Montgomery curves and the Montgomery ladder

Daniel J. Bernstein and Tanja Lange

Technische Universiteit Eindhoven, The Netherlands
University of Illinois at Chicago, USA

Abstract. The Montgomery ladder is a remarkably simple method of computing scalar multiples of points on a broad class of elliptic curves. This article surveys a wide range of topics related to the Montgomery ladder, both from the historical perspective of Weierstrass curves and from the modern perspective of Edwards curves. New material includes a full proof of a complete constant-time ladder algorithm suitable for cryptographic applications.

This article is planned to appear as Chapter 4 of the book *Topics in Computational Number Theory inspired by Peter L. Montgomery*, edited by Joppe W. Bos and Arjen K. Lenstra. Two cross-references are to descriptions of ECM in *FFT extension for algebraic-group factorization algorithms*, by Richard P. Brent, Alexander Kruppa, and Paul Zimmermann, which is planned to appear as Chapter 9 of the same book.

Author list in alphabetical order; see `https://www.ams.org/profession/leaders/culture/CultureStatement04.pdf`. This work was supported by the Commission of the European Communities through the Horizon 2020 program under project ICT-645421 (ECRYPT-CSA), by the U.S. National Science Foundation under grants 1018836 and 1314919, and by the Netherlands Organisation for Scientific Research (NWO) under grant 639.073.005. "Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation" (or other funding agencies). Permanent ID of this document: `8d6889ca8534ce31cdc989005f2b27fb25c301cf`. Date: 2017.03.30.

Contents

4

Montgomery curves and the Montgomery ladder

Daniel J. Bernstein and Tanja Lange

4.1 Introduction

The **Montgomery ladder** is the following remarkably simple method of computing scalar multiples of points on a broad class of elliptic curves. Define sequences $(X_1, X_2, \ldots)$ and $(Z_1, Z_2, \ldots)$, starting from X_1, Z_1, A, by the equations

$$X_{2n} = (X_n^2 - Z_n^2)^2, \qquad X_{2n+1} = 4(X_n X_{n+1} - Z_n Z_{n+1})^2 Z_1,$$
$$Z_{2n} = 4X_n Z_n (X_n^2 + AX_n Z_n + Z_n^2), \qquad Z_{2n+1} = 4(X_n Z_{n+1} - Z_n X_{n+1})^2 X_1$$

for $n \geq 1$. Then the points

$$\left(\frac{X_n}{Z_n}, \pm\sqrt{\frac{1}{B}\left(\frac{X_n^3}{Z_n^3} + A\frac{X_n^2}{Z_n^2} + \frac{X_n}{Z_n}\right)}\right)$$

are, under minor hypotheses, the nth multiples of the points

$$\left(\frac{X_1}{Z_1}, \pm\sqrt{\frac{1}{B}\left(\frac{X_1^3}{Z_1^3} + A\frac{X_1^2}{Z_1^2} + \frac{X_1}{Z_1}\right)}\right)$$

on the **Montgomery curve** $By^2 = x^3 + Ax^2 + x$. The Montgomery ladder is also remarkably fast: the optimized formulas

$$X_{2n} = (X_n - Z_n)^2 (X_n + Z_n)^2,$$
$$Z_{2n} = ((X_n + Z_n)^2 - (X_n - Z_n)^2)\left((X_n + Z_n)^2 + \frac{A-2}{4}((X_n + Z_n)^2 - (X_n - Z_n)^2)\right),$$
$$X_{2n+1} = ((X_n - Z_n)(X_{n+1} + Z_{n+1}) + (X_n + Z_n)(X_{n+1} - Z_{n+1}))^2 Z_1,$$
$$Z_{2n+1} = ((X_n - Z_n)(X_{n+1} + Z_{n+1}) - (X_n + Z_n)(X_{n+1} - Z_{n+1}))^2 X_1$$

compute (X_n, Z_n) using just 11 multiplications per bit of n.

Montgomery introduced these curves and optimized formulas in a classic

1987 paper "Speeding the Pollard and elliptic curve methods of factorization" [Mon87]. See Chapter 9 for more information about ECM, the elliptic-curve method of factorization.

The advent of ECM prompted further applications of elliptic-curve computations, notably elliptic-curve primality proving (ECPP) and elliptic-curve cryptography (ECC). It is easy to see that these applications can also use the Montgomery ladder. Extensive research has produced a wide range of more complicated scalar-multiplication methods (for pointers see, e.g., [BDL+11], [BCLS14], and [BL16]), outperforming the Montgomery ladder for tasks such as computing $n \mapsto nP$ for a *fixed* point P, or computing nth multiples of points on certain special curves, but the Montgomery ladder seems practically unbeatable for the core task of computing $n, P \mapsto nP$ on typical curves.

In ECC it is important to avoid failure cases, so the minor hypotheses mentioned above are worrisome. Fortunately, a careful analysis shows that the Montgomery ladder *always* computes a modified x-coordinate function that identifies ∞ with 0. Working correctly for *all* inputs is an unusual feature of elliptic-curve formulas: one expects scalar-multiplication methods to have failure cases that require constant attention from implementors.

Twenty years later the introduction of "complete Edwards curves" allowed algebraic computations of arbitrary sums $n_1P_1 + \cdots + n_kP_k$ by the "Edwards addition law" without failure cases. It turned out that complete Edwards curves are birationally equivalent to Montgomery curves with points of order 4 and unique points of order 2, and vice versa. More generally, "twisted Edwards curves" are birationally equivalent to Montgomery curves, and vice versa. The Montgomery ladder is closely related to the Edwards addition law, as we show in Sections 4.4 and 4.5.

The United States National Institute of Standards and Technology (NIST) issued ECC standards fifteen years ago. These standards recommended various non-Montgomery curves that had been selected by the National Security Agency. The only justification provided for the curve shape was an incorrect claim that the standards provided "the fastest arithmetic on elliptic curves". The simplicity, speed, and completeness of the Montgomery ladder have led to widespread deployment of "Curve25519" [Ber06a], the Montgomery curve $y^2 = x^3 + 486662x^2 + x$ over the prime field $\mathbb{F}_p$ where $p = 2^{255} - 19$; see Section 4.3.7 for details.

4.2 Fast scalar multiplication on the clock

We define the **clock** as the curve $u^2 + v^2 = 1$ with specified point $(0, 1)$. More generally, we define a **twisted clock** as a curve $au^2+v^2 = 1$ with specified point $(0, 1)$, where a is nonzero. This section introduces the **group of points** on this curve and relates the computation of point multiples to Lucas sequences. The Lucas ladder can be viewed as a "degeneration" of the Montgomery ladder.

Fix a field k and fix a nonzero $a \in k$. Define $\mathrm{Clock}_a(k)$ as the set of k-points on the twisted clock, i.e., the set of pairs $(u, v) \in k \times k$ satisfying the curve equation $au^2 + v^2 = 1$. Then $\mathrm{Clock}_a(k)$ is an abelian group under the following operations. The neutral element is the specified point $(0, 1)$. The negative $-(u, v)$ of a point (u, v) is $(-u, v)$. The sum (u_5, v_5) of (u_2, v_2) and (u_3, v_3) is given by

$$(u_5, v_5) = (u_2, v_2) + (u_3, v_3) = (u_2v_3 + u_3v_2, v_2v_3 - au_2u_3).$$

The difference is $(u_1, v_1) = (u_3, v_3) - (u_2, v_2) = (-u_2v_3 + u_3v_2, v_2v_3 + au_2u_3)$.

For the special case $(k, a) = (\mathbb{R}, 1)$ the addition operation can be visualized as adding times on a conventional clock, using $12{:}00 = (0, 1)$ as neutral element. For example, $2{:}00 + 3{:}00 = 5{:}00$, and $9{:}00 + 4{:}00 = 1{:}00$.

The addition can be computed with just 4 multiplications using the sequence of intermediate steps $A = u_2u_3$, $B = v_2v_3$, $C = (u_2 + v_2)(u_3 + v_3)$ to get $(u_5, v_5) = (C - A - B, B - aA)$. We denote the cost of a general multiplication by $\mathbf{M}$ and the cost of multiplication by a curve constant (such as a) by $\mathbf{C}$, so one point addition costs $3\mathbf{M} + \mathbf{C}$. Additions and subtractions are usually not counted: they are significantly cheaper than multiplications for typical fields.

Doubling means adding a point to itself, giving $(u_4, v_4) = (u_2, v_2)+(u_2, v_2) = (2u_2v_2, v_2^2 - au_2^2) = (2u_2v_2, 2v_2^2 - 1)$, costing $\mathbf{M} + \mathbf{S}$, where $\mathbf{S}$ denotes the cost of a squaring.

Note that, in doubling, v_4 is computed purely from v_2 and does not involve u_2. Similarly, $v_5 = v_2v_3 - au_2u_3 = 2v_2v_3 - v_1$, showing that the v-coordinate of the sum $P + Q$ can be computed given the v-coordinates of P, Q, and $Q - P$.

For each $n \geq 0$, the **scalar multiple** nP is $P + P + \cdots + P$, adding n copies of P together. For example, (u_4, v_4) above is $2(u_2, v_2)$. Computing nP in a naive way takes $n - 1$ additions for $n \geq 1$, i.e., $3(n - 1)\mathbf{M} + (n - 1)\mathbf{C}$, but using the binary expansion $n = \sum_{i=0}^{c} n_i 2^i$ (with $n_i \in \{0, 1\}$ and $n_c = 1$) to compute

$$nP = 2(2(\cdots 2(2P + n_{c-1}P) + n_{c-2}P \cdots) + n_1P) + n_0P$$

takes only $c = \lfloor \log_2 n \rfloor$ doublings and at most c additions. This **double-and-add method** takes on average $c(\mathbf{M} + \mathbf{S}) + 0.5c(3\mathbf{M} + \mathbf{C}) = 2.5c\mathbf{M} + c\mathbf{S} + 0.5c\mathbf{C}$ to compute nP.

4.2.1 The Lucas ladder

Fix $(u_1, v_1) \in \mathrm{Clock}_a(k)$. Define $(u_n, v_n) = n(u_1, v_1)$ for each $n \geq 2$. Then

$$v_{2n} = 2v_n^2 - 1, \qquad v_{2n+1} = 2v_n v_{n+1} - v_1.$$

Recursively applying these two formulas computes (v_n, v_{n+1}) using just $c\mathbf{M}+c\mathbf{S}$ if n has c bits. Specifically, to compute (v_n, v_{n+1}), first recursively compute (v_m, v_{m+1}) where $m = \lfloor n/2 \rfloor$, and then compute v_n and v_{n+1} using two out of the three formulas

$$v_{2m} = 2v_m^2 - 1, \qquad v_{2m+1} = 2v_m v_{m+1} - v_1, \qquad v_{2m+2} = 2v_{m+1}^2 - 1,$$

namely the first two if n is even, and the last two if n is odd. Either way costs $\mathbf{M} + \mathbf{S}$. The base case is $n = 0$, where $(v_n, v_{n+1}) = (1, v_1)$.

As an example, Figure 4.1 shows the indices used in computing (v_{73}, v_{74}). A double arrow from m to $2m$ indicates that v_{2m} is computed from v_m. Single arrows from m and $m + 1$ to $2m + 1$ indicate that v_{2m+1} is computed from v_m and v_{m+1}. One can save time in the first few lines by skipping recomputations of v_1 and v_2.

The cost $c\mathbf{M} + c\mathbf{S}$ here is significantly less than the cost of the double-and-add method. This comparison might seem unfair: if the objective is to compute nth multiples then the double-and-add method produces (u_n, v_n) while this recursion does not seem to produce u_n. However, v_n is sufficient for many applications. Furthermore, the recursion is best understood as producing both v_n and v_{n+1}, and solving for u_n in the addition formula $v_{n+1} = v_1 v_n - a u_1 u_n$ produces a "u-recovery" formula $u_n = (v_1 v_n - v_{n+1})/(a u_1)$, assuming $u_1 \neq 0$.

The sequence $(2v_1, 2v_2, 2v_3, \dots)$ is an example of a **Lucas sequence of the second kind** over k, i.e., a sequence of the form $(\alpha + \beta, \alpha^2 + \beta^2, \alpha^3 + \beta^3, \cdots)$ where $\alpha + \beta, \alpha\beta \in k$; specifically, take $\alpha = v_1 + u_1\sqrt{-a}$ and $\beta = v_1 - u_1\sqrt{-a}$. A **Lucas sequence of the first kind** has nth entry $(\alpha^n - \beta^n)/(\alpha - \beta)$. The special case $\alpha\beta = 1$ used here was introduced by Chebyshev before Lucas: v_n, viewed as a polynomial in v_1, is the nth **Chebyshev polynomial of the first kind**.

4.2.2 Differential addition chains

Montgomery in [Mon92b] introduced an even faster method of computing v_n. Experiments show this method taking only about $1.55\mathbf{M}$ per bit of n, as already announced in [Mon87]. The idea of this method applies to computing the nth term x_n of any sequence $(x_0, x_1, \dots)$ that satisfies a recurrence of the form

$$x_{m+n} = f(x_m, x_n, x_{n-m}),$$

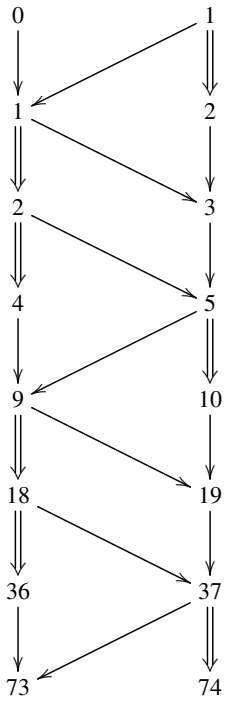

Figure 4.1 A uniform double-add ladder.

for some function f, starting from some initial values x_0 and x_1. The clock example has $f(v_2, v_3, v_1) = 2v_2v_3 - v_1$, starting with $x_0 = 1$ and with x_1 as the v-coordinate of an input point.

