WITT'S CANCELLATION THEOREM SEEN AS A CANCELLATION

SUNIL K. CHEBOLU, DAN MCQUILLAN, AND JÁN MINÁČ

Dedicated to the late Professor Amit Roy

Abstract. Happy birthday to the Witt ring! The year 2017 marks the 80th anniversary of Witt's famous paper containing some key results, including the Witt cancellation theorem, which form the foundation for the algebraic theory of quadratic forms. We pay homage to this paper by presenting a transparent, algebraic proof of the Witt cancellation theorem, which itself is based on a cancellation. We also present an overview of some recent spectacular work which is still building on Witt's original creation of the algebraic theory of quadratic forms.

1. Introduction

The algebraic theory of quadratic forms will soon celebrate its 80th birthday. Indeed, it was 1937 when Witt's pioneering paper [24] – a mere 14 pages – first introduced many beautiful results that we love so much today. These results form the foundation for the algebraic theory of quadratic forms. In particular they describe the construction of the Witt ring itself. Thus the cute little baby, "The algebraic theory of quadratic forms" was healthy and screaming with joy, making his father Ernst Witt very proud. Grandma Emmy Noether,

Figure 1. Shortly after the birth of the Algebraic Theory of Quadratic Forms.

had she still been alive, would have been so delighted to see this little tyke! [1] Almost right

[1]Emmy Noether's male Ph.D students, including Ernst Witt, were often referred to as "Noether's boys." The reader is encouraged to consult [6] to read more about her profound influence on the development of mathematics.

near the beginning, this precocious baby was telling us the essential fact needed to construct the Witt ring of quadratic forms over an arbitrary field. This result, originally Satz 4 in [24], is now formulated as the Witt cancellation theorem, and it is the technical heart of Witt's brilliant idea to study the collection of *all* quadratic forms over a given field as a single algebraic entity. Prior to Witt's paper, quadratic forms were studied one at a time. However Witt showed that a certain collection of quadratic forms under an equivalence relation can be equipped with the structure of a commutative ring. Indeed, Satz 6 says:

"*Die Klassen ähnlicher Formen bilden einen Ring*"

which means, "The classes of similar forms, form a ring." In order to honor Witt's contributions, this ring is now called the *Witt ring*.

The Witt ring remains a central object of study, even 80 years after its birth. Building on Voevodsky's Fields medal winning work from 2002, Orlov, Vishik and Voevodsky recently settled Milnor's conjecture [14] on quadratic forms, which is a deep statement about the structure of the Witt ring. This work uses sophisticated tools from algebraic geometry and homotopy theory to provide a complete set of invariants for quadratic forms, extending the classical invariants known to Witt [24], including dimension, discriminant and the Clifford invariant.

In addition to its crucial role in defining the Witt ring, the Witt cancellation theorem also has other important applications, such as establishing Sylvester's law of Inertia, which classifies quadratic forms over the field of real numbers. Clearly the Witt cancellation theorem is special and therefore deserves further analysis. The main goal of this paper is to present a transparent and algebraic proof to complement the classical geometric proof, and then carefully compare the two approaches.

The paper is organized as follows. In Section 2 we state the Witt cancellation theorem, guide the reader towards our proof of the cancellation theorem, and then present the proof itself. A geometric approach to Witt cancellation, based on hyperplane reflections, is presented in Section 3. In Section 4 we provide a "homotopy" (a gentle deformation) between the algebraic and geometric approaches. Using Witt cancellation as the key, we review the construction of the Witt ring of quadratic forms in Section 5. In Section 6 we present an informal overview of the Milnor conjectures on quadratic forms and some recent related developments. In Section 7 we reveal an interesting surprise. Section 8, the epilogue, is a tribute to several great mathematicians connected with our story. The epilogue also contains a challenge for our readers. All sections, except possibly Section 6, can be read profitably by any undergraduate student who is familiar with basic linear algebra.

We begin with some preliminaries. Throughout the paper we assume that our base field F has characteristic not equal to 2. There are several equivalent definitions of a quadratic form. The following is probably the most commonly used definition. An n-ary *quadratic form* q over F is a homogeneous polynomial of degree 2 in n variable over F:

$$q = \sum_{i,j=1}^{n} a_{ij} x_i x_j \quad \text{for} \quad a_{ij} \text{ in } F.$$

It is customary to render the coefficients symmetric by writing

$$q = \sum_{i,j=1}^{n} b_{ij} x_i x_j, \quad \text{where } b_{ij} = \frac{a_{ij} + a_{ji}}{2},$$

therefore $b_{ij} = b_{ji}$. (This is possible because the characteristic of our field is not 2.)

If we view $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ as a column vector, and its transpose $\mathbf{x}^t$ as a row vector, then we can write

$$q(\mathbf{x}) = \mathbf{x}^t B\mathbf{x},$$

where $B = (b_{ij})$ is an $n \times n$ matrix. In other words, we associate q with a symmetric matrix B which also defines a symmetric bilinear form on $V \times V$, where $V = F^n$. Two n-ary quadratic forms q_a and q_b are *equivalent*, or *isometric* if for some non-singular $n \times n$ matrix M we have

$$q_a(\mathbf{x}) = q_b(M\mathbf{x}).$$

In this case we write $q_a \cong q_b$. Recall that two symmetric matrices A and B are said to be *congruent* if there exist and invertible matrix M such that $A = M^t BM$. Equivalence classes of quadratic forms thus correspond to congruence classes of symmetric matrices [2].

The following useful result is well known and can be found in any standard textbook on quadratic forms; see for example [8] or [19].

Theorem 1.1. *An n-ary quadratic form over a field F of characteristic not equal to 2 is equivalent to a diagonal form, i.e., a form that is equal to $a_1x_1^2 + \cdots + a_nx_n^2$ for some field elements $a_1, \ldots, a_n$.*

For brevity we shall denote the diagonal quadratic form $a_1x_1^2+\cdots+a_nx_n^2$ by $\langle a_1, \ldots, a_n\rangle$. In view of this theorem it is enough to study diagonal forms over F. Furthermore, we assume that our diagonal quadratic forms are *non-degenerate*, i.e., $a_i \neq 0$ for $i = 1, \ldots, n$. The number n is called the dimension of q.

2. Witt Cancellation: algebraic approach

In this section we will present a transparent and algebraic proof of the Witt cancellation theorem to complement the classical geometric proof. The following is the simplest form of the Witt cancellation theorem. Other general statements can be easily derived from this simple form.

