Efficient Zero-Knowledge Arguments in the Discrete Log Setting, Revisited

Max Hoffmann
max.hoffmann@rub.de
Ruhr-University Bochum
Horst Görtz Institute for IT-Security
Bochum, Germany

Michael Klooß
michael.klooss@kit.edu
Karlsruhe Institute of Technology
Department of Informatics
Karlsruhe, Germany

Andy Rupp
andy.rupp@kit.edu
Karlsruhe Institute of Technology
Department of Informatics
Karlsruhe, Germany

ABSTRACT

Zero-knowledge arguments have become practical, and widely used, especially in the world of Blockchain, for example in Zcash.

This work revisits zero-knowledge proofs in the discrete logarithm setting. First, we identify and carve out basic techniques (partly being used implicitly before) to optimise proofs in this setting. In particular, the *linear combination of protocols* is a useful tool to obtain zero-knowledge and/or reduce communication. With these techniques, we are able to devise zero-knowledge variants of the logarithmic communication arguments by Bootle et al. (EUROCRYPT '16) and Bünz et al. (S&P '18) thereby introducing almost no overhead. We then construct a conceptually simple commit-and-prove argument for satisfiability of a set of *quadratic equations*. Unlike previous work, we are not restricted to rank 1 constraint systems (R1CS). This is, to the best of our knowledge, the first work demonstrating that general quadratic constraints, not just R1CS, are a natural relation in the dlog (or ideal linear commitment) setting. This enables new possibilities for optimisation, as, e.g., *any* degree n^2 polynomial $f(X)$ can now be "evaluated" with at most $2n$ quadratic constraints.

Our protocols are modular. We easily construct an efficient, logarithmic size shuffle proof, which can be used in electronic voting.

Additionally, we take a closer look at quantitative security measures, e.g. the efficiency of an extractor. We formalise *short-circuit extraction*, which allows us to give tighter bounds on the efficiency of an extractor.

CCS CONCEPTS

• **Theory of computation → Communication complexity**; **Interactive proof systems**; *Cryptographic protocols.*

KEYWORDS

zero-knowledge, argument system, quadratic equations, arithmetic circuit satisfiability, discrete logarithm assumption,

CCS '19, November 11–15, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6747-9/19/11…$15.00
https://doi.org/10.1145/3319535.3354251

ACM Reference Format:
Max Hoffmann, Michael Klooß, and Andy Rupp. 2019. Efficient Zero-Knowledge Arguments in the Discrete Log Setting, Revisited. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3319535.3354251

1 INTRODUCTION

Zero-knowledge arguments (of knowledge) (ZKAoK) allow a party $\mathcal{P}$, the prover, to convince another party $\mathcal{V}$, the verifier, of the truth of a statement (and knowledge of a witness) without revealing any other information. For example, one may prove knowledge of a valid signature on some message, without revealing the signature. The ability to ensure *correctness* without compromising *privacy* makes zero-knowledge arguments a powerful tool, which is ubiquitous in theory and application of cryptography. Since the first *practical* construction of succinct non-interactive arguments of knowledge (SNARK) [21], and their application to Blockchain and related areas, research in theory and applications of efficient ZKAoKs has progressed significantly, see the works [3, 7, 10, 14, 18, 21–23, 39] to name a few. Arguably, zero-knowledge proofs have become *practical* for many applications. As efficiency improved and demand for privacy increased, possible use cases and applications have grown explosively.

In this paper, we revisit a line of works [10, 13, 25] in the discrete logarithms setting. From an abstract point of view, in terms of [29], one part of our work is in the world of ideal linear commitments (ILC). That is, our verifier can do "matrix-vector queries" on a committed value $\boldsymbol{w}$, e.g. request an opening for a matrix-vector product $\boldsymbol{\Gamma w}$. A priori, this is more powerful than other settings like PCP or IOP, where the verifier's queries are restricted to point or inner-product queries [29]. Nonetheless, the ILC-arguments in [10, 13, 25] only work for the language R1CS "natively", which is also covered by more restricted verifiers. We show that with ILC, one can directly handle *systems of quadratic equations*, of which R1CS is a special case. This broadens possible optimisation from (already used [1]) R1CS-friendly to quadratics-friendly cryptography. Yet, even for R1CS, our performance improves upon Bulletproofs [13].

Another part of this work treats proofs of knowledge of preimages of group homomorphisms. For example, proving knowledge of the decryption of an ElGamal ciphertext. This does not fit into the ILC setting. Hence we do not use the ILC abstractions in this work. Combining this with our proofs for quadratic equations is efficient. Thus one can generically construct primitives such as shuffle proofs, which are important for electronic voting.

1.1 Basic techniques

We identify and present basic design principles underlying most efficient zero-knowledge arguments in the group setting.

In the following, we use implicit representation for group elements, see Section 2. Let us recall (a slight variant of) the standard Σ-Protocol (Σ_{std}) for proving knowledge of a preimage $\boldsymbol{w}$ for $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ for $[\boldsymbol{A}] \in \mathbb{G}^{m\times n}$. This proof covers a large class of statements, including dlog relations, knowing the opening of a commitment, etc. The protocol works as follows:

- Prover: Pick $\boldsymbol{r} \leftarrow \mathbb{F}_p^n$, let $[\boldsymbol{a}] := [\boldsymbol{A}]\boldsymbol{r}$, send $[\boldsymbol{a}]$.
- Verifier: Pick and send $\boldsymbol{x} = (x_1, x_2) \leftarrow \mathbb{F}_p^2$ (with $x_2 \neq 0$).
- Prover: Send $\boldsymbol{z} := x_1\boldsymbol{w} + x_2\boldsymbol{r}$.
- Verifier: Accept iff $[\boldsymbol{A}]\boldsymbol{z} = x_1[\boldsymbol{t}] + x_2[\boldsymbol{a}]$.

Intuitively, this is zero-knowledge since $\boldsymbol{r}$ completely masks $\boldsymbol{w}$ in $\boldsymbol{z} = x_1\boldsymbol{w} + x_2\boldsymbol{r}$ (since $x_2 \neq 0$), and finding $\boldsymbol{r}$ from $[\boldsymbol{a}]$ is hard. It is extractable, since two linearly independent challenges $\boldsymbol{x}_1, \boldsymbol{x}_2$ with answers $\boldsymbol{z}_1, \boldsymbol{z}_2$ (for fixed $[\boldsymbol{a}]$) allow to reconstruct $\boldsymbol{w}, \boldsymbol{r}$. But Protocol Σ_{std} is not particularly communication-efficient, as it sends the full masked witness $\boldsymbol{z} \in \mathbb{F}_p^n$ as well as $[\boldsymbol{a}] \in \mathbb{G}^m$. Using probabilistic verification, one can often improve this.

1.1.1 Probabilistic verification. The underpinning of *efficient* arguments of knowledge (without zero-knowledge) is probabilistic verification of the claim. For instance, instead of verifying $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ directly, the verifier could send a random $y \leftarrow \mathbb{F}_p$. Both parties compute $\boldsymbol{y} = (y^i)_i \in \mathbb{F}_p^m$ and verify $[\widehat{\boldsymbol{A}}]\boldsymbol{w} = [\widehat{t}]$ for $[\widehat{\boldsymbol{A}}] = \boldsymbol{y}^\top[\boldsymbol{A}] \in \mathbb{G}^{1\times n}$ and $[\widehat{t}] = \boldsymbol{y}^\top[\boldsymbol{t}] \in \mathbb{G}$ instead. This would result in a communication complexity which is independent of m as $[\widehat{\boldsymbol{a}}] = [\widehat{\boldsymbol{A}}]\boldsymbol{r} \in \mathbb{G}$.

Not all probabilistic verifications are alike. To work well with zero-knowledge, we need "suitable" verification procedures, so that techniques to attain zero-knowledge can be applied. This essentially means that the verification should be *linear*, i.e. all tested equations should be linear. (Abstract groups only allow linear operations anyway.)

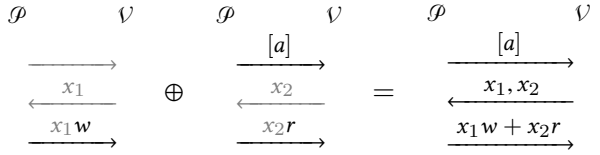

Figure 1: Linear Combination of Protocols. *Left:* The trivial proof of knowledge: Send the witness. *Middle:* Send a random statement. Then send the witness. *Grayed out:* Terms for linear combination. *Right:* The linear combination with verifier's randomness.

1.1.2 Linear combinations of protocols. A core insight for achieving zero-knowledge (and reducing communication) in our setting *efficiently* is that protocols can often be linearly combined, see Fig. 1 for an illustration. This exploits the *linearity* of the computations and checks of verifier and prover in each round. By running an "umasked *non*-zero-knowledge argument" (Fig. 1, left) and linearly combining it with an argument for a "masking randomness" (middle), one can achieve zero-knowledge (right). All of our zero-knowledge compilations rely on this strategy. We typically consider *random* linear combinations of protocols, where the verifier picks the randomness (x_1, x_2 in Fig. 1), as this often achieves extractability. In fact, this kind of linear combination recovers the batch proofs of [38], see Appendix C. Non-randomised linear combinations are also used, e.g. Protocol B.1, or [13].

1.1.3 Uniform(-or-unique) responses. In our setting, for simulation it is typically enough to ensure that the prover's messages are distributed uniformly at random. More concretely, the responses should be either uniformly distributed (conditioned on all *later* messages, *not* previous messages), such as $\boldsymbol{z}$ in Protocol Σ_{std}. Or they should be uniquely determined and *efficiently* computable from the challenges and all *later* messages, such as $[\boldsymbol{a}]$ in Protocol Σ_{std}. This allows to construct a trivial simulator, which constructs the transcript *in reverse*: Starting with the final messages, and working its way towards the beginning, the simulator picks the uniformly distributed messages itself, and computes the uniquely determined ones. All simulators in this paper work like this.

1.1.4 Kernels and redundancy. Many interesting statements are non-linear. For example, for polynomial commitments [12], we want to show that $[\boldsymbol{c}] \in \mathbb{G}^m$ is a commitment to a polynomial $f \in \mathbb{F}_p[X]$ (of degree at most $d-1$) and $f(x) = t$, where $x \in \mathbb{F}_p$ is a random challenge. Naively, one commits to the coefficients of the polynomial with monomial basis X^i for $i = 0, \ldots, n-1$. Suppose we have a (linear) protocol which proves $f(x) = t$. We could hope that running a random linear combination as in Fig. 1 should give us uniform-or-unique responses (and hence zero-knowledge). However, we are in a predicament: For random $g \in \mathbb{F}_p[X]$, we have $(f+g)(x) \neq f(x)$ and thus we have to let $\mathcal{V}$ know $y = g(x)$ somehow. To ensure the prover does not send arbitrary y, we have to rely on a proof again! But if this proof leaks information we cannot use it to randomise the response. We can escape by having a way to randomise *without changing the statement*. In other words, we need some g with $g(x) = 0$ for all $x \in \mathbb{F}_p$. Clearly, that means $g = 0$, and there's nothing random anymore! Another dead end.

One solution is to add *redundancy*, which does not "influence" soundness: Here, we artificially create a non-trivial kernel of the "evaluate at x"-map. We can do so by representing $f(X)$ as $\sum_i(\alpha_i + \beta_i)X^i$ and commit to all α_i and β_i. Now we can mask with $g(X)$ where $\alpha_i \leftarrow \mathbb{F}_p$ and $\beta_i = -\alpha_i$. Thus, we successfully injected randomness into the response. Generally, adding just enough redundancy to achieve uniformly random responses is our goal.

1.1.5 Composition of arguments systems. By committing to and then sharing intermediate results in multiple argument systems, one can combine the most efficient arguments for each task.

Example 1.1. In our logarithmic communication zero-knowledge inner product argument $\text{IPA}_{\text{almZK}}$ for $\exists \boldsymbol{x}, \boldsymbol{y}\colon \langle \boldsymbol{x}, \boldsymbol{y}\rangle = t$, we randomise as $\langle \boldsymbol{x} + \boldsymbol{r}, \boldsymbol{y} + \boldsymbol{s}\rangle = t$ so that $\langle \boldsymbol{r}, \boldsymbol{y}\rangle = \langle \boldsymbol{r}, \boldsymbol{s}\rangle = \langle \boldsymbol{x}, \boldsymbol{s}\rangle = 0$ with only logarithmically many (specially chosen) random components in $\boldsymbol{r}, \boldsymbol{s}$. This is an application of the "redundancy/kernel" technique. The "uniform-or-unique" guideline ensures that it is enough that each response is random. Hence a logarithmic number of (well-chosen) random components in $\boldsymbol{r}$, $\boldsymbol{s}$ does suffice.

On the other hand, our logarithmic communication linear map preimage argument $\mathsf{LMPA}_{\mathsf{ZK}}$ for $\exists \boldsymbol{w}\colon [\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ uses a linear combination of a *non*-zero-knowledge argument for $[\boldsymbol{A}]$, plus a similar argument for a *different* $[\boldsymbol{A}]$ and $[\boldsymbol{t}]$ (of the same size). Finally, for our logarithmic communication shuffle argument Π_{shuffle} (Appendix D), we compose $\mathsf{QESA}_{\mathsf{ZK}}$ (our quadratic equation argument) and $\mathsf{LMPA}_{\mathsf{ZK}}$ by sharing a commitment to the witness.

1.2 Contribution

To the best of our knowledge, there is no work which presents these techniques, in particular linear combination of protocols, as *unifying* guidelines. Implicitly, these techniques are used in many works [10, 12, 13, 25, 38]. We follow the above guidelines for constructing and explaining our zero-knowledge arguments.

See the full version [31] for protocol diagrams.

1.2.1 Linear map preimage argument (LMPA). We give in two steps an argument for $\exists \boldsymbol{w}\colon [\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ for $[\boldsymbol{A}] \in \mathbb{G}^{m\times n}$ with communication $O(\log(n))$. The idea is to first use batch verification. Essentially, $\mathsf{LMPA}_{\mathsf{batch}}$ multiplies the equation with a random vector $\boldsymbol{y} \in \mathbb{F}_p^m$ from the left to obtain $[\widehat{\boldsymbol{A}}] = \boldsymbol{y}^\top[\boldsymbol{A}] \in \mathbb{G}^{1\times n}$ and $[\hat{t}] = \boldsymbol{y}^\top[\boldsymbol{t}] \in \mathbb{G}$. Thus, communication is independent of m. Now, we prove $\exists \boldsymbol{w}\colon [\widehat{\boldsymbol{A}}]\boldsymbol{w} = [\hat{t}]$ using $\mathsf{LMPA}_{\mathsf{ZK}}$, which is derived from [10]. It is enhanced with zero-knowledge for overhead which is constant w.r.t. communication and logarithmic w.r.t. computation (in n).

1.2.2 Quadratic equation commit-and-prove. First of all, we derive a(n almost) zero-knowledge inner product argument $\mathsf{IPA}_{\mathsf{almZK}}$ from [10, 13], again with constant communication and logarithmic computational overhead compared to [10, 13]. From $\mathsf{IPA}_{\mathsf{almZK}}$ we obtain an argument for proving $\exists \boldsymbol{w}\colon \forall i\colon \langle \boldsymbol{w}, \boldsymbol{\Gamma}_i \boldsymbol{w}\rangle = 0$, where $\boldsymbol{\Gamma}_i \in \mathbb{F}_p^{n\times n}$ and $\boldsymbol{w}$ is committed to. For efficiency, we carry out a batch proof, i.e. we prove $\langle \boldsymbol{w}, \boldsymbol{\Gamma}\boldsymbol{w}\rangle$ with $\boldsymbol{\Gamma} \coloneqq \sum_i r_i \boldsymbol{\Gamma}_i$ for random $r_i \in \mathbb{F}_p$. The resulting argument, $\mathsf{QESA}_{\mathsf{ZK}}$, is "adaptive commit-and-prove", i.e. the statement $\boldsymbol{\Gamma}_i$ may be chosen after the commitment to $\boldsymbol{w}$.

The commit-and-prove system $\mathsf{QESA}_{\mathsf{ZK}}$ is conceptually simple and can be efficiently combined with other arguments. We leave as an open question whether its strategies can be adapted by linear IOPs or whether they are unique to ILC.

1.2.3 Sets of quadratic equations. Being able to prove arbitrary quadratic equations instead of R1CS equations, i.e. equations of the form $(\sum a_i x_i)(\sum b_i x_i) + \sum c_i x_i = 0$, gives much flexibility. To the best of our knowledge, expressing the quadratic equation $\langle \boldsymbol{x}, \boldsymbol{x}\rangle = \sum x_i^2 = t$ as R1CS requires n equations: $y_i = x_i^2$ $(i = 1, \ldots, n-1)$ and $x_n^2 = t - \sum_i y_i$, where y_i are additionally introduced variables. Requiring n equations is surprising for [10, 13] which build on an *inner product argument*. Obviously, $\mathsf{QESA}_{\mathsf{ZK}}$ needs one (quadratic) equation to express $\langle \boldsymbol{x}, \boldsymbol{x}\rangle = t$.

Using general quadratic equations, one can evaluate any (univariate) polynomial $f(X) = \sum_{i=0}^{d^2-1} a_i X^i$ of degree $d^2 - 1$ with $2d$ equations and intermediate variables. Concretely, let $y_i = x^i = y_{i-1}x$, $z_i = x^{di} = z_1 z_{i-1}$, for $i = 2, \ldots d-1$ and $z_1 = y_{d-1}x$ and $z_0 = 1$. Then $f(x) = \sum_{i,j=0}^{d} a_{i+jd} y_i z_j$. This can speed up "table lookups", which are typically encoded as polynomial evaluation.

For S(N)ARK-friendly cryptography [33], supporting quadratic equations is very useful. Matrix-vector multiplications are efficient even when both matrix and vector are *secret*. "Embedding" an elliptic curve (see Jubjub [1]), is also more efficient than for R1CS. For general point addition in a (twisted) Edwards curve, we need 5 instead of 8 constraints per bit.

Similar to SNARK- and MPC-friendly cryptography, quadratics-friendly cryptography may enable significant speedups. A prime candidate is multi-variate quadratic cryptography, where suitably adapted schemes might integrate very well with our proof system.

1.2.4 Correctness of a shuffle. By instantiating the shuffle proof of Bayer and Groth [5] with $\mathsf{LMPA}_{\mathsf{ZK}}$ and $\mathsf{QESA}_{\mathsf{ZK}}$ as sub-protocols, we obtain an argument Π_{shuffle} for correctness of a shuffle. To the best of our knowledge, this is the first efficient argument with proof size $O(\log(N))$. Our computational efficiency is comparable to [5], which has proof size $O(\sqrt{N})$. More concretely, we (very roughly) estimate at worst 2–3× the computation.

1.2.5 Knowledge errors, tightness and short-circuit extraction. From a quantitative perspective, our notion of testing distributions and their soundness errors, are useful to separate study of knowledge errors and extraction in the setting of special soundness. Testing distributions have associated soundness errors, which (up to technical difficulties we state as open problems) translate to knowledge errors of the protocol. Explicit knowledge errors achieve tuneable levels of soundness, e.g. 2^{-120} instead of 2^{-256}, which impacts runtime positively.

Short-circuit extraction. We give a definition of *short-circuit extraction*. This treats extraction assertions such as "Ext either finds a witness or it solves a hard problem". It formalises the (common) behaviour of an extractor to either find a witness with *few* transcripts, or solve the hard problem (e.g. equivocating a commitment). Without distinguishing these cases, the bounds on the necessary number of transcripts for extraction is much higher. For example, we show that the extractor for the $\mathsf{LMPA}_{\mathsf{ZK}}$ and $\mathsf{IPA}_{\mathsf{almZK}}$ (and also [10, 13]) needs a tree of transcripts of size $O(\log(n)n)$ in the worst case. For $\mathsf{QESA}_{\mathsf{ZK}}$, extracting a proof for N quadratic equations in n variables requires $O(\log(n)nN)$ transcripts. The extractor in [13] needs $O(n^3N)$ transcripts, which for $n, N \approx 2^{16}$ implies a security loss of $\approx 2^{64}$ instead of $\approx 2^{34}$. This also opens the possibility for using *small exponents* as challenges, further improving our argument systems performance. Note that, due to their structure, Bulletproofs [13] are not well-suited for small exponents.

In the full version [31], we give a conjectured relation between communication efficiency and extraction efficiency, which implies that extraction from $O(\frac{n}{\log(n)})$ transcripts would be optimal. We also elaborate on a loophole in above security estimates, namely how to *efficiently obtain* the transcripts.

1.2.6 Dual testing distributions. Dual testing distributions are a technical tool which allow us to sample a "new" commitment key from a given one, such that knowledge (e.g. commitment opening) cannot be transferred. This turns out to be more communication efficient than letting the verifier send a new commitment key. To the best of our knowledge, this is a new technique.

1.2.7 Theoretical comparison to [13] and improved inner product argument (IPA). In Table 1, we compare our argument systems with

related work in the group setting. In Table 2, we give precise efficiency measures for $\mathsf{LMPA}_{\mathsf{ZK}}$ and $\mathsf{QESA}_{\mathsf{ZK}}$. In any case, $n = |\boldsymbol{w}|$ is the size of the witness $\boldsymbol{w} \in \mathbb{F}_p^n$. Since it is statement dependent, we ignore that QE is more powerful than R1CS, possibly allowing smaller witness size (as seen in the example $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = t$ above). In Table 2, we omit the verifier's computation, since after optimisations [13], both are almost identical. For the prover, we do not optimise (e.g. we use no multi-exponentiations), and are not aware of non-generic optimisation. Much of our theoretical improvement is due to our improvements to the IPA. Using [13] with our IPA yields identical asymptotics. Even then, $\mathsf{QESA}_{\mathsf{ZK}}$ covers *general* quadratic equations, while Bulletproofs [13] which only cover R1CS.

