(u))$ e $d_2 = \deg_u(b(u))$. Con la definizione di curva iperellittica, $\deg_u(h(u)) \leq g$, e $\deg_u(f(u)) = 2g + 1$. Ora ci sono due casi da considerare:

   - Se $2d_1 > 2g + 1 + 2d_2$ allora $2d_1 \geq 2g + 2 + 2d_2$, e quindi $d_1 \geq g + 1 + d_2$. Quindi

     $$\deg_u(a^2) = 2d_1 \geq d_1 + g + 1 + d_2 > d_1 + d_2 + g \geq \deg_u(abh)$$

   - Se $2d_1 < 2g + 1 + 2d_2$ allora $2d_1 \leq 2g + 2d_2$, e quindi $d_1 \leq g + d_2$. Così,

     $$\deg_u(abh) \leq d_1 + d_2 + g \leq 2g + 2d_2 + 1 = \deg_u(b^2 f)$$

     Segue quindi che:

     $$\deg_u(N(G)) = \max(2d_1, 2g + 1 + 2d_2) = \deg(G)$$

2. Abbiamo:

   $\deg(GH) = \deg_u(N(GH))$, da 1) $= \deg_u(N(G)N(H))$, da parte (3) del lemma (4.2.5) $= \deg_u(N(G)) + \deg_u(N(H)) = \deg(G) + \deg(H)$.

3. Da $N(G) = N(\overline{G}$, segue che $\deg(G) = \deg_u(N(G)) = \deg_u(N(\overline{G})) = \deg(\overline{G})$.

□

Definizione 4.2.10. Sia $R = G/H \in \overline{F}(C)$ una funzione razionale.

1. Se $\deg(G) < \deg(H)$ allora $R(\infty) = 0$.
2. Se $\deg(G) > \deg(H)$ allora R non è definito in ∞.
3. Se $\deg(G) = \deg(H)$ allora $R(\infty)$ si definisce come il rapporto del coefficiente di testa (per quanto riguarda la funzione grado), di G ed H.

4.3 Zeri e poli

In questa sezione vengono introdotti i concetti di ordine di zeri e poli di funzioni razionali.

Definizione 4.3.1. Sia $R \in \overline{F}(C)$ una funzione razionale non zero, e sia $P \in C$. Se $R(P) = 0$ allora si dice che R ha uno zero in P. Se R non è definita in P allora si dice che R ha un polo in P, in tal caso si scrive $R(P) = \infty$.

Lemma 4.3.2. *Sia $G \in \overline{F}[C]$ una funzione polinomiale non zero, e sia $P \in C$ un punto. Se $G(P) = 0$, allora $\overline{G}(\widetilde{P}) = 0$.*

Lemma 4.3.3. *Sia $P = (x, y)$ un punto speciale di C. Allora $(u - x)$ può essere scritto nella forma $(v - y)^2 S(u, v)$, dove $S(u, v) \in \overline{F}(C)$ non ha né uno zero né un polo in P.*

Dimostrazione.
Sia $G = a(u) - b(u)v$ e $P = (x, y)$. Allora $\overline{G} = a(u) + b(u)(v + h(u))$, $\widetilde{P} = (x, -y - h(x))$, e $\overline{G}(\widetilde{P}) = a(x) + b(x)(-y - h(x) + h(x)) = a(x) - yb(x) = G(P) = 0$.

□

Teorema 4.3.4. *Sia $P \in C$. Allora esiste una funzione $U \in \overline{F}(C)$ con $U(P) = 0$ tale che valga la seguente proprietà: per ogni funzione polinomiale non zero $G \in \overline{F}[C]$, esiste un intero d e una funzione $S \in \overline{F}(C)$ tale che $S(P) \neq 0, \infty$, e $G = U^d S$. Inoltre, il numero d non dipende dalla scelta di U. La funzione U è chiamata parametro uniforme per P.*

Dimostrazione.
Sia $G(u, v) \in \overline{F}[C]$ una funzione polinomiale non zero. Se P è un punto finito, supponiamo che $G(P) = 0$; se $P = \infty$, supponiamo che $G(P) = \infty$. (Se $G(P) \neq 0, \infty$, allora possiamo scrivere $G = U^0 G$ dove U è ogni polinomio in $\overline{F}[C]$ che soddisfi $U(P) = 0$.) Dimostriamo il teorema cercando un parametro uniforme per ognuno dei casi seguenti:

1. $P = \infty$;
2. P è un punto ordinario; e
3. P è un punto speciale.

1. Mostriamo che un parametro uniforme per il punto $P = \infty$ è $U = u^g/v$. Per prima cosa notiamo che $U(\infty) = 0$ per cui $\deg(u^g) < \deg(v)$. Poi, scriviamo

$$G = (u^g/v)^d (v/u^g)^d G$$

dove $d = -\deg(G)$. Sia $S = (v/u^g)^d G$. Da cui $\deg(v) - \deg(u^g) = 2g + 1 - 2g = 1$ e $d = -\deg(G)$, segue che $\deg(u^{gd}G) = \deg(v^{-d})$. Quindi $S(\infty) \neq 0, \infty$.

2. Assumiamo ora che $P = (x, y)$ è un punto ordinario. Mostriamo che un parametro uniforme per P è $U = (u - x)$; osserviamo che $U(P) = 0$. Scriviamo $G = a(u) - b(u)v$. Sia $u - x)^r$ la massima potenza di $(u - x)$ che divida sia $a(u)$ che $b(u)$, e scriviamo

$$G(u, v) = (u - x)^r (a_0(u) - b_0(u)v)$$

Segue che possiamo scrivere $(a_0(u) - b_0(u)v) = (u - x)^s S$ per qualche intero $s \geq 0$ e qualche $S \in \overline{F}(C)$ tale che $S(P) \neq 0, \infty$. Quindi $G = (u - x)^{r+s} S$ soddisfa la conclusione del teorema con $d = r + s$.

3. Assumiamo ora che $P = (x, y)$ è un punto speciale. Mostriamo che un parametro uniforme per P è $U = (v - y)$; osserviamo che $U(P) = 0$. Replicando una potenza di u maggiore di $2g$ con l'equazione della curva, possiamo scrivere

$$G(u, v) = u^{2g} b_{2g}(v) + u^{2g-1} b_{2g-1}(v) + ... + u b_1(v) + b_0(v)$$

dove ogni $b_i(v) \in \overline{F}[v]$.

Replicando tutte le occorrenze di u con $((u - x) + x)$ e espandendo, otteniamo

$$G(u, v) = (u-x)^{2g} \overline{b}_{2g}(v) + (u-x)^{2g-1} \overline{b}_{2g-1}(v) + ... + (u-x)\overline{b}_1(v) + \overline{b}_0(v)$$

$$= (u - x)B(u, v) + \overline{b}_0(v)$$

dove ogni $\overline{b}_i(v) \in \overline{F}[v]$, e $B(u, v) \in \overline{F}[C]$. Ora $G(P) = 0$ implica $\overline{b}_0(y) = 0$, e quindi possiamo scrivere $\overline{b}_0(v) = (v - y)c(v)$ per qualche

$c \in \overline{F}[v]$. Con la dimostrazione del lemma (4.3.3), possiamo scrivere $(u-x) = (v-y)^2/A(u,v)$, dove $A(u,v) \in \overline{F}[C]$ e $A(P) \neq 0, \infty$. Quindi

$$G(u,v) = (v-y)[\frac{(v-y)B(u,v)}{A(u,v)} + c(v)] =$$

$$= \frac{(v-y)}{A(u,v)}[(v-y)B(u,v) + A(u,v)c(v)] = \frac{(v-y)}{A(u,v)}G_1(u,v)$$

Ora se $G_1(P) \neq 0$, allora abbiamo fatto, altrimenti possiamo prendere $S = G_1/A$. D'altra parte, se $G_1(P) = 0$, allora $c(y) = 0$ e possiamo scrivere $c(v) = (v-y)c_1(v)$ per qualche $c_1 \in \overline{F}[v]$. Quindi:

$$G = (v-y)^2[\frac{B(u,v)}{A(u,v)} + c_1(v)] = \frac{(v-y)^2}{A(u,v)}[B(u,v) + A(u,v)c_1(v)] = \frac{(v-y)^2}{A(u,v)}G_2(u,v)$$

Ancora, se $G_2(P) \neq 0$, abbiamo fatto. Altrimenti, il processo può essere ripetuto. Per vedere quanto il processo termina, supponiamo che noi abbiamo estratto k fattori $(v-y)$. Ci sono due casi da considerare:

- Se k è pari, diciamo $k = 2l$, e scriviamo

  $$G = \frac{(v-y)^{2l}}{A(u,v)^l}D(u,v)$$

  dove $D \in \overline{F}[C]$. Quindi, $A^l G = (v-y)^{2l}D = (u-x)^l A^l D$, quindi $G = (u-x)^l D$. Prendendo la norma da entrambe le parti, abbiamo $N(G) = (u-x)^{2l}N(D)$. Quindi $k \leq \deg_u(N(G))$.

- Se k è dispari, diciamo $k = 2l+1$, e scriviamo

  $$G = \frac{(v-y)^{2l+1}}{A(u,v)^{l+1}}D(u,v)$$

  dove $D \in \overline{F}[C]$. Quindi, $A^{l+1}G = (v-y)^{2l+1}D = (u-x)^l A^l(v-y)D$, segue che $AG = (u-x)^l(v-y)D$. Prendendo la norma da entrambe le parti, abbiamo $N(AG) = (u-x)^{2l}N(v-y)N(D)$. Quindi $2l < \deg_u(N(AG))$, e così $k \leq \deg_u(N(AG))$.

  In altri casi, k è limitato da $\deg_u(N(AG))$, e così il processo deve terminare.

  Per vedere come d è indipendente dalla scelta di U, supponiamo che U_1 è un altro parametro uniforme per P.

Poiché $U(P) = U_1(P) = 0$, possiamo scrivere $U = U_1^a A$ e $U_1 = U^b B$, dove $a \geq 1$, $b \geq 1$, A, $B \in \overline{F}(C)$, $A(P) \neq 0, \infty$, $B(P) \neq 0, \infty$. Così $U = (U^b B)^a A = U^{ab} B^a A$. Dividendo entrambe le parti per U, otteniamo $U^{ab-1} B^a A = 1$. Se sostituiamo P in entrambe le parti di questa equazione, vediamo che $ab - 1 = 0$. Quindi $G = U^d S = U_1^d (A^d S)$, dove $A^d S$ non ha né zeri né poli in P.

□

4.4 I divisori

Le definizioni (4.1.6) e (4.1.9) si applicano su ogni curva C. Perché allora noi insistiamo a lavorare con il gruppo jacobiano di una curva ellittica? La prima ragione è che la definizione (4.1.9) è piuttosto astratta. J è infatti definito come il quoziente di un infiniti altri gruppi.

Al fine di impostare calcoli uno ha bisogno di un facile insieme di divisori che rappresentano la classe di equivalenza di D^0 modulo P. Nel caso delle curve iperellittiche uno può verificare che ogni elemento di J può essere rappresentato unicamente con un divisore *ridotto*.

Definizione 4.4.1. Un divisore $D = \sum m_i P_i - (*)\infty \in D^0$ è detto ridotto se:

1. tutti gli m_i sono non negativi, e $m_i \leq 1$ se P_i è uguale al suo opposto.
2. Se $P_i \neq \widehat{P}_i$, allora P_i e $\widehat{P}_i$ non occorrono entrambi nella somma.
3. $\sum m_i \leq g$.

Ogni divisore ridotto $D = \sum m_i P_i - (*)\infty \in D^0$ può essere unicamente rappresentato come il massimo comune divisore del divisore della funzione $a(u) = \prod (u - x_i)^{m_i}$ ed il divisore della funzione $b(u) - v$, dove $P_i = (x_i, y_i)$ è l'unico polinomio di grado minimo di $\deg(a(u))$ tale che $b(x_i) = y_i$, per ogni i e $b(u)^2 + h(u)b(u) - f(u)$ è divisibile per $a(u)$. Se D è rappresentato in questo modo, possiamo scrivere $D = div(a, b)$.

La seconda ragione per la quale si lavora con le curve iperellittiche piuttosto che altre curve generali, è che è relativamente semplice sommare due elementi di J. Più precisamente, dati due divisori ridotti $D_1 = div(a_1, b_1)$ e

$D_2 = div(a_2, b_2)$, non è difficile calcolare il divisore ridotto $D_3 = D_1 + D_2$ nel gruppo J. Questo algoritmo per la somma arriva da Gauss.

C'è un collegamento tra l'esistenza di un algoritmo per addizioni sul jacobiano di una curva iperellittica e l'algoritmo per la composizione di forme quadratiche. Da un punto di vista moderno, le classi di equivalenza di forme quadratiche binarie sono elementi di un gruppo di classe di divisori (chiamate abitualmente *gruppo di classe ideale* del campo quadratico immaginario $Q(\sqrt{d})$ (dove d è il discriminante di una forma quadratica). Allo stesso modo, le curve ellittiche portano la funzione campo consistente in funzioni razionali $G(u, v)/H(u, v)$ considerate modulo la relazione quadratica $v^2 + h(u)v = f(u)$. Questa funzione campo è un'estensione quadratica del campo di base $K(u)$, così come $Q(\sqrt{d})$ è un'estensione quadratica del campo base Q. Comunque, la definizione di jacobiano, il quoziente di divisore di grado 0 per un divisore di funzioni razionali, è analogo alla definizione del gruppo di classe ideale di $Q(\sqrt{d})$ come il quoziente dei divisori (ideali) per il principali ideali generati da elementi di $Q(\sqrt{d})$.

Definizione 4.4.2. Un divisore D è una somma formale di punti su C

$$D = \sum_{P \in C} m_P P \tag{4.4}$$

con $m_P \in Z$, dove solo un numero finito di interi m_P sono diversi da 0. Il *grado* di D, denotato con $\deg D$, è l'intero $\sum_{P \in C} m_P$. L'*ordine* di D su P è l'intero m_P; e si scrive $ord_P(D) = m_P$.

L'insieme di tutti i divisori, denotato con D, forma un gruppo addittivo sotto la regola di addizione:

$$\sum_{P \in C} m_P P + \sum_{P \in C} n_P P = \sum_{P \in C} (m_P + n_P) P \tag{4.5}$$

L'insieme di tutti i divisori di grado 0, denotato con D^0, è un sottogruppo di D.

Definizione 4.4.3. Siano $D_1 = \sum_{P \in C} m_P P$ e $D_2 = \sum_{P \in C} n_P P$ due divisori. Il massimo comune divisore di D_1 e D_2 è definito come:

$$\gcd(D_1, D_2) = \sum_{P \in C} \min(m_p, n_P) P - (\sum_{P \in C} \min(m_P, n_P)) \infty \tag{4.6}$$

Si noti che $\gcd(D_1, D_2) \in D^0$.

Definizione 4.4.4. Sia $R \in \overline{F}(C)$ una funzione razionale non zero. Il divisore di R è

$$div(R) = \sum_{P \in C} (ord_P R)P$$

Si noti che se $R = G/H$ allora $div(R) = div(G) - div(H)$.

Definizione 4.4.5. Un divisore $D \in D^0$ è chiamato *divisore principale* se $D = div(R)$ per qualche funzione razionale non zero $R \in \overline{F}(C)$. L'insieme di tutti i divisori principali, denotato con P, è un sottogruppo di D^0. Il gruppo quoziente $J = D^0/P$ è chiamato *jacobiano* della curva C. Se D_1 e $D_2 \in D^0$ allora scriviamo $D_1 \sim D_2$ se $D_1 - D_2 \in P$; D_1 e D_2 sono chiamati divisori *equivalenti*.