To compute x_8, starting from x_0 and x_1, we compute $x_2 = f(x_1, x_1, x_0)$, $x_4 = f(x_2, x_2, x_0)$, and $x_8 = f(x_4, x_4, x_0)$. To compute x_9 we cannot simply extend this chain using x_8 and x_1 because we do not have $x_{8-1} = x_7$. Instead we can compute it via $x_2 = f(x_1, x_1, x_0)$, $x_3 = f(x_1, x_2, x_1)$, $x_4 = f(x_2, x_2, x_0)$, $x_5 = f(x_2, x_3, x_1)$, $x_9 = f(x_4, x_5, x_1)$.

The indices $0, 1, 2 = 1 + 1, 3 = 2 + 1, 4 = 2 + 2, 5 = 3 + 2, 9 = 5 + 4$ form a **differential addition chain**. This means a sequence that starts $0, 1$ and that continues with sums $n + m$ where $n, m, n - m$ all appear earlier in the sequence. Montgomery calls these chains "Lucas chains"; other names in the literature include "strong addition chain" and "Chebyshev chain".

The simplest way to build a differential addition chain is to allow only 0 and 1 as differences $n - m$: i.e., to compute x_{2m} as $f(x_m, x_m, x_0)$ for $m \geq 1$ and to

compute x_{2m+1} as $f(x_m, x_{m+1}, x_1)$ for $m \geq 1$. Montgomery calls this the "binary method"; we follow common naming and call it a "ladder". This method takes two evaluations of f per bit of n to compute x_n and x_{n+1} from x_0 and x_1. In the clock example, f costs **S** for $n = m$ and **M** for $n = m + 1$, for a cost of $\mathbf{M} + \mathbf{S}$ per bit to compute the v-coordinate of nP from the v-coordinate of P, as mentioned above.

Shorter chains exist for many numbers: e.g., 9 can be reached via the chain $0, 1, 2, 3, 6, 9$, taking 4 steps instead of 5. It might also be helpful to allow a **differential addition-subtraction chain**: this means that one allows not only sums $n + m$ after $n, m, n - m$, but also differences $n - m$ after $n, m, n + m$, using some f' with $x_{n-m} = f'(x_m, x_n, x_{m+n})$. For the clock one can take $f' = f$.

In [Mon92b], Montgomery studied lower bounds for the lengths of these chains, and systematic methods to find short chains. See Section 4.8 below for several such methods. Montgomery's PRAC method (short for "Practical Algorithm") achieves the 1.55**M** per bit mentioned above.

4.3 Montgomery curves

Fix a field k not of characteristic 2, and fix $A, B \in k$ with $B(A^2 - 4) \neq 0$. The curve $By^2 = x^3 + Ax^2 + x$ is a **Montgomery curve**. This section introduces the **group of points** on this curve, both from the historical perspective of Weierstrass curves and from the modern perspective of Edwards curves.

4.3.1 Montgomery curves as Weierstrass curves

A **short Weierstrass curve** is a curve of the form $y^2 = x^3 + ax + b$ where $4a^3 + 27b^2 \neq 0$. A small calculation (relying on the hypothesis that $2 \neq 0$ in k) shows that this curve is geometrically nonsingular: this means that the equation $y^2 = x^3 + ax + b$, its x-derivative $0 = 3x^2 + a$, and its y-derivative $2y = 0$ have no common solutions (x, y) over any extension of k. Indeed, any common solution has $y = 0$ so $x^3 + ax + b = 0$ so $b = -x^3 - ax = 2x^3$ so $4a^3 + 27b^2 = -108x^6 + 108x^6 = 0$, contradiction.

More generally (even in characteristic 2), a **Weierstrass curve** is a geometrically nonsingular curve of the form $a_0y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$ with $a_0 \neq 0$. The Montgomery curve $By^2 = x^3 + Ax^2 + x$ has this form with $(a_0, a_1, a_3, a_2, a_4, a_6) = (B, 0, 0, A, 1, 0)$, and is geometrically nonsingular, so it is a Weierstrass curve. The nonsingularity calculation boils down to the calculation that the cubic polynomial $x^3 + Ax^2 + x$ has discriminant $A^2 - 4$, which was hypothesized to be nonzero. Concretely, if $By^2 = x^3 + Ax^2 + x$ and $2By = 0$

and $3x^2 + 2Ax + 1 = 0$ then $y = 0$ so $x^3 + Ax^2 + x = 0$, so the discriminant identity

$$A^2 - 4 = \Big((-6A^2 + 18)x + (-4A^3 + 15A)\Big)(x^3 + Ax^2 + x) + \Big((2A^2 - 6)x^2 + (2A^3 - 7A)x + (A^2 - 4)\Big)(3x^2 + 2Ax + 1)$$

implies that $A^2 - 4 = 0$, contradiction.

The name "Weierstrass curve" arises, historically, from an identity of the form $a_0(\wp')^2 = \wp^3 + a_4\wp + a_6$ satisfied by the Weierstrass $\wp$ function and its derivative $\wp'$, specifically with $a_0 = 1/4$. In other words, $(\wp, \wp')$ are points (x, y) on the Weierstrass curve $a_0y^2 = x^3 + a_4x + a_6$.

We are deviating slightly from standard terminology here. The standard definition of "Weierstrass curve" in the literature assumes $a_0 = 1$, so it allows the Montgomery curve $By^2 = x^3 + Ax^2 + x$ only in the case $B = 1$. Dropping the restriction $a_0 = 1$ allows $(\wp, \wp')$ to be points (x, y) on a "Weierstrass curve" without any rescaling, allows Montgomery curves without any rescaling, and makes the theory of Weierstrass curves only negligibly more complicated.

4.3.2 The group law for Weierstrass curves

The **set of k-points** on a Weierstrass curve W, written $W(k)$, is the set of pairs $(x, y) \in k \times k$ satisfying the curve equation, together with an extra point ∞. The points (x, y) are called **affine points**. Define a unary operation $-$ on $W(k)$ as follows:

- $-\infty = \infty$.
- $-(x, y) = (x, -(y + (a_1/a_0)x + (a_3/a_0)))$.

Define a binary operation $+$ on $W(k)$ as follows:

- $\infty + \infty = \infty$.
- $\infty + (x, y) = (x, y)$.
- $(x, y) + \infty = (x, y)$.
- $(x, y) + (-(x, y)) = \infty$.
- If $2a_0y + a_1x + a_3 \neq 0$ then $(x, y) + (x, y) = -(x'', y + \lambda(x'' - x))$, where $\lambda = (3x^2 + 2a_2x + a_4)/(2a_0y + a_1x + a_3)$ and $x'' = a_0\lambda^2 + a_1\lambda - a_2 - 2x$.
- If $x' \neq x$ then $(x, y) + (x', y') = -(x'', y + \lambda(x'' - x))$ where $\lambda = (y' - y)/(x' - x)$ and $x'' = a_0\lambda^2 + a_1\lambda - a_2 - x - x'$.

One can prove that these definitions cover all cases; that the ouptuts are in $W(k)$; and that $W(k)$ is a commutative group with ∞ as neutral element, $-$ as negation, and $+$ as addition.

For ease of reference we repeat the rules in the special case of Montgomery curves, using the simplifications $a_1 = 0$, $a_3 = 0$, $a_0 = B$, $a_2 = A$, $a_4 = 1$, $a_6 = 0$, and $2B \neq 0$:

- $-\infty = \infty$.
- $-(x, y) = (x, -y)$.
- $\infty + \infty = \infty$.
- $\infty + (x, y) = (x, y)$.
- $(x, y) + \infty = (x, y)$.
- $(x, y) + (x, -y) = \infty$.
- If $y \neq 0$ then $(x, y) + (x, y) = -(x'', y + \lambda(x'' - x))$, where $\lambda = (3x^2 + 2Ax + 1)/(2By)$ and $x'' = B\lambda^2 - A - 2x$.
- If $x' \neq x$ then $(x, y) + (x', y') = -(x'', y + \lambda(x'' - x))$ where $\lambda = (y' - y)/(x' - x)$ and $x'' = B\lambda^2 - A - x - x'$.

4.3.3 Other views of the group law

The **projective k-points** on W are all points $(X : Y : Z) \in \mathbb{P}^2(k)$ satisfying the homogeneous equation

$$a_0ZY^2 + a_1ZXY + a_3Z^2Y = X^3 + a_2ZX^2 + a_4Z^2X + a_6Z^3.$$

Here $\mathbb{P}^2(k) = \{(X : Y : Z) : (X, Y, Z) \in k^3 - \{(0, 0, 0)\}\}$. Sometimes $(X : Y : Z)$ is defined as the subspace $\{(\lambda X, \lambda Y, \lambda Z) : \lambda \in k\}$ of the k-vector space k^3; sometimes it is defined as the set $\{(\lambda X, \lambda Y, \lambda Z) : \lambda \in k^*\}$. With either definition, $(X' : Y' : Z') = (X : Y : Z)$ if and only if $(X', Y', Z') = (\lambda X, \lambda Y, \lambda Z)$ for some $\lambda \in k^*$.

For each affine point $(x, y) \in W(k)$ there is a corresponding projective k-point $(x : y : 1)$ on W. The point $\infty \in W(k)$ corresponds to the projective k-point $(0 : 1 : 0)$ on W. These cover all projective k-points on W: if $Z \neq 0$ then $(X : Y : Z) = (x : y : 1)$ where $x = X/Z$ and $y = Y/Z$, and then the homogeneous equation implies $(x, y) \in W(k)$; if $Z = 0$ then the homogeneous equation forces $(X : Y : Z) = (0 : 1 : 0)$. Taking projective coordinates thus unifies the two cases in the definition of $W(k)$.

As for the addition law, the whole group definition can be understood as just two rules:

- ∞ is the neutral element in the group.
- There is a standard definition of the multiset of intersection points of a line with the curve; if this multiset consists of three points P, Q, R then $P+Q+R = 0$ in the group.

There are several reasons that the second rule splits into cases. First, the multiset is not always a set; for example, if a line is tangent to the curve at P then P appears at least twice in the multiset. Second, the multiset is defined in terms of projective points, so it does not always consist of affine points; for example, a vertical line intersects the curve at ∞. Third, *if* k is algebraically closed then the multiset always has size exactly 3 by Bézout's theorem, but for more general fields k a line can intersect the curve in fewer points.

Sometimes "elliptic curve" is defined more generally as

- a nonsingular cubic curve C in two-dimensional projective space with a specified inflection point I (such as ∞ for Weierstrass curves); or
- a nonsingular cubic curve in two-dimensional projective space with a specified point (not necessarily an inflection point); or
- a nonsingular genus-1 curve in n-dimensional projective space with a specified point.

With the first definition, one can use the same $P + Q + R = 0$ rule to define a group law on $C(k)$ with I as neutral element. With the second definition, slightly more work is required; see, e.g., [Hus04, Chapter 3, Theorem 1.2]. With the third definition, the standard approach is to abandon the chord-and-tangent approach and instead declare that the zeros and poles of *any* algebraic function on C, not just a linear function, have sum 0. The main work is then to show that points P of $C(k)$ map bijectively to elements $P - I$ of this "divisor class group", see [ACD+05, Chapter 4] for details.

4.3.4 Edwards curves and their group law

An **Edwards curve** is a curve of the form $u^2 + v^2 = 1 + du^2v^2$ with $d \notin \{0, 1\}$ over a field k not of characteristic 2. Edwards curves were introduced in a slightly less general form by Edwards in [Edw07], who also defined a group operation on them. Bernstein and Lange in [BL07] introduced this form and defined efficient formulas for the group operation. A **twisted Edwards curve** [BBJ+08] is a curve of the form $au^2 + v^2 = 1 + du^2v^2$ with $a \neq d$ and $a, d \neq 0$.

A twisted Edwards curve E where a is square in k and d is non-square in k is called k**-complete**, or simply **complete** if k is clear from context. In this case the group of k-points of E, written $E(k)$, is defined as the set of $(u, v) \in k \times k$ satisfying the curve equation, with the following operations. The neutral element is $(0, 1)$. The negative of (u, v) is $(-u, v)$. The sum of two points (u_2, v_2) and (u_3, v_3) is defined as

$$(u_2, v_2) + (u_3, v_3) = \left(\frac{u_2v_3 + u_3v_2}{1 + du_2u_3v_2v_3}, \frac{v_2v_3 - au_2u_3}{1 - du_2u_3v_2v_3} \right). \tag{4.1}$$

The denominators are never 0; see [BL07]. The Edwards addition law (4.1) is a **complete addition law**, i.e., an addition law that holds for all inputs.

To define $E(k)$ for general a and d, without the requirements of a being square and d being non-square, one needs more work: the addition law (4.1) is defined almost everywhere but can produce divisions by 0. The general definition is as follows.

Define $E(k)$ to have the following elements: all $(u, v) \in k \times k$ satisfying the curve equation; $(\pm 1/\sqrt{d}, \infty)$ if d is a square; and $(\infty, \pm\sqrt{a/d})$ if a/d is a square. In other words, $E(k)$ is the set of $(u, v) \in (k \cup \{\infty\}) \times (k \cup \{\infty\})$ satisfying the curve equation, with a careful definition of arithmetic on ∞. Formally, consider the projective embedding of E into $\mathbb{P}^1 \times \mathbb{P}^1 = \{((U : Z), (V : T))\}$, namely

$$aU^2T^2 + V^2Z^2 = Z^2T^2 + dU^2V^2.$$

Each affine point (u, v) corresponds to the projective point $((u : 1), (v : 1))$. Additional projective points are $((1 : \pm\sqrt{d}), (1 : 0))$ and $((1 : 0), (\pm\sqrt{a/d} : 1))$ if those are defined over k. We identify $(1 : 0)$ with ∞ and identify the rest of $\mathbb{P}^1(k)$ with k, so each point is a pair of coordinates in $k \cup \{\infty\}$.

As before, the neutral element of $E(k)$ is $(0, 1)$, and the negative of (u, v) is $(-u, v)$, where $-\infty$ means ∞. The sum of two points (u_2, v_2) and (u_3, v_3) is defined by the Edwards addition law (4.1) together with the **dual addition law**

$$(u_2, v_2) + (u_3, v_3) = \left(\frac{u_2v_2 + u_3v_3}{au_2u_3 + v_2v_3}, \frac{u_2v_2 - u_3v_3}{u_2v_3 - u_3v_2}\right). \tag{4.2}$$

For each pair of points $((u_2, v_2), (u_3, v_3)) \in E(k) \times E(k)$, at least one of these laws is defined. Here "defined" allows divisions by 0, producing ∞ as output, but does not allow $0/0$. If both laws are defined then the results are identical. This is true for each coordinate separately: if both laws have a defined u-coordinate then those u-coordinates are identical; if both laws have a defined v-coordinate then those v-coordinates are identical. $E(k)$ forms an abelian group under these operations.