	Setup	Ass.	Moves	Comm.	Comp. $\mathcal{P}$	Comp. $\mathcal{V}$	Nat. $\mathcal{R}$
[21]	✗	KoE	1	$O(1)$	$O(n)$	$\leq \lvert\boldsymbol{w}\rvert$	R1CS
[13]	✓	dlog	$O(\log(n))$	$O(\log(n))$	$O(n)$	$\lvert\boldsymbol{w}\rvert$	R1CS
This	✓	dlog	$O(\log(n))$	$O(\log(n))$	$O(n)$	$\lvert\boldsymbol{w}\rvert$	QE

Table 1: Comparison of [21], Bulletpoofs [13] and this work. Setup: Common *random* string sufficient? Security Ass(umption): Knowledge of exponents (KoE); Hardness of dlogs. Moves: The number of messages sent. Comm(unication) in group elements. Comp(upation) in group exponentiations. Nat(ive) relation $\mathcal{R}$.

	Comm. $\mathbb{G}$	Comm. $\mathbb{F}_p$	Comp. $\mathcal{P}$	$\mathcal{R}$
$\mathsf{LMPA}_{\mathsf{ZK}}$	$\approx 4m\log_2(n)$	$4m$	$\approx 4mn$	LMP
$\mathsf{QESA}_{\mathsf{ZK}}$	$2\lceil\log(n+2)\rceil + 3$	2	$\approx 8n$	QE
[13]	$2\lceil\log(n)\rceil + 8$	5	$\approx 12n$	R1CS

Table 2: Estimates in terms of the group elements and exponentiations. By "$\approx f$" we denote $f + O(\log(f))$.

1.2.8 Comparison with other proof systems. It is hard to make a fair comparison of proof systems. There are many relevant parameters, such as setup, assumptions, quantum resistance, native languages, etc. beyond mere proof size and performance. See Section 1.3 for a high-level discussion. To draw (non-trivial) conclusions from comparisons on an implementation level, one should compare fully optimised implementations. Thus, we restrict ourselves to a comparison with Bulletproofs (which we reimplemented with the same optimisation level as our proof systems). For somewhat concrete numbers regarding (implementation) performance, as well as other factors relevant to the comparison of proof systems, we refer to [7, Figure 2]. Our proof systems are similar enough to Bulletproofs for these comparisons to still hold.

1.2.9 Implementation. In Section 5, we compare our implementations of (aggregate) range proofs. The theoretical prediction of $0.75\times$ prover runtime compared to [13] is close to measurements, which suggest $0.7\times$. Using 140bit exponents, we experimentally attain $\approx 0.63\times$ compared to [13] on the same platform. As an important remark, we compare the dedicated range proofs of [13] with our generic instantiation of $\mathsf{QESA}_{\mathsf{ZK}}$.

1.3 Related work

Due to space constraints, we only elaborate on the most important concepts and related work. We refer to [29] for an overview and a general taxonomy.

The dlog setting and ILC. Very closely related works are [10, 12, 13, 25], which are efficient proofs in the dlog setting. Many zero-knowledge proofs in the group setting are instantiations of [16, 36]. The possibilities of our setting, namely ability to apply linear transformations to a committed witness has been abstracted in the ideal linear commitment model [11]. (Our techniques for $\mathsf{QESA}_{\mathsf{ZK}}$ are amenable to ILC.)

Knowledge assumptions. Another line of work [9, 18, 21, 26, 27, 35] gives non-interactive arguments using knowledge of exponent assumptions. They attain *constant size* proofs for arithmetic circuits and sublinear verification costs, but require a *trusted setup*.

PCPs, IOPs, MPC-in-the-head. Techniques, such as probabilistically checkable proofs (PCP), MPC-in-the-head [32], interactive oracle proofs (IOP) and more, construct efficient zero-knowledge proofs without relying on public key primitives. The possible performance gain (and quantum resistance) is interesting from a practical point of view. There is much progress on improving these techniques [3, 7, 14, 22, 39], which until recently suffered from relatively large proof size or unacceptable constants. In [7], Ben-Sasson et al. present a logarithmic communication IOP for R1CS, which, by avoiding public key primitives, likely outperforms our $\mathsf{QESA}_{\mathsf{ZK}}$ by order(s) of magnitude. Still, according to [7], proof sizes for R1CS statements of size $N = 10^6$ are about 130kb whereas our proofs, like Bulletproofs, stay well below 2kb. For combining proofs in the "symmetric key" setting with efficient proofs for "public key" algebraic statements, [2] can be used. Our proofs can be directly combined with algebraic statements over the same group $\mathbb{G}$.

2 PRELIMINARIES

For a set S and probability distribution χ on S we write $s \leftarrow \chi$ for drawing s according to χ. We write $s \leftarrow S$ for a uniformly random element. We also write $y \leftarrow \mathcal{A}(x; r)$ for running an algorithm $\mathcal{A}$ with randomness r and $y \leftarrow \mathcal{A}(x)$ for running $\mathcal{A}$ with (uniformly) random $r \leftarrow R$ (where R is the randomness space). We let κ denote the security parameter and note that almost all objects are *implicitly* parameterised by it. By negl we denote some (fixed) negligible function, i.e. a function with $\lim_{\kappa\to\infty} \kappa^c \mathsf{negl}(\kappa) = 0$ for any $c \in \mathbb{N}$. We assume we can sample uniformly random from any $\{1, \ldots, n\}$. The number $p \in \mathbb{N}$ will always denote a prime, $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$, and $\mathbb{G}$ is a (cyclic abelian) group of order p. We use additive implicit notation for $\mathbb{G}$ as introduced in [20]. That is, we write $[1]$ for some (fixed public) generator associated with $\mathbb{G}$ and $[x] := x[1]$. We extend this notation to vectors and matrices, i.e. for compatible $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ over $\mathbb{F}_p$, we write $\boldsymbol{A}[\boldsymbol{B}]\boldsymbol{C} = [\boldsymbol{ABC}]$. Matrices are bold, e.g. $[\boldsymbol{a}]$, components not, e.g. $[a_i]$. By $\boldsymbol{e}_i$ we denote the i-th standard basis vector. We write $\mathrm{diag}(\boldsymbol{M}_1, \ldots, \boldsymbol{M}_n)$ for a block-diagonal matrix. By id_n we denote the $n \times n$ identity matrix.

2.1 Matrix kernel assumptions and Pedersen commitments

Instead of discrete logarithm assumptions, the generalisation of hard (matrix) kernel assumptions [37], but for right-kernels, better suits our needs.

Definition 2.1. Let $\mathbb{G} \leftarrow \mathsf{GrpGen}(1^\kappa)$ be a group generator (we let $[1]$ and p be implicitly given by $\mathbb{G}$). Let $\mathcal{D}_{m,n}$ be a (efficiently samplable) distribution over $\mathbb{G}^{m\times n}$ (where m and n may depend on κ). We say $\mathcal{D}_{m,n}$ has a **hard kernel assumption** if for all efficient adversaries $\mathcal{A}$, we have

$$\mathbb{P}\left(\begin{array}{c}\mathbb{G} \leftarrow \mathsf{GrpGen}(1^\kappa); [\boldsymbol{A}] \leftarrow \mathcal{D}_{m,n};\\ \boldsymbol{x} \leftarrow \mathcal{A}(1^\kappa, \mathbb{G}, [\boldsymbol{A}]) \colon [\boldsymbol{A}]\boldsymbol{x} = 0 \ \wedge \ \boldsymbol{x} \neq 0\end{array}\right) \leq \mathsf{negl}(\kappa)$$

For simplicity, we will often only implicitly refer to $\mathcal{D}_{m,n}$ and just say $[\boldsymbol{A}]$ has hard kernel assumption. Matrix kernel assumptions generalise DLOG assumptions: A non-trivial kernel element of $[h, 1] \in \mathbb{G}^2$ immediately yields the discrete logarithm h of $[h]$.

If $\mathcal{D}_{m,n}$ is a matrix distribution with hard kernel assumption, then $[\boldsymbol{A}] \leftarrow \mathcal{D}_{m,n}$ is a (Pedersen) commitment key ck. Commit to $\boldsymbol{x} \in \mathbb{F}_p^n$ via $\mathsf{Com}_{\mathsf{ck}}(\boldsymbol{x}) = [\boldsymbol{c}] \in \mathbb{G}^m$. Breaking the binding property of the commitment is equivalent to finding non-trivial elements in $\ker([\boldsymbol{A}])$. The common case will be $[\boldsymbol{g}] \in \mathbb{G}^{1\times(n+1)}$ drawn uniformly as commitment key ck. Breaking the hard kernel assumption for $[\boldsymbol{g}]$ is tightly equivalent to breaking the dlog assumption in $\mathbb{G}$. Write $\boldsymbol{x} = (r_{\boldsymbol{w}}, \boldsymbol{w})$ with $r_{\boldsymbol{w}} \in \mathbb{F}_p$, $\boldsymbol{w} \in \mathbb{F}_p^n$. If $r_{\boldsymbol{w}} \leftarrow \mathbb{F}_p$ is drawn uniformly, it is evident that $[c] = [\boldsymbol{g}]\boldsymbol{x}$ perfectly hides $\boldsymbol{w}$, i.e. $[c]$ is uniformly distributed in $\mathbb{G}$.

2.2 Interactive arguments, extractability and zero-knowledge

Our setting will be the common reference string model, i.e. there is some CRS crs, typically a commitment key, set up by a trusted party. In the following $\mathcal{R}$ denotes a binary relation for which $(st, w) \in \mathcal{R}$ is efficiently decidable. We call st the statement and w the witness. ($\mathcal{R}$ does depend on crs, i.e. actually we consider (crs, st, w) tuples, but we suppress this.) The (NP-)language $\mathcal{L}$ defined by $\mathcal{R}$ is the language of statements in $\mathcal{R}$, i.e. $\mathcal{L} = \{st \mid \exists w \colon (st, w) \in \mathcal{R}\}$.

Definition 2.2. An **(interactive) argument system** for a relation $\mathcal{R}$ is a protocol between two parties, a **prover** $\mathcal{P}$ and a **verifier** $\mathcal{V}$. We use the name **(interactive) proof system** interchangably. The transcript of the interaction of $\mathcal{P}$ and $\mathcal{V}$ on inputs x and y is denoted $\langle \mathcal{P}(x), \mathcal{V}(y)\rangle$ where both parties have a final "output" message. We write $b = \langle \mathcal{P}(x), \mathcal{V}(y)\rangle$, for the bit b indicating whether an (honest) verifier accepts ($b = 1$) the argument.

Definition 2.3 (Completeness). An interactive argument system for $(st, w) \in \mathcal{R}$ is (computationally) **complete** if for all efficient adversaries $\mathcal{A}$, the probability

$$\mathbb{P}\left(\begin{array}{c}\mathrm{crs} \leftarrow \mathsf{GenCRS}(1^\kappa); (st, w) \leftarrow \mathcal{A}(\mathrm{crs}) \colon (st, w) \notin \mathcal{R} \text{ or}\\ \langle \mathcal{P}(st, w), \mathcal{V}(st)\rangle = 1\end{array}\right)$$

is overwhelming, i.e. bounded below by $1 - \mathsf{negl}(\kappa)$. It is **perfectly complete** if $\mathsf{negl} = 0$.

In the full version [31], we give a definition of witness-extended emulation [10, 28] with extraction error (i.e. knowledge error). It turns out that preserving a good extraction error over multiple rounds is non-trivial. See Sections 2.3 and 2.4.

Definition 2.4 (Public coin). An interactive argument system for $\mathcal{R}$ is **public coin** if all of the verifier's challenges are independent of any other messages or state (essentially $\mathcal{V}$ makes his random coins public). Furthermore, $\mathcal{V}$'s verdict is a function $\mathsf{Verify}(tr)$ of the transcript.

Honest verifier zero-knowledge guarantees the existence of a simulator which, without the witness, generates transcripts which are indistinguishable from transcripts of a real interaction with an honest verifier. Hence, an honest verifier learns nothing from the proof.

Definition 2.5. Let $(\mathcal{P}, \mathcal{V})$ be an interactive argument system for $\mathcal{R}$. We call $(\mathcal{P}, \mathcal{V})$ **(ε-statistical) honest-verifier zero-knowledge** (HVZK), if there exists an expected polynomial-time simulator Sim such that for all expected polynomial-time $\mathcal{A}$ the probabiliy distributions of $(\mathrm{crs}, tr, state)$, where

- $\mathrm{crs} \leftarrow \mathsf{GenCRS}(1^\kappa); (st, w) \leftarrow \mathcal{A}(\mathrm{crs}); tr \leftarrow \langle \mathcal{P}(st, w), \mathcal{V}(st)\rangle$
- $\mathrm{crs} \leftarrow \mathsf{GenCRS}(1^\kappa); (st, w) \leftarrow \mathcal{A}(\mathrm{crs}); tr \leftarrow \mathsf{Sim}(st, \rho);$

are indistinguishable (have statistical distance at most ε), assuming $tr := \perp$ if $(st, w) \notin \mathcal{R}$.

Remark 2.6. We focus on HVZK, not *special* HVZK, The latter states that even if the adversary chooses statement, witness and the verifier's *randomness* (ρ in Definition 2.5), the *special* simulator will "succeed". Our security proofs make use of *honest* challenges. Different (more complex) security proofs may be possible.

2.2.1 Full-fledged zero-knowledge. To obtain security against dishonest verifiers, i.e. full-fledged zero-knowledge, simple transformations exist [15, 17, 24, 34] for public coin HVZK arguments. The most straightforward one is to use an equivocable coin toss between prover and verifier to generate the challenge.

2.2.2 The Fiat–Shamir heuristic. In the random-oracle model (ROM), public coin arguments can be converted to non-interactive arguments by computing the (verifier's) challenges as the hash of the transcript (and relevant "context") up to that point. The statement of the argument should be part of the "context" [8].

2.3 Testing distributions

Intuitively, testing distributions are a special form of probabilistic verification where one can *efficiently* recover the "tested" value given enough "tests". Thus, they are used to recover the witness in proofs of knowledge. We only define testing distributions over $\mathbb{F}_p^m$.

Example 2.7. To test if a vector $[\boldsymbol{c}] \in \mathbb{G}^m$ is $[\boldsymbol{0}]$, test if $\boldsymbol{x}^\top[\boldsymbol{c}] \stackrel{?}{=} [0]$ for random $\boldsymbol{x} \in \mathbb{F}_p^m$. The soundness error is $1/p$.

Definition 2.8 (Subdistribution). Let χ and ψ be distributions on $\mathbb{F}_p^m$. We call ψ a **subdistribution of χ of weight ε** if

- there exists a **subdensity** $\rho_\psi : \mathbb{F}_p^m \to [0, 1]$. (It is important that $\rho_\psi(x) \leq 1$.)
- $\varepsilon = \sum_{x\in\mathbb{F}_p^n} \rho_\psi(x)\chi(x)$, and
- ψ has probability $\psi(x) = \frac{1}{\varepsilon}\rho_\psi(x)$ to pick x. (That is, ψ has density $\frac{1}{\varepsilon}\rho_\psi$ w.r.t. χ.)

The definition of a subdistribution is constructed to deal with adversaries. As a concrete example consider extraction by rewinding: It may happen that the adversary does not correctly answer a challenge. Thus, the challenges which are answered are a subset, or more generally a subdistribution. An adversary with success probability ε must succeed on a subdistribution of weight ε.

Definition 2.9. A **testing distribution** χ_m for $\mathbb{F}_p^m$ with soundness $\delta_{\mathsf{snd}}(\kappa)$ is a distribution over $\mathbb{F}_p^m$ with following property: For all subdistributions ψ of χ_m with weight $\varepsilon \geq \delta_{\mathsf{snd}} := \delta_{\mathsf{snd}}(\chi_m)$, we have

$$\mathbb{P}(\boldsymbol{x}_i \leftarrow \psi, \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \colon \det(\boldsymbol{X}) = 0) \leq \tfrac{1}{\varepsilon}\delta_{\mathsf{snd}}.$$

We write $\delta_{\mathsf{snd}}(\chi_m)$ for some (fixed) soundness error of χ_m.

Note that $\det(\boldsymbol{X}) \neq 0$, is equivalent to all $\boldsymbol{x}_i$ being linearly independent, and to $\bigcap_{i=1}^m \ker(\boldsymbol{x}_i^\top) = \{0\}$. These interpretations allow to generalise in different directions. For more about testing distributions, see the full version [31]. Typically, we want that $\delta_{\mathsf{snd}}(\chi)$ is very small, e.g. 2^{-100} in practice.

For our examples, we need a minor generalisation of the lemma of Schwartz–Zippel. For details, see Appendix A.1.

Example 2.10 (Polynomial testing). The distribution induced by $\boldsymbol{x} = (x^0, \ldots, x^{m-1})$, where $x \leftarrow \mathbb{F}_p$, is a testing distribution. This follows from $\boldsymbol{X}$ being a Vandermonde matrix, hence invertible except if the same x was chosen twice. It is easy to see that $\delta_{\mathsf{snd}}(\chi) \leq \frac{m}{p}$.

Example 2.11. For the special case $m = 2$, and testing distribution with $\boldsymbol{x} = (\alpha, 1)$ where $\alpha \leftarrow \mathcal{S}$ for some $\mathcal{S} \subseteq \mathbb{F}_p$ we write $\chi^{(\beta)}$ and $\alpha \leftarrow \chi^{(\beta)}$. If $\mathcal{S} \subseteq \mathbb{F}_p^\times$, i.e. $\alpha \neq 0$, we write $\chi^{(\beta \neq 0)}$.

Example 2.12 (Random testing). The uniform distribution over $\mathbb{F}_p^m$ is a testing distribution. The Lemma of Schwartz–Zippel immediately yields $\delta_{\mathsf{snd}}(\chi) \leq \frac{m}{p}$. Drawing from a set $\mathcal{S}$ of "small exponents", e.g. from $\mathcal{S} = \{1, \ldots, \ell\}$, still has soundness $\delta_{\mathsf{snd}}(\chi) \leq \frac{m}{\ell}$.

Example 2.13 (Pseudo-random testing). The verifier can replace truly random choices, e.g. $\boldsymbol{x} \leftarrow \mathbb{F}_p^m$ as above, by pseudorandom choices, e.g. $\boldsymbol{x} \leftarrow \mathsf{PRG}(s)$ for $s \leftarrow \{0,1\}^\kappa$. This allows the verifier to compress such challenges to a random seed s.

It is heuristically plausible, that any non-pathological PRG has distribution with soundness error (negligibly close) to that of the respective uniform distributions. For more, see the full version [31].

Note that soundness of testing distributions is a combinatorial property. No pseudorandomness property is required, as illustrated by Example 2.10. Thus, there may be better options to use "small exponents" than (pseudo)random testing.

2.3.1 Dual testing distributions. Due to space constraints, dual testing distributions are not explicitly used in the main body, so we omit their definition. Morally, testing distributions *test* whether some $\boldsymbol{z} \in \mathbb{F}_p^m$ is $\boldsymbol{0}$. Dual testing distributions *enforce* $\boldsymbol{z} = \boldsymbol{0}$. Dual testing distributions can be used to derive fresh (Pedersen) commitment keys, and guarantee that no commitment generated prior can be opened (except to $\boldsymbol{0}$).

2.4 Special soundness

In the main body, we only consider special soundness and give extractors which produce a witness given a suitable tree of accepting transcripts, see also [10].

Definition 2.14 (μ-special soundness (over $\mathbb{F}_p$)). Let $(\mathsf{GenCRS}, \mathcal{P}, \mathcal{V})$ be a public coin argument system for $\mathcal{R}$. Suppose the verifier sends n challenges and $\mu = (\mu_0, \ldots, \mu_{n-1}) \in \mathbb{N}^n$. Furthermore, suppose the challenges are vectors in $\mathbb{F}_p^{n_i}$. Then the protocol is μ-special sound if there exists an extractor Ext such that given any good μ-tree tree_μ of transcripts, $\mathsf{Ext}(\mathit{st}, \mathit{tree}_\mu)$ returns a witness w with $(\mathit{st}, w) \in \mathcal{R}$. A μ-tree of transcripts is a (directed) tree where nodes of depth i have μ_i children, with edges labelled with the i-th challenge, and nodes labelled with the prover's i-th answer, and every path along the tree constitutes an *accepting* transcript. We call a μ-tree *good* if for every node, all its challenges (i.e. outgoing edges) are in general position.[1]

Given a TreeFind algorithm, which produces good μ-trees (with oracle access to a successful prover), and an extractor as above, one obtains witness extended emulation by plugging the tree into the extractor. To be able to speak about the security of the resulting protocol, one needs *success* and *runtime guarantees* of TreeFind. We do not deal with this here as it is a separate issue. See [10, 40] for TreeFinders and the full version [31] for more details.

2.4.1 Short-circuit extraction. Suppose TreeFind produces the tree's nodes and leaves on demand, and Ext queries TreeFind *as an oracle*, and *traverses the tree in depth-first order.* Moreover, suppose Ext either extracts a witness for some statement, or a solution to a (supposedly) hard problem, or both. Concretely, we have statements like "we extract $\boldsymbol{w}$ such that either $[\boldsymbol{g}]\boldsymbol{w} = [c]$ is a valid commitment opening, or $[\boldsymbol{g}]\boldsymbol{w} = [0]$ breaks the hard kernel assumption for $[\boldsymbol{g}]$." In such a situation, short-circuit extraction with $\mu' \leq \mu$ asserts that, extraction either finds the opening $\boldsymbol{w}$ using only the μ'-subtree of tree_μ, or for *one* layer i, all μ_i children are examined, and the extractor finds a non-trivial $\boldsymbol{w}$ in $\ker([\boldsymbol{g}])$.

