RoK and Roll – Verifier-Efficient Random Projection for $\tilde{O}(\lambda)$-size Lattice Arguments

(Full Version)

Michael Klooß[1], Russell W. F. Lai[2], Ngoc Khanh Nguyen[3], and Michał Osadnik[2]

[1] KASTEL, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] Aalto University, Espoo, Finland
[3] King's College London, London, UK

Abstract. Succinct non-interactive arguments of knowledge (SNARKs) based on lattice assumptions offer a promising post-quantum alternative to pairing-based systems, but have until now suffered from inherently quadratic proof sizes in the security parameter. We introduce RoK and Roll, the first lattice-based SNARK that breaks the quadratic barrier, achieving communication complexity of $\tilde{O}(\lambda)$ together with a succinct verification time. The protocol significantly improves upon the state of the art of fully-succinct argument systems established by "RoK, Paper, SISsors" (RPS) [ASIACRYPT'24] and hinges on two key innovations, presented as reductions of knowledge (RoKs):

- *Structured random projections:* We introduce a new technique for structured random projections that allows us to reduce the witness dimensions while approximately preserving its ℓ_2 norm and maintaining the desired tensor structure. In order to maintain succinct communication and verification, the projected image is further committed and adjoined to the original relation. This procedure is recursively repeated until dimension of the intermediate witness becomes $\mathsf{poly}(\lambda)$, i.e. independent of the original witness length.
- *Unstructured random projection:* When the witness is sufficiently small, we let the unstructured projection (over coefficients $\mathbb{Z}_q$) be sent in plain, as in LaBRADOR [CRYPTO'23]. We observe, however, that the strategy from prior works to immediately lift the projection claim to $\mathcal{R}_q$, and into our relation, would impose a quadratic communication cost. Instead, we gradually batch-and-lift the projection a the tower of intermediate ring extensions. This reduces the communication cost to $\tilde{O}(\lambda)$ while maintaining a succinct verification time.

These two techniques, combined with existing RoKs from RPS, yield a succinct argument system with communication complexity $\tilde{O}(\lambda)$ and succinct verification for structured linear relations.

1 Introduction

Succinct Non-interactive Arguments of Knowledge (SNARKs) [Kil92, Mic94] are a powerful cryptographic primitive with applications in verifiable computation, digital identity and e-wallet, and blockchain. They enable a prover to prove knowledge of a witness $\mathbf{w}$ of length m, which satisfies certain properties, with communication complexity significantly smaller than m. Recently, there has been tremendous progress in constructing succinct arguments based on lattice assumptions [AFLN24, BS23, CMNW24, KLNO24, NS24], which offer attractive proof sizes and prover running time, making them a promising post-quantum alternative to hash-based SNARKs [ACFY25, ACFY24, BBHR18, ZCF24].

In terms of efficiency, lattice-based SNARKs still lag behind the pre-quantum counterparts. Indeed, the pairing- and group-based state-of-the-art constructions [Gro16, Lip24, BCC+16, BBB+18] achieve proofs of size $\tilde{O}(\lambda)$ bits, where λ is denoted as the security parameter[4]. Despite recent advancements in lattice-based SNARKs showing promise, the newest constructions are ultimately stuck with $\tilde{O}(\lambda^2)$ communication complexity. Hence, asymptotically breaking the "quadratic barrier" in the lattice setting[5] and obtaining comparable concrete efficiency remains an open problem.

[4] We use the $\tilde{O}$-notation to hide the polylogarithmic factors in λ and m for clarity.
[5] We emphasise that the quadratic barrier is also present in hash-based SNARKs [ACFY25].

Lattice relation. At the core of lattice-based succinct arguments lies an efficient proof of knowledge of a low-norm witness vector $\mathbf{w} \in \mathcal{R}^m$ such that

$$\mathbf{F}\mathbf{w} = \mathbf{y} \bmod q \quad \text{and} \quad \|\mathbf{w}\| \leq \beta$$

where $\mathcal{R} = \mathbb{Z}[\zeta]$ is the ring of integers of the $\mathfrak{f}$-th cyclotomic field $\mathcal{K} := \mathbb{Q}(\zeta)$ of degree $\varphi = \varphi(\mathfrak{f})$. Although the linear relation can be proven efficiently by adapting ideas from Bulletproofs [BCC+16, BBB+18] and FRI [BBHR18] to the lattice setting, the main difficulty lies in proving the latter part, which we call the "shortness proof". There have been two main approaches in the literature to prove the shortness of the witness vector $\mathbf{w}$, described as follows.

1.1 Prior Works

Subtractive sets. The first one relies on so-called subtractive sets [AL21, ACK21, BLNS20, CLM23, FMN24, CMNW24, KLNO24]. A set $\mathcal{C} \subset \mathcal{R}$ is called subtractive if for any two distinct elements $c, c' \in \mathcal{C}$, the inverse of $c - c'$ exists over the ring $\mathcal{R}$ and is of low-norm. The motivation for subtractive sets comes from the fact that when arguing knowledge soundness (e.g. via special soundness) of lattice-based protocols, the extracted witness vector $\mathbf{w}$ is often defined as

$$\mathbf{w} := \frac{\mathbf{z} - \mathbf{z}'}{c - c'}$$

where $\mathbf{z}, \mathbf{z}'$ are short vectors, which are parts of the extracted transcripts, and $c \neq c'$ are taken from a challenge set $\mathcal{C}$ that is subtractive. Hence, we can deduce that $\mathbf{w}$ itself is also short (but slightly larger than $\mathbf{z}, \mathbf{z}'$).

The soundness error of the protocols, which follow this methodology, is asymptotically inverse to the cardinality of the set, i.e. $1/|\mathcal{C}|$ – hence ideally one would like to pick $\mathcal{C}$ of a large size. However, in the p^k-th cyclotomic ring all subtractive sets can have only $p = \mathsf{poly}(\lambda)$ number of elements (see e.g. [AL21]). Similarly, in the non-prime-power case, the only identified subtractive set [KLNO24] has size $O(\mathfrak{f})$, which is still only polynomial in λ.

As a consequence, some form of repetition is needed to obtain $2^{-\lambda}$ soundness error, which increases the proof size by a factor of $\lambda/\log\lambda$. Combined with the fact that a single ring element is of size $\Omega(\lambda \cdot \log q)$ bits to ensure the hardness of the Module/Ring-SIS assumption, the total asymptotic proof size can be bounded below by $\tilde{O}(\lambda^2)$ bits. Nevertheless, this strategy allows to build argument systems with succinct verification time, i.e. $\mathsf{poly}(\lambda, \log m)$ [AFLN24, CMNW24, KLNO24].

Random projections. The second approach moves away from subtractive sets, and relies on so-called random projections [BL17, LNS21, LNP22, BS23]. Namely, the verifier starts by sending a random *integer* matrix $\mathbf{J} \in \mathbb{Z}^{r \times m\varphi}$ with small entries, such as $\{-1, 0, 1\}$. The prover then outputs the image of the random projection $\mathbf{v} := \mathbf{J}\mathsf{cf}(\mathbf{w}) \bmod q \in \mathbb{Z}_q^r$, where $\mathsf{cf}(\mathbf{w}) \in \mathbb{Z}^{m\varphi}$ is the standard coefficient embedding of the witness vector $\mathbf{w}$. The verifier can check whether $\mathbf{v}$ is of low-norm. Finally, the prover needs to prove well-formedness of $\mathbf{v}$, namely

$$\mathbf{J}\mathsf{cf}(\mathbf{w}) = \mathbf{v} \bmod q. \tag{1}$$

From the correctness point of view, we use the variant of Johnson-Lindenstrauss Lemma [JL84] (and more recent variations [Ach03, FL20, HV11, GHL22, Mat08]) to argue that with high probability over the choice of $\mathbf{J}$ we have $\|\mathbf{v}\| \approx \|\mathbf{w}\|$. To argue knowledge soundness, however, one relies on the crucial fact that if the extracted *candidate witness* $\mathbf{w}$ had a large norm then with probability $\approx 2^{-r}$ over the choice of $\mathbf{J}$ so would $\mathbf{v}$, which leads to a contradiction.

Recent works showed how to efficiently prove arbitrary $\mathbb{Z}_q$-linear relations, such as (1), by utilising the structure of the underlying cyclotomic ring [LNP22, BS23, KLNO24]. Roughly speaking, proving (1) can be reduced to proving that certain r ring elements $a_0, \ldots, a_{r-1} \in \mathcal{R}$, related to $\mathbf{w}$, $\mathbf{v}$ and the rows of $\mathbf{J}$, satisfy

$$\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(a_i) = 0 \bmod q \quad \text{for } i = 0, \ldots, r-1 \tag{2}$$

or a similar relation. A naive solution would be for the prover to directly send all of $a_0, \ldots, a_{r-1}$ to the verifier, who could manually compute the trace of each ring element. However, if $r = \Omega(\lambda)$ to ensure

soundness, and each ring element is of size $\Omega(\lambda \cdot \log q)$ bits, then this message alone would be of quadratic size in λ. A cleverer solution introduced in [ENS20] is to use the homomorphic property of the trace function and first let the verifier output integer challenges $\gamma_0, \ldots, \gamma_{r-1} \leftarrow\!\$\, \mathbb{Z}_q$. Then, the prover returns the linear combination

$$a^\star := \gamma_0 a_0 + \ldots + \gamma_{r-1} a_{r-1} \bmod q.$$

Finally, the verifier can directly check whether $\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(a^\star) = 0 \bmod q$.

Compared to the subtractive set approach, there is no restriction on the challenge set $\mathcal{C}$, as long as it consists of short ring elements. Consequently, one can find an exponential-sized challenge set $\mathcal{C}$ and avoid repeating the expensive parts of the protocol. Unfortunately, the random projection methodology suffers from other asymptotic and concrete efficiency drawbacks. Firstly, the verifier runtime becomes linear in the witness length m, since ultimately it has to process the whole matrix $\mathbf{J}$. Secondly, proving $\mathbb{Z}_q$-linear relations described above only achieves $1/q$ soundness error, and therefore this procedure has to be repeated for soundness amplification. In total, the prover would output $k := \Omega(\lambda/\log q)$ such $a^\star$ field elements to achieve $\approx 2^{-\lambda}$ soundness error. Then, sending k full-fledged ring elements still makes the communication complexity at least quadratic in λ.

Alternatives. Very recently, a new approach to prove shortness has been proposed by Boneh and Chen [BC24] (and further improved in the follow-up [BC25]) in the context of folding schemes [KST22]. In short, the authors apply the sumcheck protocol [LFKN92] to prove that the committed vector has binary (NTT) entries[6]. We expect that this strategy could be adapted in the context of succinct arguments, and thus offer an alternative approach to prove shortness compared to subtractive sets and random projections.

1.2 Our Contributions

In this work, we present the first lattice-based argument system with succinct verification and communication complexity $\lambda \cdot \mathsf{polylog}(\lambda, m)$, thus breaking the quadratic barrier described above. Our protocol can be directly applied to prove knowledge of a vanishing SIS (vSIS) commitment opening defined in [CLM23], as well as to prove multilinear polynomial evaluation statements over $\mathcal{R}_q$. Combining both together immediately yields a multilinear polynomial commitment scheme, which is a key building block in constructing SNARKs via the Polynomial IOP framework [CHM+20, SZ22].

Our asymptotic results directly translate to a more concretely efficient scheme. Indeed, our protocol offers 6× smaller proof sizes (for 128-bit security) compared to the state-of-the-art lattice argument [KLNO24] with succinct verification. We provide an asymptotic comparison with prior works in Figure 1, and refer to Table 2 for more details on concrete parameters and proof sizes.

Towards our main result, we developed several technical contributions which could be of independent interest.

Structured random projections. We propose a novel shortness proof based on *structured* random projections , which unlike prior works [BS23], achieves succinct verification time. The core idea is, instead of the verifier picking a uniformly random short matrix $\mathbf{J} \in \mathbb{Z}^{r \cdot m\varphi}$, we let the matrix $\mathbf{J} := \mathbf{I} \otimes \mathbf{J}'$ carry certain tensor structure, where $\mathbf{J}'$ is a short random challenge matrix with $O(\lambda)$ columns and rows, and $\mathbf{I}$ is the identity matrix. Consequently, the verifier does not need to process the whole matrix $\mathbf{J}$, but instead only operates on $\mathbf{J}'$ – which, after a careful design of the folding procedure, enables succinct verification. We note that this strategy still inherits the benefits of using random projections, such as avoiding the subtractive restrictions on the challenge sets.

Efficient proof of trace via towers of rings. We show a new approach to prove trace-related statements, such as (2), using the tower structure of the underlying cyclotomic rings. As a result, we manage to reduce the communication complexity from $\tilde{O}(\lambda^2)$, as in the prior works [BS23], to $\tilde{O}(\lambda)$ bits.

Modular analysis and general results. We describe our protocols in a modular way, which can be instantiated to (practically) any choice of a security parameter and the underlying cyclotomic ring. In particular, our scheme can be instantiated natively over the widely-used in practice power-of-two cyclotomic rings, where the subtractive set methodology struggles the most[7].

[6] This can be easily generalised to prove infinity norm bounds.

[7] Indeed, the largest subtractive set in power-of-two cyclotomic has cardinality 2, e.g. consisting of binary challenges $\{0, 1\}$.

scheme	assumptions	shortness proof	verifier time	proof size
[BLNS20, AL21, ACK21]	R-SIS	SS	$O_\lambda(m)$	$\tilde{O}(\lambda^2)$
[CLM23]	vSIS	SS	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda^2)$
[BCS23]	M-SIS	SS	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda^2)$
[BS23]	M-SIS	RP	$O_\lambda(m)$	$\tilde{O}(\lambda^2)$
[FMN24]	PowerBASIS	SS	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda^2)$
[AFLN24]	M-SIS	SS	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda^2)$
[CMNW24]	SIS	SS	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda^2)$
[NS24]	M-SIS	RP	$O_\lambda(\sqrt{m})$	$\tilde{O}(\lambda^2)$
[KLNO24]	vSIS	SS	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda^2)$
This work	vSIS	RP	$\mathsf{poly}(\log m, \lambda)$	$\tilde{O}(\lambda)$

Fig. 1: Asymptotic efficiency of our commitment opening proof (in bits) and comparison with prior works which support succinct $\mathsf{poly}(\log m, \lambda)$ communication complexity. Here, λ is the security parameter and m is the length of the committed vector. For presentation, we use the $\tilde{O}(\cdot)$ (resp. $O_\lambda(\cdot)$) notation to hide the $\mathsf{polylog}(m, \lambda)$ (resp. $\mathsf{poly}(\log m, \lambda)$) factors. We denote by SS (resp. RP) the "subtractive sets strategy" (resp. "random projection strategy"). For each scheme, the proof size corresponds to the soundness error $\mathsf{poly}(\lambda, \log m) \cdot 2^{-\lambda}$. The SIS-related parameters were chosen with respect to the methodology from [MR09] for running BKZ on block size $b = O(\lambda)$. For [BCS23, CLM23, FMN24], which only achieve inverse-polynomial soundness in one-shot, we applied a standard soundness amplification by parallel-repeating the protocol by a factor of $O(\lambda/\log \lambda)$. We note that for [AFLN24, CMNW24, KLNO24], super-polynomial knowledge extraction runtime $O(m^{\log \lambda})$ is required.

2 Technical Overview

Throughout this work, we assume that $\mathcal{K} = \mathbb{Q}(\zeta)$ is a cyclotomic field with conductor $\mathfrak{f}$ and degree $\varphi = \varphi(\mathfrak{f}) = \tilde{O}(\lambda)$, and $\mathcal{R} = \mathbb{Z}[\zeta]$ is its ring of integers. Let q be a proof system modulus and $\mathcal{R}_q := \mathcal{R}/(q)$. We write $\tilde{O}(\lambda)$ (resp., $\tilde{O}(\lambda^2)$) to emphasize the linear (resp., quadratic) dependency on λ, while ignoring all polylogarithmic factors. Similar notation follows for other complexity classes.

2.1 Split-and-Fold Paradigm

Our starting point is the line of works [FMN24, AFLN24, CMNW24, KLNO24] which construct lattice-based interactive arguments with succinct verification using the "split-and-fold" paradigm [BBHR18]. Namely, consider the following core relation[8] Ξ^{lin}_m, which captures bounded-norm satisfiability for (structured) linear relations defined as

$$\mathbf{F}\mathbf{w}^{(0)} = \mathbf{y} \pmod q \quad \text{and} \quad \|\mathbf{w}^{(0)}\| \leq \beta,$$

where $\mathbf{w}^{(0)} \in \mathcal{R}_q^m$ is the initial witness vector of length $m := d^\mu$ and $\mathbf{F} \in \mathcal{R}_q^{n \times m}$ carries a row-tensor structure, i.e. each row $\mathbf{f}_i^\intercal := \mathbf{f}_{i,0}^\intercal \otimes \ldots \otimes \mathbf{f}_{i,\mu-1}^\intercal \in \mathcal{R}_q^m$ of $\mathbf{F}$ can be written as a tensor product of small vectors of length d each. Naturally, Ξ^{lin} captures the Vanishing SIS (vSIS) commitment opening relation [CLM23], as well as the multilinear polynomial evaluation relations. Hence, an efficient argument system for Ξ^{lin} immediately yields a multilinear polynomial commitment scheme based on the vSIS assumption.

As the name suggests, the split-and-fold strategy to prove relation Ξ^{lin} is a composition of subprotocols: Π^{split} and Π^{fold}.

Π^{split} protocol. The prover starts by sending the partial commitments

$$\mathbf{u}_i := \left(\mathbf{I}_d \otimes \left(\mathbf{f}_{i,1}^\intercal \otimes \ldots \otimes \mathbf{f}_{i,\mu-1}^\intercal\right)\right) \mathbf{w}^{(0)} \in \mathcal{R}_q^d \quad \text{for} \quad i = 0, 1, \ldots, n-1. \tag{3}$$

[8] We will omit the subscript if the witness dimension is clear from the context.

The verifier can already check the consistency between $\mathbf{y} := (y_0, \ldots, y_{n-1})$ and $\mathbf{u}_0, \ldots, \mathbf{u}_{n-1}$ by verifying:

$$y_i \stackrel{?}{=} \mathbf{f}_{i,0}^{\mathsf{T}} \mathbf{u}_i \quad \text{for} \quad i = 0, 1, \ldots, n-1.$$

Completeness holds directly via the mixed product property of tensor products. What is now left to do is to prove knowledge of $\mathbf{w}$ such that Equation 3 holds and $\|\mathbf{w}\| \leq \beta$. To this end, the Π^{fold} protocol is introduced.

Π^{fold} protocol. The verifier first sends a vector of uniformly random challenges $\mathbf{c} \leftarrow\!\!\$\, \mathcal{C}^d$ over a properly chosen challenge set $\mathcal{C} \subseteq \mathcal{R}$ of short ring elements. Then, the folded witness

$$\mathbf{w}^{(1)} := (\mathbf{c}^{\mathsf{T}} \otimes \mathbf{I}_{m/d}) \mathbf{w}^{(0)} \in \mathcal{R}_q^{m/d}$$

satisfies

$$\left(\mathbf{f}_{i,1}^{\mathsf{T}} \otimes \ldots \otimes \mathbf{f}_{i,\mu-1}^{\mathsf{T}}\right) \mathbf{w}^{(1)} = \mathbf{c}^{\mathsf{T}} \mathbf{u}_i \quad \text{for} \quad i = 0, 1, \ldots, n-1.$$

Moreover, we can bound $\|\mathbf{w}^{(1)}\| \leq \gamma \cdot \beta$, where $\gamma > 1$ is dependent on the norm of the challenge vector $\mathbf{c}$. Therefore, by applying

$$\Pi^{\mathsf{split}} \to \Pi^{\mathsf{fold}}$$

we managed to reduce the relation Ξ_m^{lin} to the relation $\Xi_{m/d}^{\mathsf{lin}}$. By applying split-and-fold recursively μ times, and sending the last folded witness $\mathbf{w}^{(\mu)} \in \mathcal{R}_q$ in the clear, we obtain a succinct argument with communication complexity $\mu \cdot (n+1) \cdot d + 1$ elements over $\mathcal{R}_q$.[9]

Correctness and soundness gaps. We discuss the two major bottlenecks of the split-and-form approach described above. The first one, called *correctness gap*, refers to the phenomenon that, due to recursive reductions (i.e., applying Π^{split} and Π^{fold}), the norm of the intermediate witnesses $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(\mu)}$ grows geometrically. Concretely, if a protocol applies μ reduction steps, each increasing the witness norm by a factor γ, then the final witness norm $\mathbf{w}_\mu$ can be as large as $\beta^* := \gamma^\mu \cdot \beta$, where β was the original bound. Ensuring correctness then requires the modulus q to exceed this inflated bound, i.e. $q \geq \beta^*$. Thus, the gap corresponds to the need for an unnecessarily large modulus to preserve shortness.

A more critical issue is a so-called *soundness gap* which captures the fact that the norm of the extracted witness, used to show knowledge soundness property, also grows geometrically. Namely, if a protocol applies μ invocations of split-and-fold, then extraction in each step increases the witness norm by a factor $\theta = \mathsf{poly}(\lambda)$ and the final extracted witness norm can be as large as $\theta^\mu \cdot \beta^*$. While this "relaxed extraction" suffices for some cryptographic primitives, it prevents from proving *exact* norm bounds, which are required for privacy-preserving applications such as set membership or range proofs. In the recursive setting, such relaxed extraction requires conservative parameter choices, and often super-polynomial modulus sizes.