Definizione 4.4.6. Sia $\sum_{P \in C} m_P P$ un divisore. Il *supporto* di D è l'insieme

$$supp(D) = \{P \in C \mid m_P \neq 0\}$$

Definizione 4.4.7. Un divisore *semi-ridotto* è un divisore della forma $D = \sum m_i P_i - (\sum m_i)\infty$, dove ogni $m_i \geq 0$, ed i P_i sono punti finiti tali che quando $P_i \in supp(D)$ uno ha $\widetilde{P}_i \notin supp(D)$, tranne $P_i = \widetilde{P}_i$, nel caso che $m_i = 1$.

Lemma 4.4.8. *Per ogni divisore $D \in D^0$ esiste un divisore semi-ridotto $D_1 \in D^0$ tale che $D \sim D_1$.*

Dimostrazione.

Sia $D = \sum_{P \in C} m_P P$. Sia (C_1, C_2) una partizione dell'insieme di punti ordinari su C tali che:

1. $P \in C_1$ se e solo se $\widetilde{P} \in C_2$; e
2. se $P \in C_1$ allora $m_P \geq m_{\widetilde{P}}$.

Sia C_0 l'insieme dei punti speciali su C. Allora si può scrivere:

$$D = \sum_{P \in C_1} m_P P + \sum_{P \in C_2} m_P P + \sum_{P \in C_0} m_P P - m\infty \tag{4.7}$$

Considerando il seguente divisore:

$$D_1 = D - \sum_{P=(x,y) \in C_2} m_P div(u - x) - \sum_{P=(x,y) \in C_0} [\frac{m_P}{2}] div(u - x) \tag{4.8}$$

Allora $D_1 \sim D$.

Infine, abbiamo:

$$D_1 = \sum_{P \in C_1} (m_P - m_{\widetilde{P}})P + \sum_{P \in C_0} (m_P - 2[\frac{m_P}{2}])P - m_1\infty \tag{4.9}$$

per qualche intero $m_1 \geq 0$, e quindi D_1 è un divisore semi-ridotto. □

4.5 Rappresentazione semi-dirette di classi di equivalenza di divisori

Questa sezione descrive la rappresentazione polinomiale per divisori semi-ridotti del jacobiano. L'obiettivo è arrivare ad un algoritmo per sommare gli elementi del jacobiano.

Lemma 4.5.1. *Sia $P = (x, y)$ un punto ordinario di C, e sia $R \in \overline{F}(C)$ una funzione razionale che non abbia un polo in P. Allora per ogni $k \geq 0$, esistono degli elementi unici $c_0, c_1, ..., c_k \in \overline{F}$ e $R_k \in \overline{F}(C)$ tali che $R = \sum_{i=0}^{k} c_i(u-x)^i + (u-x)^{k+1}R_k$, dove R_k non ha un polo in P.*

Dimostrazione.
Esiste un unico $c_0 \in \overline{F}$, chiamato $c_0 = R(x, y)$, tale che P è uno zero di $R - c_0$. Poichè $(u - x)$ è una parametrizzazione uniforme per P, e possiamo scrivere $R - c_0 = (u-x)R_1$ per qualche (unico) $R_1 \in \overline{F}(C)$ con $ord_P(R_1) \geq 0$. Quindi $R = c_0 + (u - x)R_1$. Il lemma ora segue per induzione.

□

Nel prossimo lemma, quando scriviamo $\mod (u-x)^k$, intendiamo modulo l'ideale generato da $(u-x)^k$ nel sottoanello di $\overline{F}(C)$ consistente nelle funzioni razionali che non hanno un polo in P. Così, la conclusione nel lemma (4.5.1) può essere riscritto:

$$R \equiv \sum_{i=0} kc_i(u-x)^k(\mod (u-x)^{k+1}) \tag{4.10}$$

Lemma 4.5.2. *Sia $P = (x, y)$ un punto ordinario di C. Allora per ogni $k \geq 1$, esiste un unico polinomio $b_k(u) \in \overline{F}[u]$ tale che*

1. $\deg_u b_k < k$
2. $b_k(x) = y$; *e*

3. $b_k^2 + b_k(u)h(u) \equiv f(u)(\mod (u-x)^k)$.

Dimostrazione.

Applicando il lemma (4.5.1) a $R(u,v) = v$. Sia $v = \sum_{i=0} k-1 c_i(u-x)^i + (u-x)^k R_{k-1}$, dove $c_i \in \overline{F}$ e $R_{k-1 \in \overline{F}(C)}$. Definiamo $b_k(u) = \sum_{i=0} k-1 c_i(u-x)^i$. Dalla dimostrazione del lemma (4.5.1), sappiamo che $c_0 = y$, e quindi $b_k(x) = y$. Infine, da $v^2 + h(u)v = f(u)$, se riduciamo entrambi le parti modulo $(u-x)^k$ otteniamo $b_k(u)^2 + b_k(u)h(u) \equiv f(u)(\mod (u-x)^k)$. L'unicità è facilmente provata per induzione su k. □

Il seguente teorema mostra come un divisore semi-ridotto può essere rappresentato come il massimo comune divisore dei divisori di due funzioni polinomiali.

Teorema 4.5.3. *Sia* $D = \sum m_i P_i - (\sum m_i)\infty$ *un divisore semi-ridotto, dove* $P_i = (x_i, y_i)$. *Sia* $a(u) = \prod (u - x_i)^{m_i}$. *Allora esiste un unico polinomio* $b(u)$ *che soddisfi:*

1. $\deg_u b < \deg_u a$;
2. $b(x_i) = y_i$ *per tutti gli* i *per i quali* $m_i \neq 0$; *e*
3. $a(u)$ *divide* $(b(u)^2 + b(u)h(u) - f(u))$.

Allora $D = \gcd(div(a(u)), div(b(u) - v))$.

Notazione: $\gcd(div(a(u)), div(b(u) - v))$ *è abitualmente abbreviato come* $div(a(u), b(u) - v)$ *o, ancora più semplicemente, come* $div(a, b)$.

Dimostrazione.

Sia C_1 l'insieme dei punti ordinari in $supp(D)$, e sia C_0 l'insieme dei punti speciali in $supp(D)$. Sia $C_2 = \left\{ \widetilde{P} : P \in C_1 \right\}$. allora possiamo scrivere

$$D = \sum_{P_i \in C_0} P_i + \sum_{P_i \in C_1} m_i P_i - m\infty$$

dove m_i, m sono interi positivi. Per prima cosa proviamo che esiste un unico polinomio $b(u)$ che soddisfi le condizioni del teorema. Con il lemma (4.5.2), per ogni $P_i \in C_1$ esiste un unico polinomio $b_i(u) \in \overline{F}[u]$ che soddisfi:

1. $\deg_u b_i < m_i$;
2. $b_i(x_i) = y_i$; e
3. $(u - x_i)^{m_i}$ divida $b_i^2(u) + b_i(u)h(u) - f(u)$.

Si può facilmente verificare che per ogni $P_i \in C_0$, $b_i(u) = y_i$ è l'unico polinomio che soddisfi:

1. $\deg_u b_i < 1$;
2. $b_i(x_i) = y_i$; e
3. $(u - x_i)$ divida $b_i^2(u) + b_i(u)h(u) - f(u)$.

Dal teorema del resto cinese per i polinomi segue che esiste un unico polinomio $b(u) \in \overline{F}[u]$, con $\deg_u b < \sum m_i$, tale che

$$b(u) \equiv b_i(u)(\mod (u - x_i)^{m_i})$$

per ogni i.

Può ora essere verificato che $b(u)$ soddisfa le condizioni 1, 2 e 3 del teorema.

Poi,

$$div(a(u)) = div(\prod (u - x_i)^{m_i}) = \sum_{P_i \in C_0} 2P_i + \sum_{P_i \in C_1} m_i P_i + \sum_{P_i \in C_1} m_i \widetilde{P}_i - (*)\infty$$

In aggiunta

$$div(b(u) - v) = \sum_{P_i \in C_0} t_i P_i + \sum_{P_i \in C_1} s_i P_i + \sum_{P_i \in C \setminus C_0 \cup C_1 \cup C_2 \cup \{\infty\}} m_i P_i - (*)\infty$$

dove ogni $s_i \geq m_i$, da qui $(u - x_i)^{m_i}$ divide $N(b - v) = b^2 + hb - f$. Ora se $P = (x, y) \in C_0$, allora $(u - x)$ divide $b^2 + bh - f$. La derivata di questo polinomio valutato in $u = x$ è

$$2b(x)b'(x) + b'(x)h(x) + b(x)h'(x) - f'(x)$$

$$= b'(x)(2y + h(x)) + (h'(x)y - f'(x))$$

$$h'(x)y - f'(x) \neq 0$$

da cui $2y + h(x) = 0$.

Così, $u = x$ è una radice semplice di $N(b - v) = b^2 + bh - f$, e quindi $t_i = 1$ per ogni i.

Infine

$$\gcd(a(u), b(u) - v) = \sum_{P_i \in C_0} P_i + \sum_{P_i \in C_1} m_i P_i - m\infty = D$$

come richiesto.

□

Si noti che il divisore zero è rappresentato come $div(1, 0)$. Il prossimo risultato segue dalla dimostrazione del teorema (4.5.3).

Lemma 4.5.4. *Siano $a(u)$, $b(u) \in \overline{F}[u]$ tali che $\deg_u b < \deg_u a$. Se a divide $(b^2 + bh - f)$, allora $div(a, b)$ è semi-ridotto.*

4.6 Divisori ridotti

Definizione 4.6.1. Sia $D = \sum m_i P_i - (\sum m_i)\infty$ un divisore semi-ridotto. Se $\sum m_{\leq} g$ (g è il genere di C), allora D è chiamato *divisore ridotto.*

Definizione 4.6.2. Sia $D = \sum_{P \in C} m_P P$ un divisore. La *norma* di D è definita come

$$|D| = \sum_{P \in C \setminus \{\infty\}} |m_P|$$

Si noti che dato un divisore $D \in D^0$, l'operazione descritta nella dimostrazione del lemma (4.4.8) produce un divisore semi-ridotto D_1 tale che $D_1 \sim D$ e $|D_1| \leq |D|$.

Lemma 4.6.3. *Sia R una funzione razionale non zero in $\overline{F}(C)$. Se R non ha poli finiti, allora R è una funzione polinomiale.*

Dimostrazione.

Sia $R = G/H$, dove G e H sono funzioni polinomiali non zero in $\overline{F}(C)$. Allora $R = \frac{G}{H}\frac{\overline{H}}{\overline{H}} = G\overline{H}/N(H)$, e allora possiamo scrivere $R = (a - bv)/c$, dove a, b, $c \in \overline{F}([u]$, con $c \neq 0$. Sia $x \in \overline{F}$ una radice di c. Sia $P = (x, y) \in C$, dove $y \in \overline{F}$, e sia $d \geq 1$ la massima potenza di $u - x$ che divide c. Se P è ordinario, allora $ord_P(c) = ord_{\widetilde{P}} = d$. Poiché R non ha poli finiti, $ord_P(a - bv) \geq d$ e $ord_P(a - bv) \geq d$. Ora poiché P e $\widetilde{P}$ sono entrambi zero di $a - bv$, abbiamo $a(x) = 0$ e $b(x) = 0$. Segue che $ord_P(a) \geq d$ e $ord_P(b) \geq d$. Quindi $(u - x)^d$ è un divisore comune di a e b, e può essere eliminato con il fattore $(u - x)^d$ di c. Supponiamo ora che P è speciale. Allora $ord_P(c) = 2d$. Poiché R non ha poli finiti, $ord_P(a - bv) \geq 2d$. Allora, come nella parte (3) della dimostrazione del teorema (4.3.4), possiamo scrivere

$$a - bv = \frac{(v-y)^{2d}D}{A^d}$$

dove A e D sono funzioni polinomiali non zero in $\overline{F}[C]$, e A soddisfa $v-y)^2 = (u-x)A$. Quindi $a - bv = (u-x)^d D$. Ancora, il fattore $(u-x)^d$ di $a - bv$ può essere eliminato con il fattore $(u-x)^d$ di c.

Questo può essere ripetuto per tutte le radici di c; segue che R è una funzione polinomiale.

□

Teorema 4.6.4. *Per ogni divisore $D \in D^0$ esiste un unico divisore ridotto D_1 tale che $D \sim D_1$.*

4.7 Addizione di divisori ridotti

Sia C una curva iperellittica di genere g definita su un campo finito F, e sia J il jacobiano di C. Sia $P = (x, y) \in C$, e sia σ un automorfismo di $\overline{F}$ su F. Allora $P^\sigma = (x^\sigma, y^\sigma)$ è anche un punto su C.

Definizione 4.7.1. Un divisore $D = \sum m_P P$ è detto essere *definito* su F se $D^\sigma = \sum m_P P^\sigma$ è uguale a D per tutti gli automorfismi σ di $\overline{F}$ su F.

Un divisore principale è definito su F se e solo se il divisore di una funzione razionale ha una rappresentazione che ha coefficienti in F.

L'insieme J(F) di tutte le classi di divisori in J che hanno un rappresentante che è definito su F è un sottogruppo di J. Ogni elemento di $J(F)$ ha un'unica rappresentazione come un divisore ridotto $div(a, b)$, dove a, $b \in F[u]$, $\deg_u a \leq g$, $\deg_u b < deg_u a$; e quindi $J(F)$ è infatti un gruppo abeliano finito. Questa sezione presenta un efficiente algoritmo per sommare elementi in questo gruppo.

Sia $D_1 = div(a_1, b_1)$ e $D_2 = div(a_2, b_2)$ due divisori ridotti definiti su F (cioè, $a_1, a_2, b_1, b_2 \in F[u]$). L'algoritmo (1) trova un divisore semi-ridotto $D = div(a, b)$ con a, $b \in F[u]$, tale che $D \sim D_1 + D_2$. L'algoritmo (2) riduce D a un divisore ridotto equivalente D'.

Notazione: $b \mod a$ denota il resto polinomiale quando b è diviso da a.

I due algoritmi sono stati presentati in [2](Appendix. An elementary introduction to Hyperelliptic curves (Alfred J. Menezes, Yi-Hong Wu, Robert J. Zuccherato)). Generalizzano l'algoritmo di Cantor 1987, in quanto era assunto che $h(u) = 0$ e $char(F) \neq 2$.

4.7.1 Algoritmo 1

INPUT: Siano $D_1 = div(a_1, b_1)$ e $D_2 = (a_2, b_2)$ due divisori semi-ridotti, definiti entrambi su F.

OUTPUT: $D = div(a, b)$ divisore semi-ridotto definito su F tale che $D \sim D_1 + D_2$.

1. Si usi l'algoritmo di Euclide per trovare i polinomi d_1, e_1, $e_2 \in F[u]$ dove $d_1 = \gcd(a_1, a_2)$ e $d_1 = e_1 a_1 + e_2 a_2$.

2. Si usi l'algoritmo di Euclide per trovare i polinomi d, c_1, $c_2 \in F[u]$, dove $d = \gcd(d_1, b_1 + b_2 + h)$ e $d = c_1 d_1 + c_2(b_1 + b_2 + h)$.