The dual addition law was introduced by Hisil, Wong, Carter, and Dawson in [HWCD08]. The completeness of the set of two addition laws was shown by Bernstein and Lange in [BL11].

The v-coordinate in the dual addition law (4.2) is undefined if and only if $(u_3, v_3) = (u_2, v_2)$ or $(u_3, v_3) = (-u_2, -v_2)$. By completeness, the Edwards addition law (4.1) is defined in these cases. In particular, the Edwards addition law is a valid formula for doubling any point. Write $(u_4, v_4) = 2(u_2, v_2)$; then $u_2^2 = (1 - v_2^2)/(a - dv_2^2)$, so

$$v_4 = \frac{v_2^2 - au_2^2}{1 - du_2^2v_2^2} = \frac{v_2^2(a - dv_2^2) - a(1 - v_2^2)}{a - dv_2^2 - d(1 - v_2^2)v_2^2} = \frac{2av_2^2 - a - dv_2^4}{a - 2dv_2^2 + dv_2^4}.$$

Note for future reference that this formula expresses v_4 purely in terms of v_2. This formula has no exceptional cases: in particular, it works for $u_2 = \infty$ and for $v_2 = \infty$.

4.3.5 Montgomery curves as Edwards curves

Edwards curves and Montgomery curves are examples of elliptic curves with some special properties. In particular, Edwards curves have a point of order 4 at $(1, 0)$ and a point of order 2 at $(0, -1)$. Montgomery curves have a point of order 2 at $(0, 0)$ and, over finite fields, at least one of the following: a point of order 4 doubling to $(0, 0)$ or two more points of order 2. The same conditions hold for twisted Edwards curves.

In fact, Montgomery curves and twisted Edwards curves cover the same set of elliptic curves. More precisely, for each Montgomery curve there is a birationally equivalent twisted Edwards curve, and vice versa. Here a **birational equivalence** between two elliptic curves M, E is a pair of rational maps $M \to E$ and $E \to M$ that are defined almost everywhere, that are each other's inverses when both are defined, and that map specified neutral element to specified neutral element. One can show that a birational equivalence preserves addition wherever it is defined, and can be extended to a group isomorphism.

Specifically, the transformation formulas from the twisted Edwards curve $au^2 + v^2 = 1 + du^2v^2$ to the Montgomery curve $By^2 = x^3 + Ax^2 + x$ are

$$x = \frac{1+v}{1-v} \text{ and } y = \frac{1+v}{u(1-v)} = \frac{x}{u},$$

where the curve parameters have the relationship

$$A = 2\frac{a+d}{a-d} \text{ and } B = \frac{4}{a-d}.$$

Likewise, the formulas from the Montgomery curve to the twisted Edwards curve are

$$u = \frac{x}{y} \text{ and } v = \frac{x-1}{x+1}$$

and the curve parameters satisfy

$$a = \frac{A+2}{B} \text{ and } d = \frac{A-2}{B}.$$

The map from v to $x = (1 + v)/(1 - v)$ is defined for all $v \in k \cup \{\infty\}$ if ∞ is handled carefully as input and output. If $v \in k - \{1\}$ then $x \in k - \{-1\}$. If $v = 1$ then $x = \infty$; here the input point is $(0, 1)$, the neutral element, and the output point is ∞ as required. If $v = \infty$ then $x = -1$; here the input is an order-4

point $(\pm 1/\sqrt{d}, \infty)$ whose double is the order-2 point $(0, -1)$, and the output is an order-4 point $(-1, \mp\sqrt{d})$ whose double is the order-2 point $(0, 0)$.

The inverse map from x to $v = (x-1)/(x+1)$ is similarly defined for all $x \in k \cup \{\infty\}$. If $x \in k - \{-1\}$ then $v \in k - \{1\}$. If $x = -1$ then $v = \infty$. If $x = \infty$ then $v = 1$. These are inverse maps since $v = ((1+v)/(1-v)-1)/((1+v)/(1-v)+1)$ and $x = (1 + (x-1)/(x+1))/(1 - (x-1)/(x+1))$.

4.3.6 Elliptic-curve cryptography (ECC)

Miller in [Mil86], and independently Koblitz in [Kob87], proposed an elliptic-curve variant of the Diffie–Hellman key exchange method [DH76]. Miller in [Mil86, page 425] suggested exchanging just x-coordinates instead of (x, y)-coordinates: i.e., sending $\mathbf{x}(P)$ rather than an entire point P, where $\mathbf{x}(x, y) = x$.

The **Diffie–Hellman key exchange** with x-coordinates works as follows. One user, say Alice, has a secret key s and a public key $\mathbf{x}(sP)$. Here s is an integer, and P is a standard point on a standard Weierstrass curve. Another user, say Bob, has a secret key t and a public key $\mathbf{x}(tP)$. Alice and Bob then both know a shared secret $\mathbf{x}(stP) = \mathbf{x}(s(tP)) = \mathbf{x}(t(sP))$, which seems quite difficult for an attacker to predict.

Note that $\mathbf{x}(stP)$ is entirely determined by s and $\mathbf{x}(tP)$. Indeed, the only possible ambiguity in recovering tP from $\mathbf{x}(tP)$ is the possible distinction between tP and $-tP$, and this distinction has no effect on $\mathbf{x}(stP)$: the x-coordinate is invariant under point negation, so $\mathbf{x}(s(-tP)) = \mathbf{x}(-stP) = \mathbf{x}(stP)$. The same argument applies if x-coordinates on Weierstrass curves are replaced by v-coordinates on twisted Edwards curves.

The bottleneck here is elliptic-curve scalar multiplication: Alice first has to compute her public key $\mathbf{x}(sP)$ given her secret key s, and then has to compute the shared secret $\mathbf{x}(stP)$ given her secret key s and Bob's public key $\mathbf{x}(tP)$. For any Weierstrass curve, Alice can compute a square root to obtain $\pm tP$ from $\mathbf{x}(tP)$, can use the double-and-add method with the respective doubling and addition formulas to obtain $\pm stP$, and can then discard the y-coordinate to obtain $\mathbf{x}(stP)$. For Montgomery curves, Alice can use the more efficient Montgomery ladder to compute $\mathbf{x}(stP)$ from $\mathbf{x}(tP)$, using the doubling and differential-addition formulas developed in this chapter.

Electronic signatures use secret key s and public key sP, i.e., they require both coordinates of sP. This makes short Weierstrass curves or (twisted) Edwards curves the more common choice. Verification of a signature typically involves a double-scalar multiplication $mP + nQ$, of which only $\mathbf{x}(mP + nQ)$ is used. This can be computed using a 2-dimensional addition chain (Section 4.7).

4.3.7 Examples of noteworthy Montgomery curves

Montgomery introduced Montgomery curves for more efficient factorization of integers in ECM, the elliptic-curve method of factorization; see Chapter 9. Montgomery's thesis [Mon92a] includes several Montgomery curves that are particularly suitable for ECM because the curves are guaranteed to have large $\mathbb{Q}$-torsion. These curves form parameterized families; the curve with "Suyama parameter $\sigma = 11$" was further analyzed in [BBB+13].

In cryptography, Bernstein's Curve25519 [Ber06a] has found widespread use. The curve is the Montgomery curve with $A = 486662$ and $B = 1$ defined over $\mathbb{F}_p$ with $p = 2^{255} - 19$. This prime satisfies $p \equiv 1 \bmod 4$ and so for a Montgomery curve over $\mathbb{F}_p$ either the curve or its quadratic twist has order divisible by 8. The curve is chosen to have $(A-2)/4$ minimal among the curves satisfying that the group order is 8ℓ and that the order of the twist is $4\ell'$, where ℓ and ℓ' are prime numbers, and that the curve satisfies all standard security criteria. See [BL14] for more security details, and [Cho15] for recent Curve25519 performance results. The WhatsApp messaging system now uses Curve25519 to encrypt all messages from end to end, and the TLS protocol for secure web access has recently added Curve25519 as an option. The EdDSA signature scheme [BJL+15] uses twisted Edwards curves, and in particular the Ed25519 signature scheme [BDL+11] uses a twisted Edwards curve birationally equivalent to Curve25519.

Curve41417 [BCL14] and Curve448 [Ham15] are two newer examples of Montgomery curves (equivalently, twisted Edwards curves) designed for efficient cryptography over larger prime fields. Curve41417 has been deployed in the BlackPhone, and Curve448 is being considered as an alternative for TLS.

4.4 Doubling formulas without y

Let P be a point on a Weierstrass curve. Then $\mathbf{x}(nP)$, the x-coordinate of nP, is entirely determined by n and $\mathbf{x}(P)$, as mentioned above.

Even better, similar to the clock example in Section 4.2, $\mathbf{x}(nP)$ is a rational function of $\mathbf{x}(P)$ for each n. One can compute the numerator and denominator with ring operations, and then divide; there is no need for an initial square-root computation to recover y.

In the case of short Weierstrass curves it is easy to find literature stating explicit "division polynomial" recurrences for nP. Miller's original ECC paper [Mil86] repeated these recurrences, and also mentioned the possibility of

avoiding y-coordinates in ECC. However, Miller reported "$26 \log_2 n$ multiplications" to compute nP using these recurrences.

The Montgomery ladder is much simpler and almost three times faster. The structure of Montgomery curves is important for this simplicity and speed: from the modern Edwards perspective, Montgomery takes advantage of having a point of order 4 on the curve or its twist.

In the above description we have ignored exceptional cases: e.g., dividing a 0 numerator by a 0 denominator. We start by handling the generic case. We return to exceptional cases in Sections 4.4.4, 4.5.3, and 4.6.3.

This section begins with the simplest case $n = 2$: computing $\mathbf{x}(2P)$ from $\mathbf{x}(P)$. Sections 4.5 and 4.6 handle larger values of n.

4.4.1 Doubling: the Weierstrass view

Theorem 4.1 *Fix a field k not of characteristic 2. Fix $A, B \in k$ with $B(A^2 - 4) \neq 0$. Define M as the Montgomery curve $By^2 = x^3 + Ax^2 + x$. Define $\mathbf{x} : M(k) \to k \cup \{\infty\}$ as follows: $\mathbf{x}(x, y) = x$; $\mathbf{x}(\infty) = \infty$.*

Let P be an element of $M(k)$. If $\mathbf{x}(P) = \infty$ then $\mathbf{x}(2P) = \infty$. If $\mathbf{x}(P) \neq \infty$ and $\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P) = 0$ then $\mathbf{x}(2P) = \infty$. If $\mathbf{x}(P) \neq \infty$ and $\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P) \neq 0$ then

$$\mathbf{x}(2P) = \frac{(\mathbf{x}(P)^2 - 1)^2}{4(\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P))}.$$

Proof If $\mathbf{x}(P) = \infty$ then $P = \infty$ so $2P = \infty$ so $\mathbf{x}(2P) = \infty$ as claimed. Assume from now on that $\mathbf{x}(P) \neq \infty$. Then $P = (x, y)$ for some $x, y \in k$ satisfying $By^2 = x^3 + Ax^2 + x$. By definition $\mathbf{x}(P) = x$.

If $x^3 + Ax^2 + x = 0$ then $y = 0$ so $2P = (x, 0) + (x, 0) = (x, 0) - (x, 0) = \infty$ so $\mathbf{x}(2P) = \infty$ as claimed. Assume from now on that $x^3 + Ax^2 + x \neq 0$. Then $y \neq 0$.

By definition (see Section 4.3.2) $2P = (B\lambda^2 - A - 2x, \ldots)$ where $\lambda = (3x^2 +$

$2Ax + 1)/(2By)$. Consequently

$$\begin{aligned}
\mathbf{x}(2P) &= B\lambda^2 - A - 2x = B\frac{(3x^2 + 2Ax + 1)^2}{4B^2y^2} - A - 2x \\
&= \frac{(3x^2 + 2Ax + 1)^2}{4By^2} - A - 2x = \frac{(3x^2 + 2Ax + 1)^2}{4(x^3 + Ax^2 + x)} - A - 2x \\
&= \frac{(3x^2 + 2Ax + 1)^2 - 4(x^3 + Ax^2 + x)(2x + A)}{4(x^3 + Ax^2 + x)} \\
&= \frac{9x^4{+}12Ax^3{+}(4A^2{+}6)x^2{+}4Ax{+}1 - 4(2x^4{+}3Ax^3{+}(A^2{+}2)x^2{+}Ax)}{4(x^3 + Ax^2 + x)} \\
&= \frac{x^4 - 2x^2 + 1}{4(x^3 + Ax^2 + x)} = \frac{(x^2 - 1)^2}{4(x^3 + Ax^2 + x)}
\end{aligned}$$

as claimed. □

4.4.2 Optimized doublings

Divisions are slow. To avoid divisions, the Montgomery ladder represents x-coordinates as fractions. This also means that doublings inside the Montgomery ladder take their inputs $\mathbf{x}(P)$ as fractions. (Small exception: the first doubling in the ladder can be sped up in the normal case that its input is provided with denominator 1.)

This requires extra multiplications in Theorem 4.1: for example, computing $\mathbf{x}(P)^2$ requires squaring both the numerator and the denominator. These extra operations appear inside the simple formulas for X_{2n} and Z_{2n} shown in Section 4.1. A straightforward operation count would suggest that there are six multiplications here (not counting the final multiplication by 4, which can be done with two additions): X_n^2, Z_n^2, X_nZ_n, AX_nZ_n, X_{2n}, Z_{2n}.

Montgomery's optimized formulas, also shown in Section 4.1, save a multiplication as follows. Start with $(X_n + Z_n)^2$ and $(X_n - Z_n)^2$. Compute X_{2n} as the product of these squares, and compute $4X_nZ_n$ as the difference of these squares. Multiply by $(A - 2)/4$ to obtain $(A - 2)X_nZ_n$, add $(X_n + Z_n)^2$ to obtain $X_n^2 + AX_nZ_n + Z_n^2$, and multiply by $4X_nZ_n$ to obtain Z_{2n}. In total there are two squarings, one multiplication by $(A - 2)/4$, two more multiplications, two additions, and two subtractions.