Theorem 2.1. (Witt cancellation) *Let $q_a = \langle a_1, a_2, \ldots, a_n\rangle$ and $q_b = \langle b_1, b_2, \ldots, b_n\rangle$ be non-degenerate n-ary quadratic forms over a field F of characteristic not equal to 2, with $n > 1$, and assume that $a_1 = b_1$. If there is an isometry $q_a \cong q_b$, then there is another isometry $\langle a_2, \ldots, a_n\rangle \cong \langle b_2, \ldots, b_n\rangle$.*

Before presenting our proof, we will explain the key idea in such a way that the reader may build the proof before even reading it – a *guided self-discovery* approach. Witt's cancellation theorem essentially says that we may "cancel" a common term, a_1, from both sides of a given

[2]To illustrate the notion of equivalence of quadratic forms, consider the quadratic form $q_b(z_1, z_2) = 5z_1^2 - 2z_1z_2 + 5z_2^2$ over the field $\mathbb{R}$ of real numbers. What is the conic section that is represented by the equation $q_b(z_1, z_2) = 1$? The given quadratic form is equivalent over $\mathbb{R}$ to the form $q_a(x_1, x_2) = 4x_1^2 + 6x_2^2$. The equivalence is given by the equations

$$\begin{aligned} z_1 &= \frac{1}{\sqrt{2}}x_1 - \frac{1}{\sqrt{2}}x_2, \text{ and} \\ z_2 &= \frac{1}{\sqrt{2}}x_1 + \frac{1}{\sqrt{2}}x_2. \end{aligned}$$

It is clear that the new equation $4x_1^2 + 6x_2^2 = 1$ represents an ellipse, and therefore so does the original equation.

isometry, in order to obtain a new isometry. We want a proof that reflects this cancellation directly. To this end, recall that by the definition of isometry, there is an invertible linear transformation

$$z_i = m_{i1}x_1 + \cdots + m_{in}x_n, \quad i = 1, \ldots, n, \quad m_{ir} \in F, \tag{1}$$

which takes q_b to q_a. This means that the isometry

$$a_1x_1^2 + a_2x_2^2 + \cdots + a_nx_n^2 \cong b_1z_1^2 + b_2z_2^2 + \cdots + b_nz_n^2 \tag{2}$$

becomes a polynomial identity in the n variables $x_1, x_2, \ldots, x_n$ after using the n transformations in Equation (1). Our idea then is to simply take this one step further, by substituting x_1 with a carefully chosen linear combination of the remaining $n-1$ variables $x_2, \ldots, x_n$ so that in Equation (2), the first term on the left hand side will cancel with the first term on the right hand side. This will then give us our desired isometry. So we now ask: what is this magical substitution? In other words, which linear combination do we use for x_1? If x is the answer to this question, then it should satisfy the equation

$$a_1x^2 = b_1(mx+y)^2, \quad \text{where we set } m := m_{11} \quad \text{and} \quad y := m_{12}x_2 + \cdots + m_{1n}x_n.$$

However, since $a_1 = b_1$, it is sufficient that our x satisfy

$$x = mx + y. \tag{3}$$

Note that this last equation reminds us of exciting, good old times from high school where we learned how to solve linear equations:

$$x = mx + y \quad \Longrightarrow \quad x = \frac{y}{1-m} \text{ if } m \neq 1.$$

After this motivational warm-up, it is now time to give a formal proof. The reader will see that our proof will be quite transparent and will be based on the simple identity:

$$\boxed{\frac{y}{1-m} = \frac{my}{1-m} + y} \tag{4}$$

Proof (of the Witt cancellation theorem). Since $q_a \cong q_b$, we can write

$$a_1x_1^2 + \cdots + a_nx_n^2 = q_a(\mathbf{x}) = q_b(M\mathbf{x}) = b_1z_1^2 + \cdots + b_nz_n^2, \tag{5}$$

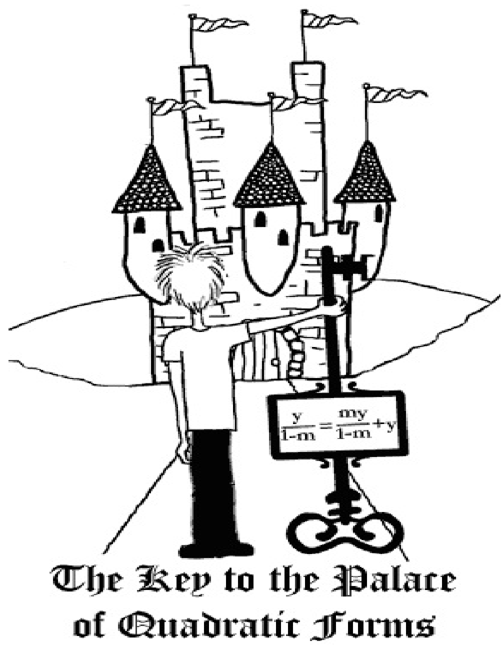
where $M = (m_{ij})$ is an $n \times n$ invertible matrix over F and $z_i = m_{i1}x_1 + \cdots + m_{in}x_n$ for i from 1 to n. We first argue that m_{11} can be assumed without loss of generality to be not equal to 1. As a matter of fact, if $m_{11} = 1$, then we replace m_{1k} with $-m_{1k}$ for all k. This changes z_1 to $-z_1$. However, that does not effect Equation (5). So we assume without loss of generality that $m_{11} \neq 1$.

To prove our theorem we would like to cancel the first terms ($a_1x_1^2$ and $b_1z_1^2$) on either sides of Equation (5). To do this, in Equation 5 we make the substitution

$$x_1 = \frac{y}{1-m_{11}}, \tag{6}$$

where

$$y := z_1 - m_{11}x_1 = m_{12}x_2 + \cdots + m_{1n}x_n. \tag{7}$$



Note that this is a valid substitution because $m_{11} \neq 1$. Moreover, this substitution expresses x_1 as a linear combination of $x_2, \ldots, x_n$. This substitution, in conjunction with the assumption $a_1 = b_1$ and our identity (4), gives the following equations.

$$\begin{aligned} a_1 x_1^2 &= a_1 \left(\frac{y}{1 - m_{11}} \right)^2 \\ &= b_1 \left(\frac{y}{1 - m_{11}} \right)^2 \\ &= b_1 \left(y + m_{11} \left(\frac{y}{1 - m_{11}} \right) \right)^2 \qquad \text{from identity (4)} \\ &= b_1 (y + m_{11} x_1)^2 \\ &= b_1 z_1^2 \end{aligned}$$