Decomposition and exact norm proofs. The recent work of [KLNO24] circumvents these limitations by introducing two new subprotocols: $\Pi^{b\text{-}\mathsf{decomp}}$ and Π^{norm}. In $\Pi^{b\text{-}\mathsf{decomp}}$, the prover, who wants to prove knowledge of the witness $\mathbf{w}$ for the relation Ξ^{lin}, first performs a b-ary decomposition of $\mathbf{w}^{(0)}$, breaking it into ℓ lower-norm components $\mathbf{w}_0^{(0)}, \ldots, \mathbf{w}_{\ell-1}^{(0)}$ such that

$$\mathbf{w}^{(0)} = \sum_{j=0}^{\ell-1} b^j \cdot \mathbf{w}_j^{(0)} \quad \text{and} \quad \|\mathbf{w}_j^{(0)}\|_\infty < b \quad \text{for} \quad j = 0, 1, \ldots, \ell-1.$$

Then, sends partial instances $\mathbf{y}_j$ defined below:

$$\mathbf{y}_j := \mathbf{F} \mathbf{w}_j^{(0)} \quad \text{for } j = 0, \ldots, \ell-1$$

and the verifier can manually check that $\mathbf{y} = \sum_{j=0}^{\ell-1} b^j \cdot \mathbf{y}_j$. The new witness becomes $(\mathbf{w}_0^{(0)}, \ldots, \mathbf{w}_{\ell-1}^{(0)})$.[10]

[9] Using simple batching strategies, the proof size is reduced to $2\mu d + 1$ ring elements.

[10] Similar decomposition strategies were already implemented in LaBRADOR [BS23], and in the context of folding schemes [BC24, FKNP24].

Further, to prove well-formedness of all $(\mathbf{y}_j)_j$, the prover performs the (parallelised) split-and-fold protocol described above, where the folded witness becomes of the form

$$\mathbf{w}^{(1)} := (\mathbf{c}^\intercal \otimes \mathbf{I}_{m/d}) \begin{bmatrix} \mathbf{w}_0^{(0)} \\ \vdots \\ \mathbf{w}_{\ell-1}^{(0)} \end{bmatrix} \in \mathcal{R}_q^{m/d}$$

where the challenge vector $\mathbf{c} \leftarrow\!\!\$\, \mathcal{C}^{\ell d}$ is now of dimension ℓd. By carefully picking the challenge set $\mathcal{C}$ and the decomposition base b, one can make sure that norm of the folded witness $\mathbf{w}^{(1)}$ can be bounded β, identically as the initial witness $\mathbf{w}^{(0)}$. In conclusion, by adapting $\Pi^{b\text{-decomp}}$ into the split-and-fold protocol, one mitigates the norm blow-up issue during recursive folding, thereby eliminating the correctness gap.

On the other hand, to eliminate the soundness gap, [KLNO24] proves $\|\mathbf{w}^{(0)}\| \leq \beta$ in a two-step approach:[11]

(i) $\mathbf{w}^{(0)}$ is *approximately* short w.r.t. the modulus q
(ii) $\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\langle \mathbf{w}^{(0)}, \bar{\mathbf{w}}^{(0)} \rangle) \leq \beta$[12].

Note that (i) comes for free as long as the challenge set $\mathcal{C}$ is subtractive. As for the latter statement, [KLNO24] defines a protocol Π^{norm} where the prover first sends $t := \langle \mathbf{w}^{(0)}, \bar{\mathbf{w}}^{(0)} \rangle$– which is used for the verification check $\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(t) \leq \beta$ – and then proves the inner-product relation $\langle \mathbf{w}^{(0)}, \bar{\mathbf{w}}^{(0)} \rangle = t$.

In summary, [KLNO24] introduces a series of reductions of knowledge (RoK) [KP23] which iteratively compresses high-dimensional witnesses while preserving norm bounds and structural integrity. A single "folding round" defined in [KLNO24] can be represented as follows:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-decomp}} \to \Pi^{\mathsf{split}} \to \Pi^{\mathsf{fold}}. \tag{4}$$

Although the protocol does achieve succinct verification time, the overall communication complexity is $\tilde{O}(\lambda^2)$ bits. Indeed, to prove approximate shortness (i), one requires the set $\mathcal{C}$ to be subtractive. Unfortunately, known constructions of subtractive sets have at most polynomial size, and thus the protocol must be repeated at least $\Omega(\lambda/\log \lambda)$ times to achieve inverse-exponential soundness error. Combined with the fact that a single ring element is of size $\Omega(\lambda)$ bits, the total proof size can be bounded by $\tilde{O}(\lambda^2)$ bits.

Our effort. Motivated by the limitations of the prior works, we present a novel way of proving approximate shortness in a succinct manner, avoiding subtractive sets altogether. This allows us to select an exponentially large challenge space $\mathcal{C}$ of short elements, and consequently obtain negligible soundness error in one shot.

Inspired by the line of works [BL17, LNS21, GHL22, LNP22, BS23], we adopt a strategy based on random projections, wherein a witness $\mathbf{w}$ can be proven to be (approximately) short by opening to a random projection of $\mathbf{w}$. However, as opposed to prior works, we introduce a *tensor-structured* random projection approach which enables us to get succinct verification time. Concretely, we split our results into two separate parts, which are outlined in the following subsections:

- In Section 2.2, we consider the case of folding a large structured relation, where the projection must preserve certain tensor-related structure. Due to its size, the projection itself cannot be sent in plain and must instead be committed to. The full description is given in Section 4.
- In Section 2.3, we examine the case where the witness length is sufficiently small so the projection can be sent explicitly, albeit at the expense of dismissing the structure of the relation. The full description is given in Section 5.

2.2 Structured Projections

Our first technical contribution is to introduce two reductions of knowledge $\Pi^{\otimes\mathsf{RP}}$ and Π^{join}. Thus, as opposed to (4), our single folding round now takes the form:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-decomp}} \to \Pi^{\mathsf{split}} \to \Pi^{\otimes\mathsf{RP}} \to \Pi^{\mathsf{fold}} \to \Pi^{\mathsf{join}}. \tag{5}$$

[11] This approach was earlier applied in [LNP22], however only in the power-of-two cyclotomic setting.
[12] Here, the authors of [KLNO24] rely on the relationship between the canonical norm and the trace function.

The reduction $\Pi^{\otimes\mathsf{RP}}$ is responsible for applying a random projection to the witness. Specifically, given a witness vector $\mathbf{w}$, we define its projected form as $\widetilde{\mathbf{w}} \coloneqq \mathbf{J}\mathbf{w} \bmod q$, where the projection matrix $\mathbf{J}$ is chosen to be block-diagonal with $\mathbf{J} \coloneqq \mathbf{I} \otimes \mathbf{J}'$. Here, the base matrix $\mathbf{J}' \in \mathcal{R}^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}$ is sampled with entries from the probability distribution χ, defined as $\chi(0) = 1/2$ and $\chi(1) = \chi(-1) = 1/4$, and $\mathbf{I} \in \mathcal{R}^{(m/m_{\mathsf{rp}}) \times (m/m_{\mathsf{rp}})}$ is the identity matrix.

For correctness, we rely on the Johnson-Lindenstrauss Lemma to argue that $\|\widetilde{\mathbf{w}}\| \approx \|\mathbf{J}\mathbf{w}\|$. As for soundness, we show that if the image $\widetilde{\mathbf{w}}$ was short, then with probability $1 - \exp(-n_{\mathsf{rp}})$, the original witness $\mathbf{w}$ would also be (approximately) short. Hence, for this strategy to be statistically sound, the matrix $\mathbf{J}'$ must have $n_{\mathsf{rp}} = \Omega(\lambda)$ rows.

Choosing the number of columns, however, is more delicate. On the one hand, the projection must sufficiently reduce the dimension of the witness so that it matches the size required by the folded witness. On the other hand, since $\mathbf{J}'$ is unstructured, we must avoid making it too wide, as this would incur an undesirable overhead in commitment and verification. These constraints jointly determine the "limits of structured foldability" of the relation. In our construction, we set the width of $\mathbf{J}'$ to be $m_{\mathsf{rp}} := O(1) \cdot n_{\mathsf{rp}} = O(\lambda)$, i.e. only a constant factor larger than its height.

Since the projected image $\widetilde{\mathbf{w}}$ of size $m \cdot (n_{\mathsf{rp}}/m_{\mathsf{rp}})$ is not succinct, we commit to $\widetilde{\mathbf{w}}$ in a similar manner as in Ξ^{lin} instead. Hence, the protocol $\Pi^{\otimes\mathsf{RP}}$ generates claims for the following two relations:

$$\begin{pmatrix} \mathbf{F} \\ \mathbf{I} \otimes \mathbf{J}' \end{pmatrix} \mathbf{w} = \begin{pmatrix} \mathbf{y} \\ \mathbf{t} \end{pmatrix} \bmod q \quad \text{and} \quad \begin{pmatrix} \widetilde{\mathbf{F}} \\ \mathbf{I} \end{pmatrix} \widetilde{\mathbf{w}} = \begin{pmatrix} \widetilde{\mathbf{y}} \\ \mathbf{t} \end{pmatrix} \bmod q,$$

where both statements crucially share the same auxiliary vector $\mathbf{t}$, and $\mathbf{F}$, $\widetilde{\mathbf{F}}$ are commitments keys. Clearly, $\mathbf{t}$ is not succinct immediately – it has length $\Theta(m)$ due to the selected dimensions of the projection matrix $\mathbf{J}'$. To ensure succinctness, we immediately compress $\mathbf{t}$ using a structured challenge vector $\mathbf{c} \in \mathcal{R}_q^{m \cdot (n_{\mathsf{rp}}/m_{\mathsf{rp}})}$ sent by the verifier[13]. Specifically, rather than sending the full vector, the prover computes and outputs only $\mathbf{c}^{\mathsf{T}}\mathbf{t} \in \mathcal{R}_q$. This allows the projection to be enforced using a single additional row. Due to the structure of $\mathbf{c}$, the statement, even after batching, still maintains its tensor structure needed in Ξ^{lin}.

At the very end of the round, the projected witness $\widetilde{\mathbf{w}}$ is horizontally concatenated with the the folded witness from Π^{fold} via the Π^{join} protocol. In particular, we reduced the witness size from m to $m \cdot (n_{\mathsf{rp}}/m_{\mathsf{rp}} + 1/d)$.

Limits of structured foldability. While the structured folding round significantly reduces dimension of the witness, the resulting witness after $\mu = O(\log m)$ rounds still consists of $\tilde{\Theta}(\lambda)$ $\mathcal{R}$ elements[14]. At this point, further folding is no longer possible using the structured variant of Π^{split}, as the associated matrix $\mathbf{F}$ becomes unstructured due to the lack of structure of $\mathbf{J}'$.

On the other hand, the witness is still too large to be sent in the plain, since each entry is a ring element of size $\tilde{\Theta}(\lambda)$, i.e. the total communication cost would be $\tilde{\Theta}(\lambda^2)$ bits (due to the size of the cyclotomic ring element) if the witness were sent in the clear after structured folding. This marks the limit of structured foldability, and motivates the transition to the unstructured rounds, where we continue the compression using alternative techniques.

2.3 Unstructured Projections

Following the structured folding phase, the witness remains of length $\Theta(\lambda)$ (in ring elements). Sending it in plain would incur a communication cost of $\Theta(\lambda^2)$ bits, defeating our purpose of achieving a proof size subquadratic in λ.

To overcome this limitation, we introduce a second projection-based reduction of knowledge, denoted Π^{RP}. This RoK is essentially very similar to the one from LaBRADOR [BS23], i.e. a projection of the coefficient vector $\mathsf{cf}(\mathbf{w}) \in \mathbb{Z}_q^{m\varphi}$ of the witness $\mathbf{w}$ is sent in plain. This comes at the cost of losing the tensor structure of the relation, and therefore verifier runtime becomes linear in the witness dimension. Nevertheless, since in the current setting the witness size is (almost) independent of the original witness dimension m, the verifier runtime still remains $\mathsf{poly}(\lambda, \log m)$.

[13] The challenge vector $\mathbf{c}$ consists of powers of some random value. The verifier actually does not need to send all the $m \cdot (n_{\mathsf{rp}}/m_{\mathsf{rp}})$ ring elements thanks to the structure of $\mathbf{c}$.

[14] This follows by the assumption that m_{rp} divides m, and the inequality $m \cdot (n_{\mathsf{rp}}/m_{\mathsf{rp}} + 1/d) \geq n_{\mathsf{rp}} = \tilde{\Theta}(\lambda)$.

As a starting point, we again consider a single folding round defined in Eq. (4). We augment it with an additional reduction Π^{RP}, so that the new round becomes:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-}\mathsf{decomp}} \to \Pi^{\mathsf{split}} \to \Pi^{\mathsf{RP}} \to \Pi^{\mathsf{fold}}. \tag{6}$$

We now provide an overview of the protocol Π^{RP}.

Let $\mathbf{J}' \in \mathbb{Z}_q^{n_{\mathsf{rp}} \times m\varphi}$ be a matrix sampled according to the distribution χ defined before. As before, the number of rows n_{rp} has to be at least linear in λ to guarantee soundness. Define $\mathbf{J} \coloneqq \mathsf{cf}_\vee^{-1}(\mathbf{J}')$, where $\mathsf{cf}_\vee$ is the coefficient embedding of $\mathcal{R}_q^\vee$ (the trace dual ring of $\mathcal{R}_q$) into $\mathbb{Z}_q$. We observe that $\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\mathbf{Jw}) = \mathbf{J}'\mathsf{cf}(\mathbf{w})$. Consequently, the projection $\mathbf{t}' \coloneqq \mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\mathbf{Jw}) \bmod q$ approximately preserves the 2-norm of the witness, i.e., $\|\mathbf{t}'\|_2 \approx \|\sigma(\mathbf{w})\|_2$, and is used to argue that $\mathbf{w}$ is short in the extraction.

The remaining challenge lies in proving that $\mathbf{t}'$ is correct, i.e. that it is indeed a valid projection of the witness $\mathbf{w}$ satisfying $\mathbf{t}' = \mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\mathbf{Jw}) \bmod q$. This is not a standard Ξ^{lin} relation, as it holds over $\mathbb{Z}_q$ rather than $\mathcal{R}_q$. A natural strategy would be to "lift" the relation to $\mathcal{R}_q$ by sending the remaining coefficients in plain. However, this approach would incur a communication cost of $n_{\mathsf{rp}} = O(\lambda)$ ring elements, which is not acceptable.

To address this, we use the fact that the trace $\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}$ is a $\mathbb{Z}$-linear map. This allows us to batch the relation into a smaller number of rows using a challenge matrix $\mathbf{C}$ generated by the verifier, i.e.

$$\mathbf{Ct}' = \mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\mathbf{CJw}).$$

The number of rows of $\mathbf{C}$ required for negligible soundness becomes now $O(\lambda/\log q)$. Note that sending $\mathbf{CJw}$ takes $O(\lambda/\log q)$ ring elements, and since each $\mathcal{R}_q$ element is of size $O(\lambda \cdot \log q)$, we still obtain a quadratic proof size in λ.

Here lies the technical core of our solution to circumvent the quadratic dependency. Instead of "lifting" the relation directly to $\mathcal{R}_q$, we lift it to an intermediate extension ring $\mathcal{R}_i$ in the tower $\mathcal{R} = \mathcal{R}_\ell \supset \mathcal{R}_{\ell-1} \supset \cdots \supset \mathcal{R}_0 = \mathbb{Z}$. Assuming that each extension $[\mathcal{R}_{i+1} : \mathcal{R}_i]$ has constant degree, sending the remaining coefficients to the lifted relation requires only $O(\lambda)$ elements of $\mathbb{Z}_q$. This process is repeated for each level of the tower. Importantly, the "batching matrix" $\mathbf{C}_i$ requires progressively fewer rows to satisfy the soundness condition. This phenomenon relies on the existence of "relatively large" subfields $\mathbb{F}_{q^{e_i}}$ of $\mathcal{R}_i$ at each level i. If, at each level, the cardinality of the subfield grows, i.e. $e_{i+1} \geq e_i \cdot d$ for some constant d, then the number of rows of $\mathbf{C}_i$ can be reduced by roughly the same factor. As a result, sending the remaining coefficients of the "lifted relation" maintains approximately the same cost—the increasing degree of the extensions is counterbalanced by the decreasing number of rows in the batching matrix. This ensures that the total number of rows across all levels remains $O(\lambda)$, while still achieving negligible soundness error.

The unstructured folding round is applied repeatedly until the witness length becomes constant. At that point, its description is solely determined by the ring element size, which is itself $\tilde{O}(\lambda)$. The final witness is then sent in the clear.

In more practical terms, consider how we batch a relation through a tower of cyclotomic rings with concrete parameters. Assume $q \approx 2^{32}$ and that $\mathcal{R}_q$ does not split, meaning $\mathcal{R}_q$, and all of its cyclotomic subrings, are fields. We embed the projection image in the constant term of $\mathcal{R}_q$. However, instead of immediately sending the complete image over $\mathcal{R}_q$, the prover first sends only the projection over $\mathbb{Z}_q$. The verifier then sends a challenge (over $\mathbb{Z}_q$) to batch the relation. To lift the batched relation to $\mathcal{R}_q$, the prover needs to send $O(\lambda \log \lambda)$ ring elements. Concretely, for $\lambda = 256$, this requires sending approximately 8 elements from $\mathcal{R}_q$. To reduce this cost, we progressively lift the relation through a tower of cyclotomic rings. When lifting to a ring of slightly larger degree (e.g. of degree 2), the verifier needs to send only about 4 challenges (over the intermediate ring) to achieve negligible soundness error. This process repeats until the relation is fully lifted to a single image over $\mathcal{R}_q$. With appropriate choices for the tower structure, the total communication can be upper-bounded by approximately 2 elements from $\mathcal{R}_q$.

2.4 Combining Everything Together

The combination of both folding loops (5) and (6) yields a complete protocol that achieves succinct verification and near-optimal communication. The structured loop (5) is responsible for compressing the initial high-dimensional witness down to $\Theta(\lambda)$ ring elements, while maintaining the tensor structure of the relation to enable efficient verification. Once this compression limit is reached, the unstructured loop

(6) is applied, further reducing the witness size using unstructured random projections and unstructured folding. This two-phase design ensures that the verifier remains efficient, proof size succinct, and that the final witness can be sent directly in plain without violating the subquadratic communication bound.

3 Preliminaries

Let $\mathbb{N} = \{1, 2, \ldots\}$ denote natural numbers and $\lambda \in \mathbb{N}$ be the security parameter. For $m, n \in \mathbb{N}$, we write $[n] \coloneqq \{0, \ldots, n-1\}$ counting from 0 and $[m : n] \coloneqq [n] \setminus [m]$. For multidimensional ranges, we use the shorthand $(i, j, k) \in [n, m, \ell]$ for $i \in [n], j \in [m]$, and $k \in [\ell]$. The logarithm log is base-2. To represent $\mathbb{Z}_q$ we use the balanced representation, i.e. $\{-\lceil q/2 \rceil + 1, \ldots, \lfloor q/2 \rfloor\}$. We use bold lower-case letters for vectors $\mathbf{v}$ and bold upper-case letters for matrices $\mathbf{M}$. For matrices (or vectors) $\mathbf{M}_0, \ldots, \mathbf{M}_{k-1}$ of appropriate dimensions, we write $(\mathbf{M}_i)_{i \in [k]}$ for horizontal concatenation.

3.1 Algebraic Number Theory

Cyclotomic Fields. Throughout this work, we let $\mathcal{K} = \mathbb{Q}(\zeta)$ be a cyclotomic field, where ζ is a root of unity of order $\mathfrak{f} \neq 2 \bmod 4$ and degree $\varphi = \varphi(\mathfrak{f})$ (where φ is Euler's totient function). We let $\mathcal{R} = \mathcal{O}_{\mathcal{K}} = \mathbb{Z}[\zeta]$ be the ring of integers of $\mathcal{K}$ and call $\mathfrak{f}$ the *conductor* of $\mathcal{K}$.

For a prime modulus $q \in \mathbb{N}$, we write $\mathcal{R}_q \coloneqq \mathcal{R}/q\mathcal{R}$. We denote by $\mathcal{R}^\times$ and $\mathcal{R}_q^\times$ the sets of units in $\mathcal{R}$ and $\mathcal{R}_q$ respectively. We endow $\mathcal{R}$ with two geometries via the coefficient embedding $\mathtt{cf}_{\mathbf{b}} : \mathcal{R} \to \mathbb{Z}^\varphi$ (for a given basis $\mathbf{b}$) and the canonical embedding $\sigma : \mathcal{K} \to \mathbb{C}^\varphi$ (of $\mathcal{K}$). Specifically, for a given $\mathbb{Z}$-basis $\mathbf{b} = (b_i)_{i \in [\varphi]}$ of $\mathcal{R}$ and an element $x = \sum_{i \in [\varphi]} x_i b_i \in \mathcal{R}$, we write

$$\mathtt{cf}_{\mathbf{b}}(x) \coloneqq (x_i)_{i \in [\varphi]} \qquad \text{and} \qquad \sigma(x) \coloneqq (\sigma_j(x))_{j \in [\varphi]}$$

where $\sigma_j \in \mathsf{Gal}(\mathcal{K}/\mathbb{Q})$. We define the powerful basis of $\mathcal{R}$ as $\mathbf{b} = (1, \zeta, \ldots, \zeta^{\varphi - 1})$ for prime-power conductor $\mathfrak{f}$. The basis generalises to the composite conductor $\mathfrak{f} = \prod_{i \in [k]} \mathfrak{f}_i^{e_i}$ for prime $\mathfrak{f}_i$ via tensor product, $\mathbf{b} = \bigotimes_{i \in [k]} \left(1, \zeta_{\mathfrak{f}_i}, \ldots, \zeta_{\mathfrak{f}_i}^{\varphi(\mathfrak{f}_i) - 1}\right)$. If $\mathbf{b}$ is the standard powerful basis, we may omit $\mathbf{b}$ from the subscript of $\mathtt{cf}_{\mathbf{b}}$. We extend the notation of $\mathtt{cf}_{\mathbf{b}}$ and σ naturally to vectors, i.e. if $\mathbf{x} = (x_i)_{i \in [m]} \in \mathcal{R}^m$, then

$$\mathtt{cf}_{\mathbf{b}}(\mathbf{x}) \coloneqq (\mathtt{cf}_{\mathbf{b}}(x_i))^{\mathsf{T}}_{i \in [\varphi]} \qquad \text{and} \qquad \sigma(\mathbf{x}) \coloneqq (\sigma_j(x_i))^{\mathsf{T}}_{j \in [\varphi]}$$

are defined as concatenations.