Definition 2.15. Consider the situation in Definition 2.14. Suppose $\mathcal{R}$ is $\mathsf{OR}(\mathcal{R}_1, \mathcal{R}_2)$, i.e.

$$\mathcal{R} = \{((\mathit{st}_1, \mathit{st}_2), (w_1, w_2)) \mid (\mathit{st}_i, w_i) \in \mathcal{R}_i \text{ for } i = 1 \text{ or } i = 2\}.$$

Suppose there is some $\mu' \leq \mu$ (i.e. $\mu'_i \leq \mu_i$ for all i) such that extractor Ext has following property. For any good μ-tree tree_μ, $\mathsf{Ext}(\mathit{st}, \mathit{tree}_\mu)$ we have either:

- Ext finishes after exploring a μ'-subtree of tree_μ and returns a witness for st_1. We call this *quick-extraction.*
- If in layer ℓ of the tree, Ext must explore more than μ'_ℓ children of some node, then after exploring all μ_ℓ children, Ext returns a witness for st_2 (and perhaps st_1). (That is, Ext *short-circuits* in layer ℓ.)

We say that such an Ext has **short-circuit** extraction with $\mu' \leq \mu$ for finding a witness to st_1 or to st_2. (Note that order of the statements matters!)

Our definition is ad-hoc and tailored to our needs. We leave a solid definition and precise treatment of short-circuit extraction for future work.

[1] Vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{F}_p^n$ are in *general position* if any subset of size n is a basis.

Corollary 2.16. *If* Ext *as in Definition 2.15 traverses a good tree tree$_\mu$ in depth-first order, we have following "runtime" guarantees: Let $\mu' = (\mu'_0, \dots, \mu'_{n-1}) \le (\mu_0, \dots, \mu_{n-1}) = \mu$. In case of quick-extraction, at most $\prod_{i=0}^{n-1} \mu'_i$ leafs are explored. In case of short-circuit extraction, at most $s_0 + 1$ leaves are explored, where $s_0 = \sum_{i=0}^{n-1}(\mu_i - 1)\prod_{j=i+1}^{n-1}\mu'_j$. In particular, $s_0 \le (\sum_{i=0}^{n-1}\mu_i)(\prod_{i=0}^{n-1}\mu'_i)$.*

We note that since the tree *tree$_\mu$* is randomised (or Ext might explore children in random order), the above worst-case analysis is very conservative.

3 HVZK ARGUMENTS FOR $[A]w = [t]$

Let $\mathsf{ck} := [g] = [g_0, \overline{g}] \leftarrow \mathbb{G}^{1\times n+1}$ be a Pedersen commitment key, where $[g_0] \in \mathbb{G}$ and $[\overline{g}] \in \mathbb{G}^n$. Define $\mathsf{Com}_g(w; r) := [g_0]r + [\overline{g}]w$ for $r \in \mathbb{F}_p$, $w \in \mathbb{F}_p^n$. In the whole section, we work with matrices $[A] \in \mathbb{G}^{m\times n}$, and vectors $w \in \mathbb{F}_p^n$ and $[t] \in \mathbb{G}^m$. The dimensions are as above, unless otherwise specified. Our witness relation $\mathcal{R}$ is $st = ([A], [t])$ and $w = w$ such that $(st, w) \in \mathcal{R} \iff [A]w = [t]$.

3.1 Intuition

In this section, we devise communication efficient public-coin HVZK arguments for knowledge of a preimage of a linear map, i.e. $\exists w\colon [A]w = [t]$. We follow two principles: "Use probabilistic (batch) verification to check many things at once" and "If messages are too long, replace them by a shorter proof (of knowledge)." For this, we use shrinking commitments, to keep the messages small.

Our strategy is as follows: First, we recall the well-known general HVZK protocol [16, 36] for proving $\exists w\colon [A]w = [t]$ where $[A] \in \mathbb{G}^{m\times n}$. Then, we show how to apply batch verification to reduce the argument for $([A], [t])$ to another an argument for some $([B], [u])$ with $[B] \in \mathbb{G}^{2\times n}$. This makes communication independent of the number m of rows of $[A]$. After this, we revisit the arguments from [10] which recursively batch statement and witness, i.e. they reduce the number n of columns of $[A]$. Unlike [10, 13], we need a zero-knowledge version of these arguments. We provide a very efficient conversion with constant communication and logarithmic computational overhead. Taken together, we can for any $[A]$ prove knowledge of w in communication $O(\log(n))$ now.

3.2 Step 0: A standard Σ-protocol for $[A]w = [t]$

We recall the prototypical Σ-protocol in a group setting [16, 36].

Protocol 3.1 (Σ_{std}). The following is a protocol to prove $\exists w\colon [t] = [A]w$, using testing distribution $\chi^{(\beta)}$ for challenges, c.f. Example 2.11. Common input is $([A], [t]) \in \mathbb{G}^{m\times n} \times \mathbb{G}^n$. The prover's witness is some $w \in \mathbb{F}_p^n$.

- $\mathcal{P} \to \mathcal{V}$: Pick $r \leftarrow \mathbb{F}_p^n$ and let $[a] = [A]r$. Send $[a] \in \mathbb{G}^m$.
- $\mathcal{V} \to \mathcal{P}$: Pick $\beta \leftarrow \chi^{(\beta)}$. Send $\beta \in \mathbb{F}_p$.
- $\mathcal{P} \to \mathcal{V}$: Compute $z = \beta w + r$. Sends $z \in \mathbb{F}_p^n$.
- $\mathcal{V}$: Check if $[A]z = \beta[t] + [a]$. (Accept/reject if true/false.)

It is straightforward to show that any $(x_1, x_2) \leftarrow \chi_2$ can be used instead of $\chi^{(\beta)}$, as long as $x_2 \neq 0$, so that $x_1 w + x_2 r$ is uniformly distributed, c.f. Section 1.1.

Lemma 3.2. *Protocol Σ_{std} is a HVZK-PoK for $\exists w\colon [A]w = [t]$. It is perfectly complete, has perfect HVZK and is 2-special sound.*

Proof. **Completeness** is straightforward. **Extraction:** We are given two accepting transcripts $([a], \beta, z)$, and $([a], \beta', z')$ with $\beta - \beta' \neq 0$. Due to the final check of the verifier, we obtain $\frac{1}{\beta-\beta'}[A](z - z') = [t]$. Consequently, $w := \frac{1}{\beta-\beta'}(z - z')$ is a witness.

HVZK: Pick $\beta \leftarrow \chi^{(\beta)}$ and $z \leftarrow \mathbb{F}_p^m$. Then $[a] := [A]z - \beta[t]$ is uniquely defined. Since the distribution of β and z is as in an honest execution, this yields a perfect simulation. □

Now, we improve communication efficiency. We apply the techniques mentioned in the introduction, using shrinking commitments to keep messages small. Composition of proof systems is implicit due the following remark.

Remark 3.3. AND-proofs for statements of the form $\exists w\colon [A]w = [t]$ are trivial. Namely, to prove $\exists w\colon [A_1]w = [t_1] \wedge [A_2]w = [t_2]$, it suffices to define $[A] = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ and $[t] = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$ and prove $\exists w\colon [A]w = [t]$. This AND-compilation technique will be used without explicit mention. Evidently, many trivial optimisations are possible, e.g. removing duplicate rows.

3.3 Step 1: Batching all equations together

In this step, we devise a HVZK-AoK for $\exists w\colon [A]w = [t]$, where $\mathcal{P}$'s communication is independent of m, the "number of equations". Thus, we have to shrink the message $[a] \in \mathbb{G}^m$ somehow. We would like to batch all m linear equations (given by $[A]$) into a single linear equation, i.e. replace $[A]$ by a random linear combination of its rows. We do not know whether this is sound or not. Nevertheless, if $\mathcal{P}$ has explicitly committed to the witness w (or $[a]$), the statement — excluding the commitment — can be batched, as $\mathcal{P}$ cannot change his mind anymore. Note that the value $[t]$ is in general *not* a commitment, since the adversary may supply (parts of) $[A]$ in the soundness experiment. Thus, he may know dlogs and generate preimages of $[t]$ freely. By adding a commitment to w, we get around this problem. Using a shrinking (Pedersen) commitment to w, keeps the communication overhead small. Now, the verifier can send batching randomness, and a HVZK-AoK for the batched statement is carried out. Details are in the full version [31] We thus reduced general $[A]$ to $[B] \in \mathbb{F}_p^{2\times n}$, where the (say top) row of $[B]$ is a commitment key.

Remark 3.4 (Commitment extending). When working with adversarial $[A]$ (and $[t]$), one can not rely on hardness assumptions. Extending $[A]$ to, for example, $[B] = \begin{bmatrix} g \\ A \end{bmatrix}$ with commitment key $[g]$ is one way to circumvent problems. For the sake of referencing, we call this *commitment extending* $[A]$.

3.4 Step 2: "Batching" the witness

In this section, we show how to "batch" the witness, i.e. proving $\exists w\colon [A]w = [t]$ for $[A] \in \mathbb{G}^{m\times n}$ with communication sublinear in n. For the introduction, one may assume $m = 1$, e.g. $[A] = [g]$.

3.4.1 The general idea. We present the technique of [10], but in our situation and notation. For the motivation, let us ignore zero-knowledge, and only construct an argument (of knowledge). We add zero-knowledge later.

In general, one can achieve a size reduction of $k \in \mathbb{N}$ recursive step. For proof size, $k = 2$ is optimal, so we restrict ourselves to

that. The full version [31] deals with general k. Assume for simplicity that $2|n$, i.e. $n/2 \in \mathbb{N}$. We will reduce the equation $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ to $[\widehat{\boldsymbol{A}}]\widehat{\boldsymbol{w}} = [\widehat{\boldsymbol{t}}]$, where $[\widehat{\boldsymbol{A}}] \in \mathbb{G}^{m\times n/2}$, $\widehat{\boldsymbol{w}} \in \mathbb{F}_p^{n/2}$, $[\widehat{\boldsymbol{t}}] \in \mathbb{G}^m$. To do so, divide $[\boldsymbol{A}]$ and $\boldsymbol{w}$ into 2 equal blocks,[2] obtaining vectors/matrices of vectors/matrices i.e. $[\boldsymbol{A}] = [\boldsymbol{A}_1|\boldsymbol{A}_2] \in (\mathbb{G}^{m\times n/2})^{1\times 2}$ with $[\boldsymbol{A}_i] \in \mathbb{G}^{m\times n/2}$, and likewise $\boldsymbol{w} = \left(\begin{smallmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\end{smallmatrix}\right) \in (\mathbb{G}^{n/2})^2$. We want to prove $\sum_{i=1}^2 [\boldsymbol{A}_i]\boldsymbol{w}_i = [\boldsymbol{t}]$.

Still, the techniques from Section 3.3 are not applicable, because $[\boldsymbol{t}] \in \mathbb{G}$ (if $m = 1$). The trick of [10] is to embed our problem into a different one which can be batch-verified. Namely, we exploit that the scalar product is the sum of the diagonal entries (i.e. the trace) of the outer product:

$$\begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}(\boldsymbol{w}_1, \boldsymbol{w}_2) = \begin{bmatrix}\boldsymbol{A}_1\boldsymbol{w}_1 & \boldsymbol{A}_1\boldsymbol{w}_2\\ \boldsymbol{A}_2\boldsymbol{w}_1 & \boldsymbol{A}_2\boldsymbol{w}_2\end{bmatrix} \in \mathbb{G}^{2\times 2} \tag{3.1}$$

Now we can send all terms $[\boldsymbol{A}_i]\boldsymbol{w}_j$ to the verifier. Our probabilistic test has to map both $[\boldsymbol{A}]$ and $\boldsymbol{w}$ to a new (smaller) statement. We can do that by multiplying from the left by $\boldsymbol{x} \in \mathbb{F}_p^2$ and from the right by $\boldsymbol{y} \in \mathbb{F}_p^2$ where $\boldsymbol{x}, \boldsymbol{y} \leftarrow \chi_2$. Consequently, we obtain

$$\begin{aligned}&\boldsymbol{x}^\top \left(\begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}(\boldsymbol{w}_1, \boldsymbol{w}_2)\right)\boldsymbol{y}\\ =\ &\underbrace{\left(\boldsymbol{x}^\top\begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}\right)}_{:=\sum_i x_i[\boldsymbol{A}_i] =: [\widehat{\boldsymbol{A}}]} \underbrace{((\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k)\boldsymbol{y})}_{:=\sum_i y_i\boldsymbol{w}_i =: [\widehat{\boldsymbol{w}}]} = \underbrace{\sum_{i,j} x_i y_j [\boldsymbol{A}_i]\boldsymbol{w}_j}_{=:[\widehat{\boldsymbol{t}}]}\end{aligned}$$

The prover thus sends the (purported) $[\boldsymbol{A}_i\boldsymbol{w}_j]$, denoted $[\boldsymbol{u}_{i,j}]$, and $\widehat{\boldsymbol{w}}$, the shrunk witness. The verifier checks $\sum_i [\boldsymbol{u}_{i,i}] \stackrel{?}{=} [\boldsymbol{t}]$ and $[\widehat{\boldsymbol{A}}]\widehat{\boldsymbol{w}} \stackrel{?}{=} [\widehat{\boldsymbol{t}}] = \sum_{i,j} x_i y_j [\boldsymbol{u}_{i,j}]$.

If each $[\boldsymbol{A}_i]$ satisfies a hard kernel assumption, the prover is committed to $\boldsymbol{w}_1, \boldsymbol{w}_2$. It is not hard to see that given enough (linearly independent) challenges, one can extract $\boldsymbol{w}$. We will show this for a more efficient special case. All in all, we reduced the statement $([\boldsymbol{A}], [\boldsymbol{t}])$ to $([\widehat{\boldsymbol{A}}], [\widehat{\boldsymbol{t}}])$ which is smaller by a factor of $k = 2$. This can be applied recursively.

3.4.2 Refining the testing distribution. It turns out, that by a good choice of testing distribution, we can reduce communication. Namely, we can pick testing distributions with $x_i y_j = z_{j-i}$ for all i, j. Then it is sufficient for the verifier to know the sum of the off-diagonals i.e. $[\boldsymbol{A}_2]\boldsymbol{w}_1$ and $[\boldsymbol{A}_1]\boldsymbol{w}_2$ (and $[\boldsymbol{t}]$). We denote the (purported) $[\boldsymbol{A}_i]\boldsymbol{w}_j$, sent by the prover, as $[\boldsymbol{u}_\ell]$, i.e. $[\boldsymbol{u}_{-1}] \coloneqq [\boldsymbol{A}_2]\boldsymbol{w}_1$ and $[\boldsymbol{u}_1] \coloneqq [\boldsymbol{A}_1]\boldsymbol{w}_2$. Note that $[\boldsymbol{u}_0] = [\boldsymbol{t}]$ need not be sent. From the testing distribution $\widetilde{\chi}_3$ we require that $\boldsymbol{z} = (z_{-1}, z_0, z_1) \leftarrow \widetilde{\chi}_3$, belongs to a pair $(\boldsymbol{x}, \boldsymbol{y})$.

One testing distribution with this property comes from monomials ξ^i, e.g. $\boldsymbol{x} = (1, \xi)$ and $\boldsymbol{y} = (1, \xi^{-1})$.[3] In this case, $z_\ell = \xi^{-\ell}$.

For efficiency, picking $\boldsymbol{x}$ as above, but $\boldsymbol{y} = (\xi, 1)$ is useful, since this preserves small ξ. In this case, $\boldsymbol{z} = (z_{-1}, z_0, z_1) = (\xi^2, \xi, 1)$.

Protocol 3.5 ($\mathsf{LMPA}_{\mathrm{noZK}}$). The following is a protocol to prove $\exists \boldsymbol{w}\colon [\boldsymbol{t}] = [\boldsymbol{A}]\boldsymbol{w}$. Let $\widetilde{\chi}_3$ be a testing distributions with the properties described above. Common input is $([\boldsymbol{A}], [\boldsymbol{t}]) \in \mathbb{G}^{m\times n} \times \mathbb{G}^m$. We assume $n = 2^d$. The prover's witness is some $\boldsymbol{w} \in \mathbb{F}_p^n$.

Recursive step. Suppose $n = 2^d > 2$.

- Notation: Let $[\boldsymbol{A}] = [\boldsymbol{A}_1, \boldsymbol{A}_2]$ and $\boldsymbol{w} = \left(\begin{smallmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\end{smallmatrix}\right)$ be as above.
- $\mathcal{P} \to \mathcal{V}$: Compute $[\boldsymbol{u}_{-1}] \coloneqq [\boldsymbol{A}_2]\boldsymbol{w}_1$ and $[\boldsymbol{u}_1] \coloneqq [\boldsymbol{A}_1]\boldsymbol{w}_2$. Send $[\boldsymbol{u}_\ell]$ for $\ell = \pm 1$. ($[\boldsymbol{u}_0] \coloneqq [\boldsymbol{t}]$ is known to the verifier.)
- $\mathcal{V} \to \mathcal{P}$: Pick $\boldsymbol{z} \leftarrow \widetilde{\chi}_3$ with corresponding $\boldsymbol{x}, \boldsymbol{y}$. Send $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$.
- Both parties compute $[\widehat{\boldsymbol{A}}] = x_1[\boldsymbol{A}_1] + x_2[\boldsymbol{A}_2] \in \mathbb{G}^{m\times n/2}$ and $[\widehat{\boldsymbol{t}}] = \sum_{\ell=-1}^1 z_\ell[\boldsymbol{u}_\ell] \in \mathbb{G}$ as the new batched statement. Moreover, $\mathcal{P}$ computes $\widehat{\boldsymbol{w}} = \boldsymbol{w}_1 y_1 + \boldsymbol{w}_2 y_2$. The protocol may then be (recursively resumed), setting $n \leftarrow n/2$, $\boldsymbol{w} \leftarrow \widehat{\boldsymbol{w}}$, $[\boldsymbol{t}] \leftarrow [\widehat{\boldsymbol{t}}]$, $[\boldsymbol{A}] \leftarrow [\widehat{\boldsymbol{A}}]$.

Base case. Suppose $n \le 2$.

- $\mathcal{P} \to \mathcal{V}$: Send $\boldsymbol{w}$.
- $\mathcal{V}$: Tests if $[\boldsymbol{A}]\boldsymbol{w} \stackrel{?}{=} [\boldsymbol{t}]$.

See the full version [31] for a sketch of the protocol. For efficiency, our base case could also start at $n = 4$, as this saves one round-trip.

Lemma 3.6 (Recursive extraction). *Consider the situation above. Let $\widetilde{\chi}_3$ be a testing distribution with $x_i y_j = z_{j-i}$ as above.[4] Let $[\boldsymbol{u}_\ell]$, $[\boldsymbol{A}_i]$, $[\boldsymbol{t}]$, $\boldsymbol{w}_j$ and $[\widehat{\boldsymbol{A}}]$, $[\widehat{\boldsymbol{t}}]$ be defined as above. Then:*

- *Given a non-trivial kernel element of $[\widehat{\boldsymbol{A}}]$, we (efficiently) find a non-trivial kernel element of $[\boldsymbol{A}]$.*
- *Given 3 linearly independent challenges (with accepting transcripts), i.e. an invertible matrix $\boldsymbol{Z}$, one can extract (unconditionally) a witness $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$.*
- *Given 4 challenges in general position,[5] if the witness from above does not fit w.r.t. the $[\boldsymbol{u}_\ell]$, i.e. if an honest prover would send different $[\boldsymbol{u}_\ell]$ for $\boldsymbol{w}$, then we find (additionally) a non-trivial kernel element $\boldsymbol{v}$, i.e $[\boldsymbol{A}]\boldsymbol{v} = 0$.*

Moreover, we have short-circuit extraction*: From 2 independent challenges, one can compute a candidate witness $\boldsymbol{w}'$ for quick-extraction. If $[\boldsymbol{A}_i]\boldsymbol{w}'_j \neq [\boldsymbol{u}_\ell]$ for some $\ell = \pm 1$, then we are guaranteed to find a non-trivial kernel element from 4 challenges in general position.*

Note that, maybe surprisingly, extraction of a witness $\boldsymbol{w}$ with $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ is unconditional, i.e. we have a *proof* of knowledge. The proof is a minor generalisation of [10, 13]. See [31] for details.

3.4.3 Going zero-knowledge. There are many variations for going zero-knowledge. The most straightforward one is to run Protocol Σ_{std} and replace sending $\boldsymbol{z}$ by proving $\exists \boldsymbol{z}\colon [\boldsymbol{A}]\boldsymbol{z} = \beta[\boldsymbol{t}] + [\boldsymbol{a}]$ via $\mathsf{LMPA}_{\mathrm{noZK}}$. This gives a *proof* of knowledge, and is quite communication efficient. But computing $[\boldsymbol{A}]\boldsymbol{r}$ for random $\boldsymbol{r}$ is expensive. This approach is similar to [10, 13], where $\mathsf{LMPA}_{\mathrm{noZK}}$ only saves communication.

We achieve zero-knowledge more carefully. Instead of blinding the witness, we note that it is enough to blind the prover's responses. For this, a *logarithmic amount of randomness* suffices. This should make the prover more efficient.

Warm-up: Proving knowledge of opening of a commitment. For simplicity, we first sketch a protocol which assumes that $[\boldsymbol{A}] = [\boldsymbol{g}] \in \mathbb{G}^{1\times n}$, and $[\boldsymbol{g}]$ is a commitment key. Thus, $[\boldsymbol{A}]$ has hard kernel assumption by construction. Later, we deal with $m > 1$ and adversarially chosen $[\boldsymbol{A}]$, which we actually solve with a different technique. But the techniques employed in this simple example help

[2] It may be helpful to think of the vector space $(\mathbb{F}_p^{n/2})^2$ as $\mathbb{F}_p^{n/2} \otimes \mathbb{F}_p^2$.