Therefore we can cancel these two terms in our original equation (5), which now reduces to one in $2(n-1)$ variables:

$$a_2 x_2^2 + \cdots + a_n x_n^2 = b_2 z_2^2 + \cdots + b_n z_n^2. \tag{8}$$

In this new equation, for $i \geq 2$, z_i is expressed as a linear combination of $x_2, x_3, \ldots, x_n$, say $z_i = w_i(x_2, x_3, \ldots, x_n)$. It remains to show that this linear transformation is invertible. To see this, let $N = (n_{ij})$ be the change-of-coordinates matrix which corresponds to our linear transformation $z_i = w_i(x_2, x_3, \ldots, x_n), i = 2, \ldots, n$. Then the transformation between the $(n-1)$-ary forms $s_a := \langle a_2, \ldots, a_n \rangle$ and $s_b := \langle b_2, \ldots, b_n \rangle$ is given by the matrix equation

$$A = N^T B N,$$

where A and B are the diagonal matrices representing the forms s_a and s_b respectively. Taking determinants on both sides of the last equation, we get

$$\det(A) = \det(B)(\det(N))^2.$$

Since s_a is non-degenerate, $\det(A)$ is non-zero and therefore $\det(N)$ is also non-zero. This shows that N is invertible. Thus we have shown that the forms s_a and s_b are isometric. □

Remark 2.2. In the above proof, we see that the Witt cancellation theorem actually follows from the formal algebraic cancellation of like terms in a polynomial identity, explaining the title of our paper.

3. Witt Cancellation: geometric approach

In this section we present the standard, coordinate-free, geometric approach to quadratic forms and the Witt cancellation theorem.

A *quadratic space* is a finite-dimensional F-vector space equipped with a symmetric bilinear form

$$B\colon V \times V \to F.$$

The associated quadratic form $q\colon V \to F$ is obtained by setting $q(\mathbf{v}) = B(\mathbf{v}, \mathbf{v})$. The bilinear form B can be recovered from q because of the identity

$$B(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(q(\mathbf{x} + \mathbf{y}) - q(\mathbf{x}) - q(\mathbf{y})),$$

as one can easily check. Therefore a quadratic space can be denoted by (V, B), or equivalently by (V, q).

Coordinate free definitions in quadratic form theory are naturally analogous to their coordinate counterparts. For instance, an *isometry* between (V, B_1) and (V, B_2) is a linear isomorphism $T\colon V \to V$ such that $B_2(\mathbf{x}, \mathbf{y}) = B_1(T(\mathbf{x}), T(\mathbf{y}))$ for all $\mathbf{x}$ and $\mathbf{y}$ in V. Vectors $\mathbf{x}$ and $\mathbf{y}$ in V are said to be *orthogonal* if $B(\mathbf{x}, \mathbf{y}) = 0$. A quadratic space (V, B) is *non-degenerate* if $B(\mathbf{v}, \mathbf{w}) = 0$ for all $\mathbf{w}$ in V implies $\mathbf{v} = \mathbf{0}$. Given two quadratic spaces (V_1, q_1) and (V_2, q_2), there is a natural quadratic from on the space $V_1 \oplus V_2$ which is defined by

$$q((\mathbf{x}_1, \mathbf{x}_2)) := q_1(\mathbf{x}_1) + q_2(\mathbf{x}_2).$$

This quadratic space is denoted by $(V_1, q_1) \perp (V_2, q_2)$.

The geometric form of the Witt cancellation theorem in its simplest form can now be stated as follows.

Theorem 3.1. *Let (V, q) be an n-dimensional non-degenerate quadratic form with $n > 1$, and let $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ be two orthogonal bases for (V, q). If $q(\mathbf{e}_1) = q(\mathbf{f}_1)$, then q restricted to Span$\{\mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is isometric to q restricted to Span$\{\mathbf{f}_2, \ldots, \mathbf{f}_n\}$.*

Given a quadratic space (V, q) and a vector $\mathbf{u}$ in V such that $q(\mathbf{u}) \neq 0$, the map

$$\tau_{\mathbf{u}}(\mathbf{z}) := \mathbf{z} - \frac{2B(\mathbf{z}, \mathbf{u})}{q(\mathbf{u})}\mathbf{u}$$

can be easily shown to be an isometry of (V, q); see [8, Page 13]. In fact, this map is the reflection in the plane perpendicular to $\mathbf{u}$. A key ingredient in the proof of Theorem 3.1 is the following hyperplane reflection lemma.

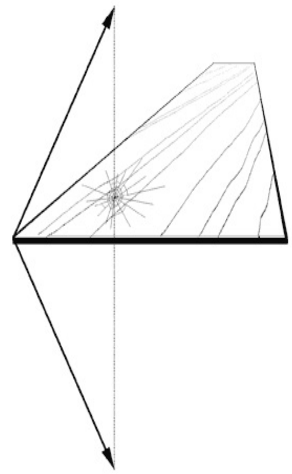


FIGURE 2. Hyperplane reflection.

Lemma 3.2. *Let (V, q) be a quadratic space and let $\mathbf{x}$ and $\mathbf{y}$ be two vectors in V such that $q(\mathbf{x}) = q(\mathbf{y}) \neq 0$. Then there exists an isometry $\rho\colon (V, q) \cong (V, q)$ which sends $\mathbf{x}$ to $\mathbf{y}$.*

Proof. Note that

$$q(\mathbf{x}+\mathbf{y}) + q(\mathbf{x}-\mathbf{y}) = B(\mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}) + B(\mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}) = 2q(\mathbf{x}) + 2q(-\mathbf{y}) = 4q(\mathbf{x}) \neq 0.$$

This means $q(\mathbf{x}+\mathbf{y})$ and $q(\mathbf{x}-\mathbf{y})$ both cannot be zero simultaneously. Suppose that $q(\mathbf{x}-\mathbf{y}) \neq 0$. Then $\tau_{\mathbf{x}-\mathbf{y}}$ is an isometry that maps $\mathbf{x}$ to $\mathbf{y}$. To see this, first note that

$$q(\mathbf{x}-\mathbf{y}) = B(\mathbf{x}, \mathbf{x}) + B(\mathbf{y}, \mathbf{y}) - 2B(\mathbf{x}, \mathbf{y}) = 2B(\mathbf{x}, \mathbf{x}) - 2B(\mathbf{x}, \mathbf{y}) = 2B(\mathbf{x}, \mathbf{x}-\mathbf{y}).$$