Trace Map. For any Galois extension $\mathcal{M}/\mathcal{L}$, the field trace can be computed as $\mathsf{Trace}_{\mathcal{M}/\mathcal{L}} : \mathcal{M} \to \mathcal{L}$, $\mathsf{Trace}_{\mathcal{M}/\mathcal{L}}(x) \coloneqq \sum_{\sigma_j \in \mathsf{Gal}(\mathcal{M}/\mathcal{L})} \sigma_j(x)$. If $x \in \mathcal{O}_{\mathcal{M}}$, then $\mathsf{Trace}_{\mathcal{M}/\mathcal{L}}(x) \in \mathcal{O}_{\mathcal{L}}$. We may therefore abuse notation and write $\mathsf{Trace}_{\mathcal{O}_{\mathcal{L}}/\mathcal{O}_{\mathcal{M}}}(x) \coloneqq \mathsf{Trace}_{\mathcal{M}/\mathcal{L}}(x)$. Furthermore, for any $q \in \mathbb{N}$, if $x = x' \bmod q\mathcal{O}_{\mathcal{M}}$, we have $\mathsf{Trace}_{\mathcal{O}_{\mathcal{L}}/\mathcal{O}_{\mathcal{M}}}(x) = \mathsf{Trace}_{\mathcal{O}_{\mathcal{L}}/\mathcal{O}_{\mathcal{M}}}(x') \bmod q$ due to linearity. We can therefore write $\mathsf{Trace}_{\mathcal{O}_{\mathcal{L}}/\mathcal{O}_{\mathcal{M}}}(y)$ for $y \in {}^{\mathcal{O}_{\mathcal{M}}}/_{q\mathcal{O}_{\mathcal{M}}}$ without ambiguity. When $\mathcal{L} = \mathbb{Q}$, we drop the subscript and write $\mathsf{Trace} = \mathsf{Trace}_{\mathcal{O}_{\mathcal{M}}/\mathbb{Z}} = \mathsf{Trace}_{\mathcal{M}/\mathbb{Q}}$.

Lemma 1. *Let q be an unramified prime, i.e. $q \nmid \mathfrak{f}$, and $\mathcal{R}$ be the ring of integers of a cyclotomic field $\mathcal{K}$. For any $\mathbb{Z}$-basis $\mathbf{b} = (b_i)_{i \in [\varphi]} \in \mathcal{R}^\varphi$ of $\mathcal{R}$, there exists $\mathbf{b}^\vee = (b_i^\vee)_{i \in [\varphi]} \in \mathcal{R}^\varphi$ such that $\mathsf{Trace}(b_i \cdot b_j^\vee) = \delta_{i,j} \bmod q$, where $\delta_{i,j}$ denotes the Kronecker delta, for all $i, j \in [\varphi]$.*

Proof. First, recall that the $\mathbb{Z}$-module dual of $\mathcal{R}$ with respect to the trace map is defined as

$$\mathcal{R}^* \coloneqq \{x \in \mathcal{K} \mid \mathsf{Trace}(x \cdot y) \in \mathbb{Z} \text{ for all } y \in \mathcal{R}\}.$$

It is well known that $\mathcal{R}^* \subseteq \mathfrak{f}^{-1}\mathcal{R}$ (see e.g. [LPR13]), where $\mathfrak{f}$ is the conductor of $\mathcal{K}$, and thus for any $\hat{x} \in \mathcal{R}^*$, we have $\hat{x} \cdot \mathfrak{f} \in \mathcal{R}$. Let $\mathbf{b}^* = (b_i^*)_{i \in [\varphi]} \in (\mathcal{R}^*)^\varphi$ be the trace-dual of $\mathbf{b}$, i.e.

$$\mathsf{Trace}(b_i \cdot b_j^*) = \delta_{i,j}.$$

Since q is an unramified prime, we have $q \nmid \mathfrak{f}$, and hence $\mathfrak{f}$ is invertible modulo q. Let $\mathfrak{f}^\ddagger \in \mathbb{Z} \subset \mathcal{R}$ be any inverse of $\mathfrak{f}$ modulo q, i.e. $\mathfrak{f} \cdot \mathfrak{f}^\ddagger = 1 \bmod q$. We define $\mathbf{b}^\vee = (b_i^\vee)_{i \in [\varphi]}$ where $b_i^\vee \coloneqq (b_i^* \cdot \mathfrak{f}) \cdot \mathfrak{f}^\ddagger \bmod q\mathcal{R}$. Recall that $b_i^* \cdot \mathfrak{f} \in \mathcal{R}$ and $\mathfrak{f}^\ddagger \in \mathcal{R}$, so the reduction modulo $q\mathcal{R}$ is well-defined. Observe that

$$\mathsf{Trace}\left(b_i \cdot b_j^\vee\right) = \mathsf{Trace}\left(b_i \cdot (b_j^* \cdot \mathfrak{f}) \cdot \mathfrak{f}^\ddagger\right) = \mathfrak{f}^\ddagger \cdot \mathfrak{f} \cdot \mathsf{Trace}\left(b_i \cdot b_j^*\right) = \mathfrak{f}^\ddagger \cdot \mathfrak{f} \cdot \delta_{i,j}$$

where the first equality is by definition of $b_j^\vee$, the second by the linearity of Trace, and the third by definition of b_j^*. Since $\mathfrak{f}^\ddagger, \mathfrak{f}, \delta_{i,j} \in \mathbb{Z}$, we can reduce both sides modulo $q\mathbb{Z}$ and obtain

$$\mathsf{Trace}\left(b_i b_j^\vee\right) = \delta_{i,j} \bmod q. \qquad \square$$

While considering the coefficient embedding with respect to the vector[15] $\mathbf{b}^\vee$, when $\mathbf{b}$ is a powerful basis, we may omit $\mathbf{b}^\vee$ and write only $\mathfrak{cf}_\vee$.

Norms and Norm Conversion. The coefficient ℓ_p-norm and canonical ℓ_p-norm of a vector $\mathbf{x} \in \mathcal{R}^m$ is denoted by $\|\mathfrak{cf}(\mathbf{x})\|_p$ and $\|\sigma(\mathbf{x})\|_p$ respectively. We will mostly use $\|\mathfrak{cf}(\cdot)\|_\infty$ and $\|\sigma(\cdot)\|_2$. For a matrix $\mathbf{M} \in \mathcal{R}^{n\times m}$, the norm is defined as $\|\mathbf{M}\| = \max_{i\in[n]}\|\mathbf{m}_i\|$, where $\mathbf{m}_i$ is an i-th column of the matrix. The operator norm of $\mathbf{M}$ is $\|\mathbf{M}\|_{\mathsf{op}} = \sup_{\mathbf{x}\neq\mathbf{0}}\|\mathbf{M}\mathbf{x}\|/\|\mathbf{x}\|$, and we write $\|\cdot\|_{\sigma\text{-}\mathsf{op}}$ if $\|\cdot\| = \|\sigma(\cdot)\|_2$. The ring expansion factor of $\mathcal{R}$ w.r.t. the coefficient ℓ_∞-norm is defined as $\gamma_\mathcal{R} := \max_{a,b\in\mathcal{R}}\|\mathfrak{cf}(a\cdot b)\|_\infty/(\|\mathfrak{cf}(a)\|_\infty \cdot \|\mathfrak{cf}(b)\|_\infty)$. Assuming balanced representation, for any $x \in \mathcal{R}_p$, we have $\|\mathfrak{cf}(x)\|_\infty \le p/2$. Note that $\|\sigma(\mathbf{x})\|_2^2 = \mathsf{Trace}(\mathbf{x}^\mathsf{T}\overline{\mathbf{x}})$, where $\bar{\cdot}$ denotes the complex conjugate. We recall following useful lemma from [KLNO24].

Lemma 2 (Relations between norms (from [KLNO24])). *Let $x \in \mathcal{K} = \mathbb{Q}(\zeta_\mathfrak{f})$ and $\varphi = \varphi(\mathfrak{f})$. Let $\hat{\mathfrak{f}}$ be $\mathfrak{f}$ if $\mathfrak{f}$ is odd and $\mathfrak{f}/2$ if it is even; let $\mathsf{rad}(\mathfrak{f})$ be the radical (i.e. the product of all primes dividing $\mathfrak{f}$). Let $\sigma : \mathcal{K} \to \mathbb{R}^\varphi$ be the canonical embedding and let $\mathfrak{cf} : \mathcal{K} \to \mathbb{R}^\varphi$ be the coefficient embedding w.r.t. the powerful basis. Then we have (i) $\|\mathfrak{cf}(x)\|_2 \le \sqrt{\frac{\mathsf{rad}(\mathfrak{f})}{\mathfrak{f}}}\,\|\sigma(x)\|_2$ (ii) $\|\sigma(x)\|_2 \le \sqrt{\hat{\mathfrak{f}}}\cdot\|\mathfrak{cf}(x)\|_2$ (iii) $\|\sigma(x)\|_\infty \le \varphi\cdot\|\mathfrak{cf}(x)\|_\infty$ (iv) $\|\mathfrak{cf}(x)\|_\infty \le \mathfrak{c}_{\mathsf{rad}(\mathfrak{f})}\cdot\|\sigma(x)\|_\infty$, where $\mathfrak{c}_j$ is a constant such that $\mathfrak{c}_{\mathsf{rad}(\mathfrak{f})} \le (4/\pi)^\ell$, where ℓ is the number of different odd prime factors in $\mathsf{rad}(\mathfrak{f})$. Moreover, $\mathfrak{c}_{f_1\cdot f_2} = \mathfrak{c}_{f_1}\cdot\mathfrak{c}_{f_2}$ for coprime f_1 and f_2 and $\mathfrak{c}_{2^e} = 1$.*

3.2 RoK, Paper, SISsors [KLNO24]

Principal relation. We recall the principal relation handled by most of the reductions of knowledge in [KLNO24].

$$\Xi^{\mathsf{lin}}_{\mathcal{R},q,m,\hat{n},r,\beta} := \left\{ \begin{array}{l} ((\mathbf{H},\mathbf{F},\mathbf{Y}),\mathbf{W}): \\ \quad \mathbf{H} \in \mathcal{R}_q^{\hat{n}\times n} \quad \mathbf{F} \in \mathcal{R}_q^{n\times m} \\ \quad \mathbf{Y} \in \mathcal{R}_q^{\hat{n}\times r} \quad \mathbf{W} \in \mathcal{R}^{m\times r} \\ \quad \|\sigma(\mathbf{W})\|_2 \le \beta \ \text{ and } \ \mathbf{HFW} = \mathbf{Y} \bmod q \end{array} \right\}$$

where the compression[16] matrix $\mathbf{H}$ is restricted to take the form $\mathbf{H} = \begin{pmatrix} \mathbf{I}_{\overline{n}} & \\ & \underline{\mathbf{H}} \end{pmatrix}$ for some $\overline{n} \le n$, i.e. it retains the top $\overline{n}$ rows and compresses the bottom $\underline{n}$ rows of $\mathbf{F}$. Throughout, we think of $n = \overline{n} + \underline{n}$ and $\mathbf{F} = \begin{pmatrix} \overline{\mathbf{F}} \\ \underline{\mathbf{F}} \end{pmatrix} \in \mathcal{R}_q^{(\overline{n}+\underline{n})\times m}$ as consisting of a top part $\overline{\mathbf{F}} \in \mathcal{R}_q^{\overline{n}\times m}$ and a bottom part $\underline{\mathbf{F}} \in \mathcal{R}_q^{\underline{n}\times m}$.

The work of [KLNO24] exclusively considers a special class of Ξ^{lin} parametrised by a vector $\mathbf{d}$ satisfying $\prod_{i\in[\mu]} d_i = m$ where $\mathbf{F}$ can be decomposed into a row-tensor $\mathbf{F} = \mathbf{F}_0 \bullet \ldots \bullet \mathbf{F}_{\mu-1}$ where $\mathbf{F} \in \mathcal{R}_q^{n\times d_i}$. In this work, we denote this class by $\Xi^{\mathsf{lin}\text{-}\otimes\mathbf{d}}_{\mathcal{R},q,m,\hat{n},r,\beta}$ which will be handled in Section 4. In Section 5, we will consider a semi-structured version, denoted $\Xi^{\mathsf{lin}\text{-}\bar{\otimes}\mathbf{d}}_{\mathcal{R},q,m,\hat{n},r,\beta}$, where the top part $\overline{\mathbf{F}}$ of $\mathbf{F}$ has a row-tensor structure but the bottom part $\underline{\mathbf{F}}$ is unstructured.

Slack. We further define a generalisation of Ξ^{lin} where the equation $\mathbf{HFW} = \mathbf{Y} \bmod q$ holds up to a (short) "slack" [AL21]. We still name this generalisation Ξ^{lin}, but introduce an additional parameter ϱ which controls the slack:

$$\Xi^{\mathsf{lin}}_{\mathcal{R},q,m,\hat{n},r,\beta,\varrho} := \left\{ \begin{array}{l} ((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S})): \\ \quad \mathbf{H} \in \mathcal{R}_q^{\hat{n}\times n} \quad \mathbf{F} \in \mathcal{R}_q^{n\times m} \\ \quad \mathbf{y} \in \mathcal{R}_q^{\hat{n}\times r} \quad \mathbf{W} \in \mathcal{R}_q^{m\times r}, \text{diagonal } \mathbf{S} \in \mathcal{R}^{r\times r} \ \text{ s.t.} \\ \quad \|\sigma(\mathbf{WS} \bmod q)\|_2 \le \beta \quad \wedge \quad \|\mathbf{S}\|_{\sigma\text{-}\mathsf{op}} \le \varrho \\ \quad \mathbf{HFW} = \mathbf{Y} \bmod q \end{array} \right\}.$$

[15] Not necessarily a basis.

[16] The compression matrix $\mathbf{H}$ is a way to verifier-succinctly batch-verify. Using random linear combination of rows of $\mathbf{F}$ is not verifier-succinct. Hence, we delay and keep track of such linear combinations via $\mathbf{H}$, until the size of $\mathbf{F}$ has been reduced to $\mathsf{poly}(\lambda)$.

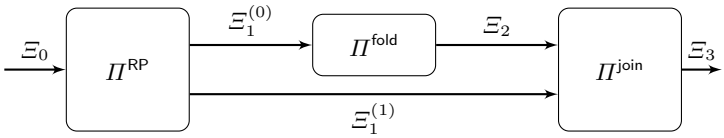

Fig. 2: Partial composition of RoKs.

where $\mathbf{H}$ is restricted as before. The diagonal matrix $\mathbf{S}$ consists[17] of the "denominators" of $\mathbf{W}$, that is, $\mathbf{WS}$ consists of (short) ring elements. The structured versions $\Xi^{\mathsf{lin}\text{-}\otimes\mathbf{d}}$ and $\Xi^{\mathsf{lin}\text{-}\bar{\otimes}\mathbf{d}}$ are defined analogously. When $\varrho = 1$, the slack $\mathbf{S}$ is restricted to consist of roots of unity, and the relation collapses back to the original Ξ^{lin} relation.

SIS Break. To catch exceptions where a prover is able to solve a (vanishing) SIS problem, we define the following relation Ξ^{sis} whose statements belong to the same space as Ξ^{lin} so that we can conveniently write $\Xi^{\mathsf{lin}} \cup \Xi^{\mathsf{sis}}$.[18]

$$\Xi^{\mathsf{sis}}_{\mathcal{R},q,m,\hat{n},r,\beta} := \left\{ \begin{array}{l} ((\mathbf{H},\mathbf{F},\mathbf{Y}),\mathbf{x}) : \\ \quad \mathbf{H} \in \mathcal{R}_q^{\hat{n}\times n} \quad \mathbf{F} \in \mathcal{R}_q^{n\times m} \\ \quad \mathbf{Y} \in \mathcal{R}_q^{\hat{n}\times r} \quad \mathbf{W} \in \mathcal{R}^{m\times r} \\ \quad \|\sigma(\mathbf{x})\|_2 \le \beta \;\text{ and }\; \overline{\mathbf{F}}\mathbf{x} = \mathbf{0} \bmod q \end{array} \right\}$$

where $\mathbf{H}$ is as above. Note that $\overline{\mathbf{F}}$ will have a row-tensor structure in our case, but that is not required for (v)SIS breaks.

Shorthands. Since the parameters $\mathcal{R}$ and q will not change throughout, we often omit them and write e.g. $\Xi^{\mathsf{lin}\text{-}\otimes\mathbf{d}}_{m,\hat{n},r,\beta,\varrho}$. For Ξ^{sis}, we skip most parameters and write e.g. $\Xi^{\mathsf{sis}}_{\beta}$. We also write e.g. $(\overline{\mathbf{F}},\mathbf{x}) \in \Xi^{\mathsf{sis}}_{\beta}$ instead of $((\mathbf{H},\mathbf{F},\mathbf{Y}),\mathbf{x}) \in \Xi^{\mathsf{sis}}_{\beta}$ since the rest of the statement is not important.

Computational assumptions and the definition of a reduction of knowledge are given in Section A.

4 Structured Projections

The principal issue we aim to address with random projection techniques is the reliance on a subtractive set as the challenge set for folding. This reliance has been a major efficiency bottleneck in prior works [BBC+18, BLNS20, AL21, CLM23, KLNO24], particularly due to the need for parallel repetitions to achieve negligible soundness error. More precisely, the proofs of special soundness (or its coordinate-wise variant [AFLN24]) in the cited works share a common pattern: During witness extraction, the witness is recovered in the relaxed form $\widetilde{\mathbf{w}} = \mathbf{w}/(c_0 - c_1)$ for c_0, c_1 sampled from the challenge set. Soundness then hinges on ensuring that $1/(c_0 - c_1)$ remains norm-bounded, a property embedded in the definition of a subtractive set. However, the size of a subtractive set is typically upper bounded by a polynomial of the security parameter, which necessitates a (near) linear number of repetitions in the security parameter to achieve negligible soundness error during folding.

A tempting alternative approach is to accept the fact that the extracted witness only satisfies a relaxed relation where the witness is expressed as a fraction with a low-norm denominator. Unfortunately, this approach is not viable because the accumulated slack grows exponentially when the extraction procedure is applied recursively.

4.1 Overview

To avoid using subtractive sets, we instead use the approach of [GHL22, BS23]. Namely, we employ Johnson–Lindenstrauss-style projections [JL84] to "upgrade" the relaxed relation to a non-relaxed one, i.e.

[17] Since $\mathbf{S}$ is diagonal, $\|\mathbf{S}\|_{\mathsf{op}} = \|\sigma(\mathbf{S})\|_\infty = \max_i \|\sigma(s_i)\|_\infty$ if $s_0, s_1, \ldots$ is the diagonal.

[18] To allow extractors to pass around SIS-breaks, we formally need knowledge reductions claims of the form $\Xi^{\mathsf{lin}} \cup \Xi^{\mathsf{sis}} \leftarrow \Xi^{\mathsf{lin}} \cup \Xi^{\mathsf{sis}}$ (cf. [KLNO24]). We often simplify to $\Xi^{\mathsf{lin}} \cup \Xi^{\mathsf{sis}} \leftarrow \Xi^{\mathsf{lin}}$ or even $\Xi^{\mathsf{lin}} \leftarrow \Xi^{\mathsf{lin}}$, when handling SIS-breaks is straightforward.

to argue that the extracted witness $\widetilde{\mathbf{w}}$ should be low-norm without requiring a subtractive challenge set. More concretely, following the approach in [BS23], the idea is for the prover to send (and prove about) a random projection $\mathbf{Jw}$ of the witness. If $\mathbf{J}$ satisfies certain properties, it holds that $\|\mathbf{Jw} \bmod q\|_2 \in \Theta(\|\mathbf{w}\|_2)$ and thus provides an approximate norm proof.

In [GHL22, BS23] the matrix $\mathbf{J}$ was chosen as $\mathbf{J} \leftarrow \chi_{\mathsf{ghl}}^{2\lambda \times m}$ where $\mathrm{Supp}(\chi_{\mathsf{ghl}}) = \{-1, 0, 1\}$. However, this is incompatible with *succinct verification*. Given the row-structure required by our protocol, structured projection matrices such as circulant or Toeplitz matrices [FL20, HV11, Mat08] are natural candidates, but their description is still not succinct (in m), making them not a right fit for further foldability.

To address this problem, we relax the compression required of $\widehat{\mathbf{J}}$ and thus to commit to the image $\widehat{\mathbf{J}}\mathbf{w}$ instead. This allows us to set $\widehat{\mathbf{J}} = \mathbf{I} \otimes \mathbf{J}$, where $\mathbf{J}' \leftarrow \chi_{\mathsf{ghl}}^{2\lambda \times m'}$ is small in dimension and unstructured. While the witness size now increases by a constant factor, this factor can be absorbed through appropriate parameter adjustments during folding, avoiding asymptotic proof size inflation.

Our proposed strategy solves the issue of verifier succinctness but introduces a new challenge. Unlike [BS23], the image $\widehat{\mathbf{J}}\mathbf{w}$ is not sent in plain. Thus, it must be obtained through extraction (of later protocol rounds). If the extracted witness has slack, which occurs due to the folding protocol, then we cannot use $\widehat{\mathbf{J}}\mathbf{w}$ to argue shortness of $\mathbf{w}$ anymore. To address this, we delay the folding of the "projection" relation. That is, we do not immediately fold $\mathbf{w}$ and $\widehat{\mathbf{J}}\mathbf{w}$ to a (smaller) claim. Instead, we first "fold" $\mathbf{w}$ to reduce its size. Only later, we combine it with $\widehat{\mathbf{J}}\mathbf{w}$. In short, we carefully set up our reductions of knowledge in a way that ensures we can extract $\widehat{\mathbf{J}}\mathbf{w}$ without slack. This strategy is illustrated in Fig. 2.