Montgomery's formulas can be viewed as expressing doubling as the composition of two 2-isogenies. Montgomery reportedly found these formulas via experiments with 2-isogeny formulas. The same idea has been productively reused to build other curve shapes with efficient formulas for doubling and tripling; see, e.g., [DIK06] and [BCKL15].

If $d = (A-2)/(A+2)$ is a square, say r^2, then one can replace $2\mathbf{M}+2\mathbf{S}+1\mathbf{C}$ with $4\mathbf{S}+3\mathbf{C}$ as follows (and with $4\mathbf{S}+2\mathbf{C}$ if one changes coordinates from X_n and Z_n to $r(X_n-Z_n)$ and X_n+Z_n). Precompute $s = (1+r)/(1-r)$. Then compute $Y = r(X_n-Z_n)^2$, $Z = (X_n+Z_n)^2$, $V = s(Z-Y)^2$, $W = (Z+Y)^2$, $Y' = W-V$, and $Z' = r(W+V)$. Now $(Z'+Y', Z'-Y')$ is (X_{2n}, Z_{2n}) times an irrelevant factor $4(d+r)$. This is a speedup if r has small enough numerator and denominator. This speedup is due in essence to Gaudry; see [Gau06], [GL09], and [BL09].

4.4.3 A word of warning: projective coordinates

Here is a different way to view representing $\mathbf{x}(P)$ as a fraction. Use a tuple (X, Y, Z) to represent a point $P = (X : Y : Z)$ on M in projective coordinates. Discard the Y-coordinate, leaving only the pair (X, Z) to represent $X/Z = \mathbf{x}(P)$.

One might think that this view smoothly generalizes from affine points to all points on M, and that the case distinctions in Theorem 4.1 are merely artifacts of working in affine coordinates. However, discarding the Y-coordinate from the point $(0 : 1 : 0)$ produces $(0 : 0)$. The definition of $\mathbb{P}^1$ excludes $(0 : 0)$; the standard notion of fractions excludes $0/0$. More importantly, converting the generic case of Theorem 4.1 into projective formulas for $\mathbf{x}(2P)$, and then applying those formulas to the case $P = (0, 0)$ with $\mathbf{x}(P) = 0/1$, does *not* produce $0/0$ as output; it produces $1/0$.

4.4.4 Completeness of generic doubling formulas

The Montgomery ladder defines X_{2n} and Z_{2n} from X_n and Z_n using formulas that match the *generic* case of Theorem 4.1: it defines, e.g., $X_4 = (X_2^2 - Z_2^2)^2$ and $Z_4 = 4X_2Z_2(X_2^2 + AX_2Z_2 + Z_2^2)$. This raises the question of what exactly these formulas do in the *other* cases. As noted in Section 4.1, the exceptional cases did not matter for Montgomery's application to ECM, but they do matter for cryptography.

The following theorem says that $\mathbf{x}(2P) = X_4/Z_4$ if $\mathbf{x}(P) = X_2/Z_2$. This was proven in [Ber06a] under the assumption that $A^2 - 4$ is non-square.

Theorem 4.2 *Fix a field k not of characteristic 2. Fix $A, B \in k$ with $B(A^2 - 4) \neq 0$. Define M as the Montgomery curve $By^2 = x^3 + Ax^2 + x$. Define $\mathbf{x} : M(k) \to k \cup \{\infty\}$ as follows: $\mathbf{x}(x, y) = x$; $\mathbf{x}(\infty) = \infty$.*

Let X_2, Z_2 be elements of k. Define

$$X_4 = (X_2^2 - Z_2^2)^2,$$
$$Z_4 = 4X_2Z_2(X_2^2 + AX_2Z_2 + Z_2^2).$$

Let P be an element of $M(k)$. If $(X_2, Z_2) \neq (0,0)$ and $\mathbf{x}(P) = X_2/Z_2$ then $(X_4, Z_4) \neq (0,0)$ and $\mathbf{x}(2P) = X_4/Z_4$.

Here X/Z means the quotient of X and Z in k if $Z \neq 0$; it means ∞ if $X \neq 0$ and $Z = 0$; it is undefined if $X = Z = 0$.

Proof If $Z_2 = 0$ then $X_4 = X_2^4 \neq 0$ and $Z_4 = 0$ so $(X_4, Z_4) \neq (0,0)$ as claimed. Also $\mathbf{x}(P) = X_2/0 = \infty$ so, by Theorem 4.1, $\mathbf{x}(2P) = \infty = X_4/Z_4$ as claimed.

Assume from now on that $Z_2 \neq 0$; then $\mathbf{x}(P) \neq \infty$.

If $\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P) = 0$ then $X_2^3 + AX_2^2Z_2 + X_2Z_2^2 = 0$ so $Z_4 = 4Z_2(X_2^3 + AX_2^2Z_2 + X_2Z_2^2) = 0$. Suppose that $X_4 = 0$; then $X_2^2 = Z_2^2$ so $\mathbf{x}(P)^2 = 1$ so $\mathbf{x}(P) = \pm 1$ so $0 = \mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P) = A \pm 2$, contradicting the hypothesis that $A^2 - 4 \neq 0$. Hence $X_4 \neq 0$ so $(X_4, Z_4) \neq (0,0)$ as claimed. Also, by Theorem 4.1, $\mathbf{x}(2P) = \infty = X_4/Z_4$ as claimed.

Assume from now on that $\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P) \neq 0$.

Multiply by $Z_2^3 \neq 0$ to obtain $X_2^3 + AX_2^2Z_2 + X_2Z_2^2 \neq 0$. In particular $X_2 \neq 0$ so $Z_4 \neq 0$ so $(X_4, Z_4) \neq (0,0)$ as claimed. Furthermore $4(\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P)) = Z_4/Z_2^4$ and $(\mathbf{x}(P)^2 - 1)^2 = X_4/Z_2^4$. By Theorem 4.1, $\mathbf{x}(2P) = (\mathbf{x}(P)^2 - 1)^2/(4(\mathbf{x}(P)^3 + A\mathbf{x}(P)^2 + \mathbf{x}(P))) = X_4/Z_4$ as claimed. □

4.4.5 Doubling: the Edwards view

We now use the Edwards addition law to give a direct proof of Theorem 4.2, without the calculations from Theorem 4.1.

Alternate proof of Theorem 4.2 M is birationally equivalent to the twisted Edwards curve $au^2 + v^2 = 1 + du^2v^2$ with $a = (A+2)/B$ and $d = (A-2)/B$.

Define (u_2, v_2) and (u_4, v_4) as the points corresponding to P and $2P$ respectively. Recall that $v_4 = (2av_2^2 - a - dv_2^4)/(a - 2dv_2^2 + dv_2^4)$. We now develop matching formulas for the Montgomery x-coordinates $x_2 = \mathbf{x}(P)$ and $x_4 = \mathbf{x}(2P)$. First

$$x_4 = \frac{1+v_4}{1-v_4} = \frac{a - 2dv_2^2 + dv_2^4 + 2av_2^2 - a - dv_2^4}{a - 2dv_2^2 + dv_2^4 - (2av_2^2 - a - dv_2^4)}$$
$$= \frac{(a-d)v_2^2}{a - av_2^2 - dv_2^2 + dv_2^4} = \frac{(a-d)v_2^2}{(1-v_2^2)(a-dv_2^2)}.$$

Use $v_2 = (x_2 - 1)/(x_2 + 1)$ and the relation of the curve coefficients:

$$x_4 = \frac{(a-d)v_2^2}{(1-v_2^2)(a-dv_2^2)} = \frac{(a-d)(x_2-1)^2(x_2+1)^2}{((x_2+1)^2-(x_2-1)^2)(a(x_2+1)^2-d(x_2-1)^2)}$$
$$= \frac{(a-d)(x_2-1)^2(x_2+1)^2}{4x_2((a-d)(x_2^2+1)+2(a+d)x_2)} = \frac{(x_2^2-1)^2}{4x_2(x_2^2+Ax_2+1)}.$$

Replace x_2 by X_2/Z_2 and clear denominators. □

4.5 Differential-addition formulas

One cannot expect to be able to compute $\mathbf{x}(P_3+P_2)$ given only $\mathbf{x}(P_3)$ and $\mathbf{x}(P_2)$. Usually there are four possibilities for (P_3, P_2), four possibilities for $P_3 + P_2$, and two possibilities for $\mathbf{x}(P_3 + P_2)$.

Montgomery's differential-addition formulas instead compute $\mathbf{x}(P_3 + P_2)$ given $\mathbf{x}(P_3)$, $\mathbf{x}(P_2)$, and $\mathbf{x}(P_3 - P_2)$, as explained in this section. Similar to the clock addition formulas from Section 4.2, these formulas produce $\mathbf{x}(3P)$ given $\mathbf{x}(2P), \mathbf{x}(P), \mathbf{x}(P)$; they produce $\mathbf{x}(7P)$ given $\mathbf{x}(4P), \mathbf{x}(3P), \mathbf{x}(P)$; they produce $\mathbf{x}(13P)$ given $\mathbf{x}(7P), \mathbf{x}(6P), \mathbf{x}(P)$.

4.5.1 Differential addition: the Weierstrass view

We begin by deriving Montgomery's **differential-addition formulas** in essentially the same way that Montgomery did, starting from the definition of addition for Weierstrass curves.

Theorem 4.3 *Fix a field k not of characteristic 2. Fix $A, B \in k$ with $B(A^2 - 4) \neq 0$. Define M as the Montgomery curve $By^2 = x^3 + Ax^2 + x$. Define $\mathbf{x} : M(k) \to k \cup \{\infty\}$ as follows: $\mathbf{x}(x, y) = x$; $\mathbf{x}(\infty) = \infty$.*

Let P_2, P_3 be elements of $M(k)$ with $P_3 \neq \infty$, $P_2 \neq \infty$, $P_3 \neq P_2$, and $P_3 \neq -P_2$. Then $\mathbf{x}(P_3) \neq \mathbf{x}(P_2)$ and

$$\mathbf{x}(P_3 + P_2)\mathbf{x}(P_3 - P_2) = \frac{(\mathbf{x}(P_3)\mathbf{x}(P_2) - 1)^2}{(\mathbf{x}(P_3) - \mathbf{x}(P_2))^2}.$$

Proof $P_3 \neq \infty$ so $P_3 = (x, y)$ for some $x, y \in k$ satisfying $By^2 = x^3 + Ax^2 + x$; and $P_2 \neq \infty$ so $P_2 = (x', y')$ for some $x', y' \in k$ satisfying $B(y')^2 = (x')^3 + A(x')^2 + x'$.

Suppose that $x = x'$. Then $By^2 = B(y')^2$ so $y = \pm y'$. If $y = y'$ then $P_3 = P_2$, contradiction. If $y = -y'$ then $P_3 = -P_2$, contradiction.

Thus $x \neq x'$, and $P_3 + P_2 = (B\lambda^2 - A - x - x', \ldots)$ where $\lambda = (y' - y)/(x' - x)$. Consequently

$$\begin{aligned}
\mathbf{x}(P_3 + P_2) &= B\lambda^2 - A - x - x' = B\frac{(y'-y)^2}{(x'-x)^2} - A - x - x' \\
&= \frac{B(y')^2 + By^2 - 2Byy'}{(x'-x)^2} - A - x - x' \\
&= \frac{(x')^3 + A(x')^2 + x' + x^3 + Ax^2 + x - 2Byy' - (A + x' + x)(x'-x)^2}{(x'-x)^2} \\
&= \frac{(x')^3 + x' + x^3 + x + 2Axx' - 2Byy' - (x' + x)(x'-x)^2}{(x'-x)^2} \\
&= \frac{(x' + x)(1 + xx') + 2Axx' - 2Byy'}{(x'-x)^2}.
\end{aligned}$$

Similarly $\mathbf{x}(P_3 - P_2) = ((x' + x)(1 + xx') + 2Axx' + 2Byy')/(x' - x)^2$. Thus

$$\begin{aligned}
&\mathbf{x}(P_3 + P_2)\mathbf{x}(P_3 - P_2)(x' - x)^4 \\
&\quad = ((x' + x)(1 + xx') + 2Axx')^2 - (2Byy')^2 \\
&\quad = ((x' + x)(1 + xx') + 2Axx')^2 - 4By^2B(y')^2 \\
&\quad = ((x' + x)(1 + xx') + 2Axx')^2 - 4(x^3 + Ax^2 + x)((x')^3 + A(x')^2 + x') \\
&\quad = (x' + x)^2(1 + xx')^2 + 4Axx'(x' + x)(1 + xx') + 4A^2x^2(x')^2 \\
&\qquad - 4(x^3+x)((x')^3+x') - 4A((x^3+x)(x')^2+((x')^3+x')x^2) - 4A^2x^2(x')^2 \\
&\quad = (x' + x)^2(1 + xx')^2 + 4Axx'(x' + x + (x')^2x + x^2x') \\
&\qquad - 4(x^3 + x)((x')^3 + x') - 4Axx'(x^2x' + x' + (x')^2x + x) \\
&\quad = ((x')^2 + 2xx' + x^2)(1 + 2xx' + x^2(x')^2) \\
&\qquad - 4(x^3(x')^3 + x^3x' + (x')^3x + xx') \\
&\quad = (x')^2 + 2xx' + x^2 + 2x(x')^3 + 4x^2(x')^2 + 2x^3x' \\
&\qquad + x^2(x')^4 + 2x^3(x')^3 + x^4(x')^2 - 4(x^3(x')^3 + x^3x' + (x')^3x + xx') \\
&\quad = (x')^2 - 2xx' + x^2 - 2x(x')^3 + 4x^2(x')^2 - 2x^3x' \\
&\qquad + x^2(x')^4 - 2x^3(x')^3 + x^4(x')^2 \\
&\quad = ((x')^2 - 2xx' + x^2)(1 - 2xx' + x^2(x')^2) \\
&\quad = (x' - x)^2(xx' - 1)^2
\end{aligned}$$

so $\mathbf{x}(P_3 + P_2)\mathbf{x}(P_3 - P_2) = (xx' - 1)^2/(x' - x)^2$. □

4.5.2 Optimized differential addition

As discussed in Section 4.4.2, the Montgomery ladder represents x-coordinates as fractions. This eliminates the division in Theorem 4.3 but uses extra multiplications. The simple formulas for X_{2n+1} and Z_{2n+1} in Section 4.1 use 4**M** for $X_nX_{n+1} - Z_nZ_{n+1}$ and $X_nZ_{n+1} - Z_nX_{n+1}$, 2**S**, and 2**M** by the numerator X_1 and denominator Z_1 of $\mathbf{x}(P_3 - P_2)$.