3. Siano $s_1 = c_1 e_1$, $s_2 = c_1 e_2$ e $s_3 = c_2$, così che

$$d = s_1 a_1 + s_2 a_2 + s_3(b_1 + b_2 + h) \tag{4.11}$$

4. Siano

$$a = \frac{a_1 a_2}{d^2} \tag{4.12}$$

e

$$b = \frac{s_1 a_1 b_2 + s_2 a_2 b_1 + s_3(b_1 b_2 + f)}{d} \mod a \tag{4.13}$$

Teorema 4.7.2. *Siano $D_1 = div(a_1, b_1)$ e $D_2 = div(a_2, b_2)$ due divisori semi-ridotti. Siano a e b definiti come nelle equazioni (4.12) e (4.13). Allora $D = div(a, b)$ è un divisore semi-ridotto e $D \sim D_1 + D_2$.*

Dimostrazione. Per prima cosa verifichiamo che b è un polinomio. Usando l'equazione (4.11), possiamo scrivere

$$\frac{s_1 a_1 b_2 + s_2 a_2 b_1 + s_3(b_1 b_2 + f)}{d}$$

$$= \frac{b_2(d - s_2 a_2 s_3(b_1 + b_2 + h)) + s_2 a_2 b_1 + s_3(b_1 b_2 + f)}{d}$$

$$= b_2 \frac{s_2 a_2(b_1 - b_2) - s_3(b_2^2 + b_2 h - f)}{d}$$

Segue che $d | a_2$, e $a_2 | (b_2^2 + b_2 h - f)$, b è infatti un polinomio.
Sia $b = (s_1 a_1 b_2 + s_2 a_2 b_1 + s_3(b_1 b_2 + f))/d + sa$, dove $s \in F[u]$. Ora

$$b - v = \frac{s_1 a_1 b_2 + s_2 a_2 b_1 + s_3(b_1 b_2 + f) - dv}{d} + sa =$$
$$\frac{s_1 a_1 b_2 + s_2 a_2 b_1 + s_3(b_1 b_2 + f) - s_1 a_1 v - s_2 a_2 v - s_3(b_1 + b_2 + h)v}{d} + sa =$$
$$\frac{s_1 a_1 (b_2 - v) + s_2 a_2 (b_1 - v) + s_3 (b_1 - v)(b_2 - v)}{sa} \tag{4.14}$$

Da questo non è difficile vedere che $a | b^2 + bh - f$.

Vale a dire, $b^2 + bh - f$ è ottenuto moltiplicando $(b - v)$ per il suo coniugato $(b - v)(b + v + h) = b^2 + bh - f$. Così, vedere che $a | b^2 + bh - f$ è sufficiente per vedere che a_1 e a_2 dividono il prodotto di $(s_1 a_1 (b_2 - v) + s_2 a_2 (b_1 - v) + s_3 (b_1 - v)(b_2 - v))$ con il suo coniugato; e questo segue perché $a_1 | b_1^2 + b_1 h - f = (b_1 - v)(b_1 + v + h)$ e $a_2 | b_2^2 + b_2 h - f = (b_2 - v)(b_2 + v + h)$.

Il lemma (4.5.4) ora implica che $div(a, b)$ è un divisore semi-ridotto.

Ora proviamo che $D \sim D_1 + D_2$. Ci sono due casi da considerare:

1. Sia $P = (x, y)$ un punto ordinario. Ci sono due sottocasi da considerare:

   - Supponiamo che $ord_P(D_1) = m_1$, $ord_{\widetilde{P}}(D_1) = 0$, $ord_P(D_2) = m_2$, e $ord_{\widetilde{P}}(D_2) = 0$, dove $m_1 \geq 0$, $m_2 \geq 0$. Ora $ord_P(a_1) = m_1$, $ord_P(a_2) = m_2$, $ord_P(b_1 - v) \geq m_1$, e $ord_P(b_2 - v) \geq m_2$. Se $m_1 = 0$ o $m_2 = 0$ (o entrambi) allora $ord_P(d_1) = 0$, da cui $ord_P(d) = 0$ e $ord_P(a) = m_1 + m_2$. Se $m_1 \geq 1$ e $m_2 \geq 1$, allora segue $(b_1 + b_2 + h)(x) = 2y + h(x) \neq 0$, abbiamo $ord_P(d) = 0$ e $ord_P(a) = m_1 + m_2$. Dall'equazione 4.14 segue che

     $$ord_P(b - v) \geq \min\{m_1 + m_2, m_2 + m_1, m_1 + m_2\} = m_1 + m_2$$

     Quindi, $ord_P(D) = m_1 + m_2$.

   - Supponiamo che $ord_P(D_1) = m_1$, e $ord_{\widetilde{P}}(D_2) = m_2$, dove $m_1 \geq m_2 \geq 1$. Abbiamo $ord_P(a_1) = m_1$, $ord_P(a_2) = m_2$, $ord_P(d_1) = m_2$, $ord_P(b_1 - v) \geq m_1$, $ord_P(b_2 - v) = 0$, e $ord_{\widetilde{P}}(b_2 - v) \geq m_2$. L'ultima disequazione implica che $ord_P(b_2 + h + v) \geq m_2$, e quindi $ord_P(b_1 + b_2 + h) \geq m_2$ o $(b_1 + b_2 + h) = 0$. Segue che $ord_P(d) = m_2$ e $ord_P(a) = m_1 - m_2$. Dall'equazione 4.14 segue che

     $$ord_P(b - v) \geq \min\{m_1 + 0, m_2 + 0, m_1 + 0\} - m_2 = m_1 - m_2$$

     Quindi, $ord_P(D) = m_1 - m_2$.

2. Si $P = (x, y)$ un punto speciale. Ci sono due sottocasi da considerare:

   - Supponiamo che $ord_P(D_1) = 1$ e $ord_P(D_2) = 1$. Allora $ord_P(a_1) = 2$, $ord_P(a_2) = 2$, e $ord_P(d_1) = 2$. Ora $(b_1+b_2+h)(x) = 2y+h(x) = 0$, da cui o $ord_P(b_1 + b_2 + h) \geq 2$ o $b_1 + b_2 + h = 0$. Segue che $ord_P(d) = 2$ e $ord_P(a) = 0$. Quindi, $ord_P(D) = 0$.
   - Supponiamo che $ord_P(D_1) = 1$ e $ord_P(D_2) = 0$. Allora $ord_P(a_1) = 2$, $ord_P(a_2) = 0$, da cui $ord_P(d_1) = ord_P(d) = 0$, e $ord_P(a) = 2$. Da $ord_P(b_1 - v) = 1$, segue dall'equazione (4.14) che $ord_P(b - v) \geq 1$. Può essere dedotto dall'equazione (4.14) che $ord_P(b - v) \geq 2$, se e solo se $ord_P(s_2a_2 + s_3(b_2 - v)) \geq 1$. Se questo è il caso, allora $ord_P(s_2a_2 + s_3(b_2 + h + v)) \geq 1$, e quindi $ord_P(s_2a_2 + s_3(b_1 + b_2 + h)) \geq 1$ (o $s_2a_2 + s_3(b_1 + b_2 + h) = 0$). Ora segue dall'equazione (4.11) che $ord_P(d) \geq 1$, una contraddizione. Quindi $ord_P(b - v) = 1$, da cui $ord_P(D) = 1$.

□

4.7.2 Algoritmo 2

INPUT: Un divisore semi-ridotto $D = div(a, b)$ definito su F.

OUTPUT: L'unico divisore ridotto $D' = div(a', b')$ tale che $D' \sim D$.

1. Posti $a' = (f - bh - b^2)/a$ e $b' = (-h - b) \mod a'$
2. Se $\deg_u a' > g$ allora si pone $a \leftarrow a'$, $b \leftarrow b'$ e si torna al passo 1.
3. Sia c il coefficiente di testa di a', e si pone $a' \leftarrow c^{-1}a'$.
4. output(a', b').

Teorema 4.7.3. *Sia $D = div(a, b)$ un divisore semi-ridotto. Allora il divisore $D' = div(a', b')$ output dell'algoritmo 2 è ridotto, e $D' \sim D$.*

Dimostrazione.
Siano $a' = (f - bh - b^2)/a$ e $b' = (-h - b) \mod a'$. Mostriamo che:

1. $\deg_u(a') < \deg_u(a)$;
2. $D' = div(a', b')$ è semi-ridotto; e
3. $D \sim D'$.

L'algoritmo ora segue da applicazioni ripetute del processo di riduzione (passo 1 dell'algoritmo 2).

1. Sia $m = \deg_u a$, $n = \deg_u b$, dove $m > n$ e $m \geq g+1$. Allora $\deg_u a' = \max\{2g+1, 2n\} \leq 2(m-1)$, da cui $\deg_u a' \leq m-2 < \deg_u a$. Se $m = g+1$, allora $\max\{2g+1, 2n\} = 2g+1$, da cui $\deg_u a' = g < \deg_u a$.

2. Ora $f - bh - b^2 = aa'$. Riducendo entrambe le parti modulo a', otteniamo

$$f + (b'+h)h - (b'+h)^2 \equiv 0(\mod a')$$

che si semplifica come

$$f - b'h - (b')^2 \equiv 0(moda')$$

Quindi $a'|(f - b'h - (b')^2)$. Segue dal lemma (4.5.4) che $div(a', b')$ è semi-ridotto.

3. Siano $C_0 = \{P \in supp(D) : P\ speciale\}$, $C_1 = \{P \in supp(D) : P\ ordinario\}$, $C_2 = \{P \in supp(D) : P \in C_1\}$, così che

$$D = \sum_{P_i \in C_0} P_i + \sum_{P_i \in C_i} m_i P_i - (*)\infty$$

Allora, come nella dimostrazione del teorema (4.5.3), possiamo scrivere

$$div(a) = \sum_{P_i \in C_0} 2P_i + \sum_{P_i \in C_1} m_i P_i + \sum_{P_i \in C_1} m_i \widetilde{P}_i - (*)\infty$$

e

$$div(b-v) = \sum_{P_i \in C_0} P_i + \sum_{P_i \in C_1} n_i P_i + \sum_{P_i \in C_1} 0\widetilde{P}_i + \sum_{P_i \in C_3} s_i P_i - (*)\infty$$

dove $n_i \geq m_i$, C_3 è un insieme di punti in $C \backslash (C_0 \cup C_1 \cup C_2 \cup \{\infty\})$, $s_i \geq 1$, e $s_1 = 1$ se P_i è speciale. Da $b^2 + bh - f = N(b-v)$, segue dal lemma (**??**) che

$$div(b^2+bh-f) = \sum_{P_i \in C_0} 2P_i \sum_{P_i \in C_1} n_i P_i + \sum_{P_i \in C_1} n_i \widetilde{P}_i + \sum_{P_i \in C_3} s_i \widetilde{P}_i + \sum_{P_i \in C_3} s_i P_i - (*)\infty$$

e quindi

$div(a') = div(b^2 + bh - f) - div(a) = \sum_{P_i \in C_1'} t_i P_i + \sum_{P_i \in C_1'} t_i \widetilde{P}_i + \sum_{P_i \in C_3} s_i P_i + \sum_{P_i \in C_3} s_i \widetilde{P}_i - (*)\infty$

dove $t_i = n_i - m_i$, e $C_1' = \{P_i \in C_1 : n_i > m_i\}$. Ora $b' = -h - b + sa'$ per qualche $s \in F[u]$. Se $P_i = (x_1, y_i) \in C_1' \cup C_3$, allora $b'(x_1) = -h(x_i) - b(x_i) + s(x_i)a'(x_i) = -h(x_i) - y_i$. Quindi, come nel teorema (4.5.3), segue che

$$div(b' - v) = \sum_{P_i \in C_1'} 0P_i + \sum_{P_i \in C_1'} r_i \widetilde{P}_i + \sum_{P_i \in C_3} 0P_i + \sum_{P_i \in C_3} w_i \widetilde{P}_i + \sum_{P_i \in C_4} z_i P_i - (*)\infty$$

dove $r_i \geq t_i$, $w_i \geq s_i$, $w_i = 1$ se $P_i \in C_3$ è speciale, e C_4 è un insieme di punti in $C \backslash (C_1' \cup C_3 \cup \{\infty\})$. Quindi

$$div(a', b') =$$

$$\sum_{P_i \in C_1'} t_i \widetilde{P}_i + \sum_{P_i \in C_3} s_i \widetilde{P}_i (*)\infty \sim - \sum_{P_i \in C_1'} t_i P_i - \sum_{P_i \in C_3} s_i P_i + (*)\infty =$$

$$= D - div(b - v)$$

da cui $D \sim D'$.

□

4.8 Sistemi crittografici iperellittici

Lo scambio di chiavi basato sulle curve ellittiche Diffie-Hellman e la trasmissione di messaggi ElGamal possono essere applicati al gruppo jacobiano di una curva iperellittica.

Per implementare un sistema crittografico logaritmo discreto con curve iperellittiche, si utilizza una curva adatta C e un campo finito F_q. E' fondamentale che l'ordine $J(F_q)$ del jacobiano di C sia divisibile per un numero primo grande. Dato l'attuale livello di tecnologia informatica, $J(F_q)$ deve essere divisibile da un numero primo l di almeno 40 cifre.

Un'altra considerazione è che per avere un implementazione efficiente è conveniente utilizzare un campo finito di caratteristica 2.

4.9 Breve storia di credenze ed intuizioni sbagliate

Questa è la traduzione del titolo dell'intervento di N. Koblitz come main speaker in un importante convegno di crittografia ([3]). Le sue parole qui riassunte sono in sintonia perfetta con le ultime 3 righe della sezione 6 (Conclusions) di [10]. Il lettore dovrebbe tener presente che ogni volta che in queste pagine Koblitz fa la figura dello stupido, be' è lui che nel suo talk usa se stesso per mostrare cosa non fare.

Nel 1980 sembrava che ogni gruppo di curve ellittiche potesse essere sicuro fino a quando il suo ordine è primo o quasi primo. Con questa conclusione, Koblitz pensava che tutte le curve fossero create uguali e fossero tutte valide con un ECDLP intrattabile.

Questo è un esercizio elementare che mostra che la curva $y^2 = x^3 - x$ su F_p con $4|(p+1)$ o la curva $y^2 + y = x^3$ su F_p con $3|(p+1)$ ha ordine di gruppo $p + 1$.

Solo scegliendo p tale che $(p + 1)/4$ o $(p + 1)/6$ è primo, e ECC è sicuro... Queste curve hanno alcuni interessati ed efficienti vantaggi per calcolare multipli di punti, specialmente su estensioni di campo di F_2 e F_3.

Nel 1991 Menezes-Okamoto-Vanstone mostrarono che l'accoppiamento di Weil dà una riduzione sull'ECDLP al DLP sul gruppo moltiplicativo di un'estensione di campo di definizione.

Per curve supersingolari, come le due scritte sopra, il grado di estensione è molto piccolo. Abitualmente è 2, come nei casi sopra.

Questo uccide le curve supersingolari per ECC ed è stato un errore averle usate come esempi illustrativi.

Quello che sorprese Koblitz tanto quanto l'attacco che uccise le curve supersingolari nel 199,1 fu che 10 anni dopo loro fecero un 'ritorno ruggente'.

Nel 1989, quando per primo Koblitz propose la crittografia iperellittica, se gli avessero chiesto avrebbe spiegato quello che ha visto come il principale vantaggio potenziale della crittografia iperellittica rispetto alla crittografia ellittica, dicendo che molto probabilmente più il genere è alto, più la sicurezza è efficace. Cioè, che un attacco avrebbe più probabilità di successo con un genere piccolo rispetto a un genere alto.