4.2 Formal Construction and Analysis

We present a set of two complimentary RoKs – $\Pi^{\otimes\mathsf{RP}}$ and Π^{join} – in Fig. 3 and Fig. 4, respectively. We specify the distribution χ_{ghl} as in Definition 1 and consider parameters m, r as usual. We let $n_{\mathsf{rp}}, m_{\mathsf{rp}} \in \mathbb{N}$ and will sample $\mathbf{J} \leftarrow \chi_{\mathsf{ghl}}^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}$ as the basic random projection matrix. For efficiency, we need to stack the r-many outputs of $\mathbf{I} \otimes \mathbf{J}$ (for the r columns of $\mathbf{W}$) into a single vector of size m. Hence, we choose parameters with $m_{\mathsf{rp}}/n_{\mathsf{rp}} = r$, see the construction of $\widehat{\mathbf{w}}$ in Fig. 3. Also for efficiency, we cannot let the verifier operate on $\mathbf{I} \otimes \mathbf{J}$ and therefore compress it via a random challenge c.

Theorem 1 ($\Pi^{\otimes\mathsf{RP}}$). *Let $0 < \delta < 1$, $n_{\mathsf{rp}} \geq \log_2(1/\delta)$, $m_{\mathsf{rp}} = r \cdot n_{\mathsf{rp}}$, $m' \coloneqq m/r$, $\beta, \beta', \beta'' < q$ and $\varrho \geq 1$. Let $\hat{\beta} \coloneqq m_{\mathsf{rp}} \cdot \beta$, $\tilde{\beta} \coloneqq 2 \cdot \sqrt{\mathsf{rad}(\mathsf{f}) \cdot m_{\mathsf{rp}}} \cdot \beta''$ and $\hat{\beta}' \coloneqq \max\{2\beta'\varrho, 2\beta''\}$. Then $\Pi^{\otimes\mathsf{RP}}$ is a perfectly correct reduction of knowledge for*

$$\Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m,\hat{n},r,\beta} \quad \rightarrow \quad \Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m,\hat{n}+1,r,\beta} \times \Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m,\overline{n}+1,1,\hat{\beta}}.$$

Moreover, $\Pi^{\otimes\mathsf{RP}}$ is knowledge sound with a knowledge error $\kappa = m \cdot (1/|\mathcal{C}_{\mathcal{R}_q}| - \varphi \cdot \delta / \log(\delta))$ for

$$\Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m,\hat{n},r,\tilde{\beta}} \cup \Xi^{\mathsf{sis}}_{\hat{\beta}'} \leftarrow \Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m,\hat{n}+1,r,\beta',\varrho} \times \Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m',\overline{n}+1,1,\beta''}.$$

The prover to verifier communication cost is $(\overline{n} + r)$ $\mathcal{R}_q$ elements.

Proof. **Correctness**: Consider an instance $((\mathbf{H}, \mathbf{F}, \mathbf{Y}), \mathbf{W}) \in \Xi^{\mathsf{lin}\text{-}\otimes(r,m')}_{m,\hat{n},r,\beta}$. For any $\mathbf{J} \in \mathbb{Z}^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}$, $\mathbf{c} = \mathbf{c}_0 \otimes \mathbf{c}_1 \in \mathcal{R}_q^m$, $\widehat{\mathbf{w}} = \mathsf{vec}(\widehat{\mathbf{J}} \cdot \mathbf{W}) = (\mathbf{I}_r \otimes \widehat{\mathbf{J}}) \cdot \mathsf{vec}(\mathbf{W})$, and $\mathbf{r}^\intercal = \mathbf{c}_1^\intercal \cdot \widehat{\mathbf{J}} \cdot \mathbf{W} \bmod q$, observe that $\mathbf{c}^\intercal \cdot \widehat{\mathbf{w}} = (\mathbf{c}_0^\intercal \otimes \mathbf{c}_1^\intercal) \cdot (\mathbf{I}_r \otimes \widehat{\mathbf{J}}) \cdot \mathsf{vec}(\mathbf{W})$
$= \mathbf{c}_0^\intercal \cdot (\mathbf{I}_r \otimes \mathbf{c}_1^\intercal \cdot \widehat{\mathbf{J}}) \cdot \mathsf{vec}(\mathbf{W}) = \mathbf{c}_0^\intercal \cdot \mathsf{vec}(\mathbf{c}_1^\intercal \cdot \widehat{\mathbf{J}} \cdot \mathbf{W}) = \mathbf{c}_0^\intercal \cdot \mathbf{r}$. It then follows by inspection that

$$\widetilde{\mathbf{H}}\widetilde{\mathbf{F}}\mathbf{W} = \widetilde{\mathbf{Y}} \bmod q \quad \text{and} \quad \widehat{\mathbf{F}}\widehat{\mathbf{w}} = \widehat{\mathbf{Y}} \bmod q.$$

It remains to bound the norm of $\widehat{\mathbf{w}}$. Recall that $\widehat{\mathbf{w}} = (\mathbf{I}_r \otimes \widehat{\mathbf{J}}) \cdot \mathsf{vec}(\mathbf{W})$ where $\widehat{\mathbf{J}} = \mathbf{I}_{m/m_{\mathsf{rp}}} \otimes \mathbf{J}$ and $\mathbf{J} \in \{-1, 0, 1\}^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}$. Since $\mathbf{I}_r \otimes \widehat{\mathbf{J}} = \mathbf{I}_r \otimes \mathbf{I}_{m/m_{\mathsf{rp}}} \otimes \mathbf{J}$ is over $\mathbb{Z}$ and has a block-diagonal structure, $\|\sigma(\widehat{\mathbf{w}})\|_2 = \|\sigma\left((\mathbf{I}_r \otimes \widehat{\mathbf{J}}) \cdot \mathsf{vec}(\mathbf{W})\right)\|_2 \leq \|\mathbf{J}\|_F \cdot \|\sigma(\mathsf{vec}(\mathbf{W}))\|_2$ where $\|\mathbf{J}\|_F$ denotes the Frobenius norm of $\mathbf{J}$ (as a matrix in $\mathbb{R}^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}$). Since $\mathbf{J}$ takes $\{-1, 0, 1\}$ values, $\|\mathbf{J}\|_F \leq \sqrt{n_{\mathsf{rp}} \cdot m_{\mathsf{rp}}}$. Observe also that $\|\sigma(\mathsf{vec}(\mathbf{W}))\|_2 \leq \sqrt{r} \cdot \|\sigma(\mathbf{W})\|_2 \leq \sqrt{r} \cdot \beta$. We conclude that $\|\sigma(\widehat{\mathbf{w}})\|_2 \leq \sqrt{n_{\mathsf{rp}} \cdot m_{\mathsf{rp}} \cdot r} \cdot \beta = m_{\mathsf{rp}}\beta = \hat{\beta}$.

Knowledge Soundness: For the ease of analysis we split the protocol into two parts at the two challenges $\mathbf{J}$ and c:

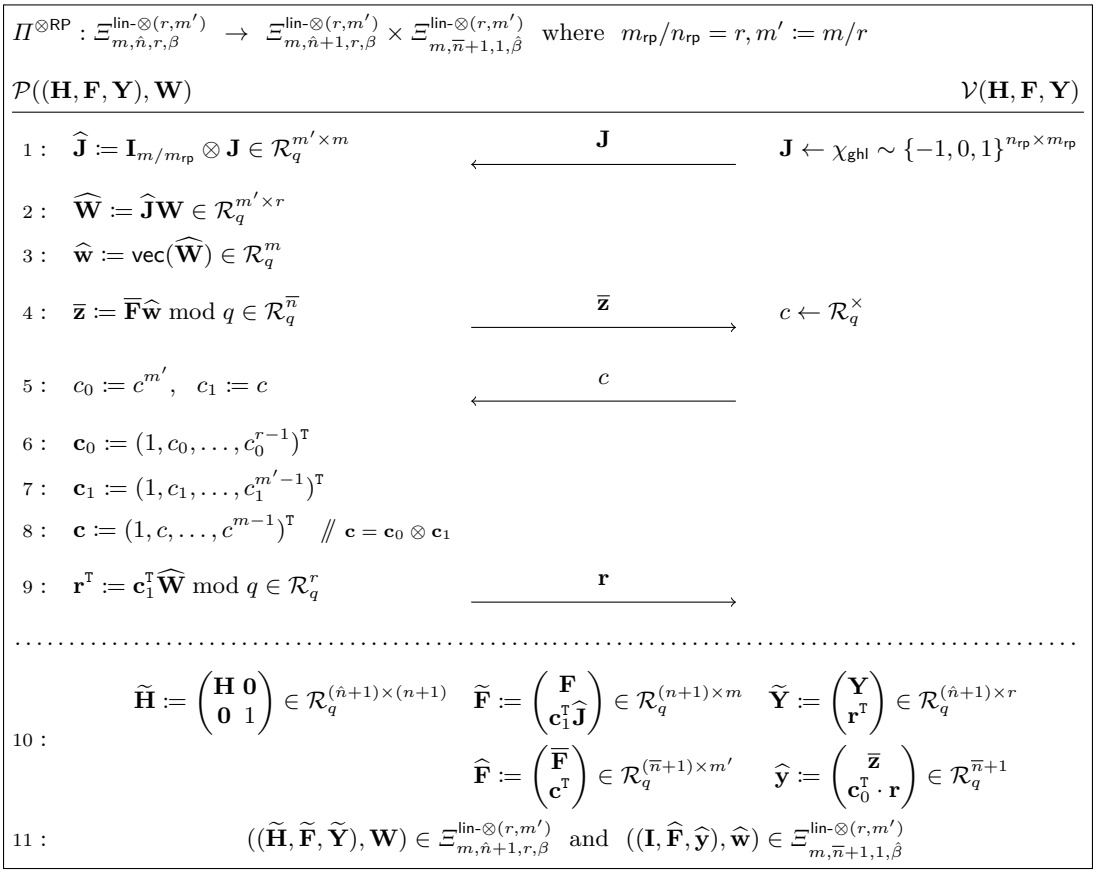

Fig. 3: RoK $\Pi^{\otimes\mathsf{RP}}$.

(i) $\Pi^{\otimes\mathsf{RP}}_{\mathsf{RP}} : \Xi^{\mathsf{lin}} \to \Xi^{\mathsf{lin}}$ turns $((\mathbf{H},\mathbf{F},\mathbf{Y}),\mathbf{W})$ into the tuple of relations

$$\left(\left(\begin{pmatrix} \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m'} \end{pmatrix}, \begin{pmatrix} \mathbf{F} & \mathbf{0} \\ \widehat{\mathbf{J}} & -\mathbf{I}_{m'} \end{pmatrix}, \begin{pmatrix} \mathbf{Y} \\ \mathbf{0}_{m'\times r} \end{pmatrix}\right), \begin{pmatrix} \mathbf{W} \\ \widehat{\mathbf{W}} \end{pmatrix}\right)$$

and

$$\left(\left(\mathbf{I}_{\overline{n}+m}, \begin{pmatrix} \mathbf{0} & \overline{\mathbf{F}} \\ \mathbf{I}_r \otimes \widehat{\mathbf{J}} & -\mathbf{I}_m \end{pmatrix}, \begin{pmatrix} \overline{\mathbf{z}} \\ \mathbf{0}_m \end{pmatrix}\right), \begin{pmatrix} \mathsf{vec}(\mathbf{W}) \\ \mathsf{vec}(\widehat{\mathbf{W}}) \end{pmatrix}\right)$$

where $\widehat{\mathbf{w}} = \mathsf{vec}(\widehat{\mathbf{W}})$. These are the first two moves and we assume $\mathbf{W}, \widehat{\mathbf{W}}$ get sent at the end. Denote the prover and the verifier of this part with $\langle \mathcal{P}^*_{\mathsf{RP}}, \mathcal{V}_{\mathsf{RP}} \rangle$.

(ii) $\Pi^{\otimes\mathsf{RP}}_{\mathsf{SZ}} : \Xi^{\mathsf{lin}} \to (\Xi^{\mathsf{lin}}, \Xi^{\mathsf{lin}})$ turns the above tuple of relations into

$$\left(\left(\begin{pmatrix} \mathbf{H} & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \mathbf{F} \\ \mathbf{c}_1^{\intercal} \cdot \widehat{\mathbf{J}} \end{pmatrix}, \begin{pmatrix} \mathbf{Y} \\ \mathbf{r}^{\intercal} \end{pmatrix}\right), \mathbf{W}\right)$$

and

$$\left(\left(\begin{pmatrix} \overline{\mathbf{F}} \\ \mathbf{c}^{\intercal} \end{pmatrix}, \begin{pmatrix} \overline{\mathbf{z}} \\ \mathbf{c}_0^{\intercal} \cdot \mathbf{r} \end{pmatrix}\right), \widehat{\mathbf{w}}\right).$$

These are moves three and four. Denote the prover and verifier of this part by $\langle \mathcal{P}^*_{\mathsf{SZ}}, \mathcal{V}_{\mathsf{SZ}} \rangle$.

We analyse the knowledge soundness of both parts separately, i.e., we construct an pair of extractors $\mathcal{E} = (\mathcal{E}_{\mathsf{RP}}, \mathcal{E}_{\mathsf{SZ}})$.

Extractor $\mathcal{E}_{\mathsf{SZ}}$. Without loss of generality, assume that the prover is deterministic. The extractor $\mathcal{E}_{\mathsf{SZ}}$ first runs the prover $\mathcal{P}^*_{\mathsf{SZ}}$ on some challenge $c \leftarrow\!\!\$\, \mathcal{C}_{\mathcal{R}_q}$. If the prover fails, $\mathcal{E}_{\mathsf{SZ}}$ aborts. Otherwise, write $c_0 = c^{m'}$, $c_1 = c$, $\mathbf{c}_0^{\intercal} = (1, c_0, \ldots, c_0^{r-1})$, $\mathbf{c}_1^{\intercal} = (1, c_1, \ldots, c_1^{m'-1})$, and $\mathbf{c} = \mathbf{c}_0 \otimes \mathbf{c}_1$ as in the construction. Hence, $\mathcal{E}_{\mathsf{SZ}}$ obtained $(\mathbf{r}, \mathbf{W}, \mathbf{S}, \widehat{\mathbf{w}})$, where $\mathbf{S} \in \mathcal{R}^{r\times r}$ is a diagonal matrix, satisfying the relations

$$\mathbf{HFW} = \mathbf{Y} \bmod q, \qquad \|\mathbf{S}\|_{\sigma\text{-}\mathsf{op}} \leq \varrho,$$

$$\begin{aligned} \overline{\mathbf{F}}\widehat{\mathbf{w}} &= \overline{\mathbf{z}} \bmod q, & \|\sigma(\mathbf{WS} \bmod q)\|_2 &\leq \beta', \\ \mathbf{c}_1^{\mathsf{T}} \cdot \widehat{\mathbf{J}} \cdot \mathbf{W} &= \mathbf{r}^{\mathsf{T}} \bmod q, & \|\sigma(\widehat{\mathbf{w}})\|_2 &\leq \beta'', \\ \mathbf{c}^{\mathsf{T}} \cdot \widehat{\mathbf{w}} &= \mathbf{c}_0^{\mathsf{T}} \cdot \mathbf{r} \bmod q & & . \end{aligned}$$

If $\mathsf{vec}(\widehat{\mathbf{JW}}) = \widehat{\mathbf{w}} \bmod q$, $\mathcal{E}_{\mathtt{SZ}}$ outputs $(\mathbf{W}, \widehat{\mathbf{JW}})$ and concludes the extraction. If not, suppose that $\mathsf{vec}(\widehat{\mathbf{JW}}) \neq \widehat{\mathbf{w}} \bmod q$. In this case, $\mathcal{E}_{\mathtt{SZ}}$ (repeatedly) re-runs the prover $\mathcal{P}^*_{\mathtt{SZ}}$ on random challenges $c' \leftarrow\!\$\, \mathcal{C}_{\mathcal{R}_q}$ until obtaining $(\mathbf{S}', \mathbf{W}', \widehat{\mathbf{w}}')$ satisfying

$$\begin{aligned} \mathbf{HFW}' &= \mathbf{Y} \bmod q, & \|\sigma(\mathbf{S}')\|_\infty &\leq \varrho, \\ \overline{\mathbf{F}}\widehat{\mathbf{w}}' &= \overline{\mathbf{z}} \bmod q, & \|\sigma(\mathbf{W}'\mathbf{S}' \bmod q)\|_2 &\leq \beta', \\ (\mathbf{c}_1')^{\mathsf{T}} \cdot \widehat{\mathbf{J}} \cdot \mathbf{W}' &= \mathbf{r}^{\mathsf{T}} \bmod q, & \|\sigma(\widehat{\mathbf{w}}')\|_2 &\leq \beta'', \\ (\mathbf{c}')^{\mathsf{T}} \cdot \widehat{\mathbf{w}} &= (\mathbf{c}_0')^{\mathsf{T}} \cdot \mathbf{r} \bmod q \end{aligned}$$

where $\mathbf{c}', \mathbf{c}_0', \mathbf{c}_1'$ are defined analogously.

If it happens that $\mathbf{W} \neq \mathbf{W}'$, then since $\mathbf{S}$ and $\mathbf{S}'$ are diagonal we have $\mathbf{HF}(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S}) = \mathbf{0} \bmod q$. Taking the top $\overline{n}$ rows yields $\overline{\mathbf{F}}(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S}) = \mathbf{0} \bmod q$. Furthermore, $\|\sigma(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \bmod q)\|_2 \leq 2\beta'\varrho \leq \hat{\beta}'$. Thus, the extractor can output any column $\mathbf{x}'$ of $\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \bmod q$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}') \in \Xi^{\mathsf{sis}}_{\hat{\beta}'}$. Similarly, if it happens that $\widehat{\mathbf{w}} \neq \widehat{\mathbf{w}}'$, then we have $\overline{\mathbf{F}}(\widehat{\mathbf{w}} - \widehat{\mathbf{w}}') = \mathbf{0} \bmod q$. Furthermore, $0 < \|\sigma(\widehat{\mathbf{w}} - \widehat{\mathbf{w}}')\|_2 \leq 2\beta'' \leq \hat{\beta}'$. Thus the extractor can output $\mathbf{x}'' \coloneqq (\widehat{\mathbf{w}} - \widehat{\mathbf{w}}')$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}'') \in \Xi^{\mathsf{sis}}_{\hat{\beta}'}$.

Now, assume that $\mathbf{W} = \mathbf{W}'$ and $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}'$ and thus $\mathsf{vec}(\widehat{\mathbf{JW}}) - \widehat{\mathbf{w}} = \mathsf{vec}(\widehat{\mathbf{JW}}') - \widehat{\mathbf{w}}' \neq \mathbf{0} \bmod q$. In this case, $\mathbf{c}'$ is uniformly random independently (conditioned on acceptance) of $\mathsf{vec}(\widehat{\mathbf{JW}}') - \widehat{\mathbf{w}}'$. Since $\mathsf{vec}(\widehat{\mathbf{JW}}') - \widehat{\mathbf{w}}' \neq \mathbf{0} \bmod q$, the event $\mathbf{c}'^{\mathsf{T}}(\mathsf{vec}(\widehat{\mathbf{JW}}') - \widehat{\mathbf{w}}') = 0 \bmod q$ happens with probability at most $\kappa_{\mathtt{SZ}}/\epsilon$, where $\kappa_{\mathtt{SZ}} \coloneqq (m-1)/|\mathcal{C}_{\mathcal{R}_q}|$ by the Schwartz-Zippel lemma. In the event that this does not happen, we have

$$\begin{aligned} (\mathbf{c}')^{\mathsf{T}}\widehat{\mathbf{w}}' &\neq (\mathbf{c}')^{\mathsf{T}}\mathsf{vec}(\widehat{\mathbf{JW}}') \\ &= (\mathbf{c}')^{\mathsf{T}}(\mathbf{I}_r \otimes \widehat{\mathbf{J}})\mathsf{vec}(\mathbf{W}') = ((\mathbf{c}_0')^{\mathsf{T}} \otimes (\mathbf{c}_1')^{\mathsf{T}} \cdot \widehat{\mathbf{J}})\mathsf{vec}(\mathbf{W}') = (\mathbf{c}_0')^{\mathsf{T}} \cdot \mathbf{r} \bmod q, \end{aligned}$$

contradicting with $(\mathbf{c}')^{\mathsf{T}}\widehat{\mathbf{w}}' = (\mathbf{c}_0')^{\mathsf{T}} \cdot \mathbf{r} \bmod q$. To summarise, with probability at least $\epsilon \cdot (1 - \kappa_{\mathtt{SZ}}/\epsilon) = \epsilon - \kappa_{\mathtt{SZ}}$, the extractor outputs either (i) $(\mathbf{W}, \widehat{\mathbf{W}})$ satisfying the starting pair of Ξ^{lin} relations, or (ii) $\mathbf{x}$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}) \in \Xi^{\mathsf{sis}}_{\hat{\beta}'}$.

Expected runtime: To find an accepting transcripts when sampling random challenges requires $\frac{1}{\epsilon}$ tries in expectation. Since $\mathcal{E}_{\mathtt{SZ}}$ only rewinds if the initial uniform challenge was successful, and then samples challenges uniformly at random, it follows by a standard argument that $\epsilon \cdot \frac{1}{\epsilon} = 1$ rewinds are required in expectation. Hence, $\mathcal{E}_{\mathtt{SZ}}$ is expected polynomial-time.