Montgomery's optimized formulas, also shown in Section 4.1, replace the first four multiplications with just two: they rewrite $2(X_nX_{n+1} - Z_nZ_{n+1})$ and $2(X_nZ_{n+1} - Z_nX_{n+1})$ as the sum and difference of $(X_n - Z_n)(X_{n+1} + Z_{n+1})$ and $(X_n + Z_n)(X_{n+1} - Z_{n+1})$. In total there are 2**S**, 1**M** by X_1, 1**M** by Z_1, 2**M** more, three additions, and three subtractions.

The same trick works for any expressions of the form $\alpha\beta - \gamma\delta$ and $\alpha\delta - \beta\gamma$: except for a rescaling by 2, these are the sum and difference of $(\alpha - \gamma)(\beta + \delta)$ and $(\alpha + \gamma)(\beta - \delta)$. In other words, to multiply the polynomials $\alpha + \gamma t$ and $\beta - \delta t$ modulo $t^2 - 1$, first multiply modulo $t - 1$ and $t + 1$, and then interpolate.

4.5.3 Quasi-completeness

The Montgomery ladder defines X_{2n+1} and Z_{2n+1} from $X_{n+1}, Z_{n+1}, X_n, Z_n, X_1, Z_1$ using formulas that match Theorem 4.3: it defines, e.g., $X_5 = 4(X_2X_3 - Z_2Z_3)^2Z_1$ and $Z_5 = 4(X_2Z_3 - Z_2X_3)^2X_1$. But there are various hypotheses in Theorem 4.3, raising the question of what happens when these hypotheses are violated, as in Section 4.4.4.

Theorem 4.4 *almost* says that $\mathbf{x}(P_3 + P_2) = X_5/Z_5$ in all cases. However, it excludes two values of $\mathbf{x}(P_3 - P_2)$, namely 0 and ∞. In other words, it excludes two values of $P_3 - P_2$, namely $(0, 0)$ and ∞. Later we will analyze what the Montgomery ladder does for these inputs.

Theorem 4.4 *Fix a field k not of characteristic 2. Fix $A, B \in k$ with $B(A^2 - 4) \neq 0$. Define M as the Montgomery curve $By^2 = x^3 + Ax^2 + x$. Define $\mathbf{x} : M(k) \to k \cup \{\infty\}$ as follows: $\mathbf{x}(x, y) = x$; $\mathbf{x}(\infty) = \infty$.*

Let $X_1, Z_1, X_2, Z_2, X_3, Z_3$ be elements of k. Define

$$X_5 = 4(X_2X_3 - Z_2Z_3)^2Z_1,$$
$$Z_5 = 4(X_2Z_3 - Z_2X_3)^2X_1.$$

Let P_2, P_3 be elements of $M(k)$. Assume that $X_1 \neq 0$; $Z_1 \neq 0$; $\mathbf{x}(P_3 - P_2) = X_1/Z_1$; $(X_2, Z_2) \neq (0, 0)$; $\mathbf{x}(P_2) = X_2/Z_2$; $(X_3, Z_3) \neq (0, 0)$; and $\mathbf{x}(P_3) = X_3/Z_3$. Then $(X_5, Z_5) \neq (0, 0)$ and $\mathbf{x}(P_3 + P_2) = X_5/Z_5$.

We emphasize that both X_1 and Z_1 are required to be nonzero individually.

Proof If $P_3 = P_2$ then $X_1/Z_1 = \mathbf{x}(P_3 - P_2) = \mathbf{x}(\infty) = \infty$ so $Z_1 = 0$, contradiction. Hence $P_3 \neq P_2$.

If $P_2 = \infty$ then $X_2/Z_2 = \mathbf{x}(P_2) = \infty$ so $Z_2 = 0$. Furthermore $X_1/Z_1 = \mathbf{x}(P_3 - P_2) = \mathbf{x}(P_3) = X_3/Z_3$ so $X_3Z_1 = X_1Z_3$. Hence $X_5 = 4(X_2X_3)^2Z_1 = 4X_2^2X_1X_3Z_3$ and $Z_5 = 4(X_2Z_3)^2X_1 = 4X_2^2X_1Z_3^2$.

By hypothesis $X_1 \neq 0$; also $X_2 \neq 0$ since $Z_2 = 0$; and $Z_3 \neq 0$ since $X_3/Z_3 = \mathbf{x}(P_3) \neq \mathbf{x}(P_2) = \infty$. Hence $Z_5 \neq 0$ and $X_5/Z_5 = X_3/Z_3 = \mathbf{x}(P_3) = \mathbf{x}(P_3 + P_2)$ as claimed.

Similarly, if $P_3 = \infty$ then $X_3/Z_3 = \mathbf{x}(P_3) = \infty$ so $Z_3 = 0$. Furthermore $X_1/Z_1 = \mathbf{x}(P_3 - P_2) = \mathbf{x}(-P_2) = \mathbf{x}(P_2) = X_2/Z_2$ so $X_2Z_1 = X_1Z_2$. Hence $X_5 = 4(X_2X_3)^2Z_1 = 4X_3^2X_1Z_2X_2$ and $Z_5 = 4(Z_2X_3)^2X_1 = 4X_3^2X_1Z_2^2$.

Again $X_1 \neq 0$; $X_3 \neq 0$ since $Z_3 = 0$; and $Z_2 \neq 0$ since $X_2/Z_2 = \mathbf{x}(P_2) \neq \infty$. Hence $Z_5 \neq 0$ and $X_5/Z_5 = X_2/Z_2 = \mathbf{x}(P_2) = \mathbf{x}(P_3 + P_2)$ as claimed.

Assume from now on that $P_2 \neq \infty$ and $P_3 \neq \infty$. Note that $Z_2 \neq 0$ and $Z_3 \neq 0$.

If $P_3 = -P_2$ then $X_2/Z_2 = \mathbf{x}(P_2) = \mathbf{x}(-P_3) = \mathbf{x}(P_3) = X_3/Z_3$ so $X_2Z_3 = Z_2X_3$ so $Z_5 = 0$. We will show in a moment that $X_5 \neq 0$, so $X_5/Z_5 = \infty = \mathbf{x}(\infty) = \mathbf{x}(P_3 + P_2)$ as claimed.

Note that $X_2 \neq 0$: if $X_2 = 0$ then $Z_2 \neq 0$ so $\mathbf{x}(P_2) = X_2/Z_2 = 0$ so $P_2 = (0, 0)$ so $P_3 = -P_2 = -(0, 0) = (0, 0) = P_2$, contradiction. Similarly $X_3 \neq 0$.

Now suppose that $X_5 = 0$. Then $4(X_2X_3 - Z_2Z_3)^2Z_1 = 0$, but $Z_1 \neq 0$, so $X_2X_3 = Z_2Z_3$. Consequently $(X_2 + Z_2)(X_3 - Z_3) = (X_2X_3 - Z_2Z_3) - (X_2Z_3 - Z_2X_3) = 0$. If $X_2 + Z_2 \neq 0$ then $X_3 - Z_3 = 0$ so $\mathbf{x}(P_2) = \mathbf{x}(-P_3) = \mathbf{x}(P_3) = X_3/X_3 = 1$. Otherwise $X_2 = -Z_2$ so $\mathbf{x}(P_2) = -1$. Either way $\mathbf{x}(P_2)^2 = 1$ so $\mathbf{x}(2P_2) = 0$ by Theorem 4.2. Hence $X_1/Z_1 = \mathbf{x}(P_3 - P_2) = \mathbf{x}(-2P_2) = \mathbf{x}(2P_2) = 0$ so $X_1 = 0$, contradiction.

Assume from now on that $P_3 + P_2 \neq \infty$. All hypotheses of Theorem 4.3 are now satisfied, so $\mathbf{x}(P_3) \neq \mathbf{x}(P_2)$ and $\mathbf{x}(P_3 + P_2)\mathbf{x}(P_3 - P_2)(\mathbf{x}(P_3) - \mathbf{x}(P_2))^2 = (\mathbf{x}(P_3)\mathbf{x}(P_2) - 1)^2$. Multiply through by appropriate powers of Z_1, Z_2, Z_3 to see that $X_3Z_2 \neq X_2Z_3$ and $\mathbf{x}(P_3 + P_2)X_1(X_3Z_2 - X_2Z_3)^2 = Z_1(X_2X_3 - Z_2Z_3)^2$; i.e., $Z_5 \neq 0$ and $\mathbf{x}(P_3 + P_2) = X_5/Z_5$ as claimed. □

4.5.4 Differential addition: the Edwards view

Alternate proof of Theorem 4.4 Write x_1, x_2, x_3, x_5 for, respectively, $\mathbf{x}(P_3 - P_2), \mathbf{x}(P_2), \mathbf{x}(P_3), \mathbf{x}(P_3 + P_2)$.

As before, M is birationally equivalent to the twisted Edwards curve $au^2 + v^2 = 1 + du^2v^2$ with $a = (A+2)/B$ and $d = (A-2)/B$. Let (u_2, v_2) and (u_3, v_3) be the points on the twisted Edwards curve equivalent to P_2 and P_3 respectively.

Write $(u_5, v_5) = (u_3, v_3) + (u_2, v_2)$ and $(u_1, v_1) = (u_3, v_3) - (u_2, v_2)$. The dual

addition law (4.2) says

$$v_5 = \frac{u_3v_3 - u_2v_2}{u_3v_2 - u_2v_3} \qquad \text{and} \qquad v_1 = \frac{u_3v_3 + u_2v_2}{u_3v_2 + u_2v_3}$$

except when $(u_3, v_3) \in \{(u_2, v_2), (-u_2, -v_2), (-u_2, v_2), (u_2, -v_2)\}$, i.e., except when $u_3^2 = u_2^2$. Assume for the moment that $u_3^2 \neq u_2^2$; the exceptional cases will be treated later.

Recall that the maps for x and v between M and E are defined on all of $k \cup \{\infty\}$. Now

$$\begin{aligned}
x_1x_5 &= \frac{1+v_1}{1-v_1} \cdot \frac{1+v_5}{1-v_5} \\
&= \frac{(u_3v_2 + u_2v_3) + (u_3v_3 + u_2v_2)}{(u_3v_2 + u_2v_3) - (u_3v_3 + u_2v_2)} \cdot \frac{(u_3v_2 - u_2v_3) + (u_3v_3 - u_2v_2)}{(u_3v_2 - u_2v_3) - (u_3v_3 - u_2v_2)} \\
&= \frac{(u_3v_2 + u_3v_3)^2 - (u_2v_3 + u_2v_2)^2}{(u_3v_2 - u_3v_3)^2 - (u_2v_3 - u_2v_2)^2} = \frac{(u_3^2 - u_2^2)(v_3 + v_2)^2}{(u_3^2 - u_2^2)(v_3 - v_2)^2} = \frac{(v_3 + v_2)^2}{(v_3 - v_2)^2} \\
&= \frac{((x_3 - 1)(x_2 + 1) + (x_2 - 1)(x_3 + 1))^2}{((x_3 - 1)(x_2 + 1) - (x_2 - 1)(x_3 + 1))^2} = \frac{(x_2x_3 - 1)^2}{(x_2 - x_3)^2}.
\end{aligned}$$

Substitute $1/x_1 = Z_1/X_1$, $x_2 = X_2/Z_2$, and $x_3 = X_3/Z_3$ to see that $x_5 = X_5/Z_5$ as claimed.

If $(u_3, v_3) = (u_2, v_2)$ then $(u_1, v_1) = (0, 1)$, corresponding to ∞ on M. If $(u_3, v_3) = (-u_2, -v_2)$ then $(u_1, v_1) = (0, -1)$, corresponding to $(0, 0)$ on M. Both of these points on M are excluded by the hypothesis that $x_1 \notin \{0, \infty\}$.

If $(u_3, v_3) = (-u_2, v_2)$ then $(u_5, v_5) = (0, 1)$ so $x_5 = \infty$. Also $(u_1, v_1) = 2(u_3, v_3)$ so $x_1 = (1 + v_1)/(1 - v_1) = (a - d)v_3^2/((1 - v_3^2)(a - dv_3^2))$ as in the alternate proof of Theorem 4.2, and $x_1 \neq 0$ by hypothesis, so $v_3 \notin \{0, \infty\}$, i.e., $x_3 \notin \{-1, 1\}$. To summarize, $x_2 = x_3$ (since $v_2 = v_3$) while $x_2x_3 \neq 1$. Multiply by Z_2Z_3 to see that $X_2Z_3 = X_3Z_2$ while $X_2X_3 \neq Z_2Z_3$, i.e., $Z_5 = 0$ while $X_5 \neq 0$, so $X_5/Z_5 = \infty = x_5$ as claimed.

If $(u_3, v_3) = (u_2, -v_2)$ then $(u_5, v_5) = (0, -1)$ so $x_5 = 0$. Also $(u_1, v_1) = 2(u_3, v_3) + (0, -1)$ so $1/x_1 = (1 - v_1)/(1 + v_1) = (a - d)v_3^2/((1 - v_3^2)(a - dv_3^2))$. Now $x_1 \neq \infty$, so $v_3 \notin \{0, \infty\}$, so $x_3 \notin \{-1, 1\}$. To summarize, $x_2x_3 = 1$ while $x_2 \neq x_3$. Hence $X_5 = 0$ while $Z_5 \neq 0$, so $X_5/Z_5 = 0 = x_5$ as claimed. □

4.6 The Montgomery ladder

This section combines Montgomery's doubling formulas with Montgomery's differential-addition formulas to obtain one step of the **Montgomery ladder**; and then iterates these steps to obtain the full Montgomery ladder.