Ma si è rivelato essere l'opposto rispetto a quello che è accaduto successivamente. L'esempio preferito da Koblitz come illustrativo, quello con curve di genere 191 su F_2, divenne totalmente insicuro a causa dell'algoritmo di Adleman-DeMarrais-Huang del 1994. Dopo il lavoro di Gaudry, Diem e altri, sembrava che qualunque cosa con genere più grande di 2 fosse meno

sicura di quella con genere 1 o 2. L'unico HCC che è competitivo con l'ECC è quello con genere $g = 2$. L'intuizione di Koblitz sul grado di sicurezza di un genere alto non sarebbe potuta essere più sbagliata.

Nel 1990, Koblitz e Mike Fellows si affezionarono alla nozione che, nonostante il fiasco, buoni sistemi sistemi crittografici potessero essere costruiti da problemi combinatori HP-hard.

Scrissero perfino un articolo con titolo: 'Combinatorial Crypsystems Galore!'.

Mike Fellows and Koblitz costruirono poi un sistema basato sugli ideali che coinvolge i polinomi, e chiamarono persone a provare a romperlo. La più grande caratteristica del sistema crittografico era il nome che Fellows pensò: Polly Cracker.

Era molto inefficiente, ed il codice è stato presto attaccato e distrutto.

Tornado all'ECC, durante i primi 15 anni di ricerca la sensazione di Koblitz era che non importa quale campo si è lavorato sopra. Si dovevano evitare algoritmi generici, lavorando in gruppi di grande ordine primo, e dopo MOV si dovevano evitare curve supersingolari.

Ma altrimenti si potrebbe usare qualsiasi campo tu puoi preferire per lavorare, e la sicurezza non è influenzata da questa scelta.

Ma ciò era sbagliato.

Negli anni 90: Frey propose la discesa di Weil per attaccare il DLP su curve in campi di estensione di grado composito. La sua idea era di trasportare l'ECDLP al DLP su una curva con genere grande ad una subestensione, dove poteva essere attaccato da un algoritmo.

Molto presto Gaudry, Hess, Smart, Galbraith, Menezes, Teske e altri trovarono curve deboli su campi $\mathbb{F}_{2^s}$ con s non primo.

Fortunatamente, altre persone (come Scott Valdstone) hanno pensato meglio di Koblitz, e tutte le implementazioni commerciali e tutti gli standards ECC usano campi primi o estensioni di grado primo di F_2.

Il lettore dovrebbe tener presente che anche qui, nelle ultime 4 frasi, è sempre Kobliz che parla.

Capitolo 5

Le curve di Edwards

5.1 Definizioni e proprietà

In questo capitolo verranno introdotte le curve di Edwards. Verranno illustrate la definizione e le principali proprietà. In particolare verrà analizzata la birazionalità che lega le curve di Edwards con le curve ellittiche e come si può passare da una curva ellittica ad una di Edwards e viceversa. Le curve di Edwards sono un particolare tipo di curve, molto simili alle curve ellittiche, ma in cui è molto facile e veloce svolgere le operazioni di somma dei punti. Inoltre la legge di addizione è unica, a differenza della somma dei punti delle curve ellittiche, dove ci sono vari casi particolari, al variare della caratteristica del campo. Molti passi sono citazioni da [13].

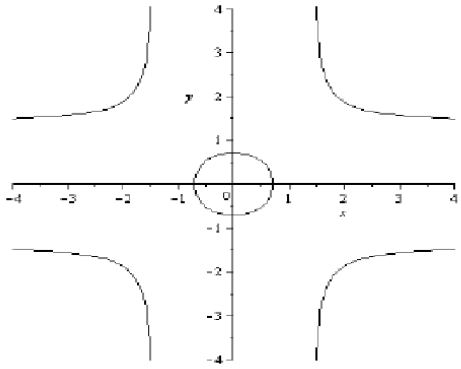
Uno dei problemi per l'uso delle curve ellittiche in crittografia viene dal fatto che quasi sempre l'operazione di somma (fatta usando le formule esplicite) non è ovunque definita: nelle formule per certi valori dei dati il denominatore diventa 0. Il primo caso in cui sono state ottenute formule di addizione ovunque definite si è avuto usando le curve di Edwards. Per uno studio del problema ed aggiornamenti bibliografici, vedi [17].

Definizione 5.1.1. Una curva di Edwards su un campo K di caratteristica diversa da 2, è definita dall'equazione

$$y^2 + x^2 = a^2 + a^2x^2y^2 \tag{5.1}$$

con $a^5 \neq a$, $a \in K$.

La cosa più interessante ed utile delle curve di Edwards è quella di avere un'unica legge di addizione dei punti.



(a) $x^2 + y^2 = \frac{1}{2} + \frac{1}{2}x^2y^2$

Figura 5.1: esempio di curva di Edwards

Teorema 5.1.2. *Sia a una costante per cui $a^5 \neq a$, siano $P = (x_1, y_1)$ e $Q = (x_2, y_2)$ appartenenti alla curva. Per calcolare $P+Q = (x_3, y_3)$ si hanno le seguenti formule:*

$$x_3 = \frac{1}{a}\frac{x_1y_2 + y_1x_2}{1 + x_1x_2y_1y_2}$$

$$y_3 = \frac{1}{a}\frac{y_1y_2 - x_1x_2}{1 - x_1x_2y_1y_2}$$

Quando si parla di curve algebriche, tra cui le curve di Edwards, su un campo che non sia algebricamente chiuso, come ad esempio Q o R, è utile abbandonare la nozione di punti di una curva e lavorare con le funzioni razionali sulla curva. Queste funzioni razionali formano un campo, le cui proprietà algebriche descrivono le proprietà geometriche della curva.

Possiamo considerare una curva ellittica come una curva nella forma $z^2 = f(x)$, dove $f(x)$ è un polinomio di terzo o quarto grado con radici distinte e coefficienti appartenenti a un campo numerico algebrico. Questo significa che per avere una curva algebrica è sufficiente conoscere:

1. un campo numerico algebrico K;
2. un polinomio $f(x)$ di grado 3 o 4 con coefficienti in K che non abbia radici multiple.

Definiamo un campo dato da una funzione ellittica. Il campo delle funzioni razionali su $z^2 = f(x)$ è il campo i cui elementi siano esprimibili nella forma $r(x) + s(x)z$, con $r(x)$ e $s(x)$ funzioni razionali di x a coefficienti in K, con denominatore non nullo.

La somma e il prodotto di polinomi sono definiti nel modo consueto. Grazie alla relazione $z^2 = f(x)$ possiamo eliminare il termine in z^2. Un campo così definito è un campo da funzione ellittica.

Il campo delle funzioni razionali sulla curva (5.1) non è esattamente un campo da funzione ellittica, ma lo diventa quando $z = y(1-a^2x^2)$ è utilizzato per scrivere la curva nella forma $z^2 = (a^2 - x^2)(1 - a^2x^2)$, cioè $z^2 = f(x)$.

Definizione 5.1.3. Due curve sono *birazionalmente equivalenti* se i loro campi di funzioni sono isomorfi.

Nel nostro caso possiamo scrivere $y = \frac{z}{1-a^2x^2}$ e creare un'equivalenza birazionale tra $x^2 + y^2 = a^2 + a^2x^2y^2$ e $z^2 = (a^2 - x^2)(1 - a^2x^2)$, nel senso che x e y possono essere espressi in termini di x e z e viceversa.

Definizione 5.1.4. Un elemento di campo da funzione ellittica è detto costante se è la radice di un polinomio a coefficienti interi.

Gli elementi del campo si possono scrivere nella forma $r(x) + s(x)z$, e tale elemento è una costante se e solo se $r(x) \in K$ e $s(x) = 0$. Possiamo estendere K aggiungendo delle costanti, ma questo cambia il campo delle funzioni determinato dalla nostra curva. Viene quindi a mancare il rapporto di birazionalità tra curva ellittica e curva di Edwards, benché resti un legame tra la curva e il campo di funzioni.

Per aggirare questo inconveniente, diremo che un campo da funzione ellittica è equivalente ad un campo ottenuto da questo campo aggiungendo il campo delle costanti in K. Più in generale due campi da funzione ellittica sono equivalenti quando possiamo trovarne un terzo che sia equivalente ai primi due. Quindi due campi da funzione ellittica sono equivalenti tra loro se possono essere resi isomorfi aggiungendo abbastanza costanti.

Proposizione 5.1.5. *Un campo da funzione ellittica è equivalente, nel senso appena enunciato, ad un campo di funzioni razionali su $x^2+y^2 = a^2+a^2x^2y^2$ per un certo a.*

Dimostrazione.

Sia K un campo numerico algebrico e sia $f(x)$ un polinomio di grado 4 a coefficienti in K a radici distinte. Aggiungiamo a K le costanti necessarie in modo che si possa fattorizzare $f(x)$ in termini di grado 1, scrivendo $f(x) = c(x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)$. Aggiungiamo al campo anche $\sqrt{c}$ se è

necessaria per scrivere $z^2 = f(x)$ in forma monica, cioè $v^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ con $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in K$ tutti distinti. Supponiamo di avere due campi da funzione ellittica $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ e $v^2 = (x-\beta_1)(x-\beta_2)(x-\beta_3)(x-\beta_4)$. La condizione sotto la quale questi campi ellittici sono equivalenti è quella che ci sia una trasformazione lineare fratta del tipo $x \mapsto \frac{Ax+B}{Cx+D}$, con $AD \neq BC$ che mandi $\alpha_i \mapsto \beta_i$ per $i = 1, 2, 3, 4$. Verifichiamo che questa è una condizione sufficiente.

Abbiamo che $u - \beta_i = \frac{(AD-BC)(x-\alpha_i)}{(Cx+D)(C\alpha_i+D)}$ per ogni i, quando $u = \frac{Ax+B}{Cx+D}$. Il prodotto di $(u-\beta_1)(u-\beta_2)(u-\beta_3)(u-\beta_4)$ è una costante moltiplicata per $\frac{(x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)}{(Cx+D)^4}$. Poiché $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ abbiamo il seguente prodotto $(u-\beta_1)(u-\beta_2)(u-\beta_3)(u-\beta_4)$ che è una costante moltiplicata per il quadrato di $\frac{z}{(Cx+D)^2}$. Aggiungendo, se necessario, la radice quadrata della costante al campo, si ottiene un cambio di variabili birazionali fra (z, x) e (v, u) mediante il quale $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ corrisponde a $v^2 = (u-\beta_1)(u-\beta_2)(u-\beta_3)(u-\beta_4)$, come volevasi dimostrare.

In particolare si ha che la trasformazione lineare fratta

$$x \mapsto \frac{\alpha_4 - \alpha_2}{(\alpha_2+\alpha_4)(x+\alpha_3) - 2\alpha_3 x - 2\alpha_2\alpha_4}$$

manda $\alpha_2 \mapsto -1$, $\alpha_3 \mapsto 0$ e $\alpha_4 \mapsto 1$.

Questo dimostra che $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ e $v^2 = (u-\phi)(u+1)u(u-1)$ definiscono dei campi di funzione equivalenti prendendo:

$$\phi = \frac{(\alpha_4-\alpha_2)(\alpha_1-\alpha_3)}{(\alpha_2+\alpha_4)(\alpha_1+\alpha_3) - 2\alpha_3\alpha_1 - 2\alpha_2\alpha_4}$$

$$= \frac{\alpha_1\alpha_4 + \alpha_2\alpha_3 - \alpha_1\alpha_2 - \alpha_3\alpha_4}{\alpha_1\alpha_2 + \alpha_2\alpha_3 + \alpha_3\alpha_4 + \alpha_4\alpha_1 - 2\alpha_1\alpha_3 - 2\alpha_2\alpha_4}$$

La relazione che definisce il campo da curva ellittica determinato da $x^2 + y^2 = a^2 + a^2x^2y^2$ può essere scritta come $(\frac{z}{a})^2 = (x-a)(a-\frac{1}{a})(x+a)(x+\frac{1}{a})$, da cui segue che il campo è equivalente a quello definito mediante $v^2 = (u-\phi)(u+1)u(u-1)$ quando $\phi = \frac{-1-1-1-1}{1-1+1-1+2a^2+2a^{-2}} = -\frac{2}{a^2+a^{-2}}$. Quanto appena visto ci permette di passare dapprima da $z^2 = (x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)$ a $v^2 = (u-\phi)(u+1)u(u-1)$ e quindi a $x^2 + y^2 = a^2 + a^2x^2y^2$ quando a è costante per cui $\phi = -\frac{2}{a^2+a^{-2}}$, cioè quando a è una soluzione di $a^4 + \frac{2}{\phi}a^2 + 1 = 0$. Questo completa la costruzione quando $f(x)$ è di grado 4.

Quando invece $f(x)$ è di grado 3, con radici distinte, possiamo considerare, se necessario, $f(x+c)$ in modo da avere il termine costante non nullo. Quindi dividendo l'equazione $z^2 = f(x)$ per x^4, possiamo riscrivere l'equazione nella forma $(\frac{z}{x^2})^2 = f_1(\frac{1}{x})$, dove f_1 è un polinomio di grado 4 a radici distinte. Possiamo quindi ricondurci al caso del polinomio di grado 4.

□

Dopo aver dimostrato che un campo da funzione ellittica è equivalente a uno della forma $x^2 + y^2 = a^2 + a^2y^2y^2$, vogliamo ora dare una condizione di equivalenza tra campi.

Proposizione 5.1.6. *Il campo da funzione ellittica determinato da* $x^2+y^2 = a^2 + a^2y^2y^2$ *è equivalente a quello determinato da* $x^2 + y^2 = b^2 + b^2y^2y^2$ *se il valore di* b *è uno dei seguenti 24:*

$$i^\epsilon a, \frac{i^\epsilon}{a}, i^\epsilon\frac{a-1}{a+1}, i^\epsilon\frac{a+1}{a-1}, i^\epsilon\frac{a-i}{a+i}, i^\epsilon\frac{a+i}{a-i}$$

dove i *è l'unità immaginaria, mentre* $\epsilon = 0, 1, 2, 3$.

Dimostrazione.

I possibili valori di b elencati rappresentano l'orbita di a sotto il gruppo delle trasformazioni lineari frazionali della sfera di Riemann, generato dalle trasformazioni $a \mapsto ia$ e $a \mapsto \frac{a-1}{a+1}$. Questo gruppo è isomorfo al gruppo di un cubo, infatti possiamo osservare che i sei punti $0, \pm1, \pm i, \infty$ sulla sfera di Riemann, che sono i valori che a non può assumere, sono permutati dal gruppo nello stesso modo in cui le facce del cubo sono permutate dai suoi movimenti.

In particolare quando mappiamo $a \mapsto ia$, permutiamo $1\ i\ -1\ -i$ ciclicamente, mentre lasciamo 0 e ∞ fissi. Invece quando associamo $a \mapsto \frac{a-1}{a+1}$ permutiamo $(1\ 0\ -1\ \infty)$ ciclicamente, lasciando invariati i e $-i$, che è il modo di permutare le facce del cubo quando le coppie ±1, $\pm i$ e la coppia $(0, \infty)$ identificando le tre coppie di facce opposte.

Quindi, non solo il gruppo contiene 24 elementi, ma l'orbita di ogni a sotto l'azione del gruppo contiene 24 elementi distinti; ad eccezione di quei valori di a che corrispondono ai vertici del cubo che costituiscono un'orbita che contiene solo 8 valori distinti di a, ed i valori di a che corrispondono ai punti medi degli spigoli del cubo, che danno luogo ad un'orbita di 12 valori distinti.