Extractor $\mathcal{E}_{\mathtt{RP}}$. Without loss of generality, assume that the prover is deterministic. Consider a knowledge extractor which runs the prover $\mathcal{P}^*_{\mathtt{RP}}$ on some challenge $\mathbf{J} \leftarrow \chi^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}_{\mathsf{ghl}}$. If the prover fails, $\mathcal{E}_{\mathtt{RP}}$ aborts. Otherwise, $\mathcal{E}_{\mathtt{RP}}$ obtains $(\mathbf{W}, \mathbf{S}, \widehat{\mathbf{w}})$ satisfying the relations

$$\begin{aligned} \mathbf{HFW} &= \mathbf{Y} \bmod q, & \|\mathbf{S}\|_{\sigma\text{-op}} &\leq \varrho, \\ \overline{\mathbf{F}} \cdot \widehat{\mathbf{w}} &= \overline{\mathbf{z}} \bmod q, \quad \text{where} \quad \widehat{\mathbf{w}} = \mathsf{vec}(\widehat{\mathbf{W}}), & \|\sigma(\mathbf{WS} \bmod q)\|_2 &\leq \beta', \\ \widehat{\mathbf{J}}\mathbf{W} - \widehat{\mathbf{W}} &= \mathbf{0} \bmod q, & \|\sigma(\widehat{\mathbf{w}})\|_2 &\leq \beta''. \end{aligned}$$

If $\|\sigma(\mathbf{W})\|_2 \leq \tilde{\beta}$, $\mathcal{E}_{\mathtt{RP}}$ outputs $\mathbf{W}$ and concludes the extraction. If not, assume that $\|\sigma(\mathbf{W})\|_2 > \tilde{\beta}$. In this case, $\mathcal{E}_{\mathtt{RP}}$ (repeatedly) re-runs the prover $\mathcal{P}^*_{\mathtt{RP}}$ on challenge $\mathbf{J}' \leftarrow \chi^{n_{\mathsf{rp}} \times m_{\mathsf{rp}}}_{\mathsf{ghl}}$ until obtaining $(\mathbf{S}', \mathbf{W}', \widehat{\mathbf{w}}')$ satisfying

$$\begin{aligned} \mathbf{HFW}' &= \mathbf{Y} \bmod q, & \|\sigma(\mathbf{S}')\|_\infty &\leq \varrho, \\ \overline{\mathbf{F}} \cdot \widehat{\mathbf{w}}' &= \overline{\mathbf{z}} \bmod q, & \|\sigma(\mathbf{W}'\mathbf{S}' \bmod q)\|_2 &\leq \beta', \\ \widehat{\mathbf{J}}\mathbf{W}' - \widehat{\mathbf{W}'} &= \mathbf{0} \bmod q, & \|\sigma(\widehat{\mathbf{w}}')\|_2 &\leq \beta'', \\ \widehat{\mathbf{w}}' &= \mathsf{vec}(\widehat{\mathbf{W}'}). \end{aligned}$$

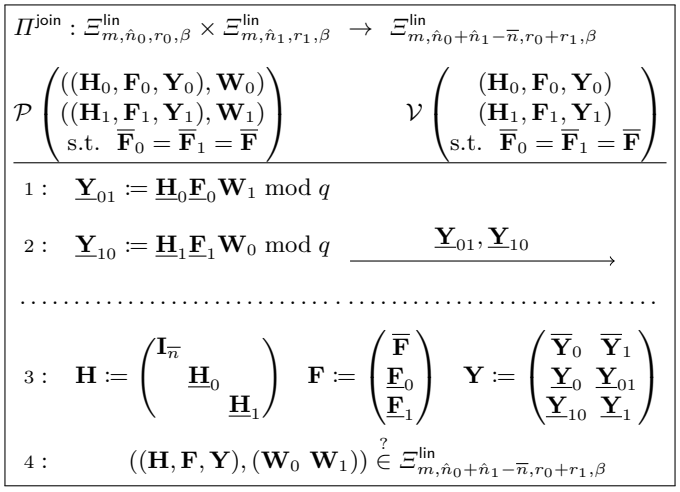

Fig. 4: RoK Π^{join}.

As before, if it happens that $\mathbf{W} \neq \mathbf{W}'$, we have $\overline{\mathbf{F}} \cdot (\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S}) = \mathbf{0} \bmod q$ and

$$0 < \|\sigma\left(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \bmod q\right)\|_2 \leq 2\beta'\varrho \leq \hat{\beta}'.$$

The extractor $\mathcal{E}_{\mathtt{RP}}$ outputs any column $\mathbf{x}$ of $\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \bmod q$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}) \in \Xi^{\mathsf{sis}}_{\hat{\beta}'}$.

Now, assume $\mathbf{W} = \mathbf{W}'$. We have $\widehat{\mathbf{w}}' = \mathsf{vec}(\widehat{\mathbf{J}'} \cdot \mathbf{W}) = (\mathbf{I}_{m/n_{\mathsf{rp}}} \otimes \mathbf{J}')\mathsf{vec}(\mathbf{W})$. Since $\mathbf{I}_{m/n_{\mathsf{rp}}} \otimes \mathbf{J}'$ is over $\mathbb{Z}$, we have $\mathsf{cf}(\widehat{\mathbf{w}}') = (\mathbf{I}_{\varphi \cdot m/n_{\mathsf{rp}}} \otimes \mathbf{J}') \cdot \mathsf{cf}(\mathsf{vec}(\mathbf{W}))$. Split $\mathsf{cf}(\widehat{\mathbf{w}}') = (\widehat{\mathbf{w}}'_i)_{i \in [\varphi \cdot m/n_{\mathsf{rp}}]}$ and $\mathsf{cf}(\mathsf{vec}(\mathbf{W})) = (\mathbf{w}_i)_{i \in [\varphi \cdot m/n_{\mathsf{rp}}]}$ into $\varphi \cdot m/n_{\mathsf{rp}}$ chunks where $\widehat{\mathbf{w}}'_i \in \mathbb{Z}^{n_{\mathsf{rp}}}$ and $\mathbf{w}_i \in \mathbb{Z}^{m_{\mathsf{rp}}}$ (by lifting from $\mathbb{Z}_q$ to $\mathbb{Z}$ with balanced representation). For each i, since $\mathbf{J}'$ is independent of $\mathbf{w}_i$, by Lemma 4 we have that with probability at least $1 - \delta$ that $\|\mathbf{w}_i\|_2 \leq 2 \cdot \sqrt{m_{\mathsf{rp}}} \cdot \|\widehat{\mathbf{w}}'_i\|_2$. It thus holds with probability at least $1 - \kappa_{\mathtt{RP}}/\epsilon$, where $\kappa_{\mathtt{RP}} \coloneqq \delta \cdot \varphi \cdot m/n_{\mathsf{rp}}$, that the above holds for all i, and therefore $\|\mathsf{cf}(\mathsf{vec}(\mathbf{W}))\|_2 \leq 2 \cdot \sqrt{m_{\mathsf{rp}}} \cdot \|\mathsf{cf}(\widehat{\mathbf{w}}')\|_2$. Using the conversion bounds from Lemma 2, we conclude that

$$\begin{aligned}\|\sigma\left(\mathbf{W}\right)\|_2 &\leq \|\sigma\left(\mathsf{vec}(\mathbf{W})\right)\|_2 \leq \sqrt{\hat{\mathfrak{f}}} \cdot \|\mathsf{cf}(\mathsf{vec}(\mathbf{W}))\|_2\\ &\leq 2 \cdot \sqrt{\hat{\mathfrak{f}} \cdot m_{\mathsf{rp}}} \cdot \|\mathsf{cf}(\widehat{\mathbf{w}}')\|_2 \leq 2 \cdot \sqrt{\mathsf{rad}(\mathfrak{f}) \cdot m_{\mathsf{rp}}} \cdot \|\sigma\left(\widehat{\mathbf{w}}'\right)\|_2\\ &\leq 2 \cdot \sqrt{\mathsf{rad}(\mathfrak{f}) \cdot m_{\mathsf{rp}}} \cdot \beta'' = \tilde{\beta}\end{aligned}$$

with probability at least $1 - \kappa_{\mathtt{RP}}/\epsilon$,

To summarise, with probability at least $\epsilon \cdot (1 - \kappa_{\mathtt{RP}}/\epsilon) = 1 - \kappa_{\mathtt{RP}}$, the extractor outputs either (i) $(\mathbf{W}, \mathbf{S})$ satisfying Ξ^{lin}, or (ii) $\mathbf{x}$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}) \in \Xi^{\mathsf{sis}}_{\hat{\beta}'}$.

Expected runtime: The extractor $\mathcal{E}_{\mathtt{RP}}$ is expected polynomial-time by an completely analogous argument to $\mathcal{E}_{\mathtt{SZ}}$.

Completed extractor. The extractor $\mathcal{E}$ for $\Pi^{\otimes\mathsf{RP}}$ runs $\mathcal{E}^{\mathcal{E}^{\mathcal{P}^*}_{\mathtt{SZ}}}_{\mathtt{RP}}$ (handling SIS breaks if needed). It succeeds probability at least $\epsilon - \kappa$, where $\kappa = \kappa_{\mathtt{SZ}} + \kappa_{\mathtt{RP}} = (m-1)/|\mathcal{C}_{\mathcal{R}_q}| + \delta \cdot \varphi \cdot m/n_{\mathsf{rp}} \leq m \cdot (1/|\mathcal{C}_{\mathcal{R}_q}| - \varphi \cdot \delta/\log(\delta)) = \kappa$ and makes at most 4 queries to $\mathcal{P}^*$ in expectation. □

4.3 Joining Relations

Protocol $\Pi^{\otimes\mathsf{RP}}$ outputs two relations sharing the same commitment key $\overline{\mathbf{F}}$. In the composition (see Fig. 2), the first relation is further folded and after this, those two relations are joined together by a horizontal concatenation of the witness. It remains to handle the technicality, that only the top part $\overline{\mathbf{F}}$ is shared by both relations. Figure 4 presents a simple protocol to move to a combined relation, by sending missing images. This step is implicit in many protocols, but in our case, it is useful to make it explicit. We leave the proof of Lemma 3 to the reader.

Lemma 3. *Π^{join} is a perfectly correct and perfectly knowledge sound reduction of knowledge for $\Xi^{\mathsf{lin}}_{m,\hat{n}_0,r_0,\beta} \times \Xi^{\mathsf{lin}}_{m,\hat{n}_1,r_1,\beta} \leftrightarrow \Xi^{\mathsf{lin}}_{m,\hat{n}_0+\hat{n}_1-\overline{n},r_0+r_1,\beta}$. The communication cost is $((\hat{n}_0-\overline{n})\cdot r_1 + (\hat{n}_1-\overline{n})\cdot r_0)\ \mathcal{R}_q$ elements.*

5 Unstructured Projections

The parameters for the structured projection RoK $\Pi^{\otimes\mathsf{RP}}$ obey $m_{\mathsf{rp}} \geq n_{\mathsf{rp}} \geq \Omega(\lambda)$ for λ bits of soundness. This imposes a lower-bound on $m \in O(\lambda)$, and hence a limitation on when $\Pi^{\otimes\mathsf{RP}}$ can be applied. Sending the witness in plain is not an option, as its size is still about $\approx \lambda^2$. In this section, we resolve these issues.

5.1 Overview

As a first step, use an unstructured projection Π^{RP} to replace $\Pi^{\otimes\mathsf{RP}}$, as in [BS23]. The idea is to send the projection of the coefficient embedding of the witness $\mathbf{W}$ in the clear. We denote the coefficient embedding as $\mathfrak{cf}(\mathbf{W})$ and send the projection $\mathbf{T}_0 := \mathbf{J}\mathfrak{cf}(\mathbf{W}) \bmod q$ for $\mathbf{J} \leftarrow \chi$, or equivalently, $\mathbf{T}_0 := \mathsf{Trace}(\mathfrak{cf}_\vee^{-1}(\mathbf{J})\mathbf{W}) \bmod q$.

To prove such a trace relation, our strategy is to lift it to an equation which conforms to our Ξ^{lin} relation. Naively, we could send $\mathfrak{cf}_\vee^{-1}(\mathbf{J})\mathbf{W}$ and let the verifier compute the trace $\mathbf{T}_0$. However, this incurs $n_{\mathsf{rp}}\varphi\log(q) \approx \lambda^2$ bits of communication, which is too much. Our idea to solve this is to assume and exploit a d-smooth tower structure of $\mathcal{R}$, i.e. assume that

$$\mathcal{R} = \mathcal{R}_\ell \supset \mathcal{R}_{\ell-1} \supset \cdots \supset \mathcal{R}_1 \supset \mathcal{R}_0 = \mathbb{Z}. \quad \text{and} \quad [\mathcal{R}_{i+1} : \mathcal{R}_i] \leq d$$

The trick is to *gradually* lift the n_{rp} trace value from $\mathcal{R}_i$ to $\mathcal{R}_{i+1}$, while *simultaneously* applying batch verification to reduce the n_{rp} elements in $\mathcal{R}_i$ to fewer elements in $\mathcal{R}_{i+1}$, and so on. That is, before sending $\mathbf{T}_{i+1}$, we batch it into $\mathbf{T}_{i+1} := \mathbf{C}_i\mathbf{T}_i$ where $\mathbf{C}_i$ is a suitable random matrix over $\mathcal{R}_i$. The matrix $\mathbf{C}_i$ is a "batching matrix", and chosen so that the knowledge error is at most $2^{-\lambda}$. Lifting an element from $\mathcal{R}_i$ to $\mathcal{R}_{i+1}$ incurs a multiplicative overhead of $[\mathcal{R}_{i+1} : \mathcal{R}_i] \leq d$ (which is $O(1)$ for our parameters). The verifier must also ensure that the claimed lift value preserves the relative trace, i.e. $\mathsf{Trace}_{\mathcal{R}_{i+1}/\mathcal{R}_i}(\mathbf{T}_{i+1}) = \mathbf{C}_i\mathbf{T}_i$.

For suitable choices of parameters, the sketched protocol has $O(\lambda\log(q)^2)$ communication (for $m \in O(\lambda)$). To ensure soundness of batching, we further require that each intermediate extension $\mathcal{R}_i$ admits a "relatively large" subfield modulo q. Concretely, we write $\mathcal{Q}_i = \mathcal{R}_i/(q\mathcal{R}_i) \cong (\mathbb{F}_{q^{e_i}})^{\varphi_i/e_i}$. Given t_{i+1} rows in $\mathbf{C}_i$, we can achieve a knowledge error of $(t_i/q^{e_i})^{t_{i+1}}$. For favourable rings, e.g. if some intermediate $\mathcal{Q}_i$ is a field of size $2^{-\lambda}$, we observe that communication for the batching phase is bounded by $\ell|\mathcal{R}_q|$.

5.2 Formal Construction and Analysis

Fig. 5 shows the protocol described in Section 5.1 with a quasilinear communication cost by employing the tower structure of $\mathcal{R}$.

Theorem 2 (Π^{RP}). *Let $0 < \delta < 1$, $n_{\mathsf{rp}} \geq \log_2(1/\delta)$, $\beta, \beta' < q$ and $\varrho \geq 1$. Let $\mathcal{R} = \mathcal{R}_\ell \supset \mathcal{R}_{\ell-1} \supset \cdots \supset \mathcal{R}_1 \supset \mathcal{R}_0 = \mathbb{Z}$ be a tower of rings of integers of cyclotomic fields where $\mathcal{R}_i$ is of conductor $\mathfrak{f}_i$ and degree $\varphi_i = \varphi(\mathfrak{f}_i)$. Denote $\mathcal{Q}_i := \mathcal{R}_i/q\mathcal{R}_i \cong (\mathbb{F}_{q^{e_i}})^{\varphi_i/e_i}$ where e_i is the multiplicative order of q modulo $\mathfrak{f}_i$ (so $e_\ell \geq e_{\ell-1} \geq \ldots \geq e_0 = 1$). Set $t_0 := n_{\mathsf{rp}}$ and let $t_1, \ldots, t_\ell \in \mathbb{N}$ be arbitrary (typically chosen such that $(t_i/q^{e_i})^{t_{i+1}} \leq 2^{-\lambda}$ for every $i \in [\ell]$).*

Then Π^{RP} is correct with a correctness error δ for

$$\Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m,\hat{n},r,\beta} \ \rightarrow\ \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m,\hat{n}+t_\ell,r,\beta}.$$

Let $\tilde{\beta} := 2m\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\beta$. Then Π^{RP} is knowledge sound with a knowledge error $\kappa \leq \delta + \sum_{i\in[\ell]}(t_i/q^{e_i})^{t_{i+1}}$ for

$$\Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m,\hat{n},r,\tilde{\beta}} \cup \Xi^{\mathsf{sis}}_{2\beta'\varrho} \ \leftarrow\ \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m,\hat{n}+t_\ell,r,\beta',\varrho}$$

Proof. Let $\hat{\beta} := \sqrt{m\varphi n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})/\mathfrak{f}}\beta$ as in Fig. 5 for the rest of the proof.

Correctness: Consider an instance $((\mathbf{H},\mathbf{F},\mathbf{Y}),\mathbf{W}) \in \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m,\hat{n},r,\beta}$. Observe that

$$\widehat{\mathbf{J}}\mathbf{W} = \mathbf{T} \bmod q \implies \mathbf{J}\mathfrak{cf}(\mathbf{W}) = \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_0}(\mathbf{T}) = \mathbf{T}_0 \bmod q.$$

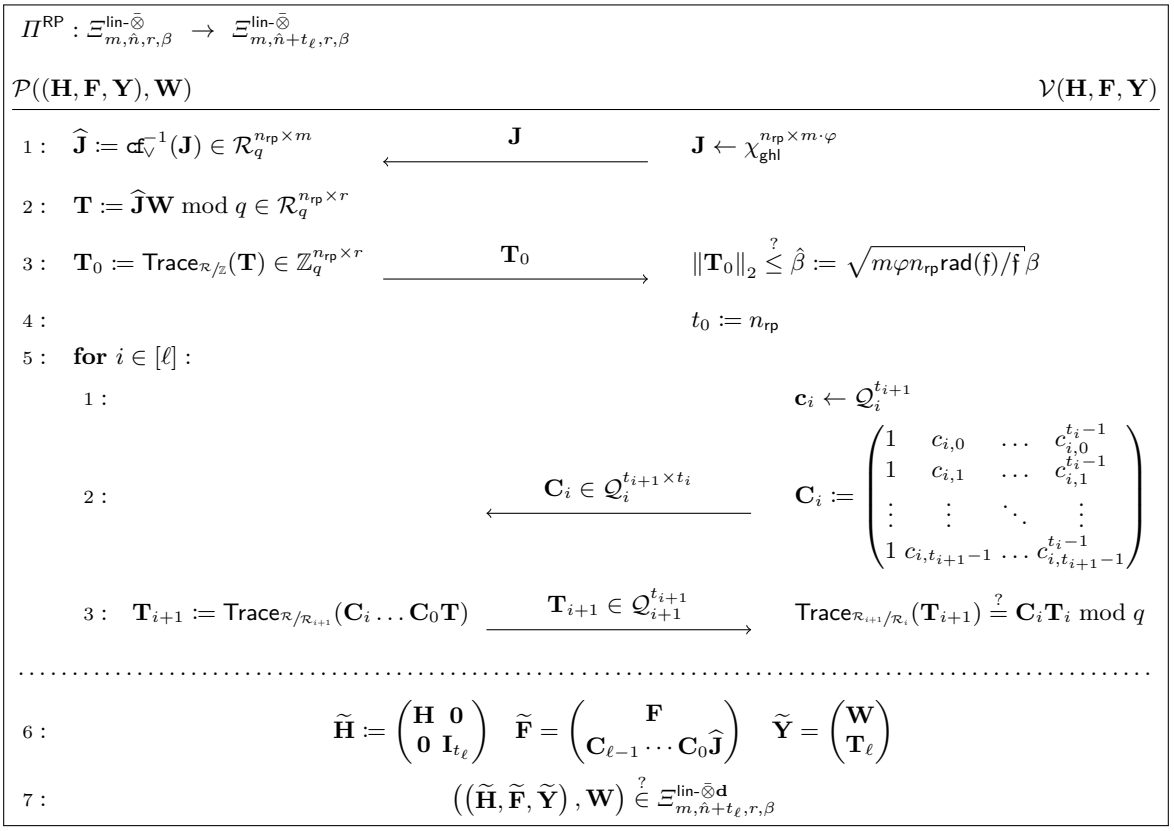

Fig. 5: RoK Π^{RP}.

As $\|\sigma(\mathbf{W})\|_2 \leq \beta$, we apply the conversion bound in Lemma 2 to obtain $\|\mathfrak{cf}(\mathbf{W})\|_2 \leq \sqrt{\mathsf{rad}(\mathfrak{f})/\mathfrak{f}}\beta$. Therefore, except with probability δ, due to Lemma 4, we have

$$\|\mathbf{T}_0\|_2 \leq \sqrt{m\varphi n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})/\mathfrak{f}}\,\beta = \hat{\beta}.$$

We next analyse the for-loop. For each $i \in [\ell]$, since $\mathbf{C}_i$ is over $\mathcal{Q}_i$, it holds that $\mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_i) = \mathbf{C}_i$, and thus

$$\begin{aligned} \mathbf{T}_{i+1} &= \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_{i+1}}(\mathbf{C}_i \dots \mathbf{C}_0\mathbf{T}), \\ \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{T}_{i+1}) &= \mathbf{C}_i\mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_{i-1} \dots \mathbf{C}_0\mathbf{T}) = \mathbf{C}_i\mathbf{T}_i \bmod q. \end{aligned}$$

Finally, since $\mathbf{T}_\ell$ is over $\mathcal{Q}_\ell = \mathcal{R}_q$, it holds as desired that

$$\mathbf{T}_\ell = \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_\ell}(\mathbf{C}_{\ell-1} \dots \mathbf{C}_0\mathbf{T}) = \mathbf{C}_{\ell-1} \dots \mathbf{C}_0\mathbf{T} = \mathbf{C}_{\ell-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W} \bmod q.$$

Knowledge-soudness: For the ease of analysis, we split the protocol into multiple parts, namely:

(i) The first reduction "$\Pi^{\mathsf{RP}}_{\mathsf{RP}} \colon \Xi^{\mathsf{lin}} \to (\Xi^{\mathsf{lin}}, \mathsf{RP})$" turns an instance $((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S}))$ into an instance which additionally bounds the trace of $\mathbf{T}_0$, namely

$$\begin{aligned} &((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S})) \\ \mapsto &\left[((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S})) \;\text{ and }\; \mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\widehat{\mathbf{J}}\mathbf{W}) = \mathbf{T}_0, \|\mathbf{T}_0\|_2 \leq \hat{\beta}\right]. \end{aligned}$$

This corresponds to the first challenge-response, denoted by $\langle \mathcal{P}^*_{\mathsf{RP}}, \mathcal{V}_{\mathsf{RP}} \rangle$.

(ii) The loop "$\Pi^{\mathsf{RP}}_{\mathsf{SZ}_i} \colon (\Xi^{\mathsf{lin}}, \mathsf{RP}_i) \to (\Xi^{\mathsf{lin}}, \mathsf{RP}_{i+1})$" transforms instances as

$$\left[((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S})) \;\text{ and }\; \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) = \mathbf{T}_i\right]$$

$$\mapsto \left[((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S})) \;\text{ and }\; \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_{i+1}}(\mathbf{C}_i \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) = \mathbf{T}_{i+1}\right].$$

This corresponds to the i-th loop challenge-response, denoted by $\left\langle \mathcal{P}^*_{\mathtt{SZ}_i}, \mathcal{V}_{\mathtt{SZ}_i}\right\rangle$.