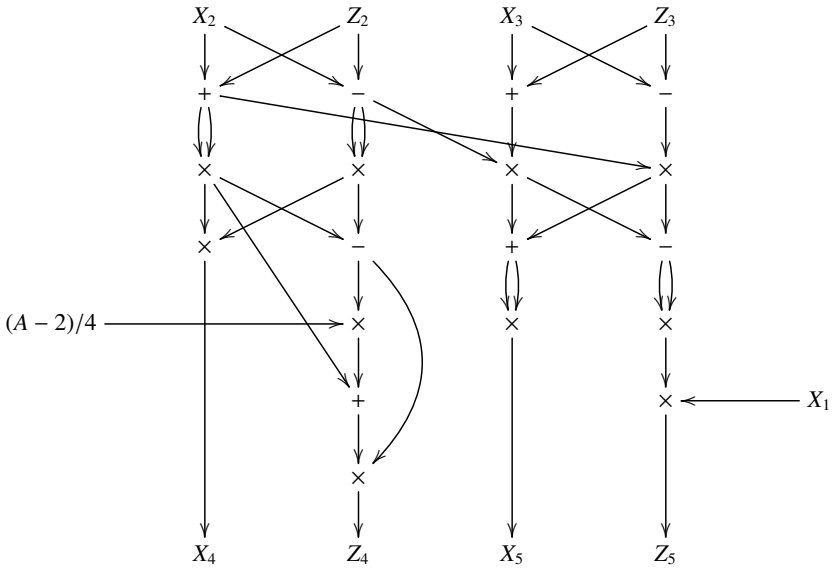

Figure 4.2 Montgomery's optimized formulas for doubling and differential addition, assuming $Z_1 = 1$.

4.6.1 The Montgomery ladder step

Theorems 4.2 and 4.4 together compute $\mathbf{x}(2P_2) = X_4/Z_4$ and $\mathbf{x}(P_3 + P_2) = X_5/Z_5$, given as input $\mathbf{x}(P_2) = X_2/Z_2$, $\mathbf{x}(P_3) = X_3/Z_3$, and $\mathbf{x}(P_3 - P_2) = X_1/Z_1$, assuming $X_1 \neq 0$ and $Z_1 \neq 0$. The merged optimized formulas are shown in Figure 4.2, under the simplifying assumption $Z_1 = 1$. In total there are $5\mathbf{M}$, $4\mathbf{S}$, $1\mathbf{C}$ by $(A-2)/4$, four additions, and four subtractions. This is cheaper than the total costs from Sections 4.4 and 4.5, for two reasons: first, a multiplication by Z_1 has been eliminated; second, $X_2 + Z_2$ and $X_2 - Z_2$ are reused between the differential addition and the doubling.

The ladder in Section 4.2.1 for computing X_n starting from X_0 and X_1 used one doubling and one differential addition per bit of n. We now combine the doubling and the differential addition into a single step to emphasize the benefits from combining.

For fixed (X_1, Z_1) define $\text{step}_0(X_2, Z_2, X_3, Z_3) = (X_4, Z_4, X_5, Z_5)$ where

$$X_4 = (X_2^2 - Z_2^2)^2, \qquad X_5 = 4(X_2X_3 - Z_2Z_3)^2Z_1,$$
$$Z_4 = 4X_2Z_2(X_2^2 + AX_2Z_2 + Z_2^2), \qquad Z_5 = 4(X_2Z_3 - Z_2X_3)^2X_1,$$

and also define $\text{step}_1(X_3, Z_3, X_2, Z_2) = (X_5, Z_5, X_4, Z_4)$. The recursive definition

of the Montgomery ladder in Section 4.1 can be abbreviated as $L_{2n} = \text{step}_0 L_n$ where $L_n = (X_n, Z_n, X_{n+1}, Z_{n+1})$, and implies $L_{2n+1} = \text{step}_1 L_n$, so in general $L_n = \text{step}_{n \bmod 2}(L_{\lfloor n/2 \rfloor})$ for $n \geq 2$.

4.6.2 Constant-time ladders

A typical problem in ECC is to compute a scalar multiple nP, where n is a secret element of $\left\{0, 1, \ldots, 2^{256} - 1\right\}$. Montgomery's original ladder is unsatisfactory in this context: anyone who observes the time taken by the ladder can deduce the position of the top bit set in n, since this position dictates the number of steps of the ladder. One fix is to always arrange for n to have a fixed top bit, for example by adding an appropriate multiple of the order of P. Another fix, which we use in Theorem 4.5 below, is to switch to a more general ladder in which the number of steps can be chosen separately from the position of the top bit set in n. It is important here that one can start the Montgomery ladder from $0P, 1P$, rather than from $1P, 2P$, and that applying a ladder step to $0P, 1P$ produces another valid representation of $0P, 1P$.

In this context it is also important for each ladder step to involve a constant sequence of operations, without splitting into cases that depend on the secret bits inside n. Notice that step_b can be computed as $\text{cswap}_b \circ \text{step}_0 \circ \text{cswap}_b$, where

$$\text{cswap}_0(X_2, Z_2, X_3, Z_3) = (X_2, Z_2, X_3, Z_3),$$
$$\text{cswap}_1(X_2, Z_2, X_3, Z_3) = (X_3, Z_3, X_2, Z_2).$$

One should compute $\text{cswap}_b(X_2, Z_2, X_3, Z_3)$ as $(b(X_3 - X_2) + X_2, b(Z_3 - Z_2) + Z_2, (1 - b)(X_3 - X_2) + X_2, (1 - b)(Z_3 - Z_2) + Z_2)$, or some equivalent constant-time arithmetic expression, rather than computing the two cases separately.

A composition of two steps produces a cswap-step-cswap-cswap-step-cswap pattern. One can merge the adjacent swaps, defining b as the xor of the two bits. A many-step ladder then follows the pattern cswap-step-cswap-step-cswap-step-cswap etc.

4.6.3 Completeness of the ladder

The end of the Montgomery ladder divides X_n by Z_n to obtain $\mathbf{x}(nP)$. Typically one computes X_n/Z_n as $X_n Z_n^{\#k-2}$ when k is finite; but for $Z_n = 0$ this computation outputs 0 rather than ∞.

A further difficulty arises when the Montgomery ladder is allowed to receive 0 or ∞ as its input $\mathbf{x}(P) = X_1/Z_1$; we excluded this case in Theorem 4.4. The ladder then produces $X_n Z_n = 0$ for each n, but it is not always true that

$\mathbf{x}(nP) = X_n/Z_n$: it is possible to have $X_n/Z_n = \infty$ while $\mathbf{x}(nP) = 0$, and it is even possible to have $(X_n, Z_n) = (0, 0)$.

Define $\mathbf{x}_0 : M(k) \to k$ as follows: $\mathbf{x}_0(x, y) = x$; $\mathbf{x}_0(\infty) = 0$. Using $\mathbf{x}_0(nP)$ in place of $\mathbf{x}(nP)$ as a ladder output merges ∞ with 0, eliminating all of the above case distinctions. It is then harmless to also use $\mathbf{x}_0(P)$ in place of $\mathbf{x}(P)$ as a ladder input, since inputs 0 and ∞ always produce the same outputs. The idea of using $\mathbf{x}_0$ in this context was introduced in [Ber06a].

Theorem 4.5 *Fix a field k not of characteristic 2. Fix $A, B \in k$ with $B(A^2 - 4) \neq 0$. Define M as the Montgomery curve $By^2 = x^3 + Ax^2 + x$. Define $\mathbf{x}_0 : M(k) \to k$ as follows: $\mathbf{x}_0(x, y) = x$; $\mathbf{x}_0(\infty) = 0$.*

Let P be an element of $M(k)$. Let X_1, Z_1 be elements of k such that $Z_1 \neq 0$ and $\mathbf{x}_0(P) = X_1/Z_1$. Let c be a nonnegative integer. Let $n_0, \ldots, n_{c-1}$ be elements of $\{0, 1\}$. Define $n = 2^{c-1}n_{c-1} + 2^{c-2}n_{c-2} + \cdots + 2^0 n_0$. Define

$$(X, Z, X', Z') = \operatorname{step}_{n_0} \operatorname{step}_{n_1} \cdots \operatorname{step}_{n_{c-1}} (1, 0, X_1, Z_1).$$

If $Z = 0$ then $\mathbf{x}_0(nP) = 0$; otherwise $\mathbf{x}_0(nP) = X/Z$.

Proof The main case is that $X_1 \neq 0$. Then $\mathbf{x}(P) = \mathbf{x}_0(P) = X_1/Z_1$. We will prove the following statement by induction on c: $(X, Z) \neq (0, 0)$; $X/Z = \mathbf{x}(nP)$; $(X', Z') \neq (0, 0)$; and $X'/Z' = \mathbf{x}((n + 1)P)$. This implies the claim: if $Z = 0$ then $\mathbf{x}(nP) = X/0 = \infty$ so $\mathbf{x}_0(nP) = 0$ as claimed; otherwise $\mathbf{x}(nP) \neq \infty$ so $\mathbf{x}_0(nP) = \mathbf{x}(nP) = X/Z$ as claimed.

If $c = 0$ then $(X, Z, X', Z') = (1, 0, X_1, Z_1)$. Evidently $(X, Z) = (1, 0) \neq (0, 0)$; $X/Z = \infty = \mathbf{x}(\infty) = \mathbf{x}(nP)$ since $n = 0$; $(X', Z') = (X_1, Z_1) \neq (0, 0)$; and $X'/Z' = X_1/Z_1 = \mathbf{x}(P) = \mathbf{x}((n + 1)P)$.

For $c \geq 1$: Write $(X_2, Z_2, X_3, Z_3) = \operatorname{step}_{n_1} \cdots \operatorname{step}_{n_{c-1}} (1, 0, X_1, Z_1)$. By the inductive hypothesis, $(X_2, Z_2) \neq (0, 0)$; $X_2/Z_2 = \mathbf{x}(mP)$ where $m = 2^{c-2}n_{c-1} + 2^{c-3}n_{c-2} + \cdots + 2^0 n_1$; $(X_3, Z_3) \neq (0, 0)$; and $X_3/Z_3 = \mathbf{x}((m + 1)P)$.

Now $(X, Z, X', Z') = \operatorname{step}_{n_0}(X_2, Z_2, X_3, Z_3)$. If $n_0 = 0$ then $(X, Z) \neq (0, 0)$ and $\mathbf{x}(nP) = \mathbf{x}(2mP) = X/Z$ by Theorem 4.2; also $(X', Z') \neq (0, 0)$ and $\mathbf{x}((n+1)P) = \mathbf{x}((2m + 1)P) = X'/Z'$ by Theorem 4.4. Similar comments apply if $n_0 = 1$. Either way the claimed statement holds, finishing the main case.

The remaining case is that $X_1 = 0$. Then $\mathbf{x}_0(P) = 0$ so $P = \infty$ or $P = (0, 0)$; in both cases $nP \in \{\infty, (0, 0)\}$ so $\mathbf{x}_0(nP) = 0$. It thus suffices to show that $Z = 0$ or $X = 0$.

The initial step input $(1, 0, X_1, Z_1)$ has the form $(*, 0, 0, *)$ since $X_1 = 0$. Note that $\operatorname{step}_0(*, 0, 0, *) = (*, 0, 0, *)$; $\operatorname{step}_1(*, 0, 0, *) = (0, *, *, 0)$; $\operatorname{step}_0(0, *, *, 0) = (*, 0, 0, *)$; and $\operatorname{step}_1(0, *, *, 0) = (0, *, *, 0)$. By induction (X, Z, X', Z') has the form $(*, 0, 0, *)$ or $(0, *, *, 0)$; so $Z = 0$ or $X = 0$ as claimed. □

4.7 A two-dimensional ladder

The way that the Montgomery ladder computes $\mathbf{x}_0(nP)$ for a particular target $n \geq 1$ is by computing $\mathbf{x}_0(n'P)$ for a sequence of scalars n', namely all integers of the form $\lfloor n/2^i \rfloor$ and $\lfloor n/2^i \rfloor + 1$. Each integer larger than 1 in the sequence is a sum of two smaller integers whose difference is 0 or 1; each use of difference 0 involves Montgomery's doubling formulas, and each use of difference 1 involves Montgomery's differential-addition formulas with difference P.

This section explains an analogous method to compute $\mathbf{x}_0(mP + nQ)$ for nonnegative integers m, n, starting from $\mathbf{x}_0(P), \mathbf{x}_0(Q), \mathbf{x}_0(P - Q)$. The method computes $\mathbf{x}_0(m'P+n'Q)$ for each (m', n') in a "two-dimensional ladder" defined below. This ladder has the following features:

- It starts from $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, -1)$.
- It has $3c$ additions if m and n fit into c bits.
- For each addition $v + w$, the difference $v - w$ is either $(0, 0)$ or $(1, 0)$ or $(0, 1)$ or $(1, -1)$ or $(1, 1)$. Consequently the only possible failure cases are $P, Q, P - Q, P + Q$ colliding with $(0, 0), \infty$.
- c of the additions are doublings, i.e., have difference $(0, 0)$. The doublings appear in a uniform pattern: add, double, add; add, double, add; etc.

Each doubling costs $4\mathbf{M}$ with Montgomery's formulas. Here, for simplicity, we are taking $\mathbf{C} = 0$, which is reasonable if $(A - 2)/4$ is small, and also taking $\mathbf{S} = \mathbf{M}$. Each differential addition costs $5\mathbf{M}$ with Montgomery's formulas if $\mathbf{x}_0(P), \mathbf{x}_0(Q), \mathbf{x}_0(P - Q), \mathbf{x}_0(P + Q)$ are each provided with denominator 1. The total cost of the chain here is $14c\mathbf{M}$.

For comparison, the Montgomery ladder costs $9c\mathbf{M}$ for a single scalar. Handling two scalars thus increases costs by a factor significantly below 2. In Section 4.8 we will see even faster double-scalar methods, although those methods no longer have a uniform pattern of additions and doublings.

It is easy to write down a similar chain using $19c\mathbf{M}$, handling each bit with a uniform double-add-add-add pattern. A 2000 algorithm by Schoenmakers, published in 2003 [Sta03, Section 3.2.3], costs on average $17.25c\mathbf{M}$, with a variable pattern of additions; an algorithm by Akishita [Aki01] costs on average $14.25c\mathbf{M}$, again with a variable pattern of additions. The chain described here is slightly faster than Akishita's chain and has the advantage of a uniform add-double-add structure, analogous to the uniform double-add structure of the Montgomery ladder. This chain was introduced by Bernstein in 2006; see [Ber06b].