I valori che a non può assumere rappresentano il centro delle facce e originano un'orbita di 6 valori distinti. L'orbita di 12 punti è l'orbita di $\sqrt{i}$, infatti il suo valore non cambia quando le applichiamo un elemento del gruppo $a \mapsto \frac{i}{a}$ che scambia fra loro 1 e i, -1 i -i, 0 e ∞, e quindi lascia invariati i punti medi di due spigoli: quello tra 1 e i e quello tra -1 e -i. L'orbita di 8 punti è l'orbita di $(1+i)\frac{\sqrt{3}-1}{2}$ perché questo numero è invariante sotto la mappa $a \mapsto \frac{1+ia}{1-ia}$ che permuta ciclicamente 0, 1 ed i così come ∞, -i, -1, e che lascia invariati 2 vertici, quello in comune fra i due insiemi di tre facce che vengono permutate.

Per quanto visto sulle permutazioni associate ai valori di a, abbiamo che se il campo della funzione $x^2 + y^2 = a^2 + a^2x^2y^2$ è equivalente ai due campi ottenuti sostituendo a una volta con ia e l'altra con $\frac{a-1}{a+1}$, avremo ottenuto quando volevamo dimostrare. Quindi è sufficiente mostrare che se $b = ia$ oppure $b = \frac{a-1}{a+1}$, allora c'è una trasformazione lineare fratta che manda l'insieme $(a, -a, \frac{1}{a}, -\frac{1}{a})$ nell'insieme $(b, -b, \frac{1}{b}, -\frac{1}{b})$, questo è vero perché $x \mapsto ix$ è la trasformazione cercata nel primo caso e $x \mapsto \frac{x-1}{x+1}$ nel secondo.

□

Consideriamo ora il polinomio $K(x) = \prod_{l=1}^{2} 4(x - a_l)$, dove a_l varia tra i 24 valori. Con un calcolo diretto si vede che il polinomio ha grado 24 in x ed i suoi coefficienti sono funzioni razionali di a. Visto che $K(a) = 0$, otteniamo

$$C(a) = \frac{(a^8 + 14a^4 + 1)^3}{a^4(a^4 - 1)^4}$$

Osserviamo che se $K(x) = 0$ per $x = a$, allora $K(x) = 0$ anche per $x = ia$ e per $x = \frac{a-1}{a+1}$. Il C appena definito, viene scritto, secondo la notazione tradizionale, come $\frac{j}{16}$, anche se si può trovare anche scritto come $J = \frac{j}{1728} = \frac{C}{108}$, che è una normalizzazione perché pone $J = 1$ sull'orbita di 12 punti e 0 sull'orbita di 8 punti.

Quando $J = 1$, l'equazione $K(x) = (x^8 + 14x^4 + 1)^3 - 108(x^5 - x)^4 = (x + 1)^2(x^8 - 34x^4 + 1)^2$ ha 12 radici distinte. Quando $J = 0$, invece $K(x) = (x^8 + 14x^4 + 1)^3$ ha 8 radici triple. Per tutti gli altri valori finiti di J, $K(x)$ ha 24 radici distinte.

Il J corrispondente ad una curva ellittica generica $z^2 = x^4 + ex^3 + fx^2 + gx + h$ è

$$\frac{C(a)}{108} = \frac{(a^4 + 14 + a^{-4})^3}{108(a^2 - a^{-2})^4} = \frac{((a^2 + a^{-2})^2 + 12)^3}{108((a^2 + a^{-2})^2 - 2)^2}$$

questa formula può essere ricavata ponendo $x^4 + ex^3 + fx^2 + gx + h = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)(x - \alpha_4)$ e notando che $\phi = -\frac{2}{a^2+a^{-2}}$ è espresso in termini delle α come nella proposizione 5.1.5. Più in generale la formula per calcolare J è la seguente:

$$J = \frac{4(2f^3 - 9efg + 27g^2 + 27e^2h - 72fh)^2}{108\Delta} + 1$$

dove Δ è il discriminante di $x^4 + ex^3 + fx^2 + gx + h$, oppure, nel caso di una curva ellittica di grado 3, $z^2 = x^3 + ax^2 + bx + c$, la formula diventa

$$J = \frac{4(2a^3 - 9ab + 27c)^2}{108\Delta} + 1$$

dove Δ questa volta è il discriminante di $x^3 + ax^2 + bx + c$.

Per una curva in forma canonica $y^2 = x^3 + ax + b$ abbiamo che $J = \frac{a^3}{a^3 - 27b^2}$.

La proposizione 5.1.6 afferma che 2 curve con lo stesso J-variante sono equivalenti, cioè si dice che sono l'una il twist dell'altra.

Teorema 5.1.7. *Siano K_a e K_b campi algebrici numerici che contengano rispettivamente a e b, con $a^5 \neq a$ e $b^5 \neq b$ e per i quali i campi da funzione ellittica sono determinati dalle equazioni $x^2 + y^2 = a^2 + a^2x^2y^2$ su K_a e $u^2 + v^2 = b^2 + b^2u^2v^2$ su K_b sono equivalenti nel senso della proposizione 5.1.5, allora il J-invariante dei due campi di funzioni sono numeri algebrici coniugati, cioè sono radici dello stesso polinomio irriducibile a coefficienti interi.*

Dimostrazione.

L'ipotesi di equivalenza tra i due campi implica l'esistenza di un terzo campo algebrico numerico K e di embedding da $K_a \mapsto K$ e $K_b \mapsto K$ per i quali i campi di funzioni su K determinate da $x^2 + y^2 = a_1^2 + a_1^2x^2y^2$ e da $u^2 + v^2 = b_1^2 + b_1^2u^2v^2$ sono isomorfi, dove con a_1 e b_1 si indicano le rispettive immagini di a e b in K sotto il rispettivo embedding. Un isomorfismo di campi di funzioni implica un isomorfismo dei relativi campi di costanti, in questo modo al prezzo di rimpiazzare b_1 con un suo coniugato sotto il gruppo di Galois di K sui razionali, possiamo assumere che a, $b \in K$ e che $x^2 + y^2 = a^2 + a^2x^2y^2$ e $u^2 + v^2 = b^2 + b^2u^2v^2$ definiscano campi di funzioni su K che siano isomorfi sotto un isomorfismo che sia anche l'identità per K. Dimostreremo che con questa assunzione più forte, i due J-invarianti sono uguali. Un tale isomorfismo tra due campi da curva ellittica su K che sia contemporaneamente l'identità su K, implica una corrispondenza iniettiva tra i punti razionali delle due curve, perché i valori x_1 e y_1 di x e y determinano i valori, per tutti gli elementi del campo di funzioni, in particolare determinano i valori di u_1 e v_1 di u e v e viceversa. Possiamo assumere, senza perdita di generalità, che l'isomorfismo manda il punto $(x, y) = (0, a)$ nel punto razionale $(u, v) = (0, b)$, perché se $(x, y) = (x_1, y_1)$ è un punto razionale che corrisponde a $(u, v) = (0, b)$ sotto il dato automorfismo, allora la formula di addizione ci fornisce un automorfismo di campi che manda $(x, y) = (x_1, y_1)$ in $(x, y) = (0, a)$ e $(u, v) = (0, b)$ corrispondono sotto l'isomorfismo. Con questa più forte supposizione, il punto razionale $(u, v) = (0, -b)$ corrisponde sotto l'isomorfismo ad uno fra i tre seguenti punti razionali $(x, y) = (0, -a), (\infty, \frac{1}{a}), (\infty, -\frac{1}{a})$. Il motivo è che i quattro punti $(0, \pm a)$, $(\infty, \pm\frac{1}{a})$ sono la preimmagine dell'identità $(x, y) = (0, a)$ sotto la mappa doubling, la mappa che manda $P \mapsto 2P$, e questa mappa è intrinseca alla curva una volta che è stata data l'identità $(x, y) = (0, a)$. Questo vuol dire che una coppia di elementi del campo $(X, Y) = (\frac{2xy}{a(1+x^2y^2)}, \frac{y^2-x^2}{a(1-x^2y^2)})$ data

dalla formula di addizione, quando si mettono in input due coppie, si genera un sottocampo, i cui elementi sono esprimibili in termini razionali di x_3 e y_3, isomorfo all'intero campo, ma con indice 4 su di esso. Questo campo è intrinseco, nel senso che deve corrispondere, sotto isomorfismo, al sottocampo del campo da funzione ellittica di $u^2 + v^2 = b^2 + b^2u^2v^2$, generato dalla funzione $(U, V) = (\frac{2uv}{b(1+u^2v^2)}, \frac{v^2-u^2}{b(1-u^2v^2)})$ perché un'espressione di una funzione razionale di x e y come funzione razionale di X e Y corrisponde sotto isomorfismo ad un'espressione di una funzione razionale di u e v come funzione razionale di U e V.

Questo sottocampo ha indice 4 perché quando x^2 è scritto come $\frac{a^2-y^2}{1-a^2y^2}$, con denominatore non nullo, l'equazione $aY(1 - x^2y^2) = y^2 - x^2$ diventa un'equazione di grado 4 in y a coefficienti nel sottocampo; ogni radice di y dà un unico valore di x grazie alla relazione $aX(1 - x^2y^2) = 2xy$. Ogni punto razionale su $x^2 + y^2 = a^2 + a^2x^2y^2$ dà valori in K di X e Y ed associa a 4 punti un solo punto perché i valori di X e Y per $(x, y) = (x_1, y_1)$ sono gli stessi anche per $(-x_1, -y_1)$ e per $(\pm\frac{1}{x_1}, \pm\frac{1}{y_1})$. In particolare i quattro punti che sotto la mappa doubling danno $(X, Y) = (0, a)$, sono i punti che corrispondono alla descrizione precedente e alle seguenti conclusioni.

Possiamo assumere, senza perdita di generalità, che l'isomorfismo tra la curva in xy e quella in uv mandi $(u, v) = (0, -b)$ in $(x, y) = (0, -a)$ e $(u, v) = (0, b)$ in $(x, y) = (0, a)$. Come abbiamo visto, se $(u, v) = (0, -b)$ non è mandato in $(x, y) = (0, -a)$ allora deve necessariamente essere mandato in uno fra $(\infty, \frac{1}{a})$ e $(\infty, -\frac{1}{a})$. Quindi è sufficiente provare che esiste un isomorfismo della curva in xy con una curva della stessa forma $X^2 + Y^2A^2 + A^2X^2Y^2$ che manda uno di questi punti in $(X, Y) = (0, -A)$ mentre manda $(x, y) = (0, a)$ in $(X, Y) = (0, A)$.

La trasformazione lineare fratta $\lambda(z) \to \frac{1+iz}{1-iz}$ ha ordine 3, come abbiamo notato nella dimostrazione della proposizione 5.1.6, e corrisponde alla permutazione $(0, 1, i)$ e $(\infty, -1, -i)$ delle facce del cubo. Quando A è definito come $\lambda(a) \in K$ e Y è definito come $\lambda(y)$, cioè un elemento del campo della funzione $x^2 + y^2 = a^2 + a^2x^2y^2$ su K, allora

$$\frac{A^2 - Y^2}{1 - A^2Y^2} = \frac{(y - a)(ay + 1)}{(y + a)(1 - ay)}$$

come si può vedere dal calcolo diretto. Questo elemento del campo della funzione $x^2 + y^2 = a^2 + a^2x^2y^2$ è un quadrato, infatti

$$\frac{(y - a)(ay + 1)}{(y + a)(1 - ay)}\frac{(ay + 1)(y + a)}{(ay + 1)(y + a)}) = \frac{(ay + 1)^2(y^2 - a^2)}{(y + a)^2(1 - a^2y^2)} = (\frac{ay + 1}{y + a}ix)^2$$

Quindi porre $X = ix\frac{ay+1}{y+a}$ dà una presentazione del campo di funzione dato da $x^2 + y^2 = a^2 + a^2x^2y^2$ su K come campo della funzione $X^2 + Y^2 = A^2 + A^2X^2Y^2$ su K. Nel punto $(x, y) = (0, a)$, Y vale $\lambda(a) = A$, quindi corrisponde al punto $(X, Y) = (0, A)$. Nel punto $(x, y) = (\infty, -\frac{1}{a})$, $Y = \lambda(y)$ vale $\lambda(-\frac{1}{a}) = \frac{1-i/a}{1+i/a} = \frac{ia+1}{ia-1} = -\lambda(A) = -A$, per cui al punto $(x, y) = (\infty, -\frac{1}{a})$ corrisponde il punto $(X, Y) = (0, -A)$. Nel punto $(x, y) = (\infty, \frac{1}{a})$, Y vale $\lambda(\frac{1}{a}) = -\frac{1}{a}$, così porre $A_1 = \lambda(A)$, $Y_1 = \lambda(Y)$, e $X_1 = iX\frac{AY+1}{Y+A}$ dà una descrizione del campo come un campo della stesa forma in cui $(x, y) = (\infty, \frac{1}{a})$ corrisponde a $(X_1, Y_1) = (0, -A_1)$ e $(x, y) = (0, a)$ corrisponde a $(X_1, Y_1) = (0, A_1)$.

Infine, quando diamo l'isomorfismo tra $u^2 + v^2 = b^2 + b^2u^2v^2$ e $x^2 + y^2 = a^2 + a^2x^2y^2$ facciamo corrispondere a $(u, v) = (o, b)$ il punto $(x, y) = (0, a)$ ed a $(u, v) = (0, -b)$ il punto $(x, y) = (0, -a)$. L'isomorfismo manda anche i punti la cui immagine sotto la mappa doubling è $(u, v) = (0, -b)$ in punti la cui immagine sotto la mappa analoga è $(x, y) = (0, -a)$. Questi punti sono rispettivamente $(u, v) = (\pm b, 0), (\pm\frac{1}{b}, \infty)$ e $(x, y) = (\pm a, 0), (\pm\frac{1}{a}, \infty)$. Quindi b deve essere uno fra i seguenti quattro valori $\pm a, \pm\frac{1}{a}$, tali valori sono nella lista dei punti della proposizione 5.1.6 e quindi il J-invariante è lo stesso.

□

5.2 La somma di due punti

In questa sezione vedremo come svolgere le operazioni con le curve di Edwards. Nella sezione precedente abbiamo già visto come fare la somma due punti, ora dimostreremo che tale operazione è lecita e si comporta come si dovrebbe comportare un'operazione di somma.