[3] It can be shown that, up to scalar multiples, these are all such testing distributions.

[4] Note that the soundness error $\delta_{\mathsf{snd}}(\widetilde{\chi}_3)$ is an upper bound for the soundness errors of the (induced) testing distributions for $\boldsymbol{x}$ and $\boldsymbol{y}$.

[5] By Footnote 3, if x_2/x_1 is different for all challenges, they are in general position.

understanding the more complex technique, and they are reused and extended in Section 4.4.

Our current problem is to prove $\exists \boldsymbol{w}\colon [\boldsymbol{g}]\boldsymbol{w} = [t]$ in *zero-knowledge*. We will employ a masked version of $\mathsf{LMPA}_{\mathrm{noZK}}$, with judiciously chosen randomness $\boldsymbol{r}$, to achieve this. In particular, we do not pick $\boldsymbol{r} \leftarrow \mathbb{F}_p^n$. We pick $\boldsymbol{r}$ so that only logarithmically many r_i are non-zero. Thus, computing $[\boldsymbol{g}]\boldsymbol{r} = [a]$ is quite cheap (unlike in Protocol Σ_{std}). By the uniform-or-unique guideline, we want that each message $[\boldsymbol{u}_{\pm 1}]$ looks uniformly random. By analysing the recursive structure of $\mathsf{LMPA}_{\mathrm{noZK}}$, one sees that picking $r_i \leftarrow \mathbb{F}_p$ for $i \in \mathbb{M}_n \subseteq \{0, \ldots, n-1\}$ with $\mathbb{M}_n$ as defined below, and $r_i = 0$ else, achieves this property.[6]

Definition 3.7 (Masking sets). We define the **masking (randomness) sets/spaces** $\mathbb{M}_n \subseteq \{0, \ldots, n-1\}$ (for $n = 2^d$) by the formulas below. The set $\mathbb{M}_n$ describes the unit vectors of $\mathbb{F}_p^n$ (with zero-based indexing) which are used for random masking. We typically treat $\mathbb{M}_n$ as a subvector space of $\mathbb{F}_p^n$ (instead of explicitly referring to its span $\langle \boldsymbol{e}_i \mid i \in \mathbb{M}_n \rangle$).

- $\mathbb{M}_1 := \{0\}$ and $\mathbb{M}_2 := \{0, 1\}$.
- $\mathbb{M}_{2^d} := \{\mathbb{M}_{2^{d-1}}\} \dot\cup \{2^{d-1}, 2^{d-1}+1\}$ for $d \geq 2$.

See Fig. 2 for a pictorial description.

By the structure of the masking sets, we have that (for $k = 2$), if $\boldsymbol{r}$ is split into $\boldsymbol{r} = \binom{\boldsymbol{r}_1}{\boldsymbol{r}_2}$ as in $\mathsf{LMPA}_{\mathrm{noZK}}$, then $[\boldsymbol{u}_{j-i}] = [\boldsymbol{g}_i]\boldsymbol{r}_j$ is uniformly distributed for $\boldsymbol{r} \leftarrow \mathbb{M}_n$. Moreover, $\widehat{\boldsymbol{r}} = y_1\boldsymbol{r}_1 + y_2\boldsymbol{r}_2$ is distributed like a fresh $\boldsymbol{r}' \leftarrow \mathbb{M}_{n/2}$. This holds even when considering the joint distribution $([\boldsymbol{u}_{-1}], [\boldsymbol{u}_1], \widehat{\boldsymbol{r}})$. Thus, masking sets exhibit a useful recursive structure. There are some minor prerequisites to use the recursive structure, which we ignore for now.

Protocol 3.8. Let $\mathrm{crs} = [\boldsymbol{g}] \in \mathbb{G}^{1\times n}$ be a uniformly random commitment key (in particular, $[\boldsymbol{g}]$ has hard kernel relation under the DLOG assumption on $\mathbb{G}$.). The following is a protocol to prove $\exists \boldsymbol{w}\colon [t] = [\boldsymbol{g}]\boldsymbol{w}$. Let $\widetilde{\chi}_3$ be a testing distribution as in Protocol 3.5. Common input is $(\mathrm{crs}, [t]) \in \mathbb{G}^{1\times n} \times \mathbb{G}$. We assume $n = 2^d$. The prover's witness is some $\boldsymbol{w} \in \mathbb{F}_p^n$.

- $\mathcal{P} \to \mathcal{V}$: Choose $\boldsymbol{r} \leftarrow \mathbb{M}_n$. Compute $[a] = [\boldsymbol{g}]\boldsymbol{r}$. Send $[a]$.
- $\mathcal{V} \to \mathcal{P}$: Choose $\beta \leftarrow \chi^{(\beta)}$. Send β.
- $\mathcal{P} \leftrightarrow \mathcal{V}$: Let $\boldsymbol{z} := \beta\boldsymbol{w} + \boldsymbol{r}$ and $[t'] := \beta[t] + [a]$. Engage in $\mathsf{LMPA}_{\mathrm{noZK}}$ for $\exists \boldsymbol{z}\colon [\boldsymbol{g}]\boldsymbol{z} = [t']$.

It is clear that this protocol is correct. Short-circuit extraction follows easily as this is a composition of Protocol Σ_{std} and $\mathsf{LMPA}_{\mathrm{noZK}}$. Thus, only zero-knowledge remains. For this, one should note that $\boldsymbol{z} = \beta\boldsymbol{w} + \boldsymbol{r}$ behaves like a linear combination throughout the protocol, because the reduced witness $\widehat{\boldsymbol{z}}$ is of the form $\beta\widehat{\boldsymbol{w}} + \widehat{\boldsymbol{r}}$. Indeed, we can view the protocol as a linear combination of protocols. Thus, to see that $[\boldsymbol{u}_{\pm 1}]$ is uniformly distributed, we can focus our attention on $\boldsymbol{r}$ and its effect alone. As explained before, due to the form of $\mathbb{M}_n$, $(\widehat{\boldsymbol{r}}, [\boldsymbol{u}_{-1}], [\boldsymbol{u}_1])$ is uniformly distributed in $\mathbb{M}_{n/2} \times \mathbb{G} \times \mathbb{G}$. Thus, each iteration outputs uniformly distributed $[\boldsymbol{u}_{\pm 1}]$, and $\widehat{\boldsymbol{r}}$ distributed as $\widehat{\boldsymbol{r}} \leftarrow \mathbb{M}_{n/2}$. For the base case, we note that by construction, $\mathbb{M}_2 = \{0, 1\}$. Thus, $\boldsymbol{r} \leftarrow \mathbb{M}_2$ is uniformly random in $\mathbb{F}_p^2$, and hence $\beta\boldsymbol{w} + \widehat{\boldsymbol{r}}$ is uniformly random for $n \leq 2$, perfectly hiding $\boldsymbol{w}$. In particular, the messages in the base case are uniformly random too. The HVZK simulator can be built as usual, since the uniform-or-unique property is satisfied.

[6] The masking sets $\mathbb{M}$ use zero-based indexing for convenience.

Difficulties arising from general $[\boldsymbol{A}]$. There are two main difficulties arising from general $[\boldsymbol{A}] \in \mathbb{G}^{m\times n}$. First, the higher dimension due to $m > 1$ makes masking sets as described not directly applicable anymore. Second, we want to deal with *adversarial* $[\boldsymbol{A}]$. In the above sketch for zero-knowledge, we ignored a detail concerning the recursion. If it ever happens that in $[\boldsymbol{g}]$, for some $i \in \mathbb{M}_n$, the element $[g_i]$ is zero, the distribution of $(\widehat{\boldsymbol{r}}, [\boldsymbol{u}_{-1}], [\boldsymbol{u}_1])$ is skewed and zero-knowledge fails. An adversary can provoke this.

Resolving these problems efficiently (for the prover) is technical. See Appendix B for the construction and security claims. We remark that the naive approach to zero-knowledge for general $[\boldsymbol{A}]$ is a simple and viable option if the computational overhead is acceptable. Considering the computational costs of $\mathsf{LMPA}_{\mathrm{noZK}}$, this is often the case. Nevertheless, we demonstrate that, by applying our design guidelines, a more efficient, but more technical, conversion to zero-knowledge (with slightly larger proofs) is possible.

3.5 Step 3: Adding (arithmetic circuit) relations to the witness

If the witness $\boldsymbol{w}$ for $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ is committed to, e.g. if the first row of $[\boldsymbol{A}]$ is a Pedersen commitment CRS $[\boldsymbol{g}]$, it is easily possible to make other (zero-knowledge) statements about $\boldsymbol{w}$ by composition of zero-knowledge protocols. Using Protocol $\mathsf{QESA}_{\mathrm{Copy}}$ from Section 4 (or [13] in special cases), it is possible to add constraints on the witness. In particular, one can use range-proofs to control $\boldsymbol{w}$.

Remark 3.9. Often, $\boldsymbol{w}$ is much larger than the part which has to satisfy some constraints. It is efficiently possible to "split" and "merge" Pedersen commitments i.e. $[c] = [c_1] + [c_2]$ where $[\boldsymbol{G}] = [\boldsymbol{G}_1|\boldsymbol{G}_2]$ and $[c_i] = [\boldsymbol{G}_i]\boldsymbol{w}_i$. (Indeed, we use this quite often. With small changes, this is possible in zero-knowledge.) With this, one can split off the relevant portion $\boldsymbol{w}_1$ of $\boldsymbol{w}$ into the commitment $[c_1]$ and prove additional relations about this portion only. Splitting is generally very cheap. See Appendix D.1 for a concrete application.

4 ARITHMETIC CIRCUIT SATISFIABILITY FROM QUADRATIC EQUATIONS

In this section, we describe quadratic gates, and relate them to rank 1 constraint systems (R1CS) and arithmetic circuits (AC). Then, we construct a proof of satisfiability of a set of quadratic equations via a (zero-knowledge) inner-product argument.

4.1 Quadratic gates

The equations our scheme is able to prove are *quadratic equations*, i.e. given a witness $\boldsymbol{w} \in \mathbb{F}_p^n$ and a matrix $\boldsymbol{\Gamma} \in \mathbb{F}_p^{n\times n}$ we wish to prove $\boldsymbol{w}^\top\boldsymbol{\Gamma}\boldsymbol{w} = 0$. We choose this description of quadratic equations for *simplicity* and *uniformity* of notation. In particular, we assume without loss of generality, that the witness $\boldsymbol{w}$ has the constant 1 as first component, i.e. $w_1 = 1$. Our notation is similar to [19], which uses such notation for Groth–Sahai proofs [30]. Indeed, our arguments are essentially *commit-and-prove* systems [19].

Consider a general quadratic equation $\boldsymbol{x}^\top\boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{a}^\top\boldsymbol{x} = t$, with $\boldsymbol{a}, \boldsymbol{x} \in \mathbb{F}_p^n$, $\boldsymbol{\Gamma} \in \mathbb{F}_p^{n\times n}$, $t \in \mathbb{F}_p$ with statement given by the constants

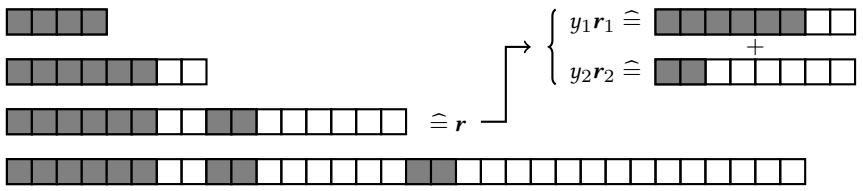

Figure 2: *Left:* The (construction of the) masking randomness sets $\mathbb{M}_4$, $\mathbb{M}_8$, $\mathbb{M}_{16}$ and $\mathbb{M}_{32}$ (for $k = 2$). The squares denote the numbers $0, \dots, n-1$ (or the respective basis vectors (with zero-based indexing)). *Right:* A demonstration of the "overlap" when a recursive step is applied to $\mathbb{M}_{16}$, i.e. $\widehat{r} = y_1\boldsymbol{r}_1 + y_2\boldsymbol{r}_2$ is computed. Note that by removing two dark squares in the overlap (i.e. the randomness being "used up" in $[u_{\pm 1}]$), the sum is still is randomised as $\mathbb{M}_8$. This "recursive property" is essential. The indices in $\mathbb{M}_n$ can also be constructed recursively via string concatenation: $m_{2n} = m_n|110^{n-2}$ and $m_1 = 1$, $m_2 = 11$.

$(\boldsymbol{a}, \boldsymbol{\Gamma}, t)$. This can be encoded via $\boldsymbol{w} = \left(\begin{smallmatrix} 1 \\ \boldsymbol{x} \end{smallmatrix}\right)$ and suitably (re)defined $\boldsymbol{\Gamma}$, namely $\boldsymbol{w}^\top \left(\begin{smallmatrix} -t & 0 \\ \boldsymbol{a} & \boldsymbol{\Gamma} \end{smallmatrix}\right) \boldsymbol{w} = 0$.

It is straightforward to encode arithmetic circuits (ACs) as systems of quadratic equations. Doing this allows for ACs built from *quadratic gates*, i.e. gates whose input-output behaviour is described by a quadratic equation.

4.2 Arithmetic circuits and rank 1 constraint systems

Rank 1 constraint systems (R1CS) are systems of equations of the form $(\boldsymbol{w}^\top \boldsymbol{a})(\boldsymbol{b}^\top \boldsymbol{w}) - \boldsymbol{c}^\top \boldsymbol{w} = 0$, where $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{F}_p^n$. Evidently, these are special cases of quadratic equations with $\boldsymbol{\Gamma} = \boldsymbol{a}\boldsymbol{b}^\top + e_1\boldsymbol{c}^\top$.[7] Arithmetic circuit satisfiability can be encoded in R1CS, c.f. [6].

The gates testable by one R1CS equation allow a single "multiplication". As we saw in the introduction, quadratic equations are more flexible. For example, the inner product $\boldsymbol{x}^\top \boldsymbol{y}$ is a single quadratic gate. To the best of our knowledge, n gates are necessary to encode this in R1CS (essentially one per $x_i y_i$ multiplication). Thus, quadratic gates enable new optimisations. Indeed, all "AC to R1CS" optimisations (and more), are applicable for "AC to QE".

4.3 The verification strategy

Verifying that a system of quadratic gates is satisfied is easy given the witness $\boldsymbol{w}$, in our case the wire assignments of the AC, and equations $\boldsymbol{\Gamma}_\mathfrak{g}$ (the gate $\mathfrak{g}$ encoded as a matrix). Just check $\boldsymbol{w}^\top \boldsymbol{\Gamma}_\mathfrak{g} \boldsymbol{w} = 0$ for all $\mathfrak{g} \in \mathfrak{G}$. By batching this can be sped up: Pick $(r_\mathfrak{g})_\mathfrak{g} \leftarrow \chi_{\#\mathfrak{G}}$ from a testing distribution. Then compute $\boldsymbol{\Gamma} := \sum_{\mathfrak{g}\in\mathfrak{G}} r_\mathfrak{g} \boldsymbol{\Gamma}_\mathfrak{g}$ as the "batched statement". Finally, check if $\boldsymbol{w}^\top \boldsymbol{\Gamma} \boldsymbol{w} = 0$.

We run this strategy in a commit-then-prove manner. First, commit to the witness $\boldsymbol{w}$. Then let the verifier pick testing randomness $(r_\mathfrak{g})_\mathfrak{g}$ and prove that $\boldsymbol{w}^\top \boldsymbol{\Gamma} \boldsymbol{w} = 0$ where $\boldsymbol{\Gamma} := \sum_{\mathfrak{g}\in\mathfrak{G}} r_\mathfrak{g} \boldsymbol{\Gamma}_\mathfrak{g}$ is the "batched statement". Note that $\boldsymbol{w}^\top \boldsymbol{\Gamma} \boldsymbol{w} = \langle \boldsymbol{w}, \boldsymbol{\Gamma}\boldsymbol{w} \rangle$ is an inner product. Hence, we require a *zero-knowledge* inner-product argument.

For technical reasons, we cannot generate a commitment to $\boldsymbol{\Gamma}\boldsymbol{w}$ efficiently (prior to knowing $\boldsymbol{\Gamma}$). Therefore, the prover first commits to $\boldsymbol{w}$ as $[c_{\boldsymbol{x}}] = \mathsf{Com}_{\mathsf{ck}_1}(\boldsymbol{w})$. Then he obtains $\boldsymbol{\Gamma}$ and commits to $\boldsymbol{\Gamma}\boldsymbol{w}$ as $[c_{\boldsymbol{y}}] = \mathsf{Com}_{\mathsf{ck}_2}(\boldsymbol{\Gamma}\boldsymbol{w})$. Then the prover carries out the inner product argument. He must also *prove* that the commitments $[c_{\boldsymbol{x}}]$ and $[c_{\boldsymbol{y}}]$ open to values $\boldsymbol{x} = \boldsymbol{w}$ and $\boldsymbol{y} = \boldsymbol{\Gamma}\boldsymbol{w}$ as promised. Again, we use (linear) batching to shorten the proof for $\boldsymbol{y} = \boldsymbol{\Gamma}\boldsymbol{x}$. Namely, to check $\boldsymbol{y} = \boldsymbol{\Gamma}\boldsymbol{x}$, the verifier picks random $\boldsymbol{s} \leftarrow \chi_n$ (after $[c_{\boldsymbol{x}}]$, $[c_{\boldsymbol{y}}]$ and hence $\boldsymbol{x}$, $\boldsymbol{y}$ are fixed) and the prover proves $0 = \langle \boldsymbol{\Gamma}\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{s} \rangle$.

Instead of two inner product arguments (for $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ and $\langle \boldsymbol{\Gamma}\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{s} \rangle = 0$) we batch verify again: The verifier picks randomness α and the prover proves knowledge of openings $\boldsymbol{x}, \boldsymbol{y}$ such that,

$$\begin{aligned} \langle \boldsymbol{x} - \alpha\boldsymbol{s}, \boldsymbol{y} + \alpha\boldsymbol{\Gamma}^\top \boldsymbol{s} \rangle &= \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \alpha \left(\langle \boldsymbol{x}, \boldsymbol{\Gamma}^\top \boldsymbol{s} \rangle - \langle \boldsymbol{s}, \boldsymbol{y} \rangle \right) - \alpha^2 \langle \boldsymbol{s}, \boldsymbol{\Gamma}^\top \boldsymbol{s} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \alpha \langle \boldsymbol{\Gamma}\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{s} \rangle - \alpha^2 \langle \boldsymbol{s}, \boldsymbol{\Gamma}^\top \boldsymbol{s} \rangle \\ &\overset{!}{=} -\alpha^2 \langle \boldsymbol{s}, \boldsymbol{\Gamma}^\top \boldsymbol{s} \rangle =: t \end{aligned} \tag{4.1}$$

where t is fixed by the random choices of the verifier. If $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Gamma}, \boldsymbol{s}$ are fixed, the lemma of Schwartz–Zippel can be applied to the polynomial in α. If $\alpha \leftarrow \mathcal{S}$, the probability that Eq. (4.1) holds but $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \neq 0$ or $\langle \boldsymbol{\Gamma}\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{s} \rangle \neq 0$ is $2/\#\mathcal{S}$. If $\boldsymbol{s}$ is chosen from a testing distribution χ_n with error $\delta_{\mathsf{snd}}(\chi_n)$, the probability that $\boldsymbol{\Gamma}\boldsymbol{x} - \boldsymbol{y} \neq 0$ is at most $\delta_{\mathsf{snd}}(\chi_n)$. Thus, this strategy is sound. To instantiate it, we need a zero-knowledge inner product argument.

4.4 Zero-knowledge inner product argument

Now, we show how to construct a zero-knowledge inner product argument (IPA). We first recall [10, 13], from a high level. We identify [13] as a linear combination of protocols. We achieve HVZK similar to Protocol 3.8 by masking the witness, but we also exploit redundancy (or kernel) guideline. Addition of zero-knowledge adds a single round, where one group element and one challenge are sent. For technical reasons we have a base case at $n = 8$.

4.4.1 Inner product argument (IPA). First, we describe the IPA following [10, 13]. For simplicity, we ignore zero-knowledge.

Our setting is as follows: We have a CRS $\mathsf{crs} = ([\boldsymbol{g}'], [\boldsymbol{g}''], [Q])$ for which finding a non-trivial kernel element of $[\boldsymbol{g}', \boldsymbol{g}'', Q] \in \mathbb{G}^{2n+1}$ is hard. In other words, these are three independent (or one large three-split) Pedersen commitment keys.

Naively, one proves knowledge of openings of $c'_{\boldsymbol{w}}$ and $c''_{\boldsymbol{w}}$ with $\langle \boldsymbol{w}', \boldsymbol{w}'' \rangle = t$. The idea and argument Protocol 3.8 allow to recursively shrink our statement. After one recursion step, we obtain $\langle \widehat{\boldsymbol{w}}', \widehat{\boldsymbol{w}}'' \rangle = \widehat{t}$. The prover sends $v_{\pm 1} = \langle \boldsymbol{w}'_i, \boldsymbol{w}''_j \rangle$ (for $j - i = \pm 1$), so that the verifier can compute $\widehat{t}$, analogous to $[u_{\pm 1}]$ in Section 3.4,

To save communication, we use a linear combination of Protocol $\mathsf{LMPA}_{\mathsf{noZK}}$ in our argument. Using the same challenge $(\boldsymbol{x}, \boldsymbol{y}, z)$ for both runs does not work. But when swapping the challenge for, say the first instance, we see that the linear combination works.