Therefore,

$$\tau_{\mathbf{x}-\mathbf{y}}(\mathbf{x}) = \mathbf{x} - \frac{2B(\mathbf{x}, \mathbf{x}-\mathbf{y})}{q(\mathbf{x}-\mathbf{y})}(\mathbf{x}-\mathbf{y}) = \mathbf{x} - (\mathbf{x}-\mathbf{y}) = \mathbf{y}.$$

If $q(\mathbf{x}+\mathbf{y}) \neq 0$, then since $q(\mathbf{x}+\mathbf{y}) = q(\mathbf{x}-(-\mathbf{y}))$, the above argument shows that $\tau_{\mathbf{x}+\mathbf{y}}(\mathbf{x}) = \tau_{\mathbf{x}-(-\mathbf{y})}\mathbf{x} = -\mathbf{y}$, and therefore $-(\tau_{\mathbf{x}+\mathbf{y}}\mathbf{x}) = \mathbf{y}$. This completes the proof our lemma. □

Proof of Theorem 3.1 We are given that $q(\mathbf{e}_1) = q(\mathbf{f}_1)$. This common value cannot be zero because q is non-degenerate. Therefore, as observed in the proof of the above lemma $q(\mathbf{e}_1 + \mathbf{f}_1)$ and $q(\mathbf{e}_1 - \mathbf{f}_1)$ both cannot be zero simultaneously. Replacing $\mathbf{f}_1$ with $-\mathbf{f}_1$ if necessary, we may assume that $q(\mathbf{e}_1 - \mathbf{f}_1) \neq 0$. Then we claim that the isometry

$$\tau_{\mathbf{e}_1-\mathbf{f}_1}$$

does the job. That is, it gives an isometry between $\mathbf{e}_1^\perp := \mathrm{Span}\{\mathbf{e}_2, \ldots, \mathbf{e}_n\}$ and $\mathbf{f}_1^\perp := \mathrm{Span}\{\mathbf{f}_2, \ldots, \mathbf{f}_n\}$. Indeed, from the above lemma, the map $\tau_{\mathbf{e}_1-\mathbf{f}_1}$ takes $\mathbf{e}_1$ to $\mathbf{f}_1$. Since $\tau_{\mathbf{e}_1-\mathbf{f}_1}$ is an isometry of (V, B), it maps $\mathbf{e}_1^\perp$ to $\mathbf{f}_1^\perp$. Thus $\tau_{\mathbf{e}_1-\mathbf{f}_1}$ restricts to a map $\mathbf{e}_1^\perp \to \mathbf{f}_1^\perp$. Since the restriction of an isometry is an isometry, we are done. □

4. A "homotopy" between the algebraic and geometric approaches

As mentioned in the introduction, our algebraic approach complements the classical geometric approach. The goal of this section is to exhibit a "homotopy" between these two approaches. More precisely, we will show that our substitution in Equation (6)

$$x_1 = \frac{y}{1 - m_{11}}$$

naturally corresponds to the hyperplane reflection mentioned in the previous section.

Let us quickly recapitulate the framework:

- (V, q) is an n-dimensional non-degenerate quadratic form.
- $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ and $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ are two orthogonal bases for (V, q).
- We let $q(\mathbf{e}_i) = a_i$ and $q(\mathbf{f}_i) = b_i$ for all i.
- $a_1 = b_1$, i.e., $q(\mathbf{e}_1) = q(\mathbf{f}_1)$.
- For all i, $w_i = w_i(x_2, x_3, \ldots, x_n)$ is obtained from $z_i = z_i(x_1, \ldots, x_n)$ after replacing x_1 with our substitution, which is a linear combination of $x_2, \ldots, x_n$.

We now have two coordinate representations

$$q_a = \langle a_1, a_2, \ldots, a_n \rangle \quad \text{and} \quad q_b = \langle b_1, b_2, \ldots, b_n \rangle$$

of the form (V, q) with respect to the bases $\{\mathbf{e}_i\}$ and $\{\mathbf{f}_i\}$ respectively. The isometry between q_a and q_b is given by an invertible matrix $M = (m_{ij})$. The change of basis matrix is then M. So we have for $j = 1, \ldots, n$,

$$\mathbf{e}_j = m_{1j}\mathbf{f}_1 + m_{2j}\mathbf{f}_2 + \cdots + m_{nj}\mathbf{f}_n.$$

For the rest of this section, we fix an integer $k \geq 2$. Before going further we explain our strategy for getting the "homotopy." We take a vector $\mathbf{e}_k$ and hit it with our hyperplane reflection $\tau_{\mathbf{e}_1 - \mathbf{f}_1}$. Then we express $\tau_{\mathbf{e}_1 - \mathbf{f}_1}(\mathbf{e}_k)$ as a linear combination of $\mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_n$. By comparing the coefficient c_{ki} of $\mathbf{f}_i$ in $\tau_{\mathbf{e}_1 - \mathbf{f}_1}(\mathbf{e}_k)$ and the coefficient d_{ki} of x_k in w_i for $i \geq 2$, and we will see the equivalence of the two approaches.

To execute this strategy, consider the vector $\mathbf{u} := \mathbf{e}_1 - \mathbf{f}_1 = (m_{11} - 1)\mathbf{f}_1 + m_{21}\mathbf{f}_2 + \cdots + m_{n1}\mathbf{f}_n$. Since $a_1 = b_1$, we have

$$\begin{aligned} q(\mathbf{u}) &= (m_{11} - 1)^2 b_1 + m_{21}^2 b_2 + \cdots + m_{n1}^2 b_n \\ &= \sum_{i=1}^{n} m_{i1}^2 b_i - 2m_{11}b_1 + b_1 = b_1 - 2m_{11}b_1 + b_1 = 2b_1(1 - m_{11}). \end{aligned}$$

(Here we are using the identity $\sum_{i=1}^{n} m_{i1}^2 b_i = b_1$ which comes from unwinding the equation $B(\mathbf{e}_1, \mathbf{e}_1) = a_1 = b_1$.) By replacing $\mathbf{f}_1$ with $-\mathbf{f}_1$ if necessary, we may assume that $m_{11} \neq 1$. Therefore, $q(\mathbf{u}) \neq 0$. Then the formula for our hyperplane reflection is given by

$$\tau_{\mathbf{u}}(\mathbf{z}) = \mathbf{z} - \frac{2B(\mathbf{z}, \mathbf{u})}{2b_1(1 - m_{11})}\mathbf{u} = \mathbf{z} - \frac{B(\mathbf{z}, \mathbf{u})}{b_1(1 - m_{11})}\mathbf{u}.$$