Siano x_3 e y_3

$$x_3 = \frac{1}{a}\frac{x_1y_2 + y_1x_2}{1 + x_1x_2y_1y_2}$$

$$y_3 = \frac{1}{a}\frac{y_1y_2 - x_1x_2}{1 - x_1x_2y_1y_2}$$

chiamiamo $W = x_1x_2y_1y_2$ per comodità. Moltiplicando l'equazione $x_3^2 + y_3^2 = a^2 + a^2x_3^2y_3^2$ per $a^2(1 - W^2)^2$ si ottiene

$$(x_1y_2+y_1x_2)^2(1-W)^2+(y_1y_2-x_1x_2)^2(1+W)^2 = a^4(1-W^2)^2+(x_1y_2+y_1+x_2)^2(y_1y_2-x_1x_2)^2$$

Si deve dimostrare che l'ultima equazione è una conseguenza dell'aver assunto che valgono $x_1^2 + y_1^2 = a^2 + a^2x_1^2y_1^2$ e $x_2^2 + y_2^2 = a^2 + a^2x_2^2y_2^2$. In altre parole dobbiamo dimostrare che il polinomio Δ definito dall'equazione

$$(x_1y_2)^2(1-W)^2+(y_1y_2-x_1x_2)^2(1+W)^2 = (x_1y_2+y_1x_2)^2(y_1y_2-x_1x_2)^2+a^4(1-W^2)^2+\Delta \tag{5.2}$$

è dato dalla somma di multipli di

$$R_1 = x_1^2 + y_1^2 - a^2 - a^2x_1^2y_1^2$$

e

$$R_2 = x_2^2 + y_2^2 - a^2 - a^2x_2^2y_2^2$$

Il termine a sinistra dell'uguale dell'equazione (5.2) può essere riscritto come

$$(x_1^2y_2^2 + 2W + y_1^2x_2^2)(1 - 2W + W^2) + (y_1^2y_2^2 - 2W + x_1^2x_2^2)(1 + 2W + W^2)$$

Raccogliendo in modo opportuno i termini con $(1+W)^2$ e quelli con $2W$, possiamo scrivere

$$(x_1^2y_2^2+2W+y_1^2y_2^2-2W+x_1^2x_2^2)(1+W^2)+(-x_1^2y_2^2-2W-y_1^2x_2^2+y_1^2y_2^2-2Wx_1^2x_2^2)2W$$

e raccogliendo i quadrati

$$(x_1^2 + y_1^2)(x_2^2 + y_2^2)(1 + W)^2 + ((x_1^2 + y_1^2)(x_2^2 + y_2^2) - 4W)(2W)$$

$$= (x_1^2 + y_1^2)(x_2^2 + y_2^2)(1 + W)^2 + (2W)(x_1^2 - y_1^2)(x_2^2 - y_2^2) - 8W^2$$

Nel secondo termine dell'equazione (5.2) possiamo invece riscrivere il primo prodotto come segue

$$(x_1y_2 + y_1x_2)^2(y_1y_2 - x_1x_2)^2 = (x_1^2y_2^2 + y_1^2x_2^2 + 2W)(y_1^2y_2^2 + x_1^2x_2^2 - 2W)$$

$$= (x_1^2y_2^2 + y_1^2x_2^2)(y_1^2y_2^2 + x_1^2x_2^2) + 2W(y_1^2y_2^2 + x_1^2x_2^2 - x_1^2y_2^2 - y_1^2x_2^2) - 4W$$

$$= x_1^2y_1^2y_2^4 + x_1^4x_2^2y_2^2 + y_1^4x_2^2y_2^2 + x_1^2y_1^2x_2^4 + 2W(x_1^2 - y_1^2)(x_2^2 - y_2^2) - 4W^2$$

Sottraendo $2W(x_1^2 - y_1^2)(x_2^2 - y_2^2) - 8W^2$ da entrambi i membri di (5.2), o meglio da (5.2) con i termini riscritti come appena detto, si ottiene

$$(x_1^2 + y_1^2)(x_2^2 + y_2^2)(1 - W^2)$$

$$= x_1^2y_1^2y_2^4 + x_1^4x_2^2y_2^2 + y_1^4x_2^2y_2^2 + x_1^2y_1^2x_2^4 + 4W^2 + a^4(1 - W^2)^2 + \Delta$$

$$= x_1^2y_1^2(y_2^4 + 2x_2^2y_2^2 + x_2^4) + x_2^2y_2^2(y_1^4 + 2x_1^2y_1^2 + x_1^4) + a^4(1 - W^2)^2 + \Delta$$

$$= x_1^2y_1^2(x_2^2 + y_2^2)^2 + x_{22}^{y2}(x_1^2 + y_1^2)^2 + a^4(1 - W^2)^2 + \Delta \qquad (5.3)$$

Possiamo riscrivere $(1 - W^2)^2$ in questo modo:

$$(1-W^2)^2 = (1+W^2)^2-4W^2 = (1+W^2)(1+x_1^2y_1^2+x_2^2y_2^2+W^2)-(1+W^2)(x_1^2y_1^2+x_2^2y_2^2)-4W^2$$

$$= (1+W^2)(1+x_1^2y_1^2)(1+x_2^2y_2^2)-x_1^2y_1^2-x_2^2y_2^2-2W^2-2W^2-x_1^2y_1^2W^2-x_2^2y_2^2W^2$$

$$= (1+W^2)(1+x_1^2y_1^2)(1+x_2^2y_2^2)-x_1^2y_1^2(1+2x_2^2y_2^2+x_2^4y_2^2)-x_2^2y_2^2(1+2x_1^2y_1^2+x_1^4y_1^4)$$

$$= (1 + W^2)(1 + x_1^2y_1^2)(1 + x_2^2y_2^2) - x_1^2y_1^2(1 + x_2^2y_2^2)^2 - x_2^2y_2^2(1 + x_1^2y_1^2)^2 \quad (5.4)$$

Ricordando i termini nell'equazione (5.3), si dimostra che

$$\Delta = ((x_1^2 + y_1^2)(x_2^2 + y_2^2) - (a^2 + a^2x_1^2y_1^2)(a^2 + a^2x_2^2y_2^2))(1 + W^2)+$$
$$x_1^2y_1^2(a^2 + a^2x_2^2y_2^2) - (x_2^2 + y_2^2)^2 + x_2^2y_2^2((a^2 + a^2x_1^2y_1^2)^2 - (x_1^2 + y_1^2)^2) \quad (5.5)$$

quindi se $x_1^2 + y_1^2 = a^2 + a^2x_1^2y_1^2$ e $x_2^2 + y_2^2 = a^2 + a^2x_2^2y_2^2$, allora abbiamo $\Delta = 0$.

Teorema 5.2.1. *Sia K un campo numerico algebrico e sia $a \in K$ con $a^5 \neq a$. Aggiungiamo al campo $K(x_1, x_2)$ di funzioni razionali in x_1 e x_2 a coefficienti in K le radici quadrate di $(a^2 - x_1^2)(1 - a^2x_1^2)$ e di $(a^2 - x_2^2)(1 - a^2x_2^2)$, e le chiamiamo rispettivamente z_1 e z_2. Le formule $y_1 = \frac{z_1}{1-a^2x_1^2}$ e $y_2 = \frac{z_2}{1-a^2x_2^2}$ definiscono gli elementi del campo esteso con le radici che generano l'estensione su $K[x_1, x_2]$ e che soddisfano $x_1^2 + y_1^2 = a^2 + a^2x_1^2y_1^2$ e $x_2^2 + y_2^2 = a^2 + a^2x_2^2y_2^2$.*

Le formule:

$$x_3 = \frac{1}{a}\frac{x_1y_2 + y_1x_2}{1 + x_1x_2y_1y_2}$$

$$y_3 = \frac{1}{a}\frac{y_1y_2 - x_1x_2}{1 - x_1x_2y_1y_2}$$

definiscono gli elementi x_3 e y_3 di tale estensione del campo che soddisfano $x_3^2 + y_3^2 = a^2 + a^2x_3^2y_3^2$.

Consideriamo la curva di Edwards $x^2 + y^2 = c^2(1 + dx^2y^2)$ su un campo K di caratteristica diversa da 2, con c, $d \in K$, con $c \neq 0$, $d \neq 0$, $dc^4 \neq 1$. La legge di addizione di Edwards ci dice che dati due punti $P(x_1, y_1)$ e $Q(x_2, y_2)$ appartenenti alla curva, la somma è un'operazione da $K \times K \mapsto K$ secondo la seguente regola:

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

Notiamo in particolare che, come visto finora, l'elemento neutro per la somma è $(0, c)$ e che il punto $-P$ ha coordinate $(x_1, -y_1)$. In particolare il punto $(0, -c)$ ha ordine 2, mentre $(c, 0)$ e $(-c, 0)$ ha ordine 4.

Teorema 5.2.2. *Sia K un campo con caratteristica diversa da 2. Siano c, $d \in K$ diversi da 0 con $dc^4 \neq 1$. Siano x_1, y_1, x_2, $y_2 \in K$ tali che $x_1^2 + y_1^2 = c^2(1 + dx_1^2y_1^2)$ e $x_2^2 + y_2^2 = c^2(1 + dx_2^2y_2^2)$. Assumiamo che $dx_1x_2y_1y_2 \neq \pm 1$.*

Definiamo

$$x_3 = \frac{x_1y_2 + x_2y_1}{c(1 + dx_1x_2y_1y_2)}$$

$$y_3 = \frac{y_1y_2 + x_1x_2}{c(1 - dx_1x_2y_1y_2)}$$

Allora $x_3^2 + y_3^2 = c^2(1 + dx_3^2y_3^2)$.

In altre parole se P e Q appartengono alla curva, allora anche $P + Q = (x_3, y_3)$ le appartiene.

Dimostrazione. Definiamo

$$T = (x_1y_2 + y_1x_2)^2(1 - dx_1x_2y_2y_1)^2 + (y_1y_2 - x_1x_2)^2(1 + dx_1x_2y_1y_2)^2 -$$

$$-d(x_1y_2 + y_1x_2)^2(y_1y_2 - x_1x_2)^2$$

Con semplici passaggi algebrici possiamo scrivere,

$$T = (x_1^2 + y_1^2 - (x_2^2 + y_2^2)dx_1^2y_1^2)(x_2^2 + y_2^2 - (x_1^2 + y_1^2)(dx_2^2y_2^2)$$

Consideriamo ora le ipotesi su (x_1, y_1) e (x_2, y_2) sottraendo l'equazione $(x_2^2 + y_2^2)dx_1^2y_1^2 = c^2(1 + dx_2^2y_2^2)dx_1^2y_1^2$ dall'equazione $x_1^2 + y_1^2 = c^2(1 + dx_1^2y_1^2)$ si ottiene $x_1^2 + y_1^2 - (x_2^2 + y_2^2)dx_1^2y_1^2 = c^2(1 - d^2x_1^1x_2^2y_1^2y_2^2)$.

Allo stesso modo possiamo ottenere $x_2^2 + y_2^2 - (x_1^2 + y_1^2)dx_2^2y_2^2 = c^2(1 - d^2x_1^2x_2^2y_1^2y_2^2$ scambiano il ruolo dei punti.

Quindi possiamo scrivere $T = c^4(1 - d^2x_1^2x_2^2y_1^2y_2^2)^2$. Applichiamo infine la legge di addizione, cioè scriviamo (x_3, y_3) in termini di x_1, x_2, y_1 e y_2, ottenendo:

$$x_3^2 + y_3^2 - c^2dx_3^2y_3^2 =$$

$$\frac{(x_1y_2 + x_2y_1)^2}{c^2(1 + dx_1x_2y_1y_2)} + \frac{(y_1y_2 - x_1x_2)^2}{c^2(1 - dx_1x_2y_1y_2)^2} - \frac{c^2d(x_1y_2 + y_1x_2)^2(y_1y_2 - x_1x_2)^2}{c^4(1 + dx_1x_2y_1y_2)^2(1 - dx_1x_2y_1y_2)} =$$

$$\frac{T}{c^2(1 + dx_1x_2y_1y_2)^2(1 - dx_1x_2y_1y_2)^2} =$$

$$\frac{T}{c^2(1 - d^2x_1^2x_2^2y_1^2y_2^2)^2}$$

$$= c^2$$

Quindi $x_3^2 + y_3^2 = c^2(1 + dx_3^2y_3^2)$.

□

Diamo ora una proposizione che sarà utile per dimostrare il prossimo teorema.

Proposizione 5.2.3. *Sia K un campo con $2 \neq 0$. Sia E una curva ellittica su K tale che il gruppo $E(K)$ ha un elemento di ordine 4.*
Allora:

1. *esiste $d \in \{0, 1\}$ tale che la curva $x^2+y^2 = 1+d+x^2y^2$ è birazionalmente equivalente su K ad un twist quadratico di E;*

2. *se $E(K)$ ha un unico elemento di ordine 2, allora esiste $d \in K$ tale che la curva $x^2 + y^2 = 1 + d + x^2y^2$ è birazionalmente equivalente su K ad un twist quadratico di E, inoltre d non è un quadrato in K;*

3. *se K è finito ed $E(K)$ ha un unico elemento di ordine 2, allora esiste un elemento $d \in K$, con d non quadrato, tale che la curva $x^2 + y^2 = 1 + d + x^2y^2$ è birazionalmente equivalente su K.*

Dimostrazione. Consideriamo la curva ellittica E scritta in forma di Weierstrass estesa:

$$s^2 + a_1rs + a_3s = r^3 + a_2r^2 + a_4r + a_6$$

Supponiamo, senza perdita di generalità, che $a_1 = 0 = a_3$. In caso contrario è sufficiente definire $\overline{s} = s + \frac{a_1r+a_3}{2}$.

Chiamiamo P il punto di ordine 4 su E. Assumiamo, ancora una volta senza perdita di generalità, che $2P = (0, 0)$, e quindi $a_6 = 0$. Se così non fosse, basta considerare $\overline{r} = r - r_2$, dove $2P = (r_2, s_2)$.

La curva ellittica è ora nella forma $s^2 = r^3 + a_2r^2 + a_4r$. Scriviamo $P = (r_1, s_1)$ e proviamo ad esprimere a_2 ed a_4 in termini di r_1 e s_1. Notiamo che $s_1 \neq 0$, altrimenti P ha ordine 2, quindi anche $r_1 \neq 0$. L'equazione $2P = (0, 0)$ implica che la retta tangente ad E in P passa attraverso il punto $(0, 0)$, il che equivale a dire che

$$s_1 - 0 = (r_1 - 0)\lambda$$

dove λ è la curva tangente $\lambda = \frac{3r_1^2+2a_2r_1+a_4}{2s_1}$.
Si ottiene allora

$$s_1^2 = 3r_1^3 + 2a_2r_1^2 + a_4r_1 \tag{5.6}$$

Se $P \in E$ notiamo facilmente che vale

$$2s_1^2 = 2r_1^3 + 2a_2r_1^2 + 2a_4r_1 \tag{5.7}$$

Sottraendo (5.7) da (5.6) abbiamo $r_1^3 = a_4 r_1$, e dividendo per $r_1 \neq 0$, $r_i^2 = a_4$.

Inoltre

$$a_2 = \frac{s_1^2 - r_1^3 - a_4 r_1}{r_1^2} = \frac{s_1^2}{r_1^2} - 2r_1$$

Ponendo $d = 1 - \frac{4r_1^3}{s_1^2}$ otteniamo $a_2 = 2r_1 \frac{1+d}{1-d}$.

Notiamo che $d \neq 1$, poiché $r_1 \neq 0$, inoltre $d \neq 1$, altrimenti il termine a destra dell'equazione di E sarebbe

$$r^3 + 2r_1 r^2 + r_1^2 = r(r + r_1)^2$$

che non è l'equazione di una curva ellittica.

Possiamo dire qualcosa su d: se d non è un quadrato, allora c'è un punto di ordine 2 in $E(K)$, il punto $(\sqrt{r_1}\frac{\sqrt{d}+1}{\sqrt{d}-1}, 0)$.

Consideriamo il twist quadratico di E, cioè le curve che chiameremo E' ed E'' definite rispettivamente da

- $(\frac{r_1}{1-d}s^2 = r^3 + a_2 r^2 + a_4 r)$;
- $(\frac{dr_1}{1-d}s^2 = r^3 + a_2 r^2 + a_4 r)$.

Se K è finito e d non è un quadrato, allora o $\frac{r_1}{1-d}$ o $\frac{dr_1}{1-d}$ è un quadrato in K, quindi E è isomorfo o a E' oppure a E''. Sostituiamo $u = \frac{r}{r_1}$ e $v = \frac{s}{r_1}$ per mostrare che le curve E' ed E'' sono isomorfe rispettivamente a

1. $\frac{1}{1-d}v^2 = u^3 + 2\frac{1+d}{1-d}u^2 + u$;
2. $\frac{d}{1-d}v^2 = u^3 + 2\frac{1+d}{1-d}u^2 + u$.