[7] The name R1CS may be misleading, since $\boldsymbol{\Gamma}$ can have (tensor) rank 2, i.e. the (tensor) rank of $\boldsymbol{\Gamma}$ is ≤ 2 for R1CS. Nevertheless, we follow this standard naming convention.

Concretely, let $\boldsymbol{x} = (1, \xi)$, $\boldsymbol{y} = (\xi, 1)$. Then

$$\langle \boldsymbol{x}^\top \boldsymbol{w}', \boldsymbol{y}^\top \boldsymbol{w}'' \rangle = \xi \langle \boldsymbol{w}', \boldsymbol{w}'' \rangle + \langle \boldsymbol{w}'_1, \boldsymbol{w}''_2 \rangle + \xi^2 \langle \boldsymbol{w}'_2, \boldsymbol{w}''_1 \rangle$$

Thus, analogous to $[\boldsymbol{u}_0] = [t]$ in $\mathsf{LMPA}_{\mathrm{noZK}}$, the term $\langle \boldsymbol{w}', \boldsymbol{w}'' \rangle = t$ is preserved. Therefore we run the first protocol for $\boldsymbol{w}'$ with flipped challenge $(\boldsymbol{y}, \boldsymbol{x})$, and the second protocol for $\boldsymbol{w}''$ with challenge $(\boldsymbol{x}, \boldsymbol{y})$. Now, as in Protocol $\mathsf{LMPA}_{\mathrm{noZK}}$, it suffices to send $v_{j-i} \coloneqq \langle \boldsymbol{w}'_i, \boldsymbol{w}''_{2-i} \rangle$ (for $i = 1, 2$).

The argument described above is a hybrid of [10] and [13]. For security, we need that "commitment merging" (see Remark 3.9), which the linear combination of protocols induces, still is *binding*. To obtain [13], we simply *commit* to v_ℓ as well (using $[Q]$), and send the combined commitment, i.e. apply again a linear combination. This "merged" commitment key is now $[\boldsymbol{g}', \boldsymbol{g}'', Q]$. Thus instead of sending two messages thrice (namely $[u'_{\pm 1}]$, $[u''_{\mp 1}]$, $[v_{\mp 1} Q]$), we only send the two "merged commitments" $[u_{\pm 1}] = [u'_{\pm 1}] + [u''_{\mp 1}] + [v_{\mp 1} Q]$. Unlike [13], which uses $\boldsymbol{x} = (\xi^{-1}, \xi)$ we prefer $\boldsymbol{x} = (1, \xi)$ since exponentiation with 1 is free.

We sketch the protocol. The CRS is $\mathrm{crs} = [\boldsymbol{g}', \boldsymbol{g}'', Q]$ where $[\boldsymbol{g}']$, $[\boldsymbol{g}''] \in \mathbb{G}^{1 \times n}$ and $[Q] \in \mathbb{G}$ are random. To prove

$$\exists \boldsymbol{w}', \boldsymbol{w}'' \in \mathbb{F}_p^n \colon [c] = [\boldsymbol{g}'] \boldsymbol{w}' + [\boldsymbol{g}''] \boldsymbol{w}'' + t[Q] \ \wedge \ \langle \boldsymbol{w}', \boldsymbol{w}'' \rangle = t,$$

the prover computes $[\boldsymbol{u}'_\ell]$, $[\boldsymbol{u}''_\ell]$ as in Protocol 3.8, but with challenges flipped as described above. For $\ell = \pm 1$, the prover sends $[\boldsymbol{u}_\ell] \coloneqq [\boldsymbol{u}'_\ell] + [\boldsymbol{u}''_{-\ell}] + v_{-\ell}[Q]$. Both parties compute the reduced statement, and another iteration (or base case) is run.

The resulting protocol is called $\mathsf{IPA}_{\mathrm{noZK}}$, see the full version [31] for a full description. It is μ-special sound (with $\mu = (2, 4, \ldots, 4)$) for finding a witness or a non-trivial element in the kernel of $[\boldsymbol{g}', \boldsymbol{g}'', Q]$. And it has short-circuit extraction with $\mu' = (1, 2, \ldots, 2)$. The proof is essentially as in [10, 13].

4.4.2 Going zero-knowledge. Making the inner-product argument zero-knowledge can be done in many ways. To be competetive with Bulletproofs [13], we directly mask the witness. This is problematic, since the scalar product is non-linear. Consequently, our (initial) approach only works under some (mild) constraints.

As mentioned above, the problem with using masking randomness and proving $\langle \boldsymbol{w}' + \boldsymbol{r}', \boldsymbol{w}'' + \boldsymbol{r}'' \rangle$ is the non-linearity: Sending only $t_{\boldsymbol{r}} = \langle \boldsymbol{r}', \boldsymbol{r}'' \rangle$ to the verifier is not enough. So we need to send also $\langle \boldsymbol{w}', \boldsymbol{r}'' \rangle$ or $\langle \boldsymbol{r}', \boldsymbol{w}'' \rangle$ or some other "error term" to correct the non-linearity. Then we have to show that these terms don't expose "information" about the witness. In particular, sending $\beta \boldsymbol{w}' + \boldsymbol{r}'$, which was possible in Section 3.3, seems impossible.

Fortunately, we already saw that the recursive argument only needs a small amount of randomness to conceal the witness. We exploit this now to show that the sketched masking *almost* yields zero-knowledge. Instead of sending the error terms, we pick randomness with the "kernel guideline" in mind:

- $\boldsymbol{r}' \in \ker(\boldsymbol{w}''^\top)$, i.e. $\langle \boldsymbol{r}', \boldsymbol{w}'' \rangle = 0$.
- $\boldsymbol{r}'' \in \ker(\boldsymbol{w}'^\top) \cap \ker(\boldsymbol{r}'^\top)$, i.e. $\langle \boldsymbol{w}', \boldsymbol{r}'' \rangle = 0 = \langle \boldsymbol{r}', \boldsymbol{r}'' \rangle$.

In other words, we pick randomness which does not induce errors. Thus, the prover only has to send $[t_{\boldsymbol{r}}] = [\boldsymbol{g}'] \boldsymbol{r}' + [\boldsymbol{g}''] \boldsymbol{r}''$ to the verifier. We first outline an almost zero-knowledge argument, using augmented masking sets $\mathbb{M}_n^+$ which are defined later.

Protocol 4.1 ($\mathsf{IPA}_{\mathrm{almZK}}$). The following is an inner product argument with the same statement, witness and notation as in $\mathsf{IPA}_{\mathrm{noZK}}$.

- $\mathcal{P} \to \mathcal{V}$: Pick $\boldsymbol{r}' \leftarrow \ker(\boldsymbol{w}''^\top) \cap \mathbb{M}_n^+$ and $\boldsymbol{r}'' \leftarrow \ker\left(\begin{pmatrix} \boldsymbol{w}'^\top \\ \boldsymbol{r}'^\top \end{pmatrix}\right) \cap \mathbb{M}_n^+$. Compute $[c_{\boldsymbol{r}}] \coloneqq [\boldsymbol{g}'] \boldsymbol{r}' + [\boldsymbol{g}''] \boldsymbol{r}''$. Send $[c_{\boldsymbol{r}}]$.
- $\mathcal{V} \to \mathcal{P}$: Pick $\beta \leftarrow \chi^{(\beta)}$. Send β.
- $\mathcal{P} \leftrightarrow \mathcal{V}$: Run Protocol $\mathsf{IPA}_{\mathrm{noZK}}$ for $\langle \beta \boldsymbol{w}' + \boldsymbol{r}', \beta \boldsymbol{w}'' + \boldsymbol{r}'' \rangle = \beta^2 t$ (with commitment $[c] = \beta [c_{\boldsymbol{w}}] + [c_{\boldsymbol{r}}] + \beta^2 t[Q]$). To compute $[c]$, the values t and $[c_{\boldsymbol{w}}]$ from the statement are used.

Correctness follows by inspection. Special soundness follows from special soundness of $\mathsf{IPA}_{\mathrm{noZK}}$ and Lemma 3.2.

Corollary 4.2. *Protocol 4.1 is special μ-sound (with $\mu = (2, 2, 4, \ldots, 4)$) for finding a witness or a non-trivial element in the kernel of $[\boldsymbol{g}', \boldsymbol{g}'', Q]$. It has short-circuit extraction with $\mu = (2, 1, 2, \ldots, 2)$.*

Showing zero-knowledge is more contrived. As for Protocol 3.8, we want to show that the prover's messages are uniformly random. Unfortunately, the constraints which must be satisfied now depend on the witness. Thus, an adversarially chosen witness may be a problem. Fortunately, we use $\mathsf{IPA}_{\mathrm{almZK}}$ with "randomised" witnesses, so this problem does not manifest.

Definition 4.3. Let k be fixed and $n \geq 8$. Define $\mathbb{M}_n^+ \coloneqq \mathbb{M}_n \dot\cup \{n-2, n-1\}$. (Recall that $\mathbb{M}_n$ indices are zero-based and $n-2, n-1 \notin \mathbb{M}_n$ for $n \geq 8$.)

We introduce $\mathbb{M}_n^+$ because satisfying the kernel constraints "consumes" one (resp. two) pieces of randomness in $\boldsymbol{r}'$ (resp. $\boldsymbol{r}''$). We compensate this in $\mathbb{M}_n^+$.

Lemma 4.4 (Informal, see [31]). *If $\boldsymbol{w}', \boldsymbol{w}''$ are of a suitable form, then the responses in $\mathsf{IPA}_{\mathrm{almZK}}$ are uniform-or-unique. More concretely, if w'_{n-1}, w'_n are random, and w''_{n-1}, w''_n also (not necessarily independent), then $\boldsymbol{w}$ is suitable.*

4.5 Quadratic equation satisfiability

We can finally instantiate our sketch of an argument system for satisfiability of a system of quadratic equations from Section 4.3. It is a commit-and-prove system as follows. The prover commits to the solution $\boldsymbol{w}$. Then $\boldsymbol{\Gamma}$ is fixed and $\langle \boldsymbol{w}, \boldsymbol{\Gamma} \boldsymbol{w} \rangle = 0$ shown to hold. The commitment scheme pads $\boldsymbol{w} \in \mathbb{F}_p^{n-2}$ with randomness and extends $\boldsymbol{\Gamma}$ in a suitable way. Intuition for soundness is given in Section 4.3.

Protocol 4.5 ($\mathsf{QESA}_{\mathrm{ZK}}$). Let $\boldsymbol{\Gamma}_i \in \mathbb{F}_p^{(n-2) \times (n-2)}$ $(i = 1, \ldots, N)$ be a system of quadratic eqations. Suppose $N \geq 2$.[8] Let $\boldsymbol{w} \in \mathbb{F}_p^{n-2}$ be a solution, i.e. $\boldsymbol{w}^\top \boldsymbol{\Gamma}_i \boldsymbol{w} = 0$ for all i. We assume that the first component w_1 of $\boldsymbol{w}$ is 1.

Let $\mathrm{crs} = [\boldsymbol{g}', \boldsymbol{g}'', Q]$, $\widetilde{\chi}_3$, $\chi^{(\beta \neq 0)}$ and $n \geq 8$ as in Protocol 4.1, and $\mathbb{M}_n^+$ as in Lemma 4.4. Let $\boldsymbol{x} \leftarrow \chi_N$ be a testing distribution with $x_1 = 1$ and $x_2 \neq 0$ for all $\boldsymbol{x}$.[9] Let $\boldsymbol{y} \leftarrow \chi_{n+1}$ be a testing distribution with $y_1 = 1$ always. The following is a protocol for proving $\exists \boldsymbol{w} \in \mathbb{F}_p^{n-2} \colon \forall i \colon \boldsymbol{w}^\top \boldsymbol{\Gamma}_i \boldsymbol{w} = 0$ where crs and $\boldsymbol{\Gamma}_i$ are common inputs and the prover's witness is $\boldsymbol{w}$.

- $\mathcal{P} \to \mathcal{V}$: (Step 0: Commitment.) Pick $\boldsymbol{r}' \leftarrow \mathbb{F}_p^2$. Let the "extended" witness be $\boldsymbol{w}' \coloneqq \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{r}' \end{pmatrix}$ and compute the commitment $[c'_{\boldsymbol{w}}] = [\boldsymbol{g}'] \boldsymbol{w}'$. Send $[c'_{\boldsymbol{w}}]$.
- $\mathcal{V} \to \mathcal{P}$: (Step 1: Batch verification.) Pick and send $\boldsymbol{x} \leftarrow \chi_N$.

[8] Otherwise, add trivial equations $\boldsymbol{\Gamma} = \boldsymbol{0}$.

[9] Restrictions on χ_N are merely to simplify protocol description and proofs.

- (Batch equations): Both parties compute $\boldsymbol{\Gamma} \coloneqq \sum x_i \boldsymbol{\Gamma}_i$.
- (Fix w_1 to 1): Both parties let $\beta \coloneqq x_2$ and redefine $[g'_1] \leftarrow \beta^{-1}[g'_1]$. Then $[c'_{\boldsymbol{w}}] \leftarrow [c'_{\boldsymbol{w}}] - (\beta-1)[g'_1]$ (with the new $[g'_1]$).
- $\mathcal{P} \to \mathcal{V}$: Let $\boldsymbol{r}'' = \boldsymbol{R}\boldsymbol{r}'$ where $\boldsymbol{R} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is a rotation by 90 degrees. Let $\boldsymbol{w}'' = \begin{pmatrix} \boldsymbol{\Gamma}\boldsymbol{w} \\ \boldsymbol{r}'' \end{pmatrix}$. Send $[c''_{\boldsymbol{w}}] = [\boldsymbol{g}'']\boldsymbol{w}''$.
- $\mathcal{V} \to \mathcal{P}$: Pick $(1, \boldsymbol{s}, \boldsymbol{b}) \leftarrow \chi_{n+1}$, where $\boldsymbol{s} \in \mathbb{F}_p^{n-2}$, $\boldsymbol{b} \in \mathbb{F}_p^2$. Send $\boldsymbol{s}' \coloneqq \left(\begin{smallmatrix} \boldsymbol{s} \\ \boldsymbol{b} \end{smallmatrix}\right)$.
- $\mathcal{P} \leftrightarrow \mathcal{V}$: Run Protocol $\mathsf{IPA}_{\mathsf{almZK}}$ for $\langle \boldsymbol{w}' - \boldsymbol{s}', \boldsymbol{w}'' + \boldsymbol{\Gamma}'^{\top}\boldsymbol{s}' \rangle = t$ with $t = -\langle \boldsymbol{s}, \boldsymbol{\Gamma}^{\top}\boldsymbol{s} \rangle$, and commitment $([c'_{\boldsymbol{w}}] - [\boldsymbol{g}']\boldsymbol{s}') + ([c''_{\boldsymbol{w}}] + [\boldsymbol{g}'']\boldsymbol{\Gamma}'^{\top}\boldsymbol{s}')$ and the modified $[\boldsymbol{g}']$ (and unmodified $[\boldsymbol{g}'']$, $[Q]$) as commitment keys. Here $\boldsymbol{\Gamma}' = \left(\begin{smallmatrix} \boldsymbol{\Gamma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{smallmatrix}\right) \in \mathbb{F}_p^{n \times n}$ where $\boldsymbol{R}$ is as in Step 1.

See the full version [31] for a sketch of this protocol.

Remark 4.6. It is not hard to see that the prover never needs to compute $[c] = ([c'_{\boldsymbol{w}}] - [\boldsymbol{g}']\boldsymbol{s}') + ([c''_{\boldsymbol{w}}] + [\boldsymbol{g}'']\boldsymbol{\Gamma}'^{\top}\boldsymbol{s}')$. (In general, $\mathcal{P}$ does not need $[\boldsymbol{u}_0]$.) While the verifier has to check $[c]$, using lazy evaluation and optimisations from [13], this hardly affects its runtime. All in all, dealing with $\boldsymbol{s}'$ is almost free.

We now state basic properties of $\mathsf{QESA}_{\mathsf{ZK}}$.

Lemma 4.7. *Protocol* $\mathsf{QESA}_{\mathsf{ZK}}$ *has perfect correctness.*

Using $\langle \begin{pmatrix} \boldsymbol{u}' \\ \boldsymbol{r}' \end{pmatrix}, \begin{pmatrix} \boldsymbol{u}'' \\ \boldsymbol{r}'' \end{pmatrix} \rangle = \langle \boldsymbol{u}', \boldsymbol{u}'' \rangle + \langle \boldsymbol{r}', \boldsymbol{r}'' \rangle$ and $\langle \boldsymbol{r}, \boldsymbol{R}\boldsymbol{r} \rangle = 0$ for all $\boldsymbol{r} \in \mathbb{F}_p^2$, this is a straightforward check.

Lemma 4.8. *Protocol* $\mathsf{QESA}_{\mathsf{ZK}}$ *has μ-special soundess (with $\mu = (N, n+1, 2, 2, 4, \ldots, 4)$) for extracting a witness or a non-trivial kernel element of $[\boldsymbol{g}', \boldsymbol{g}'', Q]$. It inherits short-circuit extraction with $\mu = (1, 1, 2, 2, 2, \ldots, 2)$.*

We did away with "α" compared to Section 4.3 to improve soundness. Extracting a challenge $(\alpha, \boldsymbol{s})$ naively requires a $(3, n-2)$ subtree. Our construction only needs an $(n+1)$ sub-"tree".

Lemma 4.9. *Protocol* $\mathsf{QESA}_{\mathsf{ZK}}$ *is ε-statistical zero-knowledge for some $\varepsilon \in O(2\log_2(n))/p$.*

For the proof, we establish that the conditions of Lemma 4.4 are met except with probability $O(2\log_2(n))/p$. This follows essentially because $\mathsf{QESA}_{\mathsf{ZK}}$ uses $\boldsymbol{w}' = \left(\begin{smallmatrix} \boldsymbol{w} \\ \boldsymbol{r} \end{smallmatrix}\right)$, where $\boldsymbol{r}$ is *random* (and similar for $\boldsymbol{w}''$). Thus, $\mathsf{IPA}_{\mathsf{almZK}}$ is statistical zero-knowledge, and consequently $\mathsf{QESA}_{\mathsf{ZK}}$ is statistical zero-knowledge as well.

4.6 Combining $\mathsf{QESA}_{\mathsf{ZK}}$ with other proof systems

As is, $\mathsf{QESA}_{\mathsf{ZK}}$ can be used to commit-and-prove quadratic equations. However, oftentimes, one wishes to prove statements about commitments which come from some other source. For example, Bulletproofs [13] were designed for confidential transaction, where the commitments are *input* to the proof system. This is not immediately feasible with $\mathsf{QESA}_{\mathsf{ZK}}$ as is, because $\mathsf{QESA}_{\mathsf{ZK}}$ is commit-and-prove only w.r.t. the solution of the set of quadratic equations.

Fortunately, extending $\mathsf{QESA}_{\mathsf{ZK}}$ is not hard. We consider following setting. There are commitment keys $\widetilde{\mathrm{ck}}^{(i)}$ for $i = 1, \ldots, M$. Each commitment key corresponds to a subset $\mathcal{I}_i \subseteq \{1, \ldots, n\}$ of the components of $[\boldsymbol{g}']$, where $\mathrm{crs} = ([\boldsymbol{g}', \boldsymbol{g}'', Q])$ is the commitment key of $\mathsf{QESA}_{\mathsf{ZK}}$. That is $\widetilde{\mathrm{ck}}^{(i)} \mathrel{\widehat{=}} \{[g'_j]\}_{j \in \mathcal{I}_i}$. Let $\mathcal{I} \coloneqq \cup_{i=1}^{M} \mathcal{I}_i$ be the set of all indices which are part of some $\widetilde{\mathrm{ck}}^{(i)}$. Let $M^{(i)} \coloneqq \#\mathcal{I}_i$ be the size of $\widetilde{\mathrm{ck}}^{(i)}$. We assume the following: Every commitment key $\widetilde{\mathrm{ck}}^{(i)}$ uses $[g'_n]$ (or $[g'_{n-1}]$) as its randomness components. Moreover, $1 \notin \mathcal{I}_i$, because the index $1 \mathrel{\widehat{=}} [g'_1]$ is reserved for the commitment to value 1 in $\mathsf{QESA}_{\mathsf{ZK}}$. A useful point of view is that $\widetilde{\mathrm{ck}}^{(i)}$ is a commitment under $[\boldsymbol{g}'] \in \mathbb{G}^n$ to a vector $\boldsymbol{v}^{(i)} \in \mathbb{F}_p^n$ with

$$\forall i \notin \mathcal{I}_i \colon v_i = 0. \tag{4.2}$$

We assume for simplicity that there is one commitment per key $\widetilde{\mathrm{ck}}^{(i)}$. To model the case of multiple commitments $[c_1], \ldots, [c_M]$ per key, e.g. all commitments are under $\widetilde{\mathrm{ck}} = \widetilde{\mathrm{ck}}^{(1)}$, we simply duplicate $\widetilde{\mathrm{ck}}$, i.e. we rewrite this as $[\widetilde{c}^{(i)}] = [c_i]$, $\widetilde{\mathrm{ck}}^{(i)} = \widetilde{\mathrm{ck}}$.

Example 4.10. In a typical range proof, with Pedersen committed value, we would have $\widetilde{\mathrm{ck}}^{(1)} \mathrel{\widehat{=}} [g'_2, g'_n]$, where $M = 1$. We write $\widetilde{\mathrm{ck}} \coloneqq \widetilde{\mathrm{ck}}^{(1)}$ for simplicity. This means $\mathcal{I} = \{2, n\}$.