Setting $\mathbf{z} = \mathbf{e}_k$, we obtain the following equations:

$$\begin{aligned} \tau_{\mathbf{u}}(\mathbf{e}_k) &= \mathbf{e}_k - \frac{B(\mathbf{e}_k, \mathbf{e}_1 - \mathbf{f}_1)}{b_1(1-m_{11})}(\mathbf{e}_1 - \mathbf{f}_1) \\ &= \mathbf{e}_k + \frac{B(\mathbf{e}_k, \mathbf{f}_1)}{b_1(1-m_{11})}(\mathbf{e}_1 - \mathbf{f}_1) \\ &= (m_{1k}\mathbf{f}_1 + \cdots + m_{nk}\mathbf{f}_n) + \frac{m_{1k}b_1}{b_1(1-m_{11})}((m_{11}-1)\mathbf{f}_1 + m_{21}\mathbf{f}_2 + \cdots + m_{n1}\mathbf{f}_n) \\ &= (m_{2k}\mathbf{f}_2 + \cdots + m_{nk}\mathbf{f}_n) + \left(\frac{m_{1k}m_{21}}{1-m_{11}}\mathbf{f}_2 + \cdots + \frac{m_{1k}m_{n1}}{1-m_{11}}\mathbf{f}_n\right) \end{aligned}$$

The coefficient of $\mathbf{f}_i$ for $i \geq 2$ in the last expression is:

$$\boxed{c_{ki} := m_{ik} + \frac{m_{1k}m_{i1}}{1-m_{11}}}$$

Now let us change gears and look at our algebraic approach. Recall that we substitute

$$x_1 \longrightarrow \frac{y}{1-m_{11}} \left(= \frac{m_{12}x_2 + \cdots + m_{1n}x_n}{1-m_{11}}\right)$$

in the equations

$$z_i = m_{i1}x_1 + \cdots + m_{in}x_n \quad \text{for } i = 1, 2, \ldots n.$$

Using our substitution for x_1, for $i \geq 2$, we get an expression for w_i:

$$w_i = m_{i1}\left(\frac{m_{12}x_2 + \cdots + m_{1n}x_n}{1-m_{11}}\right) + m_{i2}x_2 + \cdots + m_{in}x_n.$$

The coefficient of x_k in this expression is given by

$$\boxed{d_{ki} := m_{i1}\frac{m_{1k}}{1-m_{11}} + m_{ik}}$$

which agrees with the formula for c_{ki}.

In summarizing our calculations, let us show how one can see almost instantly that our substitution in Section 2 corresponds to the hyperplane reflection above. Suppose $\mathbf{z}$ is in the span of $\{\mathbf{e}_2, \ldots, \mathbf{e}_n\}$. Then plugging $\mathbf{z}$ in the formula for $\tau_{\mathbf{u}}(\mathbf{z})$, we find that

$$\tau_{\mathbf{u}}(\mathbf{z}) = \mathbf{z} + x(\mathbf{e}_1 - \mathbf{f}_1),$$

where x is our substitution $x = \frac{y}{1-m_{11}}$. But when one reflects on the corresponding map (related to our substitution)

$$\Phi\colon \mathbf{e}_1^{\perp} \to \mathbf{f}_1^{\perp},$$

one sees that

$$\Phi(\mathbf{z}) = \mathbf{z} + x\mathbf{e}_1 - t\mathbf{f}_1,$$

where t is a uniquely determined element of F such that the projection of $\Phi(\mathbf{z})$ on the line through $\mathbf{f}_1$ is 0. Since our image of reflection $\tau_{\mathbf{u}}(\mathbf{z})$ already has this property, we see that $x = t$ and $\tau_{\mathbf{u}}(\mathbf{z}) = \Phi(\mathbf{z})$.

In conclusion, we have seen that our substitution

$$x_1 \longrightarrow \frac{y}{1-m_{11}}$$

amounts to reflecting vectors in the plane orthogonal to the vector $\mathbf{u}$, i.e., sending $\mathbf{z}$ to $\tau_{\mathbf{u}}(\mathbf{z})$. Thus, we have established a "homotopy" between the algebraic and geometric approaches.

5. What is the Witt ring of quadratic forms?

In this section we will define the Witt ring of quadratic forms. As we will see, the Witt cancellation theorem will be the key for constructing the Witt ring. Some terminology is in order. We refer the reader to the excellent books by Lam [8, 9] for a thorough treatment. Other good references on this subject include [4, 5, 15, 19, 22].

Let (V, B) be a quadratic space and let q be the corresponding quadratic form. For simplicity we often drop B and q and denote a quadratic space by V. Recall that a quadratic space (V, B) is said to be *non-degenerate* if the induced map

$$B(\mathbf{v}, -) \colon V \to F$$

is the zero map only when $\mathbf{v} = 0$. It is not hard to show that any quadratic space (V, B) splits as

$$V = V_{non-deg} \perp V_{null}$$

where $V_{non-deg}$ is non-degenerate, and V_{null} is the subspace of V consisting of all vectors in V which are orthogonal to all vectors of V. In particular, the restriction of the bilinear form B on V_{null} is identically 0. Therefore there is no harm in restricting to non-degenerate quadratic spaces.

We say that a non-degenerate quadratic space (V, B) is *isotropic* if there is a non-zero vector $\mathbf{v}$ such that $q(\mathbf{v}) = 0$. It can be shown [24] that every isotropic form contains a hyperbolic plane as a summand, where, by definition, a hyperbolic plane is a two dimensional form that is equivalent to $\langle 1, -1\rangle$. Note that $\langle 1, -1\rangle$ is short for $x_1^2 - x_2^2$. This form q is isotropic as $q(1, 1) = 0$. Thus we see that a non-degenerate quadratic form V is isotropic if and only if V has a hyperbolic plane as a summand.