Mostriamo ora che la curva $x^2 + y^2 = 1 + dx^2y^2$ è birazionalmente equivalente ad E'.

Consideriamo la mappa $(u, v) \mapsto (x, y)$ definita da

$$x = \frac{2u}{v} \qquad y = \frac{u-1}{u+1}$$

i punti che richiedono una certa attenzione sono quelli per cui $v(u+1) = 0$; di questi punti ce ne sono un numero finito.

La mappa inversa $(x, y) \mapsto (u, v)$ è definita mediante

$$u = \frac{1+y}{1-y} \qquad v = 2x\frac{1+y}{1-y}$$

Anche in questo caso i punti particolari per cui $(1-y)x = 0$ sono in numero finito. Il calcolo diretto mostra che la mappa razionale inversa restituisce delle coppie (u, v) che soddisfano $\frac{1}{1-d}v^2 = u^3 + 2\frac{1+d}{1-d}u^2 + u$.

Sostituendo d con $\frac{1}{d}$ e u con $-u$, notiamo che $x^2 + y^2 = 1 + \frac{1}{d}x^2y^2$ è birazionalmente equivalente alla curva

$$\frac{1}{1-\frac{1}{d}} = (-u)^3 + 2\frac{1+\frac{1}{d}}{1-\frac{1}{d}}(-u)^2 + (-u)$$

cioè, riscritta in forma più semplice,

$$\frac{d}{1-d}v^2 = u^3 + 2\frac{1+d}{1-d}u^2 + u$$

che è equivalente ad E''. Riassumendo:

- la curva $x^2 + y^2 = 1 + dx^2y^2$ è equivalente al twist quadratico E' di E;
- se E ha un unico punto di ordine 2, allora d non è un quadrato e $x^2 + y^2 = 1 + dx^2y^2$ è equivalente ad un twist quadratico E' di E;
- se K è finito ed E ha un punto di ordine 2, allora d non è un quadrato e quindi E è isomorfa ad E' oppure ad E''; quindi E è birazionalmente equivalente a $x^2 + y^2 = 1 + dx^2y^2$ oppure a $x^2 + y^2 = 1 + \frac{1}{d}x^2y^2$.

□

Il prossimo teorema afferma che il risultato della legge di addizione di Edwards corrisponde al risultato della somma standard su una curva ellittica E birazionalmente equivalente ad una curva di Edwards.

Teorema 5.2.4. *Nelle ipotesi del teorema (5.2.2), sia $e = 1 - dc^4$, e sia E la curva ellittica di equazione $\frac{1}{e}v^2 = u^3 + (\frac{4}{e} - w)u^2 + u$. Definiamo P_i come segue per $i \in 1, 2, 3$:*

1. *$P_i = \infty$ se $(x_i, y_i) = (0, c)$;*
2. *$P_i = (0, 0)$ se $(x_i, y_i) = (0, -c)$;*
3. *$P_i = (u_i, v_i)$ se $x_i \neq 0$, dove $u_i = \frac{c+y_i}{c-y_i}$, e $v_i = 2c\frac{c+y_i}{c-y_i}x_i$.*

Allora $P_i \in E(K)$ e $P_1 + P_2 = P_3$.

Con $P_1 + P_2$ si intende la somma fra i due punti P_1 e P_2 secondo la legge di somma standard su $E(K)$. Si noti che $x_i \neq 0$ implica $y_i \neq c$, cioè i denominatori sono diversi da 0.

Dimostrazione.

Dimostriamo innanzitutto che $P_i \in E(K)$ per ogni $i = 1, 2, 3$. Se $(x_i, y_i) = (0, c)$, allora $P_i = \infty \in E(K)$.
Se $(x_i, y_i) = (0, -c)$, allora $P_i = (0, 0) \in E(K)$.

Altrimenti $P_i(u_i, v_i) \in E(K)$ grazie agli stessi conti svolti per dimostrare la proposizione (5.2.3).

Resta da dimostrare che $P_1 + P_2 = P_3$. Dobbiamo analizzare tutti i casi della legge di addizione standard per $E(K)$, quindi anche la dimostrazione si divide in diversi casi.

Se $(x_1, y_1) = (0, c)$, allora $(x_3, y_3) = (x_2, y_2)$. Sia ora P_1 il punto all'infinito e $P_2 = P_3$, allora $P_1 + P_2 = \infty + P_2 = P_3$.

Possiamo trattare analogamente il caso $(x_2, y_2) = (0, c)$. Assumiamo $(x_1, y_1) \neq (0, c)$ e $(x_2, y_2) \neq (0, c)$. Se $(x_3, y_3) = (0, c)$, allora $(x_2, y_2) = (-x_1, y_1)$, quindi se $(x_1, y_1) = (0, -c)$ avremo anche $(x_2, y_2) = (0, -c)$ e $P_1 = (0, 0) = P_2$; altrimenti x_1, x_2 sono diversi da zero e quindi abbiamo $u_1 = \frac{(c+y_1)}{(c-y_1)} = u_2$ e $v_1 = \frac{2cu_2}{x_2}$, cioè $P_1 = -P_2$. In entrambi i casi $P_1 + P_2 = \infty = P_3$.

Assumiamo d'ora in poi che $(x_3, y_3) \neq (0, c)$. Se $(x_1, y_1) = (0, -c)$ allora $(x_3, y_3) = (-x_2, -y_2)$. Poiché $(x_3, y_3) \neq (0, c)$, abbiamo che $(x_2, y_2) \neq (0, -c)$, inoltre poiché $(x_2, y_2) \neq (0, c)$ segue che $x_2 \neq 0$. Quindi $P_1 = (0, 0)$ e $P_2 = (u_2, v_2)$ con $u_2 = \frac{(c+y_2)}{(c-y_2)}$ e $v_2 = \frac{2cu_2}{x_2}$. La legge di addizione standard ci dice che $(0, 0) + (u_2, v_2) = (r_3, s_3)$, con $r_3 = (\frac{1}{e})(\frac{v_2}{u_2})^2 - (\frac{4}{e} - 2) - u_2 = \frac{1}{u_2}$ e $s_3 = (\frac{v_2}{u_2})(-r_3) = -\frac{v_2}{u_2^2}$. Inoltre $P_3 = (u_3, v_3)$ con

$$u_3 = \frac{c + y_3}{c - y_3} = \frac{c - y_2}{c + y_2} = \frac{1}{u_2} = r_3$$

$$v_3 = 2\frac{cu_3}{x_3} = -2\frac{c}{u_2 x_2} = -\frac{v_2}{u_2^2} = s_3$$

Quindi $P_1 + P_2 = P_3$. Possiamo trattare analogamente il caso in cui $(x_2, y_2) = (0, -c)$.

Assumiamo quindi $x_1 \neq 0$ e $x_2 \neq 0$ per i prossimi casi.

Allora $P_1 = (u_1, v_1)$ con $u_1 = \frac{c+y_1}{c-y_1}$ e $v_1 = \frac{2cu_1}{x_1}$, e $P = (u_2, v_2)$ con $u_2 = \frac{c+y_2}{c-y_2}$ e $v_2 = \frac{2cu_2}{x_2}$.

Se $(x_3, y_3) = (0, -c)$ allora $(x_1, y_1) = (x_2, -y_2)$ e di conseguenza

$$u_1 = \frac{c + y_1}{c - y_1} = \frac{c - y_2}{c + y_2} = \frac{1}{u_2}$$

$$v_1 = 2\frac{cu_1}{x_1} = \frac{v_2}{u_2^2}$$

Inoltre abbiamo che $P_3 = (0,0)$, per la legge di addizione standard segue che $-P_3 + P_2 = (0,0) + P_2 = (\frac{1}{u_2}, -\frac{v_2}{u_2^2}) = (u_1, -v_1) = -P_1$, cioè $P_1 + P_2 = P_3$.

Supponiamo d'ora in poi di avere $x_3 \neq 0$. Possiamo scrivere $P_3 = (u_3, v_3)$ con $u_3 = \frac{c+y_3}{c-y_3}$ e $v_3 = 2\frac{cu_e}{x_3}$. Se $P_2 = -P_1$, allora $u_2 = u_1$ e $v_2 = -v_1$, quindi $x_2 = -x_1$ e $y_2 = c\frac{u_2-1}{u_2+1} = c\frac{u_1-1}{u_1+1} = y_1$, allora $(x_3, y_3) = (0, c)$, caso già trattato in precedenza.

Supponiamo quindi per i prossimi casi che $P_2 \neq -P_1$. Se $u_2 = u_2$ e $v_2 \neq -v_1$, grazie alla legge di addizione standard possiamo scrivere che $(u_1, v_1) + (u_2, v_2) = (r_3, s_3)$ con

$$\lambda = \frac{(3u_1^2 + 2(\frac{4}{e} - 2)u_1 + 1)}{\frac{2}{e}v_1}$$

$$r_3 = \frac{1}{e}\lambda^2 - (\frac{4}{e} - 2) - 2u_1$$

$$s_3 = \lambda(u_1 - r_3) - v_1$$

Si può dimostrare che $(r_3, s_3) = (u_3, v_3)$.

Resta l'ultimo caso, $u_2 \neq u_1$. Per come è definita la somma di punti sulle curve ellittiche, abbiamo che $(u_1, v_1) + (u_2, v_2) = (r_3, s_3)$, con

$$\lambda = \frac{v_2 - v_1}{u_2 - u_1}$$

$$r_3 = \frac{1}{e}\lambda^2 - (\frac{4}{e} - 2) - u_1 - u_2$$

$$s_3 = \lambda(u_1 - r_3) - v_1$$

Anche in questo caso si può dimostrare che $(r_3, s_3) = (u_3, v_3)$ Quindi $P_3 = P_1 + P_2$ in tutti i casi possibili.

□

Per come abbiamo definito in precedenza la somma di punti e la curva di Edwards $x^2 + y^2 = c^2(1 + dx^2y^2)$, potrebbero esserci alcuni punti che rappresentano delle eccezioni. Ricordiamoci che per questi punti abbiamo che $dx_1x_2y_1y_2 = \pm 1$ e su di essi non è possibile definire la somma secondo la legge di addizione di Edwards. In questo caso è sufficiente tornare alla curva birazionalmente equivalente per trovare il punto desiderato. Tuttavia

questo sistema è poco pratico e fa perdere alle curve di Edwards uno dei motivi di maggior interesse: avere una regola unica per l'addizione dei punti contrariamente alle curve ellittiche.

Con il prossimo teorema vedremo che se d non è un quadrato, allora non ci sono punti in cui la somma non è definita e quindi il denominatore nella legge di Edwards non si può annullare. In altre parole se d non è un quadrato, la legge di addizione di Edwards è completa, cioè funziona per ogni coppia di punti sulla curva.

Teorema 5.2.5. *Sia K un campo con $2 \neq 0$. Siano c, d, $e \in K$, diversi da 0 e tali che $e = 1 - dc^4$. Assumiamo che d non sia un quadrato in K. Siano x_1, x_2, y_1 e y_2 elementi di K tali che $x_1^2 + y_1^2 = c^2(1 + dx_1^2y_1^2)$ e $x_2^2 + y_2^2 = c^2(1 + dx_2^2y_2^2)$.*

Allora $dx_1x_2y_1y_2 \neq \pm 1$.

Dimostrazione.

Poniamo $\epsilon = dx_{12}y_1y_2$.

Supponiamo per assurdo che $\epsilon \in \{-1, 1\}$. Allora $x_1,\ x_2,\ y_1,\ y_2 \neq 0$. Quindi

$$dx_1^2y_1^2(x_2^2 + y_2^2) = c^2(dx_1^2y_1^2 + d^2x_1^2y_1^2x_2^2y_2^2)$$

$$= c^2(dx_1^2y_1^2 + \epsilon^2)$$

$$= c^2(1 + dx_1^2y_1^2)$$

$$= x_1^+y_1^2$$

Da cui

$$(x_1 + \epsilon y_1)^2 = x_1^2 + y_1^2 + 2\epsilon x_1y_1$$

$$= dx_1^2y_1^2(x_2^2 + y_2^2) + 2x_1y_1x_2y_2$$

$$= dx_1^2y_1^2(x_2^2 + 2x_2^2y_2^2 + y_2^2)$$

$$= dx_1^2y_1^2(x_2 + y_2)^2$$

Se $x_2 + y_2 \neq 0$, allora $d = (\frac{(x_1+\epsilon y_1)}{x_1y_1(x_2+y_2)})^2$, cioè d è un quadrato, ma questo contraddice le ipotesi del nostro teorema.

Allo stesso modo se $x_2 - y_2 \neq 0$, allora $d = (\frac{(x_1 - \epsilon y_1)}{x_1 y_1 (x_2 - y_2)})^2$, d sarebbe ancora un quadrato, che è assurdo per le ipotesi di partenza.

Infine, sia $x_2 + y_2$ che $x_2 - y_2$ sono nulli, allora $x_2 = 0$ e $y_2 = 0$, assurdo perchè tale punto non appartiene a nessuna curva di Edwards.

□

Quindi la formula per sommare, se d non è un quadrato in K può essere utilizzata per sommare qualsiasi punti della curva e la formula si applica anche per il doubling. Ci sono altri algoritmi che permettono di calcolare la somma dei punti su una curva ellittica, per esempio il metodo delle intersezioni di Jacobi, delle quartiche di Jacobi e delle curve di Weierstrass in coordinate proiettive, ma le formule date da Edwards sono da preferirsi in quanto molto più veloci.

5.3 Le curve di Edwards twisted

Verrà ora descritto l'ultimo tipo di curve di Edwards, quelle denominate *twisted*.

Definizione 5.3.1. Sia K un campo di caratteristica diversa da 2. Fissiamo due elementi a, bin.

La curva di equazione:

$$E_{E,a,d}: \quad ax^2 + y^2 = 1 + dx^2y^2 \tag{5.8}$$

è detta curva di Edwards twisted.

Notiamo che se $a = 1$, si ha che la curva di Edwards twisted è una curva di Edwards standard. Possiamo dire quindi che le curve appena definite sono una generalizzazione delle curve di Edwards in forma standard. In particolare se a è un quadrato in K attraverso la trasformazione che manda

$x \mapsto \frac{\bar{x}}{\sqrt{a}}$ e $y \mapsto \bar{x}$

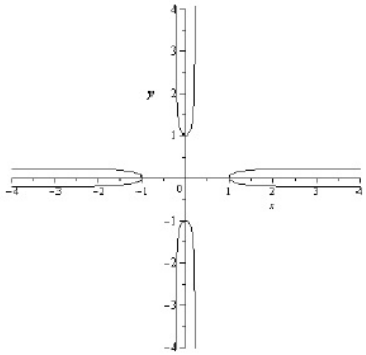
possiamo mandare $ax^2 + y^2 = 1 + dx^2y^2 \mapsto \bar{x}^2 + \bar{y}^2 = 1 + \frac{d}{a}\bar{x}^2\bar{y}^2$.

Segue quindi che le due curve sono isomorfe.

Se invece a non è un quadrato in K, allora l'isomorfismo vale, con la stessa trasformazione, ma su un campo più grande: su $K(\sqrt{a})$. Diamo ora le formule di addizione e di doubling per le curve di Edwards twisted.