Remark 4.11. Using the in n varying $[g'_n]$ in the commitment keys $\widetilde{\mathrm{ck}}^{(i)}$ is problematic and inconvenient. We want the randomness terms in $\mathsf{QESA}_{\mathsf{ZK}}$ and our commitment to "overlap". But now, running $\mathsf{QESA}_{\mathsf{ZK}}$ for a smaller or larger instance, e.g. an instance of size $n/2$ or $2n$ is incompatible. A simple solution is to fix some (random) $[g'^{,\star}_{\mathsf{rnd}1}, g'^{,\star}_{\mathsf{rnd}2}]$ (as part of crs) and construct $[\boldsymbol{g}']$ when starting Protocol $\mathsf{QESA}_{\mathsf{ZK}}$ so that $[g'_{n-1}, g'_n] = [g'^{,\star}_{\mathsf{rnd}1}, g'^{,\star}_{\mathsf{rnd}2}]$. Another solution is to permute the position of the randomness and reserve the fixed indices 2, 3 for randomness (instead of $n-1, n$).

With this setup, we can extend $\mathsf{QESA}_{\mathsf{ZK}}$ as follows: Given commitments $[\widetilde{c}^{(i)}]$ under keys $\widetilde{\mathrm{ck}}^{(i)}$, prove that the values committed in $[\widetilde{c}^{(i)}]$ satisfy some set of quadratic equations. In other words, prove that the $[\widetilde{c}^{(i)}]$ satisfy some arithmetic circuit.

Example 4.12 (Aggregate range proof). Consider $[\widetilde{c}^{(j)}], j = 1, ..., 10$. We wish to prove that the values $\boldsymbol{v}^{(j)}$ committed in $[\widetilde{c}^{(j)}]$ all lie in the range $\{0, \ldots, 2^8 - 1\}$.

Unsurprisingly, our solution to the problem is probabilistic verification. Our idea for general interoperability is as follows: The initial $\mathsf{QESA}_{\mathsf{ZK}}$ witness $\boldsymbol{w}$ (commitment $c'_{\boldsymbol{w}}$) has all components in $\mathcal{I}$ zeroed (except for randomness $n-1$, n) and also contains *copies* of the committed $\boldsymbol{v}^{(i)}$. The actual equations, i.e. the $\boldsymbol{\Gamma}_i$, only refer to the copies and the components $\mathcal{I}$. The verifier sends randomness $\boldsymbol{\alpha} \leftarrow \chi_{M+1}$ with $\alpha_0 = 0$, and we set $[c'_{\boldsymbol{w}}] \leftarrow [c'_{\boldsymbol{w}}] + \sum_i \alpha_i [\widetilde{c}^{(i)}]$, and $\boldsymbol{w}' \leftarrow \boldsymbol{w}' + \sum_i \alpha_i \boldsymbol{v}^{(i)}$ as new (extended) witness. Note that all (extended) equations $\boldsymbol{w}'\boldsymbol{\Gamma}_i'^{\top}\boldsymbol{w}'$ still hold (for an honest prover). Now we add (linear) equations $\boldsymbol{\Gamma}^{(i)}_{\mathrm{copy}}$ to the statement, which we call *copy-equations* and which depend on the randomness α_i. These equations simply assert that, if we compute $\sum_i \alpha_i \boldsymbol{v}^{(i)}$ using the committed copies in $\boldsymbol{w}$, then this equals the values committed in components $\mathcal{I}$ (again excluding the randomness components $n-1$, n). In other words, we assert that the purported copies of $\boldsymbol{v}^{(i)}$ in witness $[c'_{\boldsymbol{w},\mathrm{old}}]$ were valid copies. This "copy-based" approach is simple and modular.

The formulaic description of Protocol $\mathsf{QESA}_{\mathsf{Copy}}$ is arguably technical. However, the example in Fig. 3 demonstrates that it is actually a simple concept. In the full version [31], we give a more complex example. This extension of $\mathsf{QESA}_{\mathsf{ZK}}$ is again complete, special-sound, and statistical zero-knowledge. See the full version [31] for more details.

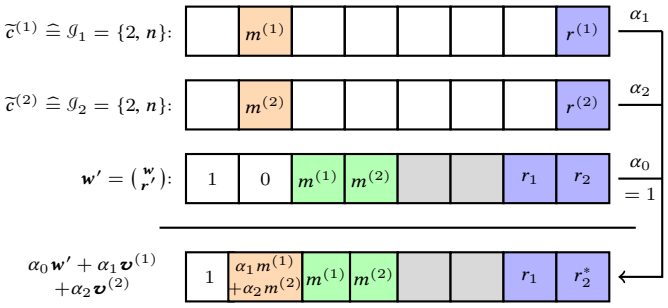

Figure 3: An example of a copying two values from two commitments. The blocks are colour-coded as follows: White blocks contain either 0 or the value indicated. Orange blocks belong to the (value-part) of commitment indices, i.e. to $\mathcal{G}$. Green blocks denote "copied" values. Gray blocks contain an arbitrary value. Blue blocks refer to randomness parts (i.e. components $n-1, n$). Note that randomness is *not* copied, as it is not a relevant part of the committed value. It is simply accumulated in $r_2^* = \alpha_0 r_2 + \alpha_1 r^{(1)} + \alpha_2 r^{(2)}$. The actual statements (i.e. the matrices $\boldsymbol{\Gamma}_i$) are statements over all variables except the orange (and blue) blocks, as these are merely "test-values" which ensure that $\boldsymbol{w}$ contains copies of (the message part of) $\boldsymbol{v}^{(i)}$, here $m^{(i)}$, as claimed.

5 IMPLEMENTATION

We implemented all protocols in C++17 using the RELIC toolkit [4] for underlying group operations. Our instantiation uses $\mathbb{G} =$ Curve25519 and thus $\mathbb{F}_p = \mathbb{F}_{2^{255}-19}$. For a fair comparison, we implemented Bulletproofs on the same architecture with equal care. The code is available on GitHub.[10]

Representing $\boldsymbol{\Gamma}$. All QESA protocols make use of sparse matrices $\boldsymbol{\Gamma}$. For efficient computation, a suitable representation is necessary. Decomposing $\boldsymbol{\Gamma}$ into a sum $\sum_i \boldsymbol{a}_i \boldsymbol{b}_i^\top$, similar to R1CS, allows for both runtime and memory optimisations. Note that vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are sparse themselves, allowing for even further optimisation via an appropriate data structure. For multiplications $\boldsymbol{\Gamma}\boldsymbol{s}$, at most $m \sum_i k_i \ell_i$ scalar multiplications are necessary, where m, k_i, ℓ_i are the number of non-zero entries in $\boldsymbol{s}$, $\boldsymbol{a}_i$, $\boldsymbol{b}_i$. Thus, all operations remain polynomial in the input size.

Results. We benchmarked our protocols on an Intel Core i7-6600U CPU at 2.6GHz running Debian Stretch 4.9.168 using a single core. A point multiplication with a random 254-bit scalar takes on average 0.28ms on this platform. Table 3 shows how our aggregate range proofs $\mathsf{QESA}_{\mathsf{RP}}$ compare to Bulletproofs. For $\mathsf{QESA}_{\mathsf{RP}}$, the internal witness $\boldsymbol{w}$ contains 4 static elements: the constant 1, the aggregate element for $\mathsf{QESA}_{\mathsf{Copy}}$, and the 2 random elements added by $\mathsf{QESA}_{\mathsf{Inner}}$, c.f. the full version [31]. Hence, we select the range as a power of 2 minus 4, in order to keep the CRS size from expanding to the next power of two. Our results show that $\mathsf{QESA}_{\mathsf{RP}}$ outperforms Bulletproofs for all tested parameters. Allowing batching randomnesses to be small further improves the performance (cf. $\mathsf{QESA}_{\mathsf{RP}}$ (short) for 140-bit random values).[11] Note that the execution times given in [13] are lower, since a highly optimised library dedicated to a single elliptic curve was used instead of a general purpose library as in this work. However, since both protocols were benchmarked on the same platform with the same underlying library, the values in Table 3 give a fair comparison.

Parameters	Bulletproofs		$\mathsf{QESA}_{\mathsf{RP}}$		$\mathsf{QESA}_{\mathsf{RP}}$ (short)	
	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{V}$
60 bit	0.26	0.17	0.16	0.07	0.15	0.06
60 bit × 2	0.47	0.29	0.32	0.15	0.30	0.10
60 bit × 32	7.4	4.5	5.1	2.4	4.6	1.7
60 bit × 128	28.9	17.9	20.6	9.4	18.4	6.7
60 bit × 512	116	78.7	82.3	37.5	73.8	27.1
124 bit	0.46	0.29	0.32	0.15	0.29	0.11
124 bit × 32	14.9	9.2	10.4	4.7	9.3	3.4
124 bit × 128	59.7	36.8	41.4	18.9	37.2	13.5
124 bit × 512	238	147	165	75.4	149	54.6
252 bit	0.95	0.59	0.65	0.30	0.57	0.22
252 bit × 32	30.2	18.6	20.8	9.5	18.9	6.8
252 bit × 128	121	74.3	83.5	37.8	76.1	27.4
252 bit × 512	484	297	358	165	302	109

Table 3: Comparison of runtime in seconds of aggregate range proofs from [13] with this work.

Shuffle size	1000		10000		100000	
	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{V}$	$\mathcal{P}$	$\mathcal{V}$
Time [s]	8.8	4.4	117	56.1	1009	491

Table 4: Evaluation of shuffle proofs via $\mathsf{QESA}_{\mathsf{Copy}}$ and $\mathsf{LMPA}_{\mathsf{simpleZK}}$.

Table 4 gives execution times for our shuffle proofs. They are an instantiation of [5], c.f. Appendix D, and we project them to be 2–3× more computationally expensive than [5], but they are size $O(\log(N))$ instead of $O(\sqrt{N})$ for N ciphertexts. Again the very high execution times compared to [5] are caused by the underlying library.

Note that we have not applied special optimisations. Using *delayed (batch) verification*, e.g. as in [13], significantly improves verifier performance. Optimised *verification* performance of Bulletproofs and our proof systems is probably almost identical. We are not aware of similar optimisations for the prover.

[10] https://github.com/emsec/QESA_ZK

[11] To justify short exponents, concrete security estimates are needed. We are not aware of results justifying any concrete instantiations. If our conjectures in the full version [31] hold, we can justify at least 80 bit security for witness size $n \leq 2^{16}$.

ACKNOWLEDGMENTS

We thank the anonymous reviewers of CRYPTO '19 and CCS '19 for helpful comments, which improved the overall quality and presentation of this work. This work is supported by the German Research Association under grants PA 587/10-1 and RU 1664/3-1.

REFERENCES

[1] [n. d.]. What is Jubjub? https://z.cash/technology/jubjub. ([n. d.]).
[2] Shashank Agrawal, Chaya Ganesh, and Payman Mohassel. 2018. Non-Interactive Zero-Knowledge Proofs for Composite Statements. In *CRYPTO 2018, Part III (LNCS)*, Hovav Shacham and Alexandra Boldyreva (Eds.), Vol. 10993. Springer, Heidelberg, Germany, Santa Barbara, CA, USA, 643–673. https://doi.org/10.1007/978-3-319-96878-0_22
[3] Scott Ames, Carmit Hazay, Yuval Ishai, and Muthuramakrishnan Venkitasubramaniam. 2017. Ligero: Lightweight Sublinear Arguments Without a Trusted Setup. In *ACM CCS 17*, Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (Eds.). ACM Press, Dallas, TX, USA, 2087–2104. https://doi.org/10.1145/3133956.3134104
[4] D. F. Aranha and C. P. L. Gouvêa. [n. d.]. RELIC is an Efficient LIbrary for Cryptography. https://github.com/relic-toolkit/relic. ([n. d.]).
[5] Stephanie Bayer and Jens Groth. 2012. Efficient Zero-Knowledge Argument for Correctness of a Shuffle. In *EUROCRYPT 2012 (LNCS)*, David Pointcheval and Thomas Johansson (Eds.), Vol. 7237. Springer, Heidelberg, Germany, Cambridge, UK, 263–280. https://doi.org/10.1007/978-3-642-29011-4_17
[6] Eli Ben-Sasson, Alessandro Chiesa, Daniel Genkin, Eran Tromer, and Madars Virza. 2013. SNARKs for C: Verifying Program Executions Succinctly and in Zero Knowledge. In *CRYPTO 2013, Part II (LNCS)*, Ran Canetti and Juan A. Garay (Eds.), Vol. 8043. Springer, Heidelberg, Germany, Santa Barbara, CA, USA, 90–108. https://doi.org/10.1007/978-3-642-40084-1_6
[7] Eli Ben-Sasson, Alessandro Chiesa, Michael Riabzev, Nicholas Spooner, Madars Virza, and Nicholas P. Ward. 2018. Aurora: Transparent Succinct Arguments for R1CS. *IACR Cryptology ePrint Archive* 2018 (2018), 828.
[8] David Bernhard, Olivier Pereira, and Bogdan Warinschi. 2012. How Not to Prove Yourself: Pitfalls of the Fiat-Shamir Heuristic and Applications to Helios. In *ASIACRYPT 2012 (LNCS)*, Xiaoyun Wang and Kazue Sako (Eds.), Vol. 7658. Springer, Heidelberg, Germany, Beijing, China, 626–643. https://doi.org/10.1007/978-3-642-34961-4_38
[9] Nir Bitansky, Ran Canetti, Alessandro Chiesa, Shafi Goldwasser, Huijia Lin, Aviad Rubinstein, and Eran Tromer. 2017. The Hunting of the SNARK. *Journal of Cryptology* 30, 4 (Oct. 2017), 989–1066. https://doi.org/10.1007/s00145-016-9241-9
[10] Jonathan Bootle, Andrea Cerulli, Pyrros Chaidos, Jens Groth, and Christophe Petit. 2016. Efficient Zero-Knowledge Arguments for Arithmetic Circuits in the Discrete Log Setting. In *EUROCRYPT 2016, Part II (LNCS)*, Marc Fischlin and Jean-Sébastien Coron (Eds.), Vol. 9666. Springer, Heidelberg, Germany, Vienna, Austria, 327–357. https://doi.org/10.1007/978-3-662-49896-5_12
[11] Jonathan Bootle, Andrea Cerulli, Essam Ghadafi, Jens Groth, Mohammad Hajiabadi, and Sune K. Jakobsen. 2017. Linear-Time Zero-Knowledge Proofs for Arithmetic Circuit Satisfiability. In *ASIACRYPT 2017, Part III (LNCS)*, Tsuyoshi Takagi and Thomas Peyrin (Eds.), Vol. 10626. Springer, Heidelberg, Germany, Hong Kong, China, 336–365. https://doi.org/10.1007/978-3-319-70700-6_12
[12] Jonathan Bootle and Jens Groth. 2018. Efficient Batch Zero-Knowledge Arguments for Low Degree Polynomials. In *PKC 2018, Part II (LNCS)*, Michel Abdalla and Ricardo Dahab (Eds.), Vol. 10770. Springer, Heidelberg, Germany, Rio de Janeiro, Brazil, 561–588. https://doi.org/10.1007/978-3-319-76581-5_19
[13] Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. 2018. Bulletproofs: Short Proofs for Confidential Transactions and More. In *2018 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, San Francisco, CA, USA, 315–334. https://doi.org/10.1109/SP.2018.00020
[14] Melissa Chase, David Derler, Steven Goldfeder, Claudio Orlandi, Sebastian Ramacher, Christian Rechberger, Daniel Slamanig, and Greg Zaverucha. 2017. Post-Quantum Zero-Knowledge and Signatures from Symmetric-Key Primitives. In *ACM CCS 17*, Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (Eds.). ACM Press, Dallas, TX, USA, 1825–1842. https://doi.org/10.1145/3133956.3133997
[15] Michele Ciampi, Giuseppe Persiano, Luisa Siniscalchi, and Ivan Visconti. 2016. A Transform for NIZK Almost as Efficient and General as the Fiat-Shamir Transform Without Programmable Random Oracles. In *TCC 2016-A, Part II (LNCS)*, Eyal Kushilevitz and Tal Malkin (Eds.), Vol. 9563. Springer, Heidelberg, Germany, Tel Aviv, Israel, 83–111. https://doi.org/10.1007/978-3-662-49099-0_4
[16] Ronald Cramer and Ivan Damgård. 1998. Zero-Knowledge Proofs for Finite Field Arithmetic; or: Can Zero-Knowledge Be for Free?. In *CRYPTO'98 (LNCS)*, Hugo Krawczyk (Ed.), Vol. 1462. Springer, Heidelberg, Germany, Santa Barbara, CA, USA, 424–441. https://doi.org/10.1007/BFb0055745
[17] Ivan Damgård. 2000. Efficient Concurrent Zero-Knowledge in the Auxiliary String Model. In *EUROCRYPT 2000 (LNCS)*, Bart Preneel (Ed.), Vol. 1807. Springer, Heidelberg, Germany, Bruges, Belgium, 418–430. https://doi.org/10.1007/3-540-45539-6_30
[18] George Danezis, Cédric Fournet, Jens Groth, and Markulf Kohlweiss. 2014. Square Span Programs with Applications to Succinct NIZK Arguments. In *ASIACRYPT 2014, Part I (LNCS)*, Palash Sarkar and Tetsu Iwata (Eds.), Vol. 8873. Springer, Heidelberg, Germany, Kaoshiung, Taiwan, R.O.C., 532–550. https://doi.org/10.1007/978-3-662-45611-8_28
[19] Alex Escala and Jens Groth. 2014. Fine-Tuning Groth-Sahai Proofs. In *PKC 2014 (LNCS)*, Hugo Krawczyk (Ed.), Vol. 8383. Springer, Heidelberg, Germany, Buenos Aires, Argentina, 630–649. https://doi.org/10.1007/978-3-642-54631-0_36
[20] Alex Escala, Gottfried Herold, Eike Kiltz, Carla Ràfols, and Jorge Villar. 2013. An Algebraic Framework for Diffie-Hellman Assumptions. In *CRYPTO 2013, Part II (LNCS)*, Ran Canetti and Juan A. Garay (Eds.), Vol. 8043. Springer, Heidelberg, Germany, Santa Barbara, CA, USA, 129–147. https://doi.org/10.1007/978-3-642-40084-1_8
[21] Rosario Gennaro, Craig Gentry, Bryan Parno, and Mariana Raykova. 2013. Quadratic Span Programs and Succinct NIZKs without PCPs. In *EUROCRYPT 2013 (LNCS)*, Thomas Johansson and Phong Q. Nguyen (Eds.), Vol. 7881. Springer, Heidelberg, Germany, Athens, Greece, 626–645. https://doi.org/10.1007/978-3-642-38348-9_37
[22] Rosario Gennaro, Michele Minelli, Anca Nitulescu, and Michele Orrù. 2018. Lattice-Based zk-SNARKs from Square Span Programs. In *ACM CCS 18*, David Lie, Mohammad Mannan, Michael Backes, and XiaoFeng Wang (Eds.). ACM Press, Toronto, ON, Canada, 556–573. https://doi.org/10.1145/3243734.3243845
[23] Irene Giacomelli, Jesper Madsen, and Claudio Orlandi. 2016. ZKBoo: Faster Zero-Knowledge for Boolean Circuits. In *USENIX Security Symposium*. USENIX Association, 1069–1083.
[24] Jens Groth. 2004. *Honest verifier zero-knowledge arguments applied.* Ph.D. Dissertation. Aarhus University.
[25] Jens Groth. 2009. Linear Algebra with Sub-linear Zero-Knowledge Arguments. In *CRYPTO 2009 (LNCS)*, Shai Halevi (Ed.), Vol. 5677. Springer, Heidelberg, Germany, Santa Barbara, CA, USA, 192–208. https://doi.org/10.1007/978-3-642-03356-8_12
[26] Jens Groth. 2010. Short Non-interactive Zero-Knowledge Proofs. In *ASIACRYPT 2010 (LNCS)*, Masayuki Abe (Ed.), Vol. 6477. Springer, Heidelberg, Germany, Singapore, 341–358. https://doi.org/10.1007/978-3-642-17373-8_20
[27] Jens Groth. 2016. On the Size of Pairing-Based Non-interactive Arguments. In *EUROCRYPT 2016, Part II (LNCS)*, Marc Fischlin and Jean-Sébastien Coron (Eds.), Vol. 9666. Springer, Heidelberg, Germany, Vienna, Austria, 305–326. https://doi.org/10.1007/978-3-662-49896-5_11
[28] Jens Groth and Yuval Ishai. 2008. Sub-linear Zero-Knowledge Argument for Correctness of a Shuffle. In *EUROCRYPT 2008 (LNCS)*, Nigel P. Smart (Ed.), Vol. 4965. Springer, Heidelberg, Germany, Istanbul, Turkey, 379–396. https://doi.org/10.1007/978-3-540-78967-3_22
[29] Jens Groth, Yael Kalai, Muthu Venkitasubramaniam, Nir Bitansky, Ran Canetti, Henry Corrigan-Gibbs, Shafi Goldwasser, Charanjit Jutla, Yuval Ishai, Rafail Ostrovsky, Omer Paneth, Tal Rabin, Mariana Raykova, Ron Rothblum, Alessandra Scafuro, Eran Tromer, and Douglas Wikström. 2018. *Security Track Proceeding*. Technical Report. ZKProof Standards, Berkeley, CA. https://zkproof.org/documents.html.
[30] Jens Groth and Amit Sahai. 2008. Efficient Non-interactive Proof Systems for Bilinear Groups. In *EUROCRYPT 2008 (LNCS)*, Nigel P. Smart (Ed.), Vol. 4965. Springer, Heidelberg, Germany, Istanbul, Turkey, 415–432. https://doi.org/10.1007/978-3-540-78967-3_24
[31] Max Hoffmann, Michael Klooß, and Andy Rupp. 2019. Efficient zero-knowledge arguments in the discrete log setting, revisited. *IACR Cryptology ePrint Archive* 2019 (2019), 944.
[32] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. 2007. Zero-knowledge from secure multiparty computation. In *39th ACM STOC*, David S. Johnson and Uriel Feige (Eds.). ACM Press, San Diego, CA, USA, 21–30. https://doi.org/10.1145/1250790.1250794
[33] Ahmed E. Kosba, Zhichao Zhao, Andrew Miller, Yi Qian, T.-H. Hubert Chan, Charalampos Papamanthou, Rafael Pass, Abhi Shelat, and Elaine Shi. 2015. $C\emptyset C\emptyset$: A Framework for Building Composable Zero-Knowledge Proofs. *IACR Cryptology ePrint Archive* 2015 (2015), 1093.
[34] Yehuda Lindell. 2015. An Efficient Transform from Sigma Protocols to NIZK with a CRS and Non-programmable Random Oracle. In *TCC 2015, Part I (LNCS)*, Yevgeniy Dodis and Jesper Buus Nielsen (Eds.), Vol. 9014. Springer, Heidelberg, Germany, Warsaw, Poland, 93–109. https://doi.org/10.1007/978-3-662-46494-6_5
[35] Helger Lipmaa. 2013. Succinct Non-Interactive Zero Knowledge Arguments from Span Programs and Linear Error-Correcting Codes. In *ASIACRYPT 2013, Part I (LNCS)*, Kazue Sako and Palash Sarkar (Eds.), Vol. 8269. Springer, Heidelberg, Germany, Bengalore, India, 41–60. https://doi.org/10.1007/978-3-642-42033-7_3