(iii) The final transformation "$\Pi^{\mathsf{RP}}_{\mathtt{finish}} : (\Xi^{\mathsf{lin}}, \mathsf{RP}_\ell) \to \Xi^{\mathsf{lin}}$" (to the intended relation) transforms instances as

$$\begin{aligned}&((\mathbf{H},\mathbf{F},\mathbf{Y}),(\mathbf{W},\mathbf{S})) \;\text{ and }\; \mathsf{Trace}_{\mathcal{R}/\mathcal{R}}(\mathbf{C}_{\ell-i} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) = \mathbf{T}_\ell\\ &\mapsto \left(\left(\begin{pmatrix}\mathbf{H} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}, \begin{pmatrix}\mathbf{F}\\ \mathbf{C}_{\ell-1}\cdots\mathbf{C}_0\widehat{\mathbf{J}}\end{pmatrix}, \begin{pmatrix}\mathbf{Y}\\ \mathbf{T}_\ell\end{pmatrix}\right), (\mathbf{W},\mathbf{S})\right).\end{aligned}$$

We analyse the knowledge soundness of $\Pi^{\mathsf{RP}}_{\mathtt{RP}}$ and $\Pi^{\mathsf{RP}}_{\mathtt{SZ},i}$ separately, and construct an extractor $\mathcal{E} = (\mathcal{E}_{\mathtt{RP}}, \mathcal{E}_{\mathtt{SZ}_0}, \dots \mathcal{E}_{\mathtt{SZ}_{\ell-1}})$ for Π^{RP}. The extractor for $\Pi^{\mathsf{RP}}_{\mathtt{finish}}$ is trivial.

Extractor $\mathcal{E}_{\mathtt{SZ}_i}$. W.l.o.g., assume that the prover is deterministic. The extractor $\mathcal{E}_{\mathtt{SZ}_i}$ runs the prover $\mathcal{P}^*_{\mathtt{SZ}_i}$ on some challenge $\mathbf{c}_i \leftarrow \mathcal{Q}_i^{t_{i+1}}$. If the prover fails, $\mathcal{E}_{\mathtt{SZ}_i}$ aborts. Otherwise, $\mathcal{E}_{\mathtt{SZ}_i}$ obtains $(\mathbf{T}_{i+1}, \mathbf{W}, \mathbf{S})$ satisfying the relations

$$\begin{aligned}\mathbf{HFW} &= \mathbf{Y} \bmod q, & \|\mathbf{S}\|_{\sigma\text{-op}} &\le \varrho,\\ \mathsf{Trace}_{\mathcal{R}_{i+1}/\mathcal{R}_i}(\mathbf{T}_{i+1}) &= \mathbf{C}_i\mathbf{T}_i \bmod q & \|\sigma\left(\mathbf{WS} \bmod q\right)\|_2 &\le \beta',\\ \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_{i+1}}(\mathbf{C}_i \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) &= \mathbf{T}_{i+1} \bmod q\end{aligned}$$

If $\mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) = \mathbf{T}_i \bmod q$, $\mathcal{E}_{\mathtt{SZ}_i}$ outputs $(\mathbf{W},\mathbf{S})$ and concludes the extraction. Else $\mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) \neq \mathbf{T}_i \bmod q$ and $\mathcal{E}_{\mathtt{SZ}_i}$ re-runs the prover $\mathcal{P}^*_{\mathtt{SZ}_i}$ on random challenges $\mathbf{c}_i \leftarrow \mathcal{Q}_i^{t_{i+1}}$ (possibly many times) until obtaining $(\mathbf{S}', \mathbf{W}', \mathbf{T}'_{i+1})$ satisfying

$$\begin{aligned}\mathbf{HFW}' &= \mathbf{Y} \bmod q, & \|\sigma\left(\mathbf{S}'\right)\|_\infty &\le \varrho,\\ \mathsf{Trace}_{\mathcal{R}_{i+1}/\mathcal{R}_i}(\mathbf{T}'_{i+1}) &= \mathbf{C}'_i\mathbf{T}_i \bmod q & \|\sigma\left(\mathbf{W}'\mathbf{S}' \bmod q\right)\|_2 &\le \beta',\\ \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_{i+1}}(\mathbf{C}'_i\mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) &= \mathbf{T}'_{i+1} \bmod q\end{aligned}$$

If it happens that $\mathbf{W} \neq \mathbf{W}' \bmod q$, then we have $\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \neq \mathbf{0} \bmod q$. Note that $\mathbf{HF}(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S}) = \mathbf{0} \bmod q$. Taking the top $\overline{n}$ rows yields $\overline{\mathbf{F}}(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S}) = \mathbf{0} \bmod q$. Furthermore, $\|\sigma\left(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \bmod q\right)\|_2 \le 2\beta'\varrho$. Thus, the extractor can output any non-zero column $\mathbf{x}'$ of $\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S} \bmod q$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}') \in \Xi^{\mathsf{sis}}_{2\beta'\varrho}$. From now, we suppose that $\mathbf{W} = \mathbf{W}' \bmod q$ and:

$$\begin{aligned}\mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) &\neq \mathbf{T}_i \bmod q\\ \mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}'_i, \mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) &= \mathbf{C}'_i\mathbf{T}_i \bmod q.\end{aligned}$$

Since $\mathbf{C}'_i \in \mathcal{Q}_i$, we have $\mathbf{C}'_i\mathsf{Trace}_{\mathcal{R}/\mathcal{R}_i}(\mathbf{C}_{i-1} \dots \mathbf{C}_0\widehat{\mathbf{J}}\mathbf{W}) = \mathbf{C}'_i\mathbf{T}_i \bmod q$. As $\mathbf{C}'_i$ is random independently of $\mathbf{W}$, we conclude that this can only happen with probability at most $(t_i/q^{e_i})^{t_{i+1}}/\epsilon$ by the Schwartz-Zippel lemma (routinely amplified by t_{i+1} rows).

To summarise, with probability at least $\epsilon - (t_i/q^{e_i})^{t_{i+1}}$, the extractor outputs either (i) $(\mathbf{W},\mathbf{S})$ satisfying the starting relation, or (ii) $\mathbf{x}'$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}') \in \Xi^{\mathsf{sis}}_{2\beta'\varrho}$.

Expected runtime: By the same argument as in the proof of Theorem 1, $\mathcal{E}_{\mathtt{SZ}_i}$ invokes $\mathcal{P}^*_{\mathtt{SZ}_i}$ 2 times in expectation.

Extractor $\mathcal{E}_{\mathtt{RP}}$. W.l.o.g., assume that the prover is deterministic. The extractor $\mathcal{E}_{\mathtt{RP}}$ runs the prover $\mathcal{P}^*_{\mathtt{RP}}$ on some challenge $\mathbf{J} \leftarrow \chi_{\mathsf{ghl}}$. If the prover fails, $\mathcal{E}_{\mathtt{RP}}$ aborts. Otherwise, $\mathcal{E}_{\mathtt{RP}}$ obtains $(\mathbf{T}_0, \mathbf{S}, \mathbf{W})$ satisfying the relations

$$\begin{aligned}\mathbf{HFW} &= \mathbf{Y} \bmod q, & \|\mathbf{S}\|_{\sigma\text{-op}} &\le \varrho, & \|\mathbf{T}_0\|_2 &\le \hat{\beta},\\ \mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\widehat{\mathbf{J}}\mathbf{W}) &= \mathbf{T}_0 \bmod q, & \|\sigma\left(\mathbf{WS} \bmod q\right)\|_2 &\le \beta'.\end{aligned}$$

If $\|\sigma\left(\mathbf{W}\right)\|_2 \le \tilde{\beta}$, $\mathcal{E}_{\mathtt{RP}}$ outputs $(\mathbf{W},\mathbf{S})$ and concludes the extraction. If not, suppose that $\|\sigma\left(\mathbf{W}\right)\|_2 \not\le \tilde{\beta}$ In this case, $\mathcal{E}_{\mathtt{RP}}$ re-run the prover $\mathcal{P}^*_{\mathtt{RP}}$ on the challenge $\mathbf{J}' \leftarrow \chi_{\mathsf{ghl}}$ (possibly many times) until obtaining $(\mathbf{S}', \mathbf{W}', \mathbf{T}'_0)$ satisfying

$$\mathbf{HFW}' = \mathbf{Y} \bmod q, \qquad \|\sigma\left(\mathbf{S}'\right)\|_\infty \le \varrho, \qquad \|\mathbf{T}'_0\|_2 \le \hat{\beta}$$

$$\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\widehat{\mathbf{J}}\mathbf{W}') = \mathbf{T}'_0 \bmod q \qquad \|\sigma\left(\mathbf{W}'\mathbf{S}' \bmod q\right)\|_2 \leq \beta'.$$

Again, if it happens that $\mathbf{W} \neq \mathbf{W}' \bmod q$, then we have $(\mathbf{WSS}' - \mathbf{W}'\mathbf{S}'\mathbf{S}) \neq \mathbf{0} \bmod q$. As before, the extracor $\mathcal{E}_{\mathtt{RP}}$ can derive a SIS break $(\overline{\mathbf{F}}, \mathbf{x}') \in \Xi^{\mathsf{sis}}_{2\beta'\varrho}$.

Now, assume that $\mathbf{W} = \mathbf{W}' \bmod q$. Since the transcripts are accepting, we have $\|\mathbf{J}\mathsf{cf}(\mathbf{W}) \bmod q\|_2 = \|\mathsf{Trace}_{\mathcal{R}/\mathbb{Z}}(\widehat{\mathbf{J}}\mathbf{W}) \bmod q\|_2 \leq \hat{\beta}$. Since $\mathbf{J}$ is independent of $\mathbf{W}$, by Lemma 4 we get $\|\mathsf{cf}(\mathbf{W})\|_2 \leq 2\sqrt{\varphi m}\hat{\beta}$ which implies

$$\|\sigma\left(\mathbf{W}\right)\|_2 \leq 2\sqrt{\varphi m \hat{\mathfrak{f}}}\,\hat{\beta} \leq 2\sqrt{\varphi m \hat{\mathfrak{f}}}\,\sqrt{m\varphi n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})/\mathfrak{f}}\,\beta \leq 2m\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\,\beta = \tilde{\beta},$$

unless with a probability δ/ϵ.

To summarise, with probability at least $\epsilon - \delta$, the extractor outputs either (i) $(\mathbf{W}, \mathbf{S})$ satisfying the $\Xi^{\mathsf{lin}}_{\tilde{\beta}}$ relation, (ii) $\mathbf{x}'$ satisfying $(\overline{\mathbf{F}}, \mathbf{x}') \in \Xi^{\mathsf{sis}}_{2\beta'\varrho}$.

Expected runtime: By the same argument as in the proof of Theorem 1, $\mathcal{E}_{\mathtt{RP}}$ invokes $\mathcal{P}^*_{\mathtt{RP}}$ 2 times in expectation.

Summary. We conclude that the extractor $\mathcal{E}$ succeeds with $\epsilon - \kappa$ probability, where $\kappa \leq \delta + \sum_{i\in[\ell]}(t_i/q^{e_i})^{t_{i+1}}$. □

6 Composition

In this section, we summarise how to compose the new and old reductions of knowledge to obtain a succinct proof system with a subquadratic (in λ) communication complexity. Below, we present the composition in the summarised form, and defer the details to Section B.

6.1 Summary of RoKs

To construct a succinct argument for Ξ^{lin}, we utilise, with minor adaptations, the RoKs $\Pi^{b\text{-decomp}}$, Π^{batch}, $\Pi^{\mathsf{split}\text{-}\otimes}$, Π^{fold} and Π^{norm} from [KLNO24], and new RoKs $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$, Π^{join}, Π^{RP}, and $\Pi^{\otimes\mathsf{RP}}$ which are introduced in this work. For a brief summary of the functionalities of these RoKs, we recall that:

(i) $\Pi^{b\text{-decomp}}$ b-ary decomposes $\mathbf{W}$, reducing β but increasing r.
(ii) Π^{batch} reduces $\hat{n}$ to $\underline{n} + 1$.
(iii) $\Pi^{\mathsf{split}\text{-}\otimes}$ reshapes $\mathbf{W}$ from $m \times r$ to $(rd) \times (m/d)$.
(iv) Π^{fold} folds $\mathbf{W}$ and reduces r to 1.
(v) Π^{norm} implements an exact norm-check.
(vi) $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$ is almost the same as $\Pi^{\mathsf{split}\text{-}\otimes}$, but for unstructured $\underline{\mathbf{F}}$, see Lemma 10.
(vii) Π^{join} joins two claims together, see Lemma 3.
(viii) Π^{RP} and $\Pi^{\otimes\mathsf{RP}}$ are approximate norm checks via (un)structured random projections, see Theorems 1 and 2. Note that $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$ is not verifier-succinct in general, but it is only intended to be used when $m \in \mathsf{poly}(\lambda)$.

Each of the RoKs above consists of constant many moves and sends at most $O(r)$ $\mathcal{R}_q$-elements, which will be $O(1)$ for our parameter choices. A more precise summary of the parameter changes induced by all "supportive RoKs", i.e. all except Π^{RP} and $\Pi^{\otimes\mathsf{RP}}$, is given in Table 1. The precise parameters for Π^{RP} and $\Pi^{\otimes\mathsf{RP}}$ are given in Theorems 1 and 2.

Although we are importing RoKs from [KLNO24] only with minor changes, we have to re-examine their soundness particularly in the presence of slack. The only RoK which *introduces* a multiplicative factor to the slack is Π^{fold}, where during witness extraction the extracted witness is a fraction with denominator given by the difference of two challenges. Other RoKs may amplify slack, i.e. raising it to some power, but do not introduce any multiplicative factors, i.e. if $\varrho = 1$ is guaranteed for the output claim then $\varrho = 1$ holds for the extracted witness. Formally, these claims can be proven by simple adaptations of the soundness proofs in [KLNO24], which we present in Section B.1.

6.2 Summary of Structured and Unstructured Loops

We overview our succinct argument for Ξ^{lin}. The protocol is divided into two phases: the structured loop and the unstructured loop.

Protocol	Reduction direction	Extraction direction
$\Pi^{b\text{-decomp}}$	$\Xi^{\mathsf{lin}}_{r,\beta} \mapsto \Xi^{\mathsf{lin}}_{r\ell,\frac{1}{2}\sqrt{m}\sqrt{\hat{\mathsf{f}}\varphi}\,b}$	$\Xi^{\mathsf{lin}}_{r,\frac{b^\ell-1}{b-1}\beta'} \mapsto \Xi^{\mathsf{lin}}_{r\ell,\beta'}$
Π^{batch}	$\Xi^{\mathsf{lin}}_{\hat{n},\beta} \mapsto \Xi^{\mathsf{lin}}_{\overline{n}+1,\beta}$	$\Xi^{\mathsf{lin}}_{\hat{n},\beta',\varrho} \mapsto \Xi^{\mathsf{lin}}_{\overline{n}+1,\beta',\varrho}$
$\Pi^{\mathsf{split}\text{-}\otimes}$	$\Xi^{\mathsf{lin}\text{-}\otimes d}_{m,r,\beta} \mapsto \Xi^{\mathsf{lin}\text{-}\otimes}_{m/d,r\cdot d,\beta}$	$\Xi^{\mathsf{lin}\text{-}\otimes d}_{m,\varrho^{d-1}\sqrt{d}\beta',\varrho^d} \mapsto \Xi^{\mathsf{lin}\text{-}\otimes}_{m/d,r\cdot d,\beta',\varrho}$
$\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$	$\Xi^{\mathsf{lin}\text{-}\bar{\otimes} d}_{m,\hat{n},r,\beta} \mapsto \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m/d,\overline{n}+\underline{n}d,r\cdot d,\beta}$	$\Xi^{\mathsf{lin}\text{-}\bar{\otimes} d}_{m,\hat{n},r,\rho^{d-1}\sqrt{d}\beta',\varrho^d} \mapsto \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m/d,\overline{n}+\underline{n}d,r\cdot d,\beta',\varrho}$
Π^{fold}	$\Xi^{\mathsf{lin}}_{m,r,\beta} \mapsto \Xi^{\mathsf{lin}}_{m,2r\Gamma\beta}$	$\Xi^{\mathsf{lin}}_{m,r,2\Gamma\varrho^{r-1}\beta',2\Gamma\varrho^r} \mapsto \Xi^{\mathsf{lin}}_{m,\beta',\varrho}$
Π^{norm}	$\Xi^{\mathsf{lin}}_{m,\hat{n},r,\beta} \mapsto \Xi^{\mathsf{lin}}_{m,\hat{n}+3,r+\ell,\beta_{\mathsf{out}}}$	$\Xi^{\mathsf{lin}}_{m,\hat{n},r,\sqrt{r}\beta} \mapsto \Xi^{\mathsf{lin}}_{m,\hat{n}+3,r+\ell,\beta_{\mathsf{out}}'}$
Π^{join}	$\Xi^{\mathsf{lin}}_{n_0,r_0} \times \Xi^{\mathsf{lin}}_{n_1,r_1} \mapsto \Xi^{\mathsf{lin}}_{n_0+n_1-\overline{n},r_0+r_1}$	$\Xi^{\mathsf{lin}}_{n_0,r_0} \times \Xi^{\mathsf{lin}}_{n_1,r_1} \mapsto \Xi^{\mathsf{lin}}_{n_0+n_1-\overline{n},r_0+r_1}$

Table 1: Summary of supportive RoKs. All knowledge errors are $\mathsf{poly}(\lambda)/|\mathcal{C}|$. For $\Pi^{b\text{-decomp}}$, we let $\ell = \lceil \log_b(2\beta+1) \rceil$. For Π^{norm}, we let $\ell = \lceil \log_b(2\beta_{\mathbf{V}}+1) \rceil$, where $\beta_{\mathbf{V}}$ is a parameter of Π^{norm} and $\beta_{\mathsf{out}} = \sqrt{\beta^2 + \beta_{\mathbf{V}}{}^2}$, see Lemma 12.

Structured Loop. The structured loop exploits the row-structure of the matrix $\mathbf{F}$ for efficient reductions. It runs the following sequence of RoKs:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-decomp}} \to \Pi^{\mathsf{split}\text{-}\otimes} \to \Pi^{\otimes\mathsf{RP}} \to \begin{matrix}\Pi^{\mathsf{fold}}\\ \Pi^{\mathsf{id}}\end{matrix} \to \Pi^{\mathsf{join}} \to \Pi^{\mathsf{batch}}$$

where Π^{id} is a dummy RoK which acts as the identity function.

Compared to [KLNO24], we have added $\Pi^{\otimes\mathsf{RP}}$ to the sequence, which is used to "upgrade" (the guarantee of) the extracted witness satisfying a Ξ^{lin} relation with slack to one without slack. This mechanism permits the instantiation of the challenge set $\mathcal{C}_{\mathcal{R}}$ in Π^{fold} with an unstructured set of short elements instead of subtractive sets. By using an exponential-size set $\mathcal{C}_{\mathcal{R}}$, we can reduce the number r of columns of the witness to $O(1)$ (compared to $\Omega(\lambda/\log(\lambda))$ needed in [KLNO24] to compensate for the use of polynomial-size subtractive sets). Consequently, the communication cost of a single round decreases to $O(1)$ ring elements.

The above sequence of RoKs is repeated $\mu = O(\log_\lambda(m))$ times, after which the witness is small enough to be handled by the unstructured loop. This switch is needed, because the structured random projection $\Pi^{\otimes\mathsf{RP}}$ requires $m \in \Omega(\lambda)$. For the unstructured loop, we simply "forget" the structure of $\underline{\mathbf{F}}$. More details on the structured loop can be found in Section B.2.

Unstructured Loop. Once the structured loop reduces the witness dimension to $\tilde{m} \leq O(\lambda)$, the unstructured loop is employed to handle the remaining unstructured components of $\mathbf{F}$. The sequence of RoKs in this phase is:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-decomp}} \to \Pi^{\mathsf{split}\text{-}\bar{\otimes}} \to \Pi^{\mathsf{RP}} \to \Pi^{\mathsf{fold}} \to \Pi^{\mathsf{batch}}$$

This phase is repeated $\tilde{\mu} = O(\log \tilde{m})$ times. Again, the communication cost of each round is $O(1)$ ring elements, therefore the overall communication cost of the unstructured loop is $O(\log \tilde{m})$ ring elements. Details on the unstructured loop can be found in Section B.3.

After the unstructured loop, the final witness is of size $O(1)$ ring elements and can be sent in plain to the verifier. Therefore, the total communication cost of the protocol is $O_\lambda(\log m)$ ring elements, or more precisely, $O\left(\lambda \cdot \log^3 m / \log \lambda\right)$.

Efficiency Comparison. To highlight the competitiveness of our proof size, we compare our protocol with that of [KLNO24], demonstrating that our method yields more compact proofs. Importantly, our enhancements offer significant scalability benefits, particularly as the security parameter increases. In Table 2 we present scalability benchmarks for typical values of the security parameters λ and κ. Section B.5 provides a more detailed discussion of the parameter selection process.

7 Random Projections modulo q

This section presents self-contained auxiliary lemmas for random projections with entries sampled from the distribution χ_{ghl} [GHL22] recalled below.

(λ, κ)	$(80, 2^{-80})$	$(128, 2^{-128})$	$(196, 2^{-196})$	$(256, 2^{-256})$
witness length in $\mathbb{Z}$-elements	$\approx 2^{32}$	$\approx 2^{32}$	$\approx 2^{32}$	$\approx 2^{32}$
$\mathfrak{f}$	60	60	120	120
m	$\approx 2^{28}$	$\approx 2^{28}$	$\approx 2^{27}$	$\approx 2^{27}$
rounds	16	16	16	15
witness size	7.69 GB	7.69 GB	7.69 GB	7.69 GB
proof size	2.46 MB	4.07 MB	5.61 MB	7.79 MB
proof size for similar instance using [KLNO24]	12.99 MB	27.57 MB	70.55 MB	112.13 MB

Table 2: Concrete parameters with proof sizes.