Now let us consider a non-degenerate quadratic space (V, B). If V is isotropic, then by the above mentioned fact we can write V as

$$V = H_1 \perp V_1,$$

where H_1 is a hyperbolic plane. If V_1 is also isotropic, we can further decompose it as

$$V = H_1 \perp (H_2 \perp V_2),$$

where H_2 is a hyperbolic plane. We proceed in this manner as far as possible, to get a decomposition:

$$V = H_1 \perp H_2 \perp \ldots \perp H_k \perp V_a$$

where H_i are hyperbolic planes and V_a is anisotropic, i.e., a form that is not isotropic. Now here is where Witt cancellation comes into play. The integer k (the number of hyperbolic planes in the above decomposition) is seen to be uniquely determined, using the Witt cancellation theorem. Furthermore, the isometry class of the anisotropic part V_a is uniquely determined, which also follows from the Witt cancellation theorem. In summary, every non-degenerate quadratic space (V, B) admits a unique decomposition called the *Witt decomposition*

$$V = H \perp V_a,$$

where H is a sum of hyperbolic planes and V_a is anisotropic. Two quadratic spaces V and W are said to be *similar* if their anisotropic parts are equivalent. Again, the Witt cancellation theorem ensures that this notion of similarity is well-defined.

With these definitions and concepts, we are now ready to define the Witt ring of quadratic forms $W(F)$ over the field F, which is a central object in the algebraic theory of quadratic forms. The elements of $W(F)$ are the similarity classes of quadratic forms. Since these classes are uniquely represented up to equivalence by anisotropic quadratic forms, we can think of $W(F)$ as the set of equivalence classes of anisotropic quadratic forms. Given two such elements (V, B_V) and (W, B_W), the ring operations of addition and multiplication are defined by

$$\begin{aligned} V+W &:= (V \perp W)_a, \quad \text{and} \\ VW &:= (V \otimes W)_a. \end{aligned}$$

Our tensor space $V \otimes W$ [3] is equipped with a bilinear form B defined by

$$B(v_1 \otimes w_1, v_2 \otimes w_2) = B_V(v_1, v_2) B_W(w_1, w_2).$$

These operations give $W(F)$ the structure of a commutative ring. The zero quadratic space is vacuously anisotropic and is the additive identity for $W(F)$, and the one dimensional form $\langle 1 \rangle$ is the multiplicative identity for $W(F)$. Further details and proofs can be found in [8, Chaper 2, Section 1].

Even though these ideas were all present in Witt's paper [24] from 1937, the algebraic theory of quadratic forms had many years of slow growth before receiving an incredible adolescent spark from the work of Pfister [16, 17] in the 1960's. It has never looked back! In particular, Pfister's work generated intense interest in powers of the so-called *fundamental ideal*, $I(F)$, defined in the next section.

6. Milnor and Bloch-Kato conjectures

Milnor, in his celebrated paper [11] indicated a close and deep connection between three central arithmetic objects: an associated graded ring of the Witt ring $W(F)$ of quadratic forms, the Galois cohomology ring $H^*(F, \mathbb{F}_2)$ of the absolute Galois group, and the reduced Milnor K-theory ring $K_*(F)/2$. In this section we will touch on these topics very briefly to show the reader the connection between the Witt ring and these topics. The interested reader is encouraged to see [13, 11] for more details. The connection between the Witt ring and Galois theory is investigated in [12].

6.1. **Associated graded Witt ring.** Let $I(F)$, or simply I, denote the ideal of $W(F)$ consisting of elements which are represented by even dimensional anisotropic quadratic forms. As an additive subgroup of $W(F)$ this is generated by forms $\langle 1, a \rangle$, and therefore I^n is additively generated by the so-called *n-fold Pfister forms* $\langle 1, a_1 \rangle \langle 1, a_2 \rangle \ldots \langle 1, a_n \rangle$ in the Witt ring; see [9, Page 36]. By convention $I^0 = W(F)$. The associated graded Witt ring is then

$$\bigoplus_{n \geq 0} \frac{I^n}{I^{n+1}} = \frac{W(F)}{I} \oplus \frac{I}{I^2} \oplus \frac{I^2}{I^3} \oplus \ldots.$$

[3] Our reader can think about tensor products as a target of some kind of "universal bilinear form" which one can define precisely. Each element in $V \otimes W$ is a sum $v_1 \otimes w_1 + \cdots + v_k \otimes w_k$ where k is in $\mathbb{N}$, and $v_i \otimes w_i$ is in the image of this bilinear form. Also, the dimension of the tensor product is a product of the dimensions of V and W. For a nice introduction to tensor products see [2, Chapter 10, Section 4].

The three classical invariants of quadratic forms, namely dimension e_0, discriminant e_1, and Clifford invariant e_2, are defined as homomorphisms on the first three summands respectively as follows:

$$e_0\colon \frac{W(F)}{I} \longrightarrow \mathbb{F}_2, \quad e_0([q]) = \dim q \,(\mathrm{mod} 2).$$
$$e_1\colon \frac{I}{I^2} \longrightarrow \frac{F^*}{(F^*)^2}, \quad e_1([q]) = [(-1)^{\frac{n(n-1)}{2}} \det q], \quad \text{where } n = \dim q.$$
$$e_2\colon \frac{I^2}{I^3} \longrightarrow B(F), \quad B(F) \text{ stands for the Brauer group of } F.$$

The definition of the Brauer group is beyond the scope of this article; see [8, Chapter 5, Section 3]. Quadratic forms would be completely classified by these classical invariants if $I^3 = 0$; see [3, Page 374]. However, that is not true in general. So one has to look for higher invariants. Milnor was able to do this by extending these classical invariants into an infinite family of invariants, taking values in the Galois cohomology ring of F. This brings us to the next object of interest.

6.2. **Galois cohomology.** Let $\overline{F}_{sep}$ denote the separable closure of a field F with characteristic not equal to 2. One of the main goals of algebraic number theory and arithmetic geometry is to understand the structure of the absolute Galois group $G_F = \mathrm{Gal}(\overline{F}_{sep}/F)$. To understand this group better one associates a cohomology theory to this group called *Galois cohomology*, which is a graded object:

$$H^*(F, \mathbb{F}_2) = H^0(F, \mathbb{F}_2) \oplus H^1(F, \mathbb{F}_2) \oplus \cdots$$

The first two groups are easy to define. $H^0(F, \mathbb{F}_2) = \mathbb{F}_2$, and $H^1(F, \mathbb{F}_2)$ is the group of continuous homomorphism from G_F to $\mathbb{F}_2$. See [21, 13] for the general definition. $H^*(F, \mathbb{F}_2)$ is also equipped with the structure of commutative ring.