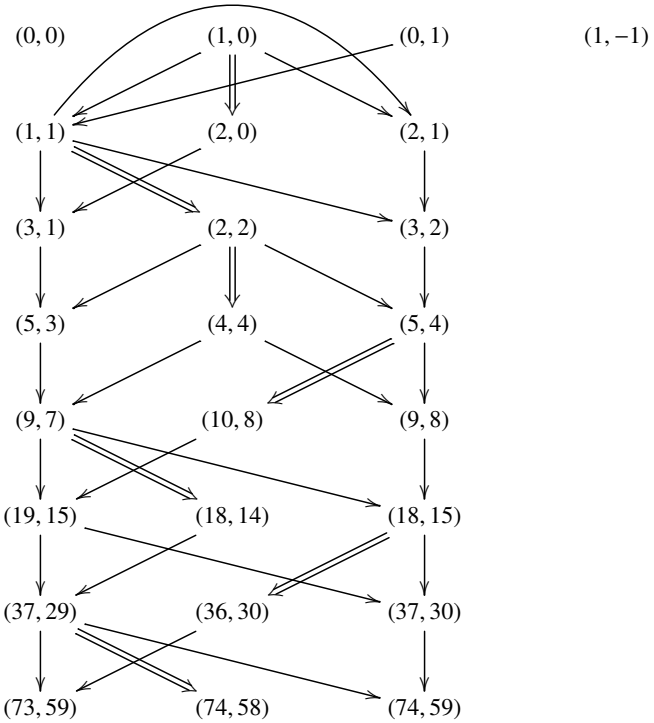

Figure 4.3 A uniform add-double-add two-dimensional ladder.

4.7.1 Introduction to the two-dimensional ladder

Figure 4.3 is an example of the differential addition chain used here. Each line after the first has three of the four pairs (a, b), $(a+1, b)$, $(a, b+1)$, $(a+1, b+1)$ for a unique (a, b). The missing element of $(a + \{0, 1\}, b + \{0, 1\})$ is always chosen as either (even, odd) or (odd, even), where the choice is related to the (A, B) for the next line:

- If $(a + A, b + B)$ is (even, odd) then the choice is (odd, even).
- If $(a + A, b + B)$ is (odd, even) then the choice is (even, odd).
- If $(a + A, b + B)$ is (even, even) then the current and next lines have the same choices.
- If $(a + A, b + B)$ is (odd, odd) then the current and next lines have opposite choices.

The pair (a, b) is also related to (A, B): it is simply $(\lfloor A/2 \rfloor, \lfloor B/2 \rfloor)$.

For comparison: The obvious way to build a two-dimensional ladder uses all four pairs (a, b), $(a + 1, b)$, $(a, b + 1)$, $(a + 1, b + 1)$. The Schoenmakers

chain [Sta03, Section 3.2.3] omits $(a+1, b+1)$. Akishita's chain [Aki01] omits $(a + 1 - (A \bmod 2), b + 1 - (B \bmod 2))$. The ladder presented here omits $(a + (a + d + 1 \bmod 2), b + (b + d \bmod 2))$ where d has a relatively complicated definition.

4.7.2 Recursive definition of the two-dimensional ladder

Define $C_D(A, B)$ recursively, for all nonnegative integers A and B and for all $D \in \{0, 1\}$, as $C_d(a, b)$ followed by the three pairs

$$\begin{aligned}
&(A + (A + 1 \mod 2), B + (B + 1 \mod 2)),\\
&(A + (A \mod 2), B + (B \mod 2)),\\
&(A + (A + D \bmod 2), B + (B + D + 1 \bmod 2)),
\end{aligned}$$

where $a = \lfloor A/2 \rfloor$, $b = \lfloor B/2 \rfloor$, and

$$d = \begin{cases} 0 & \text{if } (a + A, b + B) \bmod 2 = (0, 1) \\ 1 & \text{if } (a + A, b + B) \bmod 2 = (1, 0) \\ D & \text{if } (a + A, b + B) \bmod 2 = (0, 0) \\ 1 - D & \text{if } (a + A, b + B) \bmod 2 = (1, 1). \end{cases}$$

Exception: $C_D(0, 0)$ is defined as $(0, 0), (1, 0), (0, 1), (1, -1)$.

Here are the first few examples of this chain $C_D(A, B)$:

$C_0(0, 0)$ is $(0, 0), (1, 0), (0, 1), (1, -1)$.
$C_1(0, 0)$ is $(0, 0), (1, 0), (0, 1), (1, -1)$.
$C_0(1, 0)$ is $(0, 0), (1, 0), (0, 1), (1, -1), (1, 1), (2, 0), (2, 1)$.
$C_1(1, 0)$ is $(0, 0), (1, 0), (0, 1), (1, -1), (1, 1), (2, 0), (1, 0)$.
$C_0(0, 1)$ is $(0, 0), (1, 0), (0, 1), (1, -1), (1, 1), (0, 2), (0, 1)$.
$C_1(0, 1)$ is $(0, 0), (1, 0), (0, 1), (1, -1), (1, 1), (0, 2), (1, 2)$.
$C_0(1, 1)$ is $(0, 0), (1, 0), (0, 1), (1, -1), (1, 1), (2, 2), (2, 1)$.
$C_1(1, 1)$ is $(0, 0), (1, 0), (0, 1), (1, -1), (1, 1), (2, 2), (1, 2)$.

Note for future reference that $C_D(A, B)$ always contains the pair $(1, 1)$ if $(A, B) \neq (0, 0)$.

The rest of this section shows that $C_D(A, B)$ is a differential addition chain starting with $(0, 0), (1, 0), (0, 1), (1, -1)$ and following a uniform add-double-add pattern with all differences in $\{(0, 0), (1, 0), (0, 1), (1, 1), (1, -1)\}$. One can easily force the chain to contain any desired pair (m, n) of nonnegative integers by choosing, e.g., $(A, B) = (m, n)$ and $D = m \bmod 2$.

4.7.3 The odd-odd pair in each line: first addition

Assume that $(A, B) \neq (0, 0)$. The pair $(A + (A + 1 \bmod 2), B + (B + 1 \bmod 2))$ in $C_D(A, B)$ is equal to $(2a + 1, 2b + 1)$ where $(a, b) = (\lfloor A/2 \rfloor, \lfloor B/2 \rfloor)$ as above.

If $(a, b) = (0, 0)$ then the pair is $(1, 1)$, which can be obtained by adding $(1, 0)$ to $(0, 1)$ with difference $(1, -1)$; so assume that $(a, b) \neq (0, 0)$.

The chain already includes $C_d(a, b)$, which contains three of the four pairs $(a, b), (a + 1, b), (a, b + 1), (a + 1, b + 1)$. Consequently, $(2a + 1, 2b + 1)$ can be obtained by adding $(a + 1, b)$ to $(a, b + 1)$ with difference $(1, -1)$, or by adding $(a + 1, b + 1)$ to (a, b) with difference $(1, 1)$; recall that $(1, 1)$ is also in $C_d(a, b)$.

4.7.4 The even-even pair in each line: doubling

The next pair $(A + (A \bmod 2), B + (B \bmod 2))$ in the chain $C_D(A, B)$ is equal to $(2a + 2(A \bmod 2), 2b + 2(B \bmod 2))$.

If $(a, b) = (0, 0)$ then the pair is $(2, 0)$ or $(0, 2)$ or $(2, 2)$, so it can be obtained by doubling $(1, 0)$ or $(0, 1)$ or $(1, 1)$, all of which appear earlier in the chain; so assume that $(a, b) \neq (0, 0)$.

The chain already contains, via $C_d(a, b)$, all pairs $(a + \{0, 1\}, b + \{0, 1\})$ except $(a + (a + d + 1 \bmod 2), b + (b + d \bmod 2))$. If $(a + (A \bmod 2), b + (B \bmod 2))$ equals the missing pair then $(a + A, b + B) \bmod 2 = (2a + d + 1, 2b + d) \bmod 2 = (1 - d, d)$; but if $(a + A, b + B) \bmod 2 = (0, 1)$ then d is 0 by construction, and if $(a + A, b + B) \bmod 2 = (1, 0)$ then d is 1 by construction.

Thus $(a + (A \bmod 2), b + (B \bmod 2))$ is earlier in the chain, and doubling it produces the desired $(A + (A \bmod 2), B + (B \bmod 2))$.

4.7.5 The other pair in each line: second addition

If $D = 0$ then the pair $(A + (A + D \bmod 2), B + (B + D + 1 \bmod 2))$ is equal to $(2a + 2(A \bmod 2), 2b + 1)$. We claim that this pair can be obtained by adding $(a + (A \bmod 2), b + 1)$ and $(a + (A \bmod 2), b)$, with difference $(0, 1)$.

If $(a, b) = (0, 0)$ then $(a + (A \bmod 2), b + 1)$ is either $(0, 1)$ or $(1, 1)$, both of which are already in the chain; and $(a + (A \bmod 2), b)$ is either $(0, 0)$ or $(1, 0)$, both of which are already in the chain. So assume that $(a, b) \neq (0, 0)$.

The chain already contains, via $C_d(a, b)$, all pairs $(a + \{0, 1\}, b + \{0, 1\})$ except $(a + (a + d + 1 \bmod 2), b + (b + d \bmod 2))$. Suppose that the missing pair is equal to $(a + (A \bmod 2), b + 1)$ or $(a + (A \bmod 2), b)$. Then $a + (a + d + 1 \bmod 2) = a + (A \bmod 2)$, so $a + A \bmod 2 = 2a + d + 1 \bmod 2 = 1 - d$. If $(a + A, b + B) \bmod 2 = (0, 1)$ then $d = 0$ by construction, contradiction. If $(a + A, b + B) \bmod 2 = (1, 0)$ then $d = 1$ by construction, contradiction. If $(a + A, b + B) \bmod 2 = (0, 0)$ then

$d = D = 0$ by construction, contradiction. If $(a + A, b + B) \bmod 2 = (1, 1)$ then $d = 1 - D = 1$ by construction, contradiction.

Similarly, if $D = 1$, then the pair $(A + (A + D \bmod 2), B + (B + D + 1 \bmod 2))$ in $C_D(A, B)$ is equal to $(2a + 1, 2b + 2(B \bmod 2))$, which can be obtained by adding $(a + 1, b + (B \bmod 2))$ and $(a, b + (B \bmod 2))$, with difference $(1, 0)$.

4.8 Larger differences

Montgomery in [Mon92b] also introduced a more complicated method, called **PRAC**, to compute differential addition-subtraction chains. Recall that these are addition-subtraction chains where each sum computation $n + m$ has $n - m$ already in the chain and where each difference computation $n - m$ has $n + m$ already in the chain. A simple ladder, as in the Lucas ladder and the Montgomery ladder, uses 2 operations (1 differential addition and 1 doubling) for each bit of n; PRAC uses fewer than 1.6 operations for each bit. This section is an introduction to PRAC.

Most of the operations in PRAC are differential additions with large difference, and these are more expensive than doublings with Montgomery's formulas, but PRAC still does slightly better than 9**M** per bit. PRAC produces much larger speedups in the 2-dimensional case discussed in Section 4.7, reducing the cost of computing $mP + nQ$ below 11**M** per bit. The complicated structure of the resulting chains seems to be incompatible with constant-time ECC computations, but is not a problem for ECM.

4.8.1 Examples of large-difference chains

Let d, e be coprime integers with $0 \le d \le e$. This section reviews several ways to construct a *one-dimensional* differential addition chain that starts from $0, 1$ and that contains $e - d$ and d and e. Euclid's chain is the simplest but generally longest; Tsuruoka's chain is the most complicated but generally shortest.

Euclid's chain $E(d, e)$ is defined recursively as follows:

$$E(d, e) = \begin{cases} 0, e & \text{if } d = 0 \\ E(e - d, e) & \text{if } e/2 < d \\ E(d, e - d), e & \text{otherwise} \end{cases}$$

For example,

$$E(11, 97) = 0, 1, 2, 3, 5, 7, 9, 11, 20, 31, 42, 53, 64, 75, 86, 97.$$

One can easily prove by induction on e that $E(d, e)$ is a differential addition

chain that starts from 0, 1 and that contains $e - d$ and d and e. The point is that if $e > 1$ then e can be obtained by adding d and $e - d$, since the difference of d and $e - d$ is earlier in the chain.

A more sophisticated differential addition chain $S(d, e)$ is defined recursively as follows:

$$S(d,e) = \begin{cases} 0, e & \text{if } d = 0 \\ S(e-d, e) & \text{if } e/2 < d \\ S(d, e/2), e-d, e & \text{if } 0 < d < e/4 \text{ and } e \in 2\mathbb{Z} \\ S(d, e-d), e & \text{otherwise} \end{cases}$$

For example,

$$S(11, 97) = 0, 1, 2, 3, 4, 5, 9, 10, 11, 21, 32, 43, 75, 86, 97.$$

What's new in $S(d, e)$, compared to $E(d, e)$, is the $S(d, e/2), e - d, e$ case. In this case, e is obtained by doubling $e/2$; d appears in $S(d, e/2)$ by induction; and $e - d$ is obtained by adding $e/2 - d$ to $e/2$.

Bleichenbacher's differential addition chain $B(d, e)$, introduced in [Ble96, Section 5.3] and republished without credit as the main result of [CL00], is defined recursively as follows:

$$B(d,e) = \begin{cases} 0, e & \text{if } d = 0 \\ B(e-d, e) & \text{if } e/2 < d \\ B(d, e/2), e-d, e & \text{if } 0 < d < e/5 \text{ and } e \in 2\mathbb{Z} \\ B(d, (e+d)/2), e-d, e & \text{if } 0 < d < e/5 \text{ and } e \notin 2\mathbb{Z} \text{ and } e + d \in 2\mathbb{Z} \\ B(d/2, e-d/2), d, e & \text{if } 0 < d < e/5 \text{ and } e \notin 2\mathbb{Z} \text{ and } d \in 2\mathbb{Z} \\ B(d, e-d), e & \text{otherwise} \end{cases}$$

For example,

$$B(11, 97) = 0, 1, 2, 3, 5, 6, 10, 11, 21, 32, 43, 54, 86, 97.$$

Beware that there are two typographical errors in [Ble96, Section 5.3]: "$x - y, y/2, z$" should be "$x-y/2, y/2, z$" and "$x/2, x-y, z$" should be "$x/2, y-x/2, z$."