Definition 1. *The distribution χ_{ghl} is defined as the* $\mathsf{Bin}(2, \frac{1}{2}) - 1$*, i.e., it is* 0 *with probability* $\frac{1}{2}$*,* 1 *(resp.* -1*) with probability* $\frac{1}{4}$*.*

The main goal is to prove Lemma 4 which is in our proofs of Theorems 1 and 2.

Our aim is to establish a probabilistic upper bound on the norm of the random projection of a vector $\mathbf{x}$ sampled from χ_{ghl}^n. Specifically, we examine $\mathbf{Ax} \bmod q$, where $\mathbf{A} \in \mathcal{R}^{n \times m}$ is a random matrix with entries drawn from χ_{ghl}. Previous works, such as [GHL22] and [BS23], derived bounds under heuristic assumptions (justified by the Johnson–Lindenstrauss lemma and experiments), with the benefit of obtaining lower and upper bounds independent of m. However, our focus is on relatively small matrices $\mathbf{A}$, with dimensions $m \in O(\lambda)$ in the structured case, and $m \in O(\lambda\varphi)$ in the unstructured case. This allows us to consider less tight bounds which depend on m but still suffice for our asymptotic results.[19] We are able to prove these bounds by simpler and more direct arguments without relying on any heuristics.

Lemma 4 provides straightforward lower and upper bounds on $\mathbf{Ax} \bmod q$. These bounds play a crucial role in our asymptotic composition and in the proofs of Theorems 1 and 2. The proof of Lemma 4 relies on Lemma 5, which allows for even tighter lower bounds on the random projection. We employ these tighter bounds in our concrete parameter estimations.

Lemma 4 (Random Projections mod q). *Let $n = \log_2(1/\kappa)$ and $\mathbf{A} \in \mathbb{Z}^{n \times m}$ be an with entries sampled from $\chi_{\mathsf{ghl}} \sim \{-1, 0, 1\}$. For any $\mathbf{x} \in \mathbb{Z}^m$ s.t. $\|\mathbf{x}\|_2$ we have $\frac{1}{2\sqrt{m}} \cdot \|\mathbf{x}\|_2 \leq \|\mathbf{Ax} \bmod q\|_2 \leq \sqrt{n \cdot m} \cdot \|\mathbf{x}\|_2$. where the first inequality holds with probability at least $1 - \kappa$ and the second holds with certainty.*

Proof. The first inequality follows from Lemma 5 by setting $\alpha_\infty = 1$. For the second inequality, observe that $\mathbf{A}$ takes values in $\{-1, 0, 1\}$ and thus has Frobenius norm at most $\sqrt{n \cdot m}$, and reduction modulo q only decreases the norm.

The next lemma is a generalisation of a proof step of [BS23, Lemma 4.2].

Lemma 5. *Let $\mathbf{A} \leftarrow \chi_{\mathsf{ghl}}^{n \times m}$ and $q \in \mathbb{N}$, and let $\beta, \alpha_\infty, \delta_\infty > 0$. Let Z_n be a binomial distribution $\mathsf{Bin}(n, \frac{1}{2})$ and suppose that $\mathsf{Pr}[Z_n < \alpha_\infty^2] \leq \delta_\infty$ The we have for all $\mathbf{x} \in \mathbb{Z}_q^m$ that $\|\mathbf{x}\|_\infty > \frac{2}{\alpha_\infty}\beta \implies \mathsf{Pr}[\|\mathbf{Ax} \bmod q\|_2 \leq \beta] \leq \delta_\infty$ In other words, $\|\mathbf{x}\|_\infty > \frac{2}{\alpha_\infty}\beta$ implies that $\|\mathbf{Ax} \bmod q\|_2 > \beta$, except with probability δ_∞. Furthermore, $\|\mathbf{x}\|_2 > 2\frac{\sqrt{m}}{\alpha_\infty}\beta \implies \mathsf{Pr}[\|\mathbf{Ax} \bmod q\|_2 \leq \beta] \leq \delta_\infty$*

The main idea of Lemma 5 is simple: The distribution of $\|\chi^n\|_2^2$ is $\mathsf{Bin}(n, \frac{1}{2})^2$, and by only considering the largest component of $\mathbf{x}$, say of absolute value ξ, we can argue the lower-bound $\|\chi^n\|_2^2 \xi^2 \geq \alpha_\infty \xi^2$ except with probability δ_∞. Plugging this into the desired probabilities would yield respective constraints, but the proof is complicated by handling modulo q arithmetic. For this reason, it does not trivially generalise to any symmetric distribution χ. The proof of Lemma 5 is deferred to Section C.1.

Acknowledgments

The research of Russell W. F. Lai and Michał Osadnik are supported by Research Council of Finland grant 358951. Ngoc Khanh Nguyen was supported by the Ethereum Foundation ESP Grant FY24-1358. This work was supported by funding from the topic Engineering Secure Systems of the Helmholtz Association (HGF) and by KASTEL Security Research Labs.

[19] We have also observed that the improvements offered by the tighter bounds in [GHL22, BS23] are very modest in our setting.

References

ACFY24. Gal Arnon, Alessandro Chiesa, Giacomo Fenzi, and Eylon Yogev. STIR: Reed-solomon proximity testing with fewer queries. In Leonid Reyzin and Douglas Stebila, editors, *CRYPTO 2024, Part X*, volume 14929 of *LNCS*, pages 380–413. Springer, Cham, August 2024. 1

ACFY25. Gal Arnon, Alessandro Chiesa, Giacomo Fenzi, and Eylon Yogev. WHIR: Reed-solomon proximity testing with super-fast verification. In Serge Fehr and Pierre-Alain Fouque, editors, *EUROCRYPT 2025, Part IV*, volume 15604 of *LNCS*, pages 214–243. Springer, Cham, May 2025. 1

Ach03. Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. Special Issue on PODS 2001. 2

ACK21. Thomas Attema, Ronald Cramer, and Lisa Kohl. A compressed Σ-protocol theory for lattices. In Tal Malkin and Chris Peikert, editors, *CRYPTO 2021, Part II*, volume 12826 of *LNCS*, pages 549–579, Virtual Event, August 2021. Springer, Cham. 2, 4

AFLN24. Martin R. Albrecht, Giacomo Fenzi, Oleksandra Lapiha, and Ngoc Khanh Nguyen. SLAP: Succinct lattice-based polynomial commitments from standard assumptions. In Marc Joye and Gregor Leander, editors, *EUROCRYPT 2024, Part VII*, volume 14657 of *LNCS*, pages 90–119. Springer, Cham, May 2024. 1, 2, 4, 11

AL21. Martin R. Albrecht and Russell W. F. Lai. Subtractive sets over cyclotomic rings - limits of Schnorr-like arguments over lattices. In Tal Malkin and Chris Peikert, editors, *CRYPTO 2021, Part II*, volume 12826 of *LNCS*, pages 519–548, Virtual Event, August 2021. Springer, Cham. 2, 4, 10, 11

BBB+18. Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. Bulletproofs: Short proofs for confidential transactions and more. In *2018 IEEE Symposium on Security and Privacy*, pages 315–334. IEEE Computer Society Press, May 2018. 1, 2

BBC+18. Carsten Baum, Jonathan Bootle, Andrea Cerulli, Rafaël del Pino, Jens Groth, and Vadim Lyubashevsky. Sub-linear lattice-based zero-knowledge arguments for arithmetic circuits. In Hovav Shacham and Alexandra Boldyreva, editors, *CRYPTO 2018, Part II*, volume 10992 of *LNCS*, pages 669–699. Springer, Cham, August 2018. 11

BBHR18. Eli Ben-Sasson, Iddo Bentov, Yinon Horesh, and Michael Riabzev. Fast reed-solomon interactive oracle proofs of proximity. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *ICALP 2018*, volume 107 of *LIPIcs*, pages 14:1–14:17. Schloss Dagstuhl, July 2018. 1, 2, 4

BC24. Dan Boneh and Binyi Chen. LatticeFold: A lattice-based folding scheme and its applications to succinct proof systems. Cryptology ePrint Archive, Report 2024/257, 2024. 3, 5

BC25. Dan Boneh and Binyi Chen. LatticeFold+: Faster, simpler, shorter lattice-based folding for succinct proof systems. In Yael Tauman Kalai and Seny F. Kamara, editors, *CRYPTO 2025, Part VII*, volume 16006 of *LNCS*, pages 327–361. Springer, Cham, August 2025. 3

BCC+16. Jonathan Bootle, Andrea Cerulli, Pyrros Chaidos, Jens Groth, and Christophe Petit. Efficient zero-knowledge arguments for arithmetic circuits in the discrete log setting. In Marc Fischlin and Jean-Sébastien Coron, editors, *EUROCRYPT 2016, Part II*, volume 9666 of *LNCS*, pages 327–357. Springer, Berlin, Heidelberg, May 2016. 1, 2

BCS23. Jonathan Bootle, Alessandro Chiesa, and Katerina Sotiraki. Lattice-based succinct arguments for NP with polylogarithmic-time verification. In Helena Handschuh and Anna Lysyanskaya, editors, *CRYPTO 2023, Part II*, volume 14082 of *LNCS*, pages 227–251. Springer, Cham, August 2023. 4

BL17. Carsten Baum and Vadim Lyubashevsky. Simple amortized proofs of shortness for linear relations over polynomial rings. Cryptology ePrint Archive, Report 2017/759, 2017. 2, 6

BLNS20. Jonathan Bootle, Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. A non-PCP approach to succinct quantum-safe zero-knowledge. In Daniele Micciancio and Thomas Ristenpart, editors, *CRYPTO 2020, Part II*, volume 12171 of *LNCS*, pages 441–469. Springer, Cham, August 2020. 2, 4, 11

BS23. Ward Beullens and Gregor Seiler. LaBRADOR: Compact proofs for R1CS from module-SIS. In Helena Handschuh and Anna Lysyanskaya, editors, *CRYPTO 2023, Part V*, volume 14085 of *LNCS*, pages 518–548. Springer, Cham, August 2023. 1, 2, 3, 4, 5, 6, 7, 11, 12, 16, 21

CHM+20. Alessandro Chiesa, Yuncong Hu, Mary Maller, Pratyush Mishra, Psi Vesely, and Nicholas P. Ward. Marlin: Preprocessing zkSNARKs with universal and updatable SRS. In Anne Canteaut and Yuval Ishai, editors, *EUROCRYPT 2020, Part I*, volume 12105 of *LNCS*, pages 738–768. Springer, Cham, May 2020. 3

CLM23. Valerio Cini, Russell W. F. Lai, and Giulio Malavolta. Lattice-based succinct arguments from vanishing polynomials - (extended abstract). In Helena Handschuh and Anna Lysyanskaya, editors, *CRYPTO 2023, Part II*, volume 14082 of *LNCS*, pages 72–105. Springer, Cham, August 2023. 2, 3, 4, 11, 25

CMNW24. Valerio Cini, Giulio Malavolta, Ngoc Khanh Nguyen, and Hoeteck Wee. Polynomial commitments from lattices: Post-quantum security, fast verification and transparent setup. In Leonid Reyzin and

Douglas Stebila, editors, *CRYPTO 2024, Part X*, volume 14929 of *LNCS*, pages 207–242. Springer, Cham, August 2024. 1, 2, 4

ENS20. Muhammed F. Esgin, Ngoc Khanh Nguyen, and Gregor Seiler. Practical exact proofs from lattices: New techniques to exploit fully-splitting rings. In Shiho Moriai and Huaxiong Wang, editors, *ASIACRYPT 2020, Part II*, volume 12492 of *LNCS*, pages 259–288. Springer, Cham, December 2020. 3

FKNP24. Giacomo Fenzi, Christian Knabenhans, Ngoc Khanh Nguyen, and Duc Tu Pham. Lova: Lattice-based folding scheme from unstructured lattices. In Kai-Min Chung and Yu Sasaki, editors, *ASIACRYPT 2024, Part IV*, volume 15487 of *LNCS*, pages 303–326. Springer, Singapore, December 2024. 5

FL20. Casper Benjamin Freksen and Kasper Green Larsen. On using toeplitz and circulant matrices for johnson–lindenstrauss transforms. *Algorithmica*, 82(2):338–354, 2020. 2, 12

FMN24. Giacomo Fenzi, Hossein Moghaddas, and Ngoc Khanh Nguyen. Lattice-based polynomial commitments: Towards asymptotic and concrete efficiency. *Journal of Cryptology*, 37(3):31, July 2024. 2, 4, 29

GHL22. Craig Gentry, Shai Halevi, and Vadim Lyubashevsky. Practical non-interactive publicly verifiable secret sharing with thousands of parties. In Orr Dunkelman and Stefan Dziembowski, editors, *EUROCRYPT 2022, Part I*, volume 13275 of *LNCS*, pages 458–487. Springer, Cham, May / June 2022. 2, 6, 11, 12, 20, 21

Gro16. Jens Groth. On the size of pairing-based non-interactive arguments. In Marc Fischlin and Jean-Sébastien Coron, editors, *EUROCRYPT 2016, Part II*, volume 9666 of *LNCS*, pages 305–326. Springer, Berlin, Heidelberg, May 2016. 1

HV11. Aicke Hinrichs and Jan Vybíral. Johnson-lindenstrauss lemma for circulant matrices. *Random Struct. Algorithms*, 39(3):391–398, October 2011. 2, 12

JL84. William Johnson and Joram Lindenstrauss. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 01 1984. 2, 11

Kil92. Joe Kilian. A note on efficient zero-knowledge proofs and arguments (extended abstract). In *24th ACM STOC*, pages 723–732. ACM Press, May 1992. 1

KLNO24. Michael Klooß, Russell W. F. Lai, Ngoc Khanh Nguyen, and Michal Osadnik. RoK, paper, SISsors toolkit for lattice-based succinct arguments - (extended abstract). In Kai-Min Chung and Yu Sasaki, editors, *ASIACRYPT 2024, Part V*, volume 15488 of *LNCS*, pages 203–235. Springer, Singapore, December 2024. 1, 2, 3, 4, 5, 6, 10, 11, 19, 20, 21, 25, 26, 27, 28, 29, 30, 31, 36

KP23. Abhiram Kothapalli and Bryan Parno. Algebraic reductions of knowledge. In Helena Handschuh and Anna Lysyanskaya, editors, *CRYPTO 2023, Part IV*, volume 14084 of *LNCS*, pages 669–701. Springer, Cham, August 2023. 6

KST22. Abhiram Kothapalli, Srinath Setty, and Ioanna Tzialla. Nova: Recursive zero-knowledge arguments from folding schemes. In Yevgeniy Dodis and Thomas Shrimpton, editors, *CRYPTO 2022, Part IV*, volume 13510 of *LNCS*, pages 359–388. Springer, Cham, August 2022. 3

LFKN92. Carsten Lund, Lance Fortnow, Howard Karloff, and Noam Nisan. Algebraic methods for interactive proof systems. *J. ACM*, 39(4):859–868, October 1992. 3

Lip24. Helger Lipmaa. Polymath: Groth16 is not the limit. In Leonid Reyzin and Douglas Stebila, editors, *CRYPTO 2024, Part X*, volume 14929 of *LNCS*, pages 170–206. Springer, Cham, August 2024. 1

LNP22. Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Maxime Plançon. Lattice-based zero-knowledge proofs and applications: Shorter, simpler, and more general. In Yevgeniy Dodis and Thomas Shrimpton, editors, *CRYPTO 2022, Part II*, volume 13508 of *LNCS*, pages 71–101. Springer, Cham, August 2022. 2, 6

LNS21. Vadim Lyubashevsky, Ngoc Khanh Nguyen, and Gregor Seiler. Shorter lattice-based zero-knowledge proofs via one-time commitments. In Juan Garay, editor, *PKC 2021, Part I*, volume 12710 of *LNCS*, pages 215–241. Springer, Cham, May 2021. 2, 6

LPR13. Vadim Lyubashevsky, Chris Peikert, and Oded Regev. A toolkit for ring-LWE cryptography. In Thomas Johansson and Phong Q. Nguyen, editors, *EUROCRYPT 2013*, volume 7881 of *LNCS*, pages 35–54. Springer, Berlin, Heidelberg, May 2013. 9

Mat08. Jiří Matoušek. On variants of the johnson–lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, September 2008. 2, 12

Mic94. Silvio Micali. CS proofs (extended abstracts). In *35th FOCS*, pages 436–453. IEEE Computer Society Press, November 1994. 1

MR09. Daniele Micciancio and Oded Regev. *Lattice-based Cryptography*, pages 147–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 4

NS24. Ngoc Khanh Nguyen and Gregor Seiler. Greyhound: Fast polynomial commitments from lattices. In Leonid Reyzin and Douglas Stebila, editors, *CRYPTO 2024, Part X*, volume 14929 of *LNCS*, pages 243–275. Springer, Cham, August 2024. 1, 4

SZ22. Alan Szepieniec and Yuncong Zhang. Polynomial IOPs for linear algebra relations. In Goichiro Hanaoka, Junji Shikata, and Yohei Watanabe, editors, *PKC 2022, Part I*, volume 13177 of *LNCS*, pages 523–552. Springer, Cham, March 2022. 3

ZCF24. Hadas Zeilberger, Binyi Chen, and Ben Fisch. BaseFold: Efficient field-agnostic polynomial commitment schemes from foldable codes. In Leonid Reyzin and Douglas Stebila, editors, *CRYPTO 2024, Part X*, volume 14929 of *LNCS*, pages 138–169. Springer, Cham, August 2024. 1

A Extended Preliminaries

A.1 Computational Assumption

We state an equivalent formulation of the vanishing short integer solution (vSIS) assumption [CLM23], which has a simpler description and better aligns with the notation adopted in this work.

Definition 2 (vSIS Assumption (adapted from [CLM23])). *Let* $\mathsf{params} = (\mathcal{R}, q, \beta, \chi)$ *be parametrised by* λ*, where* $\mathcal{R}$ *is a ring,* $q \in \mathbb{N}$ *a modulus,* $\beta > 0$ *a norm bound, and* χ *a distribution over* $\mathcal{R}_q^{n \times \otimes_{i \in [\mu]} d_i}$ *for some dimensions* $n, d_0, \ldots, d_{\mu-1}, \mu \in \mathbb{N}$*. The* $\mathsf{vSIS}_{\mathsf{params}}$ *assumption states that, for any PPT adversary* $\mathcal{A}$*, the advantage function satisfies*

$$\mathsf{Adv}^{\mathsf{vsis}}_{\mathsf{params},\mathcal{A}}(\lambda) \coloneqq \Pr\left[\begin{array}{l|l} \mathbf{Fw} = \mathbf{0} \bmod q & \mathbf{F} \leftarrow\!\$\ \chi \\ \|\sigma(\mathbf{w})\|_2 \leq \beta & \mathbf{w} \leftarrow \mathcal{A}(\mathbf{F}) \end{array}\right] \leq \mathsf{negl}(\lambda).$$

A.2 Reduction of Knowledge

In [KLNO24], ternary relations $\Xi \subseteq \{0,1\}^* \times \{0,1\}^* \times \{0,1\}^*$ are considered, where a tuple $(\mathsf{pp}, \mathsf{stmt}, \mathsf{wit}) \in \Xi$ consists of public parameters pp, statement stmt, and witness wit. For simplicity, pp is omitted when it is clear from the context. The definition of a reduction of knowledge in [KLNO24] is streamlined in this work.

Definition 3 (Reduction of Knowledge (adapted from [KLNO24])). *Let* Ξ_0*,* Ξ_1 *be ternary relations. A* reduction of knowledge (RoK) Π *from* Ξ_0 *to* Ξ_1*, short* $\Pi \colon \Xi_0 \to \Xi_1$*, is defined by two PPT algorithms* $\Pi = (\mathcal{P}, \mathcal{V})$*, the prover* $\mathcal{P}$ *and the verifier* $\mathcal{V}$*, with the following interface:*

- $\mathcal{P}(\mathsf{pp}, \mathsf{stmt}_1, \mathsf{wit}_1) \to (\mathsf{stmt}_2, \mathsf{wit}_2)$*: Interactively reduce the input statement-witness tuple* $(\mathsf{pp}, \mathsf{stmt}, \mathsf{wit}) \in \Xi_0$ *to a new statement-witness tuple* $(\mathsf{pp}, \widetilde{\mathsf{stmt}}, \widetilde{\mathsf{wit}}) \in \Xi_1$ *or* $\perp$*.*
- $\mathcal{V}(\mathsf{pp}, \mathsf{stmt}) \to \widetilde{\mathsf{stmt}}$*: Interactively reduce the task of checking the input statement* $(\mathsf{pp}, \mathsf{stmt})$ *w.r.t* Ξ_0 *to checking a new statement* $(\mathsf{pp}, \widetilde{\mathsf{stmt}})$ *w.r.t.* Ξ_1*.*

A RoK Π is *correct*, if for any honest protocol run (with correct inputs), the prover outputs a witness for the reduced statement (which the verifier outputs). A RoK Π is *knowledge sound* from Ξ_0^{KS} to Ξ_1^{KS} with knowledge error $\kappa(\mathsf{pp}, \mathsf{stmt})$ if there is a *black-box* expected polynomial-time extractor $\mathcal{E}$, which succeeds with probability $\epsilon - \kappa(\mathsf{pp}, \mathsf{stmt})$ if the malicious prover outputs a valid witness for the reduced statement with probability ϵ (on verifier's input $(\mathsf{pp}, \mathsf{stmt})$).

Definition 4 (Knowledge Soundness). *A reduction of knowledge* $\Pi = (\mathcal{P}*, \mathcal{V})$ *from* Ξ_0 *to* Ξ_1 *is* knowledge sound *with* knowledge error $\kappa : \mathbb{N} \to [0,1]$ *if there exists a knowledge extractor* $\mathcal{E}$*, such that for every statement* $x \in \Xi_0$ *and any prover* $\mathcal{P}^*$*, the extractor* $\mathcal{E}^{\mathcal{P}^*}(x)$ *runs in expected time polynomial in* $|x|$ *(counting calls to* $\mathcal{P}^*$ *as unit-cost operations) and outputs a witness* w *such that*

$$\Pr\left[(x; \mathcal{E}^{\mathcal{P}^*}(x)) \in \Xi_0\right] \geq \epsilon(\mathcal{P}^*, x) - \kappa(|x|),$$

where $\epsilon(\mathcal{P}^*, x) \coloneqq \Pr\left[\Xi_1 \leftarrow (\mathcal{P}^*, \mathcal{V})(x)\right]$

B Composition

This section supplements the details missing from Section 6.