[36] Ueli Maurer. 2015. Zero-knowledge proofs of knowledge for group homomorphisms. *Des. Codes Cryptography* 77, 2-3 (2015), 663–676.
[37] Paz Morillo, Carla Ràfols, and Jorge Luis Villar. 2016. The Kernel Matrix Diffie-Hellman Assumption. In *ASIACRYPT 2016, Part I (LNCS)*, Jung Hee Cheon and Tsuyoshi Takagi (Eds.), Vol. 10031. Springer, Heidelberg, Germany, Hanoi, Vietnam, 729–758. https://doi.org/10.1007/978-3-662-53887-6_27
[38] Kun Peng, Colin Boyd, and Ed Dawson. 2007. Batch zero-knowledge proof and verification and its applications. *ACM Trans. Inf. Syst. Secur.* 10, 2 (2007), 6.
[39] Riad S. Wahby, Ioanna Tzialla, abhi shelat, Justin Thaler, and Michael Walfish. 2018. Doubly-Efficient zkSNARKs Without Trusted Setup. In *2018 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, San Francisco, CA, USA, 926–943. https://doi.org/10.1109/SP.2018.00060
[40] Douglas Wikström. 2018. Special Soundness Revisited. *IACR Cryptology ePrint Archive* 2018 (2018), 1157. https://eprint.iacr.org/2018/1157

A OMISSIONS: PRELIMINARIES

A.1 Testing distributions

Lemma A.1 (Schwartz–Zippel). *Let $f \in \mathbb{F}_p[X_1, \ldots, X_n]$ be a* non-zero *polynomial of (total) degree d. Let χ be a distribution on $\mathbb{F}_p$. Let $p_\infty(\chi) := \sup_{x \in \mathbb{F}_p} \chi(x)$, where $\chi(x) := \mathbb{P}(x = y \mid x \leftarrow \chi)$. Then $\mathbb{P}(f(\boldsymbol{x}) = 0) \le d p_\infty(\chi)$ for $\boldsymbol{x} \leftarrow \chi^n$ (i.e. $x_i \leftarrow \chi$). Moreoever, since $p_\infty(\psi) \le \frac{1}{\varepsilon} p_\infty(\chi)$ for any subdistribution ψ of weight ε, we get $\mathbb{P}_{\boldsymbol{x} \leftarrow \psi^n}(f(\boldsymbol{x}) = 0) \le \frac{d p_\infty(\chi)}{\varepsilon}$.*

The usual proof can be adapted easily.

A.1.1 Dual testing distributions. Testing distributions are essentially a stronger (and simplified) form of the general concept of probabilistic verification with efficient extraction. They allow to test if an element in $\mathbb{F}_p^n$ is 0. By dualising, we find another concept, for which an intuitive description seems harder. Instead of a distribution on $\boldsymbol{x}^\top \in \mathbb{F}_p^{1\times m}$ satisfying with high probability $\bigcap_{i=1}^m \ker(\boldsymbol{x}^\top) = \{\boldsymbol{0}\}$, we consider a distribution on $\boldsymbol{M} \in \mathbb{F}_p^{m\times m-1}$, satisfying with high probability $\bigcap_{i=1}^m \mathrm{im}(\boldsymbol{M}) = \{\boldsymbol{0}\}$, In a sense, $\boldsymbol{M}$ guarantees that for any $\boldsymbol{0} \neq \boldsymbol{z} \in \mathbb{F}_p^m$, $\boldsymbol{z} \notin \mathrm{im}(\boldsymbol{M})$ with high probability. Hence, we can use it to *enforce* $\boldsymbol{z} = \boldsymbol{0}$, instead of testing for it.

More concretely, we use this to ensure that for a Pedersen commitment $[c] = [\boldsymbol{G}|\boldsymbol{H}]\binom{\boldsymbol{w}}{\boldsymbol{z}}$ the adversary must have $\boldsymbol{z} = \boldsymbol{0}$. We do so by constructing $[\boldsymbol{H}]$ as $[\boldsymbol{H}] := [\boldsymbol{Q}]\boldsymbol{M}$. Intuitively, knowledge of some $[c'] = [\boldsymbol{G}|\boldsymbol{Q}]\binom{\boldsymbol{w}}{\boldsymbol{y}}$ cannot be transferred to $[\boldsymbol{G}|\boldsymbol{H}]$ because we must have $\boldsymbol{z} = \boldsymbol{M}\boldsymbol{y}$, i.e. $\boldsymbol{z} \in \mathrm{im}(\boldsymbol{M})$, which is unlikely (except for $\boldsymbol{z} = \boldsymbol{0}$ or if $\mathcal{A}$ breaks the binding property). Thus, we can provably "zero" a part of a commitment without an (expensive) argument. Generally, this allows to derive "fresh" commitment keys. Using this is more communication efficient than picking and sending a fresh $[\boldsymbol{H}] \leftarrow \mathbb{G}^m$.

Morally, dual testing *enforces* $\boldsymbol{z} = \boldsymbol{0}$, while "normal" testing *verifies* $\boldsymbol{z} = \boldsymbol{0}$.

Definition A.2. An (arbitrary) **dual testing distribution** $\chi_m^\vee$ is a distribution on $\mathbb{F}_p^{m\times(m-1)}$. The (computational) soundess error $\delta_{\mathsf{snd}}(\chi_m^\vee)$ is defined as before, but using $\mathbb{P}(\cap_{i=1}^m \mathrm{im}(\boldsymbol{M}_i) \neq \{\boldsymbol{0}\})$.

Let χ_m be a testing distribution on $\mathbb{F}_p^m$ such that $\boldsymbol{x} \leftarrow \chi_m$ *always* has $x_1 = 1$. Then $\chi_m^\vee$ defined as follows is a dual testing distribution: To pick $\boldsymbol{M} \leftarrow \chi_m^\vee$, pick $\boldsymbol{x}^\top = (1, \boldsymbol{x}')^\top \leftarrow \chi_m$ and let $\boldsymbol{M} := \boldsymbol{M}_{\boldsymbol{x}} := \begin{pmatrix} \boldsymbol{x}' \\ -\mathrm{id}_{m-1} \end{pmatrix}$. By construction $\ker(\boldsymbol{x}^\top) = \mathrm{im}(\boldsymbol{M}_{\boldsymbol{x}})$, and consequently $\delta_{\mathsf{snd}}(\chi^\vee) = \delta_{\mathsf{snd}}(\chi_m)$.

Note that by construction, $\boldsymbol{M}_{\boldsymbol{x}}$ is the (parity) check matrix for the linear code with generator $\boldsymbol{x}$. In particular, $\boldsymbol{x}^\top \boldsymbol{M}_{\boldsymbol{x}} = \boldsymbol{0}$. For simplicity, we only consider dual testing distributions associated to some testing distribution.

B OMISSIONS: LMPA

B.1 $\mathsf{LMPA}_{\mathsf{ZK}}$

We expand on sketches in Section 3.4 and give the description of our $\mathsf{LMPA}_{\mathsf{ZK}}$ construction with logarithmic computional overhead.

Difficulties arising from general $[\boldsymbol{A}]$. There are two main difficulties arising from general $[\boldsymbol{A}] \in \mathbb{G}^{m\times n}$. First, the higher dimension due to $m > 1$ makes masking sets as described not directly applicable anymore. Since $[\boldsymbol{u}_\ell] \in \mathbb{G}^m$, the prover now communicates mk elements, and hence we expect that $mk\log(n)$ random entries are necessary to randomise all of $[\boldsymbol{u}_\ell]$. Interestingly, the naive approach of using Protocol Σ_{std} shows that n random entries are sufficient. Note that $n < mk\log(n)$ is possible for large m. (In practice $mk\log(n) \ll n$.)

Second, we want to deal with *adversarial* $[\boldsymbol{A}]$. In the above sketch for zero-knowledge, we ignored a detail concerning the recursion. If it ever happens that in $[\boldsymbol{g}]$, for some $i \in \mathbb{M}_n$, the element g_i is zero, the distribution of $(\widehat{\boldsymbol{r}}, [\boldsymbol{u}_{-1}], [\boldsymbol{u}_1])$ is skewed and zero-knowledge fails. Note that $[\boldsymbol{g}]$ is reduced in each statement, so this can happen randomly. Thus, even the naive reduction is only statistically zero-knowledge. If $[\boldsymbol{A}]$ is chosen adversarially, it may be so that this failure case always (or often) happens. Making the definition of $\mathbb{M}_n$ dynamic and depend on $[\boldsymbol{A}]$ is inconvenient and hard. Our choice is therefore to act "dually" to commitment-extension. Remember that a commitment-extension adds a row to $[\boldsymbol{A}]$ so that $[\boldsymbol{A}]$ is "computationally injective". In contrast, we will, very roughly, add columns to $[\boldsymbol{A}]$, to ensure that $[\boldsymbol{A}]$ is surjective. Our concrete approach is detailed below.

Dealing with general $[\boldsymbol{A}]$. Our proof system separates the masking randomness from the actual witness and is a linear combination of multiple protocol instances of $\mathsf{LMPA}_{\mathsf{noZK}}$: The actual protocol for $[\boldsymbol{A}] =: [\boldsymbol{H}^{(0)}]$, and protocols for $[\boldsymbol{H}^{(i)}]$, $i = 1, \ldots, m$, where $[\boldsymbol{H}^{(i)}]$ essentially contains a Pedersen commitment key in the i-th row and is zero otherwise.

To keep things simple, we let $m = 1$ in the following discussion. Intuitively, we want to run a "randomness-extended" protocol for $[\boldsymbol{B}] = [\boldsymbol{A}|\boldsymbol{H}]\binom{\boldsymbol{w}}{\boldsymbol{r}}$. The intuition is that $\boldsymbol{r}$ will randomise all $[u_{\pm 1}]$'s (because $[\boldsymbol{H}]$ is not adversarial). Unfortunately, this intuition is wrong: $[u_1] = [\boldsymbol{H}]\boldsymbol{w}$ is certainly not zero-knowledge. The problem is how $\mathsf{LMPA}_{\mathsf{noZK}}$ divides the statement. Appropriate shuffling of $[\boldsymbol{B}]$ and $\binom{\boldsymbol{w}}{\boldsymbol{r}}$ would solve this. Instead, we work with a linear combination of $\mathsf{LMPA}_{\mathsf{noZK}}$ instances.

More precisely, we run *two* arguments, one for $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}']$ and one for $[\boldsymbol{H}]\boldsymbol{r} = [\boldsymbol{t}'']$. The messages $[u_{-1}]$ and $[u_1]$ are the sums of the messages which individual protocols would send, e.g. $[u_{-1}] = [\boldsymbol{A}_2]\boldsymbol{w}_1 + [\boldsymbol{H}_2]\boldsymbol{r}_1$. Concretely

$$\begin{bmatrix} u'_{-1} \\ u'_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1\boldsymbol{w}_2 \\ \boldsymbol{A}_2\boldsymbol{w}_1 \end{bmatrix}, \qquad \begin{bmatrix} u''_{-1} \\ u''_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{H}_1\boldsymbol{r}_2 \\ \boldsymbol{H}_2\boldsymbol{r}_1 \end{bmatrix}, \qquad \begin{bmatrix} u_{-1} \\ u_1 \end{bmatrix} = \begin{bmatrix} u'_{-1} \\ u'_1 \end{bmatrix} + \begin{bmatrix} u''_{-1} \\ u''_1 \end{bmatrix}$$

This ensures that the $[u''_{\pm 1}]$ are uniformly random in every round, because $[u''_{\pm 1}]$ is. In the base case of the recursion, i.e. small n, the

prover proves $[A]w + [H]r = [t]$ in zero-knowledge, using (for concreteness) Protocol Σ_{std}.

To keep our protocol modular and comprehensible, we split it into two steps.

Protocol B.1 ($\mathsf{LMPA}_{\text{almSnd}}$*).* The following is a protocol to prove $\exists w\colon [t^{(0)}] = [A]w$, using testing distributions χ_{m+1} resp. $\widetilde{\chi}_{2k-1}$ (resp. $\chi^{(\beta)}$). with $\widetilde{\chi}_3$, $\chi^{(\beta)}$ as in Protocol $\mathsf{LMPA}_{\text{noZK}}$. Furthermore, we require that $x \leftarrow \chi_{m+1}$ satisfies $x_i \neq 0$ for all i.

Common input is $([A], [t^{(0)}]) \in \mathbb{G}^{m\times n} \times \mathbb{G}^n$ and some $h \in \mathbb{G}^n$ (typically derived from the CRS when this protocol is used as a subprotocol). We assume $n = 2^\ell > 4$. Moreover, we let $[H^{(i)}] \in \mathbb{G}^{m\times n}$ for $i = 1, \dots, m$, be defined as the matrix with $[h]$ in the i-th row and zeroes elsewhere, i.e. $[H^{(i)}] = e_i[h]$. We use a superscript 0, e.g. $[H^{(0)}] := [A]$, for terms related to $[A]$. The prover's witness is some $w \in \mathbb{F}_p^n$ (also written $r^{(0)}$).

- $\mathcal{P} \to \mathcal{V}$: (Step 1: Prepare masking.) Pick $r^{(i)} \leftarrow \mathbb{M}_n \leq \mathbb{F}_p^n$ and compute $[t^{(i)}] = [H^{(i)}]r^{(i)}$. Send $[t^{(i)}]$ for $i = 1, \dots, m$.
- $\mathcal{V} \to \mathcal{P}$: (Step 2: Random linear combination.) Pick and send $x \leftarrow \chi_{m+1}$. The statement we prove is now effectively
$$[A|H^{(1)}|\dots|H^{(m)}]\begin{pmatrix} x_0 w \\ x_1 r^{(1)} \\ \vdots \end{pmatrix} = [t] := \sum_i x_i [t^{(i)}].$$
For simplicity, the prover redefines $r^{(i)} := x_i r^{(i)}$ for $i = 0, \dots, m$.
- $\mathcal{P} \to \mathcal{V}$: (Step 3: Begin the shrinking AoK.) Let $[H^{(i)}] = [H_1^{(i)}, H_2^{(i)}]$ with $H_j^{(i)} \in \mathbb{G}^{m\times n/2}$. Compute $[u_\ell] = \sum_{i=0}^m [u_\ell^{(i)}]$, where $[u_\ell^{(i)}]$ is computed as usual, i.e. $[u_\ell^{(i)}] = \sum_{j-i=\ell} [H_\ell^{(i)}] r_\ell^{(i)}$. Send $[u_\ell]$ for $\ell = \pm 1$.
- $\mathcal{V} \to \mathcal{P}$: Pick $z \leftarrow \widetilde{\chi}_3$ (with associated x, y). Send (x, y, z).
- $\mathcal{P} \to \mathcal{V}$: As in $\mathsf{LMPA}_{\text{noZK}}$, compute $w = x^\top \vec{w} = \sum_j x_j w_i$ and $\widehat{r}^{(i)} = x^\top \vec{r}^{(i)} = \sum_j x_j r_j^{(i)}$ and $[\widehat{A}] = x^\top [\vec{A}] = \sum_j x_j [A_j]$, $[\widehat{H}^{(i)}] = \sum_j x_j [H_j^{(i)}]$, and $[t] = z^\top u = \sum_\ell z_j u_\ell$, for the reduced statement (which $\mathcal{V}$ also computes).
If $n > 4$, engage recursively in the AoK for this statement, i.e. goto Step 3. If $n \leq 4$, engage in (for concreteness) Protocol Σ_{std} to prove the statement.

It is easy to check that Protocol B.1 is complete.

Lemma B.2. *Protocol* $\mathsf{LMPA}_{\text{almSnd}}$ *has μ-special soundness (with $\mu = (m+1, 4, \dots, 4, 2)$) for finding a preimage $\vec{v} \in (\mathbb{F}_p^n)^m$ (unconditionally) with $[A|H^{(1)}|\dots|H^{(m)}]\begin{pmatrix} v_0 \\ \vdots \\ v_m \end{pmatrix} = [t^{(0)}]$, or a non-trivial kernel element of $[A|H'^{(1)}|\dots|H'^{(m)}]$. Here, $[H'^{(i)}]$ consists only of the non-zero components of $[H^{(i)}]$. (It is easy to find non-trivial kernel elements if $[h]$ has zeroes, so we exclude them) The protocol inherits short-circuit extraction with $\mu' = (m+1, 2, \dots, 2, 2)$.*

Note Lemma B.2 *does not assert* a witness $w \in \mathbb{F}_p^n$ for $[A]w = [t^{(0)}]$. That will be assured in follow-up step.

Proof. We only sketch the proof. Let *tree*$_\mu$ be a good μ-tree of transcripts. First of all, we can extract the base subprotocol of Step 3. Using these witnesses, we can extract the linearly combined argument essentially as in Lemma 3.6.[12] Now we extract Step 2. From Step 3, we have $m+1$ preimages $\vec{v}_i \in (\mathbb{F}_p^m)^n$ with $[A|H^{(1)}|\dots|H^{(m)}]\vec{v}_i = [T]x_i$ where $[T] = [t^{(0)}, \dots, t^{(m)}]$. Arrange matrices $V = (\vec{v}_0, \dots, \vec{v}_m)$ and X as usual. (V corresponds to $\widetilde{W}$ from Lemma 3.6.) We find $[A|H^{(1)}|\dots|H^{(m)}]V_i = [t]X$. Multiplying with X^{-1}, we find preimages for each $[t^{(i)}]$, in particular a preimage for $[t^{(0)}]$. □

To prove zero-knowledge of Protocol $\mathsf{LMPA}_{\text{almSnd}}$, we first show that the prover's messages $[u_\ell]$ in the recursive steps are almost always uniformly distributed. This yields statistical HVZK via straightforward simulation.

As a preparation, note following (easy) linear algebra facts:

Lemma B.3. *Consider Protocol B.1 (*$\mathsf{LMPA}_{\text{almSnd}}$*). Suppose that (at least) all components of $[h]$ in $\mathbb{M}_n$ are distributed uniformly random (and the rest may be 0). Suppose that for any $x \leftarrow \chi_{m+1}$ we have $x_i \neq 0$ for all i.*

Then, in this argument system, with probability about $O(\log_2(n)k)/p$ the vector U consisting of messages $[u_\ell]$ of all *recursive rounds is uniformly random. The randomness is over $[h]$, the challenges and the prover's randomness.*

We give a short proof intuition for the case $m = 1$. So we have $[A], [H] \in \mathbb{G}^{1\times n}$. Intuitively, we need 2 $\mathbb{F}_p$-elements of randomness in each round to mask $[u_{\pm 1}]$. Moreover, these two terms of randomness must be split so that one is in the first half r_1 of r, and one in the second half r_2, since $[u_{j-i}] = [H_i]r_j$. The masking sets $\mathbb{M}_n$ are built exactly as such, see Fig. 2. Moreover, to allow inductive reasoning, the masking sets are built in such a way that even when "removing" two terms of randomness (say $r_{1,0}$ and $r_{2,1}$), the sum $r'_1 + r'_2$ is distributed according to $\mathbb{M}_{n/2}$. Evidently, we need $x_i \neq 0$ to prevent loss of randomness by multiplication with 0. More precisely, we want surjectivity of the "transition map", $\begin{pmatrix} x_1 \mathrm{id}_n & x_2 \mathrm{id}_n \\ H_2 & 0 \\ 0 & H_2 \end{pmatrix}\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} \widehat{r} \\ u''_{-1} \\ u''_1 \end{pmatrix}$ when restricted to $\mathbb{M}_{2n} \leq \mathbb{F}_p^{2n}$ in each step. See [31] for a full proof.

Lemma B.4. *Protocol* $\mathsf{LMPA}_{\text{ZK}}$ *is ε-statistical zero-knowledge for $\varepsilon \in O(2\log_2(n))/p$.*

We sketch HVZK simulation: For a recursive step, the HVZK simulator picks $[u_\ell] \leftarrow \mathbb{G}^m$ for $\ell \neq 0$ and computes the uniquely defined $[u_0]$ which makes the verifier accept that round. For Step 1 note that $[t^{(i)}] = [e_i t^{(i)}]$ ($i \neq 0$) and hence $[t^{(0)}]$ and $[t]$ (which is $[u_0]$ of the last recursion) uniquely define all $[t^{(i)}]$. Since the messages $[u_\ell]$ are uniformly distributed in an honest execution with probability $O(2\log_2(n)2)/p$, our claim follows.