Tsuruoka's differential addition chain $T(d, e)$, introduced in [Tsu01], is de-

fined recursively as follows:

$$T(d,e) = \begin{cases} 0,e & \text{if } d = 0 \\ T(e-d,e) & \text{if } e < 2d \\ T(d,e/2),e-d,e & \text{if } 2d \le e \le 2.09d \text{ and } e \in 2\mathbb{Z} \\ T(d,e/2),e-d,e & \text{if } 3.92d \le e \text{ and } e \in 2\mathbb{Z} \\ T(d,(e+d)/3), & \\ \quad (2e-d)/3,e-d,e & \text{if not and } 5.7d \le e \text{ and } e+d \in 3\mathbb{Z} \\ T(d,(e-d)/3),(e+2d)/3, & \\ \quad (2e-2d)/3,e-d,e & \text{if not and } 4.9d \le e \text{ and } e-d \in 3\mathbb{Z} \\ T(d,(e+d)/2),e-d,e & \text{if not and } 4.9d \le e \text{ and } d+e \in 2\mathbb{Z} \\ T(d,e/3), & \\ \quad d+e/3,2e/3,e-d,e & \text{if not and } 6.8d \le e \text{ and } e \in 3\mathbb{Z} \\ T(d/2,e-d/2),d,e & \text{if not and } 9d \le e \text{ and } d \in 6\mathbb{Z} \\ T(d,e-d),e & \text{otherwise} \end{cases}$$

For example,

$$T(11,97) = 0,1,2,3,4,7,11,14,25,36,61,86,97.$$

All of these chains were designed with the goal of minimizing length, i.e., the number of additions. Slightly different constructions should do better in other cost measures, such as the number of field multiplications in the elliptic-curve context.

4.8.2 CFRC, PRAC, etc.

There is a well-known "duality" between two-dimensional addition chains that contain the pair (d,e) and one-dimensional addition chains that contain both d and e. See, e.g., [Sta03, Section 2.2.2].

For example, one way to compute $dP + eQ$ is to first compute $P + Q$ and then compute $d(P + Q) + (e - d)Q$. This transformation reduces the problem of constructing an addition chain for (d,e) to the problem of constructing an addition chain for $(d,e-d)$. This is the dual of the following reduction, which was used repeatedly above: to build an addition chain for d and e, first build an addition chain for d and $e-d$, and then compute e as the sum of d and $e-d$.

Duality does not exactly preserve costs for differential chains; see, e.g., [Sta03, Example 3.28] and [Sta03, Section 3.4, final paragraph]. One can nevertheless see a large overlap between ideas for optimizing a two-dimensional

chain for the pair (d, e) and ideas for optimizing a one-dimensional chain for d and e. In particular, Montgomery's "CFRC" chain in [Mon92b, Section 5] is a simple construction of a two-dimensional chain for (d, e), comparable to Euclid's chain. Montgomery's "PRAC" chain in [Mon92b, Section 7] is a more complicated construction, comparable to (and predating) Tsuruoka's chain. See [Ber06b] for an "extended-gcd" chain that, in experiments, produces slightly better results than PRAC.

4.8.3 Allowing d to vary

The standard way to construct a one-dimensional differential addition chain for e is to choose some d coprime to e and use one of the above algorithms to construct a chain containing $e - d, d, e$. Here are three refinements:

- Choose d to be very close to $2e/(1 + \sqrt{5})$. This guarantees that the top half of the bits of e will be handled with about 1.44 additions per bit; see, e.g., [Sta03, Proposition 3.34]. For example, with $e = 100$, choosing $d = 61$ produces a chain ending $17, 22, 39, 61, 100$.
- Try many d's and take the shortest chain for e. One can, for example, take a range of d's around $2e/(1 + \sqrt{5})$, or around e/α for various constants α having continued fractions consisting of almost entirely 1's and a few 2's.
- If e has a known factor g, construct a chain for e by constructing a chain for e/g, multiplying it by g, and merging the result with a chain for g. This generally produces shorter chains (for a given amount of d-searching time) than handling e directly.

All of these improvements were suggested by Montgomery in [Mon92b, Section 7], in the context of Montgomery's PRAC chain.

Here is a simple experiment to illustrate the importance of trying many d's. Consider each prime number e below 10^6. For each e, try several successive d's coprime to e, starting just above $2e/(1 + \sqrt{5})$; find the shortest $E(d, e)$, the shortest $S(d, e)$, the shortest $B(d, e)$, and the shortest $T(d, e)$. Average the number of additions in these chains as e varies. The following table shows the resulting averages:

number of d's	1	2	4	8	16	32	64	128
$E((\text{best } d), e)$	47.550	34.405	31.286	29.912	29.364	28.876	28.579	28.428
$S((\text{best } d), e)$	34.125	29.630	28.758	28.371	28.194	28.048	27.950	27.899
$B((\text{best } d), e)$	30.794	29.606	28.818	28.415	28.241	28.093	27.993	27.936
$T((\text{best } d), e)$	29.159	28.723	28.431	28.220	28.105	27.996	27.919	27.875

Beware that [Tsu01, Section 4.2] uses only two d's and reports 1.61 additions per bit, and [Sta03, Algorithm 3.33] uses only one d and reports 1.64 additions per bit, while taking more d's easily reaches 1.56 additions per bit. The benefit of trying many chains, and keeping the shortest, was pointed out by Montgomery but does not seem to have been adequately emphasized in the literature. Note that, as shown by the crossover between the S and B rows in the above table, optimizing chains for many d's is not the same as optimizing them for a single d.

References

[ACD+05] Roberto M. Avanzi, Henri Cohen, Christophe Doche, Gerhard Frey, Tanja Lange, Kim Nguyen, and Frederic Vercauteren. *Handbook of Elliptic and Hyperelliptic Curve Cryptography*. Chapman & Hall/CRC Press, 2005.

[Aki01] Toru Akishita. Fast simultaneous scalar multiplication on elliptic curve with montgomery form. In Serge Vaudenay and Amr M. Youssef, editors, *SAC 2001*, volume 2259 of *LNCS*, pages 255–267, Toronto, Ontario, Canada, August 16–17, 2001. Springer, Heidelberg, Germany.

[BBB+13] Razvan Barbulescu, Joppe W. Bos, Cyril Bouvier, Thorsten Kleinjung, and Peter L. Montgomery. Finding ECM-friendly curves through a study of Galois properties. *The Open Book Series – Proceedings of the Tenth Algorithmic Number Theory Symposium*, pages 63–86, 2013.

[BBJ+08] Daniel J. Bernstein, Peter Birkner, Marc Joye, Tanja Lange, and Christiane Peters. Twisted Edwards curves. In Serge Vaudenay, editor, *AFRICACRYPT 08*, volume 5023 of *LNCS*, pages 389–405, Casablanca, Morocco, June 11–14, 2008. Springer, Heidelberg, Germany.

[BCKL15] Daniel J. Bernstein, Chitchanok Chuengsatiansup, David Kohel, and Tanja Lange. Twisted hessian curves. In Kristin E. Lauter and Francisco Rodríguez-Henríquez, editors, *LATINCRYPT 2015*, volume 9230 of *LNCS*, pages 269–294, Guadalajara, Mexico, August 23–26, 2015. Springer, Heidelberg, Germany.

[BCL14] Daniel J. Bernstein, Chitchanok Chuengsatiansup, and Tanja Lange. Curve41417: Karatsuba revisited. In Lejla Batina and Matthew Robshaw, editors, *CHES 2014*, volume 8731 of *LNCS*, pages 316–334, Busan, South Korea, September 23–26, 2014. Springer, Heidelberg, Germany.

[BCLS14] Daniel J. Bernstein, Chitchanok Chuengsatiansup, Tanja Lange, and Peter Schwabe. Kummer strikes back: New DH speed records. In Palash Sarkar and Tetsu Iwata, editors, *ASIACRYPT 2014, Part I*, volume 8873 of *LNCS*, pages 317–337, Kaoshiung, Taiwan, R.O.C., December 7–11, 2014. Springer, Heidelberg, Germany.

[BDL+11] Daniel J. Bernstein, Niels Duif, Tanja Lange, Peter Schwabe, and Bo-Yin Yang. High-speed high-security signatures. In Bart Preneel and Tsuyoshi Takagi, editors, *CHES 2011*, volume 6917 of *LNCS*, pages 124–142, Nara, Japan, September 28 – October 1, 2011. Springer, Heidelberg, Germany.

[Ber06a] Daniel J. Bernstein. Curve25519: New Diffie-Hellman speed records. In Yung et al. [YDKM06], pages 207–228.

[Ber06b] Daniel J. Bernstein. Differential addition chains, 2006. `https://cr.yp.to/papers.html#diffchain`.

[BJL+15] Daniel J. Bernstein, Simon Josefsson, Tanja Lange, Peter Schwabe, and Bo-Yin Yang. EdDSA for more curves. Cryptology ePrint Archive, Report 2015/677, 2015. `http://eprint.iacr.org/2015/677`.

[BL07] Daniel J. Bernstein and Tanja Lange. Faster addition and doubling on elliptic curves. In Kaoru Kurosawa, editor, *ASIACRYPT 2007*, volume 4833 of *LNCS*, pages 29–50, Kuching, Malaysia, December 2–6, 2007. Springer, Heidelberg, Germany.

[BL09] Daniel J. Bernstein and Tanja Lange. *YZ* coordinates with square *d* for Edwards curves, 2009. https://hyperelliptic.org/EFD/g1p/auto-edwards-yz.html.

[BL11] Daniel J. Bernstein and Tanja Lange. A complete set of addition laws for incomplete Edwards curves. *Journal of Number Theory*, 131:858–872, 2011.

[BL14] Daniel J. Bernstein and Tanja Lange. SafeCurves: choosing safe curves for elliptic-curve cryptography, 2014. https://safecurves.cr.yp.to.

[BL16] Daniel J. Bernstein and Tanja Lange. Explicit-Formulas Database, 2016. https://hyperelliptic.org/EFD.

[Ble96] Daniel Bleichenbacher. *Efficiency and security of cryptosystems based on number theory*. PhD thesis, ETH Zürich, 1996. https://cr.yp.to/bib/1996/bleichenbacher-thesis.pdf.

[Cho15] Tung Chou. Sandy2x: New Curve25519 speed records. In Orr Dunkelman and Liam Keliher, editors, *Selected Areas in Cryptography - SAC 2015 - 22nd International Conference, Sackville, NB, Canada, August 12-14, 2015, Revised Selected Papers*, volume 9566 of *Lecture Notes in Computer Science*, pages 145–160. Springer, 2015.

[CL00] S. Y. Chiou and C. S. Laih. An efficient algorithm for computing the Luc chain. *IEE Proceedings on Computers and Digital Techniques*, 147:263–265, 2000.

[DH76] Whitfield Diffie and Martin Hellman. New directions in cryptography. *IEEE Transactions on Information Theory*, 22:644–654, 1976.

[DIK06] Christophe Doche, Thomas Icart, and David R. Kohel. Efficient scalar multiplication by isogeny decompositions. In Yung et al. [YDKM06], pages 191–206.

[Edw07] Harold M. Edwards. A normal form for elliptic curves. *Bulletin of the American Mathematical Society*, 44:393–422, 2007.

[Gau06] Pierrick Gaudry. Variants of the Montgomery form based on Theta functions, 2006. https://cr.yp.to/bib/2006/gaudry-toronto.pdf.

[GL09] Pierrick Gaudry and David Lubicz. The arithmetic of characteristic 2 Kummer surfaces and of elliptic Kummer lines. *Finite Fields and Their Applications*, 15:246–260, 2009. https://hal.inria.fr/inria-00266565v2.

[Ham15] Mike Hamburg. Ed448-goldilocks, a new elliptic curve. Cryptology ePrint Archive, Report 2015/625, 2015. http://eprint.iacr.org/2015/625.

[Hus04] Dale Husemöller. *Elliptic Curves*, volume 111 of *Graduate Texts in Mathematics*. Springer, 2004.

[HWCD08] Hüseyin Hisil, Kenneth Koon-Ho Wong, Gary Carter, and Ed Dawson. Twisted Edwards curves revisited. In Josef Pieprzyk, editor, *ASIACRYPT 2008*, volume 5350 of *LNCS*, pages 326–343, Melbourne, Australia, December 7–11, 2008. Springer, Heidelberg, Germany.

[Kob87] Neal Koblitz. Elliptic curve cryptosystems. *Mathematics of Computation*, 48:203–209, 1987.

[Mil86] Victor S. Miller. Use of elliptic curves in cryptography. In Hugh C. Williams, editor, *CRYPTO'85*, volume 218 of *LNCS*, pages 417–426, Santa Barbara, CA, USA, August 18–22, 1986. Springer, Heidelberg, Germany.

[Mon87] Peter L. Montgomery. Speeding the Pollard and elliptic curve methods of factorization. *Mathematics of Computation*, 48:243–264, 1987.

[Mon92a] Peter L. Montgomery. *An FFT extension of the elliptic curve method of factorization*. PhD thesis, University of California at Los Angeles, 1992. `https://cr.yp.to/bib/1992/montgomery.pdf`.

[Mon92b] Peter L. Montgomery. Evaluating recurrences of form $X_{m+n} = f(X_m, X_n, X_{m-n})$ via Lucas chains, 1992. `https://cr.yp.to/bib/1992/montgomery-lucas.pdf`.

[Sta03] Martijn Stam. *Speeding up subgroup cryptosystems*. PhD thesis, Technische Universiteit Eindhoven, 2003. `https://dx.doi.org/10.6100/IR564670`.

[Tsu01] Yukio Tsuruoka. Computing short Lucas chains for elliptic curve cryptosystems. *IEICE Transactions on Fundamentals*, E84-A(5):1227–1233, 2001.

[YDKM06] Moti Yung, Yevgeniy Dodis, Aggelos Kiayias, and Tal Malkin, editors. *PKC 2006*, volume 3958 of *LNCS*, New York, NY, USA, April 24–26, 2006. Springer, Heidelberg, Germany.

Subject index