For certain fields F, Milnor proved [11] the existence of a well-defined map

$$e\colon \oplus_{n\geq 0} I^n/I^{n+1} \to \oplus_{n\geq 0} H^n(F, \mathbb{F}_2),$$

and he showed that it is an isomorphism. In [11] he asked if the same is true in general. For an arbitrary F, even showing that e is a well-defined map is very hard. This problem, of showing that e is a well-defined map and that it is an isomorphism for all F, is known as the *Milnor conjecture on quadratic forms*. This problem has fascinated mathematicians and was eventually settled affirmatively in [14].

6.3. **Reduced Milnor K-theory.** The ring structure on both the domain and the target of the map e is mysterious. To explain this ring structure Milnor constructed a third object, now called reduced Milnor K-theory $K_*(F)/2$, whose ring structure is far more transparent. Let F^* be the multiplicative group of non-zero elements in F. The tensor algebra $T(F^*)$ is a graded algebra defined by

$$T(F^*) := \mathbb{Z} \bigoplus F^* \bigoplus (F^* \otimes F^*) \bigoplus (F^* \otimes F^* \otimes F^*) \bigoplus \cdots .$$

The reduced Milnor K-theory $K_*(F)/2$ is the tensor algebra $T(F^*)$ modulo the two-sided ideal $\langle a \otimes b \,|\, a + b = 1, \;\; a, b \in F^* \rangle$ reduced modulo 2. That is,

$$K_*(F)/2 := \frac{T(F^*)}{\langle a \otimes b \,|\, a + b = 1, \;\; a, b \in F^* \rangle} \otimes \mathbb{F}_2$$

FIGURE 3. Left to Right : Andrei Suslin, Alexander Merkurjev, John Milnor, Vladimir Voevodsky, and Tsit-Yuen Lam, celebrating the 80th birthday party of the algebraic theory of quadratic forms.

Milnor defined two families of maps ν and η shown in the triangle below [4]. Showing that all maps in this triangle are isomorphisms was a major problem in the field and it went under the name of *The Milnor conjectures.* The map η was shown to be an isomorphism by Voevodsky, for which he won the Fields medal in 2002. As mentioned earlier, e was shown to be an isomorphism in [14], building upon the work of Voevodsky. These theorems are among the most powerful results in the algebraic theory of quadratic forms. For further details and proofs of these theorems see [10], [14] and [23].

The Milnor triangle is the triangle connecting quadratic forms, Galois cohomology and the reduced Milnor K-theory:

$$\begin{tikzcd} & K_*(F)/2 \arrow[dl, "\nu"'] \arrow[dr, "\eta"] & \\ \oplus_{n\geq 0} I^n/I^{n+1} \arrow[rr, "e"'] & & \oplus_{n\geq 0} H^n(F,\mathbb{F}_2) \end{tikzcd}$$

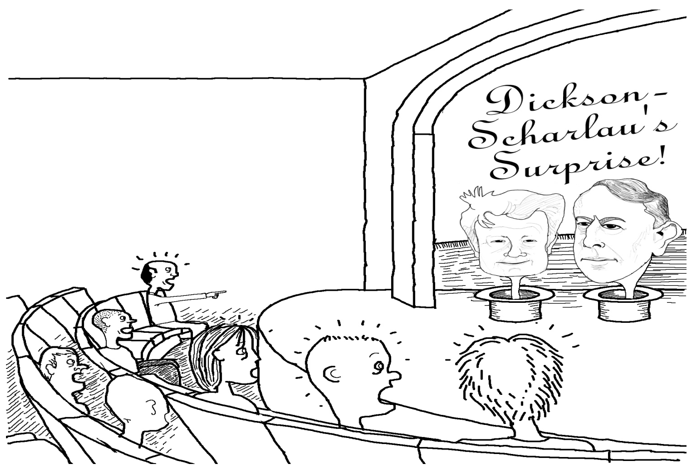
For odd primes p, a similar isomorphism was conjectured by Bloch and Kato, between the reduced Milnor K-theory $K_*(F)/p$ and the Galois cohomology ring $H^*(F,\mathbb{F}_p)$ when the field F contains a primitive p-th root of unity. This *Bloch-Kato conjecture* was proved in 2010 by Rost and Voevodsky, with a contribution from Weibel. The interested reader can consult [25, 18, 23]. The background required for these deep, very recent papers is quite extensive, so the ambitious reader will no doubt have lots of fun delving into many *extra* references, including those found in the references of the papers we cite.

7. Dickson-Scharlau's surprise

After essentially completing our article we kept searching for historical references on quadratic forms. We were astounded to find a conference proceeding article [20] by W. Scharlau entitled, "*On the history of the algebraic theory of quadratic forms.*" Scharlau explains that

[4]The map η was defined using a lemma of Bass and Tate [11, Lemma 6.1].

the algebraic theory of quadratic forms could have been born 30 years earlier! Namely, in 1907 L. Dickson published a paper [1] in which he proved a number of results on quadratic forms including the cancellation theorem which Witt proved independently 30 years later in 1937. In fact, Scharlau writes: "... *It seems that Dickson's paper went completely unnoticed; I could not find a single reference to it in the literature. However, one must admit that this paper – like most of Dickson's work – is not very pleasant to read... Nevertheless, I believe that, as far as Witt's theorem and related questions are concerned, some credit should be given to Dickson.*" Therefore, one might say that the algebraic theory of quadratic forms was *conceived* in 1907, but wasn't *born* until 1937.



8. Epilogue: The 80th birthday party in the Elysian field

If only Emmy Noether, her graduate student Ernst Witt and Leonard Dickson were here to help us celebrate this birthday. We cannot know exactly what they would say, but we may still imagine the party that is going on in the Elysian field of mathematical giants. Emmy Noether is running around full of energy, leading a lively and challenging mathematical discussion. One could hardly believe that she was born nearly 135 years ago! There are other mathematicians including David Hilbert, who are taking interest in the discussions at the party.

Noether: This Bloch-Kato conjecture is finally solved, and its proof is just beautiful. We have come so far since the early days of cyclic algebras and cross-products. Ah! Dickson, what a shame that your brilliant paper on quadratic forms from 1907 did not get the attention it deserved. Just imagine how much further we would have come had people studied it from the very beginning. Please rewrite it, with more emphasis on the concepts to illuminate the calculations.