Dati due punti (x_1, y_1) e (x_2, y_2) sulla curva, la somma di questi due punti è:

$$(\frac{x_1y_2 + y_1x_2}{1 + dx_1x_2y_1y_2}, \frac{y_1y_2 - ax_1x_2}{1 - dx_1x_2y_1y_2})$$



(a) $x^2 + y^2 = 1 + 19x^2y^2$

Figura 5.2: esempio di curva di Edwards twisted

L'elemento neutro per la somma è il punto $(0, 1)$, mentre l'inverso del punto (x_1, y_1) è il punto $-(x_1, y_1) = (-x_1, y_1)$.

La formula per la somma vale anche per il doubling, cioè se $(x_1, y_1) = (x_2, y_2)$, inoltre è completa se a è un quadrato in K e d non lo è.

Utilizzare le curve di Edwards twisted è vantaggioso perché sono di più rispetto a quelle standard e questo fatto è molto utile negli algoritmi di fattorizzazione.

Capitolo 6

Fattorizzazione con curve di Edwards

Come abbiamo già visto la fattorizzazione di numeri interi è uno dei problemi più studiati in crittografia e nella teoria dei numeri.

Si cercano in continuazione algoritmi sempre più efficienti da implementare su calcolatori per scoprire i fattori di un numero intero.

6.1 La fattorizzazione con curve ellittiche

A partire dalla metà degli anni 80, grazie ad un articolo di Lenstra [16], lo studio delle curve ellittiche si indirizzò anche nello sviluppo di algoritmi efficienti per la fattorizzazione degli interi.

Esempio:

Vogliamo fattorizzare $n = 4453$. Sia E la curva ellittica in forma di Weierstrass $y^2 = x^3 + 10x - 2 \mod 4453$ e sia $P = (1, 3)$. Calcoliamo $3P = 2P + P$.

Innanzitutto calcoliamo $2P$. La retta tangente a P è

$$\frac{3x^2 + 10}{2y} = \frac{13}{6} \equiv 3713(\mod 4453)$$

Per trovare 3713 abbiamo calcolato, grazie all'algoritmo di Euclide esteso 6^{-1}. Poiché $(6, 4453) = 1$, troviamo $6^{-1} \equiv 3711 \mod 4453$, e infine basta calcolare $13{\cdot}6^{-1} = 13{\cdot}3711 \equiv 3713 \mod 4453$. Usando questa retta troviamo $2P = (x, y)$ come

$$x \equiv 3713^2 - 2 \equiv 4332 \mod 4453$$

$$y \equiv -3713(x-1) - 3 \equiv 3230 \mod 4453$$

Calcoliamo ora $3P = 2P + P$. La retta è

$$\frac{3230 - 3}{4332 - 1} = \frac{3227}{4331}$$

Notiamo che $(4331, 4453) = 61 \neq 1$, quindi non possiamo trovare 4331^{-1} mod 4453 e non possiamo calcolare la retta. Abbiamo però trovato un fattore di 4353, cioè 61. Infatti $4453 = 61 \cdot 73$. Si ha allora che

$$E(Z/4453Z) = E(F_6 1) \oplus E(F_7 3)$$

Se cerchiamo i multipli di P mod 61 otteniamo
$P \equiv (1,3)$, $2P \equiv (1,58)$, $3P \equiv \infty$, $4P \equiv (1,3)$,...
Se invece cerchiamo i multipli di P mod 73 otteniamo
$P \equiv (1,3)$, $2P \equiv (25,18)$, $3P \equiv (28,44)$, ..., $64P \equiv \infty$...
Quindi se calcoliamo $3P$ mod 4453, otteniamo infinito modulo 61 ed un punto finito modulo 73.

Se l'ordine di P mod 73 fosse stato 3 anziché 64, la retta avrebbe avuto come denominatore 0 mod 4453 ed il massimo comune divisore sarebbe stato 4553, il che significa che non avremmo ottenuto la fattorizzazione di 4453.

La probabilità che l'ordine di un punto modulo 61 sia esattamente lo stesso di un punto modulo 73 è molto bassa, quindi questa ipotesi non crea problemi. Se proviamo ad aumentare n con un numero composto molto più grande m e lavoriamo con una curva ellittica modulo m ed un punto P su E, allora il problema maggiore sarà di trovare un intero k tale che $kP = \infty$. Effettivamente non sempre si trova un tale k, ma se usiamo un numero sufficiente di curve E, è probabile che almeno una ci permetta di trovare il k cercato.

Questa è la chiave del metodo di fattorizzazione con curve ellittiche.

Teorema 6.1.1. *Siano n_1 e n_2 due interi dispari tali che $(n_1, n_2) = 1$. Sia E una curva ellittica definita su $Z/(n_1 n_2 Z)$. Allora esiste un isomorfismo di gruppi fra*

$$E(Z/(n_1 n_2 Z) \sim E(Z/n_1 Z) \oplus E(Z/n_2 Z)$$

Per la dimostrazione si veda [15]

6.2 L'algoritmo ECM

Vedremo come le curve di Edwards sono utilizzate per migliorare il software GMP-ECM, un programma che si basa sulle curve ellittiche scritte in forma di Weierstrass, cioè nella forma $y^2 = x^3 + ax + b$ e nella forma di Montgomery, cioè $By^2 = x^3 + Ax^2 + x$.

Analizzeremo l'impatto delle curve di Edwards sul ECM, non solo nel numero di operazioni da eseguire, ma anche dal punto di vista, della velocità del software. Chiameremo EECM o GMP-EECM il nuovo programma, con la E aggiuntiva che ha il significato di Edwards (le curve considerate).

L'algoritmo sui cui si basa ECM è suddiviso in due parti. Focalizzeremo l'attenzione esclusivamente sulla prima in quando la più importante delle due, perché nella seconda le operazioni su curve ellittiche sono poche.

L'algoritmo ECM prova a fattorizzare un intero positivo n come segue. Scegliamo una curva ellittica E su Q. Prendiamo una funziona razionale $\phi : E \mapsto Q$ che abbia un polo nell'elemento neutro di E, per esempio si può scegliere ϕ come la coordinata x di Weierstrass. Scegliamo un punto $P \in E(Q)$ di non torsione. Scegliamo un intero positivo s che abbia molti fattori piccoli. Scegliamo una sequenza di addizioni, sottrazioni, moltiplicazioni e divisioni che, se eseguite su Q, diano come risultato $\phi([s]P)$, dove con $[s]P$ si indica l's-esimo multiplo di P su $E(Q)$. Calcoliamo $\phi([s]P) \mod n$, svolgendo le operazioni della sequenza scelta al punto precedente modulo n, sperando in una divisione impossibile modulo n, come nell'esempio all'inizio del capitolo. Se n ha un divisore primo q tale che $[s]P$ è l'elemento neutro per Z/qZ allora la prima parte dell'algoritmo ECM comprende una divisione impossibile modulo n, che ci rivela un fattore di n.

Questo accade quando s è un multiplo del gruppo di grandezza $E(Z/qZ)$. Al variare della curva E, varia anche il numero di elementi in $E(Z/qZ)$ all'interno dell'intervallo di Hasse.

Ciò che rende ECM interessante ed utile è che valori sorprendentemente piccoli di s, che permettono di calcolare molto velocemente $[s]P$, sono multipli di una percentuale molto ampia degli interi nell'intervallo di Hasse, inoltre s è un multiplo dell'ordine di P modulo q con una probabilità molto alta.

Facciamo un esempio. Scegliamo la curva $E : \quad y^2 = x^3 - 2$, come ϕ scegliamo la coordinata x di Weierstrass, come $P = (3, 5)$ e come intero $s = 420$.

Per calcolare $[420](3, 5)$ usiamo la formula standard di doubling e di addizione tra i punti.

Due vie veloci per trovare $[420](3, 5)$ sono:

$$[2](3,5), [4](3,5), [8](3,5), [16](3,5), [32](3,5), [64](3,5), [128](3,5),$$

$$[256](3,5), [256](3,5) + [128](3,5) = [384](3,5),$$

$$[384](3,5) + [32](3,5) = [416](3,5), [416](3,5) + [4](3,5) = [420](3,5)$$

oppure, con lo stesso numero di operazioni, possiamo trovare $[420](3,5)$ come:

$$[2](3,5), [4](3,5), [8](3,5), [8](3,5) + [4](3,5) = [12](3,5), [24](3,5),$$

$$[48](3,5), [96](3,5), [192](3,5), [384](3,5),$$

$$[384](3,5) + [24](3,5) = [408](3,5), [408](3,5) + [12](3,5) = [420](3,5)$$

Svolgiamo questi calcoli modulo n sperando in una divisione che non sia per 0 o per l'unità modulo n.

Il denominatore della coordinata x di Weierstrass di $[420](3,5)$ in $E(Q)$ ha molti fattori primi:

2, 3, 5, 7, 11, 19, 29, 31, 41, 43, 59, 67, 71, 83, 109, 163, 179, 181, 211, 223, 241, 269, 283, 383, 409, 419, 433, 523, 739,769, 811, 839,...

Se n ha uno di questi fattori primi, allora calcolando $[420](3,5)$ si incontrerà una divisione impossibile modulo n. Per controllare la presenza di uno dei fattori, nel nostro caso 769,811,839, si può osservare che $[420](3,5)$, è l'elemento neutro per i gruppi $E(Z/769Z), E(Z/811Z), E(Z/839Z)$. L'ordine di $(3,5)$ risulta essere rispettivamente 7,42,35. Possiamo notare che gli ordini dei gruppi sono 819, 756, 840, nessuno dei quali divide 420. Uno dei problemi dell'algoritmo è quindi scegliere una s adatta. Analizziamo alcuni esempi di come possa essere effettuata la scelta.

Pollard suggerisce di prendere s come il prodotto di tutti i primi $p_i < B_1$, ognuno elevato alla potenza c_i. La scelta dei c_i è libera, ma è consigliato non prendere le potenze più basse. Una possibilità, è quella di scegliere, per ogni primo $\pi < B_1$, la potenza più elevata di π tale che $\pi^{c_i} \in [1, 2\sqrt{n}+1]$. Allora $[s]P$ è l'elemento neutro di Z/qZ se e solo se l'ordine di p è B_1-liscio, cioè se e solo se l'ordine di P non ha divisori primi maggiori di B_1. Questa scelta

di P è buona da un punto di vista teorico, ma non è ottimale. Nella pratica è più conveniente scegliere la potenza maggiore di π che sta nell'intervallo $[1, B_1]$, in questo modo si riduce notevolmente la durata del processo senza ridurre significativamente la probabilità di successo.

6.3 L'algoritmo EECM

Questo algoritmo migliora quello appena enunciato cambiandolo in tre punti:

1. GMP-EECM usa le curve twisted di Edwards della forma $ax^2 + y^2 = 1 + dx^2y^2$ con coordinate di Edwards inverse e $\phi = \frac{1}{x}$, anziché le curve di Montgomery con le sue coordinate;

2. GMP-EECM tratta i fattori primi π di s in gruppi anziché uno alla volta. EECM calcola il prodotto t di un gruppo di primi, sostituisce P con $[t]P$ e quindi passa al gruppo successivo. Non è conveniente calcolare il prodotto di tutti i primi in una volta, poiché il costo delle moltiplicazioni necessarie per calcolare t sono bilanciate dal tempo risparmiato per t grandi. Notiamo inoltre che non c'è alcun motivo per il quale dato un punto P piccolo, anche $[t]P$ debba essere piccolo. Il vantaggio di avere un punto piccolo di partenza vale solo per il primo gruppo di primi;

3. GMP-EECM utilizza la catena di addizioni denominata signed sliding windows. Questa successione di operazioni permette di trovare $[t]P$ da P usando solamente un doubling ed un certo numero ϵ di addizioni per ogni bit di t. Facciamo notare che ϵ converge a 0 all'incrementare della lunghezza di t. Questo è il motivo per cui si risparmia tempo avendo un t grande. IL risparmio è amplificato dal fatto che un'addizione è un operazione più dispendiosa rispetto a doubling.

Nella scelta di s invece EECM segue l'idea di ECM. Vediamo un esempio riportato in [14].

Prendiamo $n = \frac{5^{367}+1}{2 \cdot 3 \cdot 73219364069}$. Come software usiamo GMP-ECM 6.1.3, scegliamo $B_1 = 16384$ ed utilizziamo come processore un Intel Pentium M (6b8) a 800 MHz. La prima parte dell'algoritmo richiede 210299 moltiplicazioni modulo n e richiede un totale di 2448 millisecondi.

Vediamo quanto tempo si impiega a trovare i fattori primi di n con GMP-EECM. Usiamo, per avere un confronto non falsato, lo stesso s e lo stesso B_1 del caso precedente, ma consideriamo questa volta la curva di Edwards:

$$x^2 + y^2 = 1 + \frac{161x^2y^2}{289}$$

con coordinate di Edwards inverse e signed sliding windows di lunghezza 6.

La prima parte dell'algoritmo richiede ora solamente 195111 moltiplicazioni modulo n e 2276 millisecondi per dare in output 70057995652034894429. Andando ad analizzare l'ordine del punto, scopriamo che il divisore primo più grande dell'ordine è 9103, mentre il secondo è 2459.

Bibliografia

[1] N. Koblitz, Elliptic curve cryptosystems, Math. Comp. 48 (1987), no. 177, 203–209.

[2] N. Koblitz, Algebraic aspects of cryptography. With an appendix by Alfred J. Menezes, Yi-Hong Wu and Robert J. Zuccherato. Algorithms and Computation in Mathematics, 3. Springer-Verlag, Berlin, 1998.

[3] N. Koblitz, Getting a few things right and many things wrong. Indocrypt 2010, 13 dicembre 2010.

[4] V. Miller, Use of elliptic curves in cryptography, CRYPTO 85, Advances in cryptology — CRYPTO '85 (Santa Barbara, Calif., 1985), 417–426,Lecture Notes in Comput. Sci., 218, Springer, Berlin, 1986.

[5] B. B. Brumly and N. Tuveri, Remote timing attacks are still practical, available at http://eprint.iacr.org/2011/232.

[6] Sito web http://eprint.iacr.org

[7] Sito web wikipedia, voce Elliptic Curve DSA, controllata il 20 giugno 2011.

[8] B. B. Brumley and D. Boneh, Remote timing attacks are practical. In: Proceedings of the 12th USENIX Security Symposium, 2003.

[9] B. B. Brumley and D. Boneh, Remote timing attacks are practical. Computer Networks 48 (5): 701–716, 2005.

[10] B. B. Brumly and N. Tuveri, Remote timing attacks are still practical, available at http://eprint.iacr.org/2011/232.

[11] D. Hankerson, A. Menezes and S. Vanstone, Guide to Elliptic Curve Cryptography, Springer, 2004.

[12] P. C. Kocher, Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In Neal Koblitz, editor, CRYPTO, Lect. Notes in Computer Sciences, vol. 1109, pp. 104–113, Springer, 1996.

[13] G. Benini, Le curve di Edwards e la loro utilità. Trento, tesi di laurea a.a. 2008/2009, relatore W. De Graaf

[14] D. J. Bernstein, P. Birkner, M. Joye, T. Large e C. Peters, Twisted Edwards curves. http://eem.cr.yp.to/eecm-20080120.pdf, 2009.

[15] Lawrence C. Washington, Elliptic Curves, Chapman Hall. Crc, 2003.

[16] H. W. Lenstra, Factoring integers with elliptic curves. Annals of Mathematics, 1987.

[17] D. J. Berstein, T. Lange, A complete set of addition laws for incomplete Edwards curves, J. Number Theory 131 (2011), 858–872.