Now, we finish the protocol and ensure that extraction yields a witness w for $[A]w = [t]$ as we desired. For this, we use a dual testing distribution to ensure $v_i \overset{!}{=} 0$ for $i \geq 1$ (with notation as in Lemma B.2).

Protocol B.5 ($\mathsf{LMPA}_{\text{ZK}}$*).* The following is a protocol to prove $\exists w$: $[A]w = [t]$. We use Protocol B.1 ($\mathsf{LMPA}_{\text{almSnd}}$) as a subprotocol

[12] Indeed, after suitably permuting the columns of $[A|H^{(1)}|\dots|H^{(m)}]$, witness, and randomness, the exact same reasoning as in Lemma 3.6 works for the recursive step.

with the same testing distributions χ_{m+1} resp. $\widetilde{\chi}_3$ (resp. $\chi^{(\beta)}$). By $\chi^{\vee}_{\dim(\mathbb{M}_n)+1}$, we refer to the dual testing distribution of $\chi_{\dim(\mathbb{M}_n)+1}$ as in Definition A.2. In particular, we require that the first component x_0 of $\boldsymbol{x} \leftarrow \chi_{\dim(\mathbb{M}_n)+1}$ is always 1.

Common input is $([\boldsymbol{A}], [\boldsymbol{t}]) \in \mathbb{G}^{m\times n} \times \mathbb{G}^n$ We assume $n = 2^\ell > 4$. The prover's witness is some $\boldsymbol{w} \in \mathbb{F}_p^n$ (also written $\boldsymbol{r}^{(0)}$). The CRS contains randomly (independently) chosen $[\boldsymbol{q}] \leftarrow \mathbb{G}^{1\times\dim(\mathbb{M}_n)+1}$.

- $\mathcal{V} \to \mathcal{P}$: (Step 0: Setup of a "new" crs.) $\mathcal{V}$ picks and sends $\boldsymbol{M} := \boldsymbol{M}_{\boldsymbol{x}} \leftarrow \chi^{\vee}_{\dim(\mathbb{M}_n)+1}$ (as described in Definition A.2).
- Both parties compute $[\widetilde{\boldsymbol{h}}] := [\boldsymbol{q}]\boldsymbol{M} \in \mathbb{G}^{1\times\dim(\mathbb{M}_n)}$. They define $[\boldsymbol{h}] \in \mathbb{G}^n$ so that the components $\mathbb{M}_n \subseteq \{0,\dots,n-1\}$ of $[\boldsymbol{h}]$ correspond to $[\widetilde{\boldsymbol{h}}]$ (in order). All components of $[\boldsymbol{h}]$ not in $\mathbb{M}_n$ are set to 0. See Fig. 2 for a pictorial description of (non-)zero components of $[\boldsymbol{h}]$.
- $\mathcal{P} \leftrightarrow \mathcal{V}$: Engage in Protocol $\mathsf{LMPA}_{\mathsf{almSnd}}$ for $\exists \boldsymbol{w}\colon [\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$ with parameters (in particular $[\boldsymbol{h}]$) as above.

Lemma B.6. *Protocol* $\mathsf{LMPA}_{\mathsf{ZK}}$ *has μ-special soundness (with $\mu = (\dim(\mathbb{M}_n)+1, m+1, 4, \dots, 4, 2)$) for finding a witness $\boldsymbol{w} \in \mathbb{F}_p^n$ with $[\boldsymbol{A}]\boldsymbol{w} = [\boldsymbol{t}]$, or a non-trivial kernel element of $[\boldsymbol{A}|\boldsymbol{e}_1^\top\boldsymbol{q}|\dots|\boldsymbol{e}_m^\top\boldsymbol{q}]$ (equivalently $[\boldsymbol{A}|\operatorname{diag}(\boldsymbol{q},\dots,\boldsymbol{q})]$). Moreover, the protocol has short-circuit extraction with $\mu' = (1, m+1, 2, \dots, 2, 2)$.*

There are reasons to suspect that $\mathsf{LMPA}_{\mathsf{ZK}}$ may have unconditional extraction, i.e. it is *proof* of knowledge. But we could not (dis)prove it yet.

Proof. By extracting $\mathsf{LMPA}_{\mathsf{almSnd}}$ i.e. applying Lemma B.2, we can find preimages $\vec{\boldsymbol{u}} \in (\mathbb{F}_p^n)^{m+1}$. (Also, we inherit short-circuit and unconditional extraction.) Let $[\boldsymbol{h}]$ and $[\boldsymbol{H}^{(i)}] = \boldsymbol{e}_i[\boldsymbol{h}]$ be as constructed in the protocols.

For simplicity, we first consider the case $m = 1$ and remove all 0-columns of $[\boldsymbol{H}]$. In other words, we consider $[\boldsymbol{A}|\boldsymbol{q}\boldsymbol{M}] \in \mathbb{G}^{1\times n+\dim(\mathbb{M})}$.

We know (i.e. extracted) some $\boldsymbol{w} \in \mathbb{F}_p^n, \boldsymbol{v} \in \mathbb{F}_p^{\dim(\mathbb{M})}$ such that $[\boldsymbol{A}]\boldsymbol{w} + [\boldsymbol{H}]\boldsymbol{v} = [t]$. We have to show that $[\boldsymbol{A}]\boldsymbol{w} = [t]$, or we find a non-trivial element in the kernel of $[\boldsymbol{A}|\boldsymbol{q}]$. In the case that $[\boldsymbol{H}]\boldsymbol{v} = 0$, $\boldsymbol{w}$ is the witness we want. So suppose that $[\boldsymbol{H}]\boldsymbol{v} \neq 0$. In that case, we guarantee short-circuit extraction. So, suppose we have $\dim(\mathbb{M})+1$ transcripts with "independent" challenge matrices $\boldsymbol{M}_i$. (Remember that this means $\bigcap_{i=0}^{\dim(\mathbb{M})} \operatorname{im}(\boldsymbol{M}_i) = \{0\}$, which is equivalent to $\boldsymbol{x}_i$ being linearly independent since $\boldsymbol{M}_i = \boldsymbol{M}_{\boldsymbol{x}_i}$.) By subtracting the 0-th witness from the i-th witness, we find $[\boldsymbol{A}](\boldsymbol{w}_i - \boldsymbol{w}_0) + [\boldsymbol{q}](\boldsymbol{M}_i\boldsymbol{v}_i - \boldsymbol{M}_0\boldsymbol{v}_0)$. Thus, if $\boldsymbol{M}_i\boldsymbol{v}_i - \boldsymbol{M}_0\boldsymbol{v}_0 \neq \boldsymbol{0}$, we obtain a non-trivial kernel element. The only case where we do not obtain a non-trivial kernel element of $[\boldsymbol{A}|\boldsymbol{q}\boldsymbol{M}]$ is, if for all i we have $\boldsymbol{M}_i\boldsymbol{v}_i = \boldsymbol{M}_0\boldsymbol{v}_0 =: \boldsymbol{u}$. However, this implies that $\boldsymbol{u} \in \bigcap_{i=0}^{\dim(\mathbb{M})} \operatorname{im}(\boldsymbol{M}_i)$. But, by assumption we have $\bigcap_{i=0}^{\dim(\mathbb{M})} \operatorname{im}(\boldsymbol{M}_i) = \{0\}$. Thus, the bad case is impossible.

For general m, we have $\operatorname{diag}(\boldsymbol{Q},\dots)$ and $\operatorname{diag}(\boldsymbol{M},\dots)$ instead of $\boldsymbol{Q}$ and $\boldsymbol{M}$. We obtain $\binom{\boldsymbol{w}}{\boldsymbol{v}}$ from $\mathsf{LMPA}_{\mathsf{almSnd}}$ with $[\boldsymbol{A}|\operatorname{diag}(\boldsymbol{H},\dots)]\binom{\boldsymbol{w}}{\boldsymbol{v}} = [\boldsymbol{t}]$. Evidently, $\bigcap_{i=0}^{\dim(\mathbb{M})} \operatorname{im}(\operatorname{diag}(\boldsymbol{M}_i,\dots,\boldsymbol{M}_i)) = \bigcap_{i=0}^{\dim(\mathbb{M})} \operatorname{im}(\boldsymbol{M}_i)^m = \{\boldsymbol{0}\}$. Thus, our claim follows analogously. □

Corollary B.7. *Protocol* $\mathsf{LMPA}_{\mathsf{ZK}}$ *has ε-statistical HVZK with $\varepsilon \in O(2\log_2(n))/p$.*

Proof. This is immediate from Lemma B.4. □

C BATCH PROOFS OF KNOWLEDGE

By applying the "linear combination of protocols" technique, to multiple "trivial proofs of knowledge" (c.f. Fig. 2) we obtain batch verification of statements $([\boldsymbol{A}], [\boldsymbol{t}_i])$, $i = 1,\dots,N$, i.e. the setting of [38], in a straightforward way.

Protocol C.1. The following is a protocol to prove: $\exists \boldsymbol{w}_i\colon [\boldsymbol{A}]\boldsymbol{w}_i = [\boldsymbol{t}_i]$ for $i = 1,\dots,N$. Let χ_{N+1} be a testing distribution for challenges, such that $\boldsymbol{x} \leftarrow \chi_{N+1}$ has $x_{N+1} \neq 0$ always. Common input is $([\boldsymbol{A}], ([\boldsymbol{t}_i])_i) \in \mathbb{G}^{m\times n} \times \mathbb{G}^n$. The prover's witness are some $\boldsymbol{w}_i \in \mathbb{F}_p^n$.

- $\mathcal{P} \to \mathcal{V}$: Pick $\boldsymbol{r} \leftarrow \mathbb{F}_p^n$ and let $[\boldsymbol{a}] = [\boldsymbol{A}]\boldsymbol{r}$. Send $[\boldsymbol{a}] \in \mathbb{G}^m$.
- $\mathcal{V} \to \mathcal{P}$: Pick $\boldsymbol{x} \leftarrow \chi_{N+1}$. Send $\boldsymbol{x} \in \mathbb{F}_p$.
- $\mathcal{P} \to \mathcal{V}$: Compute $\boldsymbol{z} = \boldsymbol{x}^\top \begin{pmatrix} \boldsymbol{w}_1 \\ \dots \\ \boldsymbol{w}_N \\ \boldsymbol{r} \end{pmatrix} = \sum_{i=1}^N x_i\boldsymbol{w}_i + x_{N+1}\boldsymbol{r}$. Send $\boldsymbol{z} \in \mathbb{F}_p^n$.
- $\mathcal{V}$: Check $[\boldsymbol{A}]\boldsymbol{z} \stackrel{?}{=} \sum_{i=1}^N x_i[\boldsymbol{t}_i] + x_{N+1}[\boldsymbol{a}]$, and accept/reject if true/false.

Lemma C.2. *Protocol C.1 is a HVZK-PoK for $\exists w\colon [\boldsymbol{t}] = [\boldsymbol{A}]\boldsymbol{w}$. It is perfectly complete, has perfect HVZK and is $(N+1)$-special sound.*

Proof. **Completeness** is straightforward. **Extraction** uses $N+1$ accepting transcripts $([\boldsymbol{a}], \boldsymbol{x}_j, \boldsymbol{z}_j)$. Let $[\boldsymbol{T}] := [\boldsymbol{t}_1,\dots,\boldsymbol{t}_N,\boldsymbol{a}]$ and $\boldsymbol{Z}$, $\boldsymbol{X}$ be appropriate matrices built from the $N+1$ transcripts. Since $[\boldsymbol{A}]\boldsymbol{Z} = \boldsymbol{X}$, we find $(\boldsymbol{w}_1,\dots,\boldsymbol{w}_N,\boldsymbol{r}) := \boldsymbol{Z}\boldsymbol{X}^{-1}$ is a valid witness. For **HVZK** note that $x_{N+1} \neq 0$. Hence $\boldsymbol{z}$ is uniformly distributed for any honest execution. Thus, we can pick $\boldsymbol{z} \leftarrow \mathbb{F}_p^m$ and let $[\boldsymbol{a}] := [\boldsymbol{A}]\boldsymbol{z} - [\boldsymbol{T}]\boldsymbol{x}$ as usual. □

D AN EFFICIENT PROOF OF CORRECTNESS OF A SHUFFLE

A proof of correctness of a shuffle is a proof that two (multi-)sets of ciphertexts decrypt to the same multi-set of plaintexts. This is especially interesting in settings with rerandomisable ciphertexts, because the "shuffling party" does not need to decrypt. For electronic voting, a shuffle achieves a certain unlinkability between the originally encrypted votes, and the (in a final step) decrypted votes, while the proofs of correctness of the shuffle ensure that the voting result is unaffected.

With our tools, it is possible to prove the correctness of a shuffle in logarithmic communication for ElGamal ciphertexts in a very naive manner. Namely, we commit to a permutation matrix (as part of $\boldsymbol{w}$) and rerandomisation randomness for the ElGamal ciphertexts (also part of $\boldsymbol{w}$). Then we prove that $[\boldsymbol{A}]\boldsymbol{w} = [\vec{\boldsymbol{c}}]$, where $[\boldsymbol{A}]$ is constructed from the old ciphertexts and the ElGamal public key, and $[\vec{\boldsymbol{c}}]$ is the vector of shuffled ciphertexts. We also add a proof that (the relevant part of) $\boldsymbol{w}$ commits to a permutation matrix, as sketched in Section 3.5. This all neatly fits into our framework, giving a logarithmic size proof overall. However, there is a huge drawback: The size of the permutation matrix, hence $\boldsymbol{w}$, is N^2 for N ciphertexts. Thus, the *computation* grows quadratically in N. This is unacceptable in practice.

D.1 Adapting the shuffle argument of Bayer–Groth

The shuffle argument of Bayer and Groth [5] is built from two subarguments, a "product argument" and a "multi-exponentiation argument". A generic proof of security is given in [5, Theorem 5]. The former argument can be instantiated by $\mathsf{QESA}_{\mathrm{ZK}}$, or more precisely, $\mathsf{QESA}_{\mathrm{Copy}}$. The latter argument can be instantiated by $\mathsf{LMPA}_{\mathrm{ZK}}$. Since our arguments have logarithmic communication and need linearly many exponentiations, so does the resulting shuffle argument. We give a more detailed instantiation below.

- CRS: $\mathrm{ck} = (\mathrm{ck}_Q, \mathrm{ck}_L)$, where $\mathrm{ck}_Q = ([\boldsymbol{g}', \boldsymbol{g}'', Q])$ is the commitment key for $\mathsf{QESA}_{\mathrm{ZK}}$ and $\mathrm{ck}_L = [\boldsymbol{h}]$ is the commitment key for $\mathsf{LMPA}_{\mathrm{ZK}}$ (or empty if a simple zero-knowledge LMPA version is used). Here $[\boldsymbol{g}'] \in \mathbb{F}_p^n$, where $n \geq N+2$ is a (suitably large) power of 2. Note that our commitment keys consist of random group elements.
- Common input: Old and new ciphertexts $[\mathrm{ct}_i^{\mathrm{old}}]$, $[\mathrm{ct}_i^{\mathrm{new}}] \in \mathbb{G}^2$ for $i = \{0, \dots, N-1\}$ and ElGamal public key $[\mathrm{pk}] \in \mathbb{G}^2$.
- Prover's witness: The random permutation $\boldsymbol{\pi} \in \{0, \dots, N-1\}^N$ and rerandomisation randomnesses $\rho_i \in \mathbb{F}_p$ such that $[\mathrm{ct}_i^{\mathrm{new}}] = [\mathrm{ct}_{\pi_i}^{\mathrm{old}}] + \rho_i[\mathrm{pk}]$. (Note that $\mathsf{Enc}(0; \rho_i) = \rho_i[\mathrm{pk}]$ for ElGamal.)
- $\mathcal{P} \to \mathcal{V}$: Compute and send the commitment $[c_{\boldsymbol{\pi}}]$ to π:

$$[c_{\boldsymbol{\pi}}] = \mathsf{Com}_{\mathrm{ck}_Q}(\boldsymbol{\pi}; 0, r_{\boldsymbol{\pi}}) = [g'_1 \mid g'_2, \dots, g'_{N+1} \mid g'_{N+2}, \dots, g'_{n-2} \mid g'_{n-1}, g'_n] \begin{pmatrix} 0 \\ \hline \boldsymbol{\pi} \\ \hline \mathbf{0} \\ \hline 0 \\ r_{\boldsymbol{\pi}} \end{pmatrix}$$

  (Remember that $[\boldsymbol{g}'_{n-1}]$ and $[\boldsymbol{g}'_n]$ are reserved for randomness in $\mathsf{QESA}_{\mathrm{ZK}}$ commitments, and $[\boldsymbol{g}_1]$ is also reserved (for the constant 1).)
- $\mathcal{V} \to \mathcal{P}$: Send $\boldsymbol{x} = (x_0, \dots, x_{N-1}) \leftarrow \chi_N$.
- $\mathcal{P} \to \mathcal{V}$: Send $[c_{\boldsymbol{y}}] = \mathsf{Com}_{\mathrm{ck}_Q}(\boldsymbol{y}; 0, r_{\boldsymbol{y}})$, where $[\boldsymbol{y}] \coloneqq \boldsymbol{\pi}(\boldsymbol{x}) = (\boldsymbol{x}_{\pi_i})_i = (x_{\pi_0}, \dots, x_{\pi_{N-1}})$.
- $\mathcal{V} \to \mathcal{P}$: Send $\zeta, z \leftarrow \mathbb{F}_p$.
- $\mathcal{P} \leftrightarrow \mathcal{V}$: Prove following statements using (logarithmic communication) sub-protocols $\mathsf{QESA}_{\mathrm{Copy}}$ and $\mathsf{LMPA}_{\mathrm{ZK}}$:
  - $[c_{\boldsymbol{\pi}}]$ **is a permutation and** $[c_{\boldsymbol{y}}]$ **is a commitment to** $\boldsymbol{\pi}(\boldsymbol{x})$**:** The prover shows (in zero-knowledge) that

    $$\prod_{i=0}^{N-1} (\zeta \pi_i + y_i - z) = \prod_{i=0}^{N-1} (\zeta i + x_i - z).$$

    Note that $\zeta[c_{\boldsymbol{\pi}}] + [c_{\boldsymbol{y}}]$ is a commitment to $\zeta\boldsymbol{\pi} + \boldsymbol{y}$, which can be used for $\mathsf{QESA}_{\mathrm{ZK}}$, or more precisely, $\mathsf{QESA}_{\mathrm{Copy}}$. Also note that the right-hand side is computable from public information.
  - $[\vec{\mathrm{ct}}^{\mathrm{new}}]$ **is a rerandomised permutation of** $[\vec{\mathrm{ct}}^{\mathrm{old}}]$**:** The prover shows (in zero-knowledge) that

    $$\sum_i [\mathrm{ct}_i^{\mathrm{old}}] y_i + [\mathrm{pk}] \sum_i \rho_i x_i = \sum_i [\mathrm{ct}_i^{\mathrm{new}}] x_i.$$

    This fits into our matrix multiplication proofs (with witness $\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x}^\top \rho \end{pmatrix} \in \mathbb{F}_p^{N+1}$). Concretely, the prover commits to $\sigma \coloneqq \boldsymbol{x}^\top \rho$ via $[c_\sigma] = \mathsf{Com}_{\mathrm{ck}_Q}(\binom{\mathbf{0}}{\sigma}; r_\sigma, 0) = [g'_{N+2}, g'_{n-1}]\binom{\sigma}{r_\sigma}$ for $r_\sigma \leftarrow \mathbb{F}_p$. He sends c_σ to the verifier and engages in a $\mathsf{LMPA}_{\mathrm{ZK}}$ protocol for

    $$\left[\begin{array}{c|c|cc} g'_2, \dots, g'_{N+1} & g'_{N+2} & g'_{n-1} & g'_n \\ g'_2, \dots, g'_{N+1} & 0 & 0 & g'_n \\ \mathbf{0} & g'_{N+2} & g'_{n-1} & 0 \\ \mathrm{ct}_0^{\mathrm{old}} \dots \mathrm{ct}_{N-1}^{\mathrm{old}} & \mathrm{pk} & 0 & 0 \end{array}\right] \left(\begin{array}{c} \boldsymbol{y} \\ \hline \sigma \\ \hline r_\sigma \\ r_{\boldsymbol{y}} \end{array}\right) = \begin{bmatrix} c_{\boldsymbol{y}} + c_\sigma \\ c_{\boldsymbol{y}} \\ c_\sigma \\ \boldsymbol{u} \end{bmatrix}$$

    where $[\boldsymbol{u}] \coloneqq \sum x_i [\mathrm{ct}_i^{\mathrm{new}}]$. The top row is added so one can run $\mathsf{LMPA}_{\mathrm{batch}}$, reducing to a $2 \times n$ matrix. Since $[\boldsymbol{g}']$ has hard kernel relation, so has $[\boldsymbol{A}]$. (This is a "commitment-extension", see Remark 3.4.) Also note that this LMPA proof ensures the requirements of $\mathsf{QESA}_{\mathrm{Copy}}$ on the opening of $[c_{\boldsymbol{y}}]$, hence no additional subprotocol $\mathcal{S}$ is necessary in this instance.

Honest verifier zero-knowledge of this protocol follows from honest verifier zero-knowledge of the subprotocols. Soundness (and extraction) follows from soundness (and extraction) of the subprotocols.

In [5], intuition and a detailed security argument is given. Despite our minor modifications, their proof adapts seamlessly to our setting.

A rough efficiency estimate of our scheme is $30N$ exponentiations for the prover and $10N$ exponentiations for the verifier. These are roughly twice the numbers of [5], when trading interaction for efficiency. However [5] has $O(\sqrt{N})$ size proofs, while we have $O(\log(N))$ size proofs.