Dickson: Rewrite a brilliant paper? Wow–you are just as strict as I had heard *and by the way–it has been over a century since I wrote that paper!* I hardly remember it now. I do finally have some free time on my hands to recall those ideas. In any case, it may well be a blessing that it was not popular at the beginning. Who knows if Witt would have developed his elegant geometrical approach if everyone knew about my original paper?

Witt: Oh, I know–and yes, I would have.

Noether: My earnest boy, you certainly do know how to provide a short answer. And I have missed your wit. But seriously, we should spend the next several meetings working on this Bloch-Kato conjecture. Although the proof just provided by Rost and Voevodsky is truly amazing, we can always strive towards a more elementary proof in the hopes of making it less mysterious. Perhaps we should write a book?

Dickson: Indeed this is a worthwhile and tough challenge. I will study these proofs and search for the underlying algebraic structures.

Witt: I too would love to work on this. It does indeed seem a bit mysterious that the statements of the Milnor conjecture and the Bloch-Kato conjecture can be formulated using quadratic forms, group cohomology and field theory, yet their current proofs require so much more material. We must think about what this means for Galois theory.

David Hilbert has been quietly listening to this conversation, pacing back and forth. He has something to say:

Hilbert: Wir müssen wissen. Wir werden wissen. [5]

Acknowledgements: The authors are very grateful to the remarkably gifted artist Matthew Teigen for his excellent artistic illustrations in this paper. We would like to thank David Eisenbud, Margaret Jane Kidnie, John Labute, Claude Levesque, Alexander Merkurjev, John Milnor, Raman Parimala, Andrew Ranicki, Balasubramanian Sury, Stefan Tohaneanu, Ravi Vakil and Charles Weibel for their encouragement, help with the exposition, and nice welcome of the preliminary version of our paper.

References

[1] L. E. Dickson, On quadratic forms in a general field. *Bull. Amer. Math. Soc.* 14 (1907), no. 3, 108-115.

[2] D. S. Dummit and R. M. Foote, *Abstract algebra.* Third edition. John Wiley & Sons, Inc., Hoboken, NJ, 2004.

[3] R. Elman and T. Y. Lam, Classification theorems for quadratic forms over fields. *Comment. Math. Helv.* 49 (1974), 373-381.

[4] R. Elman, N. Karpenko, and A. Merkurjev, The algebraic and geometric theory of quadratic forms *American Mathematical Society Colloquium Publications*, Vol 60 American Mathematical Society, Providence, RI, 2008.

[5] L. J. Gerstein, *Basic quadratic forms.* Graduate Studies in Mathematics, 90. American Mathematical Society, Providence, RI, 2008. xiv+255 pp.

[6] C. Kimberling, Emmy Noether and Her Influence, in James W. Brewer and Martha K. Smith, Emmy Noether: A Tribute to Her Life and Work, New York: Marcel Dekker, 1981.

[7] I. Kersten, Biography of Ernst Witt (1911-1991), pp. 155-171 Quadratic forms and their applications (Dublin, 1999), 229-259, *Contemp. Math.*, 272, Amer. Math. Soc., Providence, RI, 2000.

[8] T. Y. Lam, *Introduction to quadratic forms over fields,* American Mathematical Society, Graduate Studies in Mathematics, 67 (2005), xxii+550.

[9] T. Y. Lam, *The algebraic theory of quadratic forms*, Benjamin, New York, 1973.

[10] A. Merkurjev, Developments in Algebraic K-theory and Quadratic Forms after the work of Milnor. *John Milnor's Collected papers* Vol.5, Ed. Hyman Bass and T.-Y.Lam, AMS (2010), 399-418.

[5] We must know. We will know.

[11] J. Milnor, Algebraic K-theory and quadratic forms. *Invent. Math.* 9 (1970) 318-344.

[12] J. Mináč and M. Spira, Witt rings and Galois groups. *Ann. of Math.* (2) 144 (1996), no. 1, 35-60.

[13] J. Neukirch, A. Schmidt, and K. Wingberg, *Cohomology of number fields.* Second edition. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 323. Springer-Verlag, Berlin, 2008. xvi+825 pp.

[14] D. Orlov, A. Vishik, and V. Voevodsky, An exact sequence for $K_*^M/2$ with applications to quadratic forms. *Ann. of Math.* (2) 165 (2007), no. 1, 1-13.

[15] A. Pfister, Quadratic forms with applications to algebraic geometry and topology. London Mathematical Society Lecture Note Series, 217. *Cambridge University Press,* Cambridge, 1995. viii+179 pp.

[16] A. Pfister, Quadratische Formen in beliebigen Körpern. (German) *Invent. Math.* 1 1966 116-132.

[17] A. Pfister, Zur Darstellung von -1 als Summe von Quadraten in einem Körpern. (German). *J. London Math. Soc.* 40 (1965) 159-165.

[18] M. Rost, On the basic correspondence of a splitting variety. (September-November 2006) http://www.mathematik.uni-bielefeld.de/rost/

[19] W. Scharlau, *Quadratic and Hermitian forms.* Grundlehren der Mathematischen Wissenschaften, 270 Springer-Verlag, Berlin (1985) x+421.

[20] W. Scharlau, On the history of the algebraic theory of quadratic forms. Quadratic forms and their applications (Dublin, 1999), 229-259, *Contemp. Math.,* 272, Amer. Math. Soc., Providence, RI, 2000.

[21] J. P. Serre, *Galois cohomology.* Translated from the French by Patrick Ion and revised by the author. Corrected reprint of the 1997 English edition. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2002. x+210 pp.

[22] K. Szymiczek, *Bilinear Algebra : An Introduction to the Algebraic Theory of Quadratic Forms*, Gordon and Breach Science Publisher. Vol.7 , (1997).

[23] C. Weibel, The norm residue theorem. *J. Topology*, Vol.2, 346-372 (2009).

[24] E. Witt, Theorie der quadratischen Formen in beliebigen Körpern. *J. Reine Angew. Math.*, 176 (1937), 31-44.

[25] V. Voevodsky, On motivic cohomology with $\mathbb{Z}/l$- coefficients, *Ann. of Math.* (2) 174 (2011), no. 1, 401-438.

Department of Mathematics, Illinois State University, Normal, IL 61761, USA
E-mail address: schebol@ilstu.edu

Department of Mathematics, Norwich University, Northfield, VT 05663, USA
E-mail address: dmcquill@norwich.edu

Department of Mathematics, University of Western Ontario, London, ON N6A 5B7, Canada
E-mail address: minac@uwo.ca