B.1 Supportive RoKs

We provide the missing proofs for the "supportive RoKs", including $\Pi^{b\text{-}\mathsf{decomp}}$, Π^{batch}, $\Pi^{\mathsf{split}\text{-}\otimes}$, Π^{fold} and Π^{norm} adopted from [KLNO24] and $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$ modified from $\Pi^{\mathsf{split}\text{-}\otimes}$.

First, we recall the main characteristics of the b-ary decomposition protocol from [KLNO24]. As an important difference, we adapt the norm function from canonical Frobenius norm of $\mathbf{W}$ used in [KLNO24] to the maximum canonical 2-norm over the columns of $\mathbf{W}$ used throughout this work.

Lemma 6 (Decomposition [KLNO24, Lemma 4, adapted]). *Let $m, d, r, \in \mathbb{N}$ where $0 \leq \beta < q/2$. Protocol $\Pi^{b\text{-decomp}}$ is perfectly correct and reduces knowledge for*

$$\Xi^{\mathsf{lin}}_{r,\beta} \mapsto \Xi^{\mathsf{lin}}_{r\ell,\frac{1}{2}\sqrt{m}\sqrt{\mathring{\mathsf{f}}\varphi}\,b} .$$

It is knowledge sound for

$$\Xi^{\mathsf{lin}}_{r,\varrho^{\ell-1}\beta'_0,\varrho^\ell} \leftarrow\!\shortmid \Xi^{\mathsf{lin}}_{r\ell,\beta'_1,\varrho}$$

where $\beta'_0 = \frac{b^\ell-1}{b-1} \cdot \beta'_1$ and $\ell = \lceil \log_b(2\beta+1) \rceil$.

Proof. Only the slack terms are new, and these follow by using the (naive) common denominator of the ℓ decomposed witness parts. Hence, the new slack has norm bounded by ϱ^ℓ and the witness norms may grow by $\varrho^{\ell-1}$. See Lemma 8 where we explain this in more detail. □

Next is the batching protocol, which compresses the number of rows from $\hat{n}$ to $\overline{n}+1$, by randomly combining rows of $\mathbf{H}$ appropriately. The slack only affects the SIS-break branch of extraction, as is easy to see. We omit the proof.

Lemma 7 (Batching [KLNO24, Lemma 7, adapted]). *Let $\hat{n}, \overline{n} \in \mathbb{N}$ and $0 \leq \beta < q/2$. Protocol Π^{batch} is a perfectly correct self-reduction of knowledge for Ξ^{lin} with parameters*

$$\Xi^{\mathsf{lin}}_{\hat{n},\beta} \mapsto \Xi^{\mathsf{lin}}_{\overline{n}+1,\beta}$$

It is a relaxed knowledge sound self-reduction for Ξ^{lin} with knowledge error $\kappa = \frac{\hat{n}-\overline{n}-1}{|\mathcal{C}_{\mathcal{R}_q}|} \leq \frac{r \cdot n}{|\mathcal{C}_{\mathcal{R}_q}|}$.

$$\Xi^{\mathsf{lin}}_{\hat{n},\beta',\varrho} \cup \Xi^{\mathsf{sis}}_{2\varrho\beta'} \leftarrow\!\shortmid \Xi^{\mathsf{lin}}_{\overline{n}+1,\beta',\varrho}$$

Next is the verifier-succinct splitting protocol from [KLNO24]. This protocol significantly amplifies the slack of the extracted witness, namely from ϱ to ϱ^d when splitting into d pieces. In our succinct argument construction, we always use $\Pi^{\otimes\mathsf{RP}}$ immediately after $\Pi^{\mathsf{split}\text{-}\otimes}$ so that, in the extraction direction, witnesses fed to the extractor of $\Pi^{\mathsf{split}\text{-}\otimes}$ always have slack $\varrho = 1$.

Lemma 8 (Structured Split [KLNO24, Lemma 5, adapted]). *Let $m, d, r, \mu \in \mathbb{N}$ where $d|m$ and $0 \leq \beta < q/2$. Let $\mathbf{d} \in \mathbb{N}^\mu$. $\Pi^{\mathsf{split}\text{-}\otimes}$ is perfectly correct and reduces knowledge for*

$$\Xi^{\mathsf{lin}\text{-}\otimes d}_{m,r,\beta} \mapsto \Xi^{\mathsf{lin}\text{-}\otimes}_{m/d,r\cdot d,\beta}$$

It is knowledge sound with a knowledge error κ for

$$\Xi^{\mathsf{lin}\text{-}\otimes d}_{m,\varrho^{d-1}\sqrt{d}\beta',\varrho^d} \cup \Xi^{\mathsf{sis}}_{2\varrho\beta'} \leftarrow\!\shortmid \Xi^{\mathsf{lin}\text{-}\otimes}_{m/d,r\cdot d,\beta',\varrho}.$$

For completeness, we recall parts of the proof.

Proof (Proof sketch). The proof of correctness and extractability follows almost exactly as in [KLNO24, Lemma 5]. Namely, we establish d-special soundness. Denote by $\mathbf{W}_0^{(i)}, \ldots, \mathbf{W}_d^{(i)}$ (resp. $\mathbf{S}^{(i)}$) the witness (resp. slack) in the response of the d-th accepting transcript. In particular, we have[20] $\widetilde{\overline{\mathbf{F}}}\mathbf{W}_\ell^{(i)} = \overline{\mathbf{Y}}_\ell \bmod q$, and thus

$$\widetilde{\overline{\mathbf{F}}}\left(\mathbf{W}_\ell^{(i)}\mathbf{S}_\ell^{(i)}\mathbf{S}_\ell^{(j)} - \mathbf{W}_\ell^{(j)}\mathbf{S}_\ell^{(j)}\mathbf{S}_\ell^{(i)}\right) = \mathbf{0} \bmod q$$

for all i, j. Under the vSIS assumption for $\overline{\mathbf{F}}$ (and hence $\widetilde{\overline{\mathbf{F}}}$) for norm $2\varrho\beta'$, we conclude that $\mathbf{W}_\ell^{(i)}\mathbf{S}_\ell^{(i)}\mathbf{S}_\ell^{(j)} - \mathbf{W}_\ell^{(j)}\mathbf{S}_\ell^{(j)}\mathbf{S}_\ell^{(j)} = \mathbf{0} \bmod q$ and therefore, we can choose any fixed pair $\mathbf{W}_\ell$, $\mathbf{S}_\ell$, say the from first transcript, and we know that equality

$$\mathbf{W}_\ell = \mathbf{W}_\ell^{(i)} \quad \text{and} \quad \|\sigma\left(\mathbf{W}_\ell\mathbf{S}_\ell \bmod q\right)\|_2 \leq \beta'$$

[20] Note that $\overline{\mathbf{H}}\widetilde{\mathbf{F}} = \widetilde{\overline{\mathbf{F}}}$ are the $\overline{n}$ rows of $\widetilde{\mathbf{F}}$.

holds for any i. That is, the (potentially "fractional") witness $\mathbf{W}_\ell$ is uniquely defined. We get

$$\mathbf{F}\begin{pmatrix}\mathbf{W}_0\\ \mathbf{W}_1\\ \vdots\\ \mathbf{W}_{d-1}\end{pmatrix} = \mathbf{Y}$$

when we put the split witness together. Using $\mathbf{S} = \prod_{i=0}^{d-1}\mathbf{S}_i$ as the common denominator, we find

$$\|\sigma\left(\mathbf{W}_i\mathbf{S} \bmod q\right)\|_2 \leq \|\sigma\left(\mathbf{W}_i\mathbf{S}_i \bmod q\right)\|_2 \cdot \|\sigma\left(\prod_{j\neq i}\mathbf{S}_j\right)\|_\infty \leq \beta' \cdot \rho^{\ell-1}$$

which shows that $\rho^{\ell-1}$ is a valid bound for the slack of $\mathbf{W}_i$ and thus $\mathbf{W} = (\mathbf{W}_0, \ldots, \mathbf{W}_{d-1})$. Clearly, this shows that $\|\sigma\left(\mathbf{WS} \bmod q\right)\|_2 \leq \sqrt{d}\,\varrho^{d-1}\beta'$

Next, we use the Vandermonde matrix for challenges c, as in [KLNO24], to ensure the bottom $\underline{n}$ rows were correctly computed, and hence assert membership in $\Xi^{\mathsf{lin}\text{-}\otimes d}_{m,\varrho^{d-1}\sqrt{d}\beta',\varrho^d}$. The proof of [KLNO24] applies for $\mathbf{W}_\ell$, because we do not require shortness for this argument. □

For the unstructured random projections, we introduce an equivalent semistructured split, where only the top part $\underline{\mathbf{F}}$ of the vSIS instance is considered structured.[21] Formally, we write down following trivial reduction of knowledge which simply forgets the structure in $\underline{\mathbf{F}}$.

Lemma 9 (Forgetting structure). *Let $m, d, r, \mu \in \mathbb{N}$ where $d|m$ and $0 \leq \beta < q/2$. Let $\mathbf{d} \in \mathbb{N}^\mu$. Denote by $\Pi^{\mathsf{sf\text{-}forget}}$ the protocol which, given an instance $((\mathbf{H}, \mathbf{F}, \mathbf{Y}), (\mathbf{W}, \mathbf{S}))$ turns it into $((\mathbf{I}, \mathbf{HF}, \mathbf{Y}), (\mathbf{W}, \mathbf{S}))$. Then $\Pi^{\mathsf{sf\text{-}forget}}$ is perfectly correct and perfectly knowledge sound. Moreover, if the verifier's (and prover's) ring operations are bounded by $O(\hat{n} \cdot n \cdot m)$ (by naive matrix-multiplication).*

While Lemma 9 is trivial, we introduce it to pinpoint why structured folding is required until the witness is small enough: Unstructured folding needs to forget the structure supported by $\mathbf{H}$, and thus incurs a cost of $O(\hat{n} \cdot n \cdot m)$ ring operations. In general, this is not verifier-succinct because m is unbounded. However, if we first rely on structured folding to reduce to $m \in O(\lambda)$ and maintain $\hat{n}, n \in O(\lambda)$ as well, then verification needs at most $O(\lambda^3)$ ring operations, which is still verifier-succinct relative to the original witness size.

Next, we discuss the semistructured split protocol, which is defined in Fig. 6. Observe that, implicitly, Π^{split} runs $\Pi^{\mathsf{sf\text{-}forget}}$ to reset $\mathbf{H}$ to $\mathbf{I}$.

Lemma 10 (Semistructured Split). *Let $m, d, r, \mu \in \mathbb{N}$ where $d|m$ and $0 \leq \beta < q/2$. Let $\mathbf{d} \in \mathbb{N}^\mu$. $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$ is perfectly correct and reduces knowledge for*

$$\Xi^{\mathsf{lin}\text{-}\bar{\otimes} d}_{m,\hat{n},r,\beta} \mapsto \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m/d,\overline{n}+\underline{n}d,r\cdot d,\beta}$$

It is knowledge sound with a knowledge error κ for

$$\Xi^{\mathsf{lin}\text{-}\bar{\otimes} d}_{m,\hat{n},r,\rho^{d-1}\sqrt{d}\beta',\rho^d} \cup \Xi^{\mathsf{sis}}_{2\varrho\beta'} \leftarrow\!\shortmid \Xi^{\mathsf{lin}\text{-}\bar{\otimes}}_{m/d,\overline{n}+\underline{n}d,r\cdot d,\beta',\rho}$$

Proof. Correctness follows immediately. For extractability, notice that replacing $(\mathbf{H}, \mathbf{F})$ with $(\mathbf{I}, \mathbf{HF})$ does not change the protocol (because $\underline{\mathbf{F}}_i$ corresponds to $\underline{\mathbf{HF}}$). Hence, it suffices to consider $\mathbf{H} = \mathbf{I}$. Implicitly, $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$ applied $\Pi^{\mathsf{sf\text{-}forget}}$.

Now, consider the relaxed claim $((\mathbf{I}, \widetilde{\mathbf{F}}, \widetilde{\mathbf{Y}}), (\mathbf{W}_i, \mathbf{S}_i)_{i\in[d]})$ and observe that

$$\widehat{\mathbf{F}}\mathbf{W}_i = \overline{\mathbf{Y}}_i \qquad \overline{\mathbf{Y}} = \sum_{i\in[d]} \overline{\mathbf{r}}_i \bullet \overline{\mathbf{Y}}_i$$

$$\underline{\mathbf{F}}_i\mathbf{W}_i = \underline{\mathbf{Y}}_i \qquad \underline{\mathbf{Y}} = \sum_{i\in[d]} \underline{\mathbf{Y}}_i$$

[21] In the unstructed regime, we do not need a tensor-structured $\mathbf{F}$ per se. But it keeps consistency with structured splitting and reduces communication.

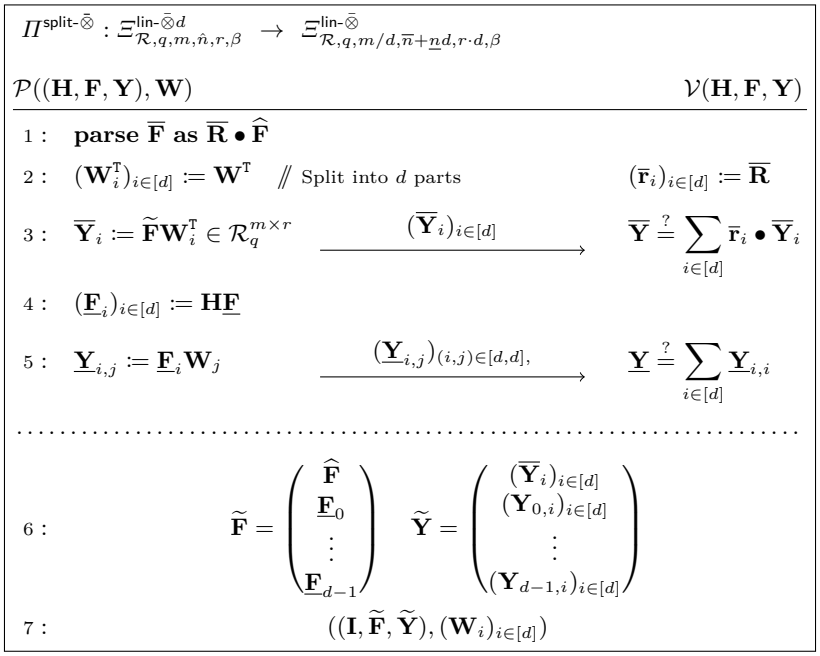

Fig. 6: RoK $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$.

holds by definition, and therefore

$$\mathbf{F}\begin{pmatrix} \mathbf{W}_0 \\ \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_{d-1} \end{pmatrix} = \mathbf{Y}.$$

Then, as in the proof of Lemma 8, find that $\mathbf{S} = \prod_{i\in[d]} \mathbf{S}_i$ is a valid slack for $\mathbf{W}$, and the norm bounds from Lemma 8 carry over as well, i.e., we have slack norm $\|\mathbf{S}\|_{\sigma\text{-op}} \leq \varrho^{d-1}$, and $\|\sigma\left(\mathbf{WS} \bmod q\right)\|_\infty \leq \sqrt{d}\,\varrho^{d-1}\beta'$. □

Next, we turn to Π^{fold}. We simplify the statement of Π^{fold} from [KLNO24] to $r_{\mathsf{out}} = 1$. Indeed, this is w.l.o.g., as the protocol Π^{fold} from [KLNO24] for $r_{\mathsf{out}} > 1$ is effectively a r_{out}-fold parallel repetition of this base case. Moreover, $r_{\mathsf{out}} > 1$ is only considered in [KLNO24] because of the polynomial-size challenge sets incurred by the need for subtractive sets, i.e., a parallel repetition (and thus $r_{\mathsf{out}} > 1$) is required to achieve negligible soundness error $r_{\mathsf{in}}/|\mathcal{C}|^{r_{\mathsf{out}}}$. In our case, $r_{\mathsf{in}}/|\mathcal{C}|$ is already negligible.

We note that Π^{fold} is the only protocol that *introduces* (a multiplicative factor of) slack. Moreover, any present slack is significantly amplified by Π^{fold}. Therefore, in the extraction direction, we need to ensure that $\varrho = 1$ before the extractor of Π^{fold} is run.

To analyse Π^{fold} in Lemma 11 below, we use the following definitions.

Definition 5 (Challenge set expansion factor). *We say that the challenge set $\mathcal{C}$ has expansion factor Γ if: For all $x \in \mathcal{R}$ and $c \in \mathcal{C}$, $\|\sigma\left(x \cdot c\right)\|_2 \leq \Gamma \cdot \|\sigma\left(x\right)\|_2$*

Definition 6 (Strong challenge set). *We say that a challenge set $\mathcal{C} \subseteq \mathcal{R}$ is a strong challenge set if For all distinct $c_1, c_2 \in \mathcal{C}$, $c_1 - c_2$ is invertible in $\mathcal{R}_q$.*

Lemma 11 (Fold [KLNO24, Lemma 6, adapted]). *Let $m, r, \in \mathbb{N}$ and $0 \leq \beta < q/2$. Π^{fold}, equipped with a strong challenge set $\mathcal{C}$ with the expansion factor Γ is perfectly correct and reduces knowledge for*

$$\Xi^{\mathsf{lin}}_{m,r,\beta} \mapsto \Xi^{\mathsf{lin}}_{m,r\Gamma\beta}\ .$$

It is coordinate-wise $(r+1)$-special sound, for

$$\Xi^{\mathsf{lin}}_{m,r,2\Gamma\varrho^{r-1}\beta',2\Gamma\varrho^r} \leftarrow\!\shortmid \Xi^{\mathsf{lin}}_{m,\beta',\varrho}\ .$$

and hence knowledge sound with a knowledge error $\kappa := \frac{r}{|\mathcal{C}|}$. □

Proof. The correctness direction is immediate and we refer to [KLNO24] for detailed reasoning. The extraction is a special case of [KLNO24], which we repeat for completeness. Recall that compared to [KLNO24], we hard-coded $r_{\mathsf{out}} = 1$, i.e., the resulting statements have vectors $\mathbf{y}$ and $\mathbf{w}$ instead of matrices. Consequently, also the slack is a single element $s \in \mathcal{R} = \mathcal{R}^{r\times r}$. We use this to simplify.

For knowledge soundness, we are given $r+1$ accepting transcripts $\mathbf{w}_0, \dots, \mathbf{w}_r$ with respective slack $s_0, \dots, s_r$ for challenges $(\mathbf{c}^{(i)})_{i\in[r+1]}$ where $\mathbf{c}^{(i)}$ and $\mathbf{c}^{(r)}$ differ exactly in row $i \in \{0, \dots r-1\}$. We now subtract the accepting equations to obtain

$$\mathbf{HF}\Big(\mathbf{w}_i - \mathbf{w}_r\Big) = \mathbf{Y}(\mathbf{c}^{(i)} - \mathbf{c}^{(r)}) = \mathbf{y}_i(c_i^{(i)} - c_i^{(r)}) \ .$$

Now, due the invertability of $c_i^{(i)} - c_i^{(r)}$, let

$$\gamma_i = c_i^{(i)} - c_i^{(r)} \qquad \text{and} \qquad \mathbf{v}_i = \frac{1}{\gamma_i} \cdot (\mathbf{w}_i - \mathbf{w}_r),$$

and set

$$\mathbf{S} = \prod_{i=0}^{r} s_i \cdot \begin{bmatrix} \gamma_0 & & \\ & \ddots & \\ & & \gamma_{r-1} \end{bmatrix} \qquad \text{and} \qquad \mathbf{W} = [\mathbf{v}_0, \dots, \mathbf{v}_{r-1}] \cdot \mathbf{S}$$

By construction, we have $\mathbf{HFW} = \mathbf{Y}$. Moreover, $\|\mathbf{S}\|_{\sigma\text{-op}} \leq \Gamma\varrho^{r-1}$. and $\|\sigma(\mathbf{W}\cdot\mathbf{S})\|_2 \leq 2\Gamma\varrho^{r-1}\beta'$ by construction, where we use that $\|\sigma(\gamma_i x)\|_2 \leq 2\Gamma\|\sigma(x)\|_2$ for $x \in \mathcal{R}$.

The extractor for coordinate-wise special soundness from [FMN24] incurs a knowledge-error of $\kappa \leq \frac{r_{\mathsf{in}}}{|\mathcal{C}|}$, which completes the proof. □

Finally, we recall the parameters of the norm-check protocol Π^{norm} from [KLNO24]. In this case, we do not consider any slack, as the norm check is completely insecure in the presence of slack.[22]

Lemma 12 (Norm check [KLNO24, Lemma 8/10, adapted]). *Let $m, r, \ell, b_{\mathsf{ip}} \in \mathbb{N}$, $0 \leq \beta < q/2$, and $\beta'^2 < q/2$. Π^{norm} is a perfectly correct reduction reduces of knowledge for*

$$\Xi^{\mathsf{lin}}_{m,\hat{n},r,\beta} \mapsto \Xi^{\mathsf{lin}}_{m,\hat{n}',r+\ell,\beta_{\mathsf{out}}}$$

where $\beta_{\mathsf{out}} = \sqrt{\beta^2 + \beta_{\mathbf{V}}{}^2} \leq \beta + \beta_{\mathbf{V}}$ and $\beta_{\mathbf{V}} \coloneqq \|\sigma(\mathbf{V})\|_2$, where $\mathbf{V}$ is a b_{ip}-ary decomposition of $\mathbf{v}$ for v_i are the coefficients of

$$L(X) = \sum_{i=-(m-1)}^{m-1} v_i X^i = \Big(\sum_{i=0}^{m-1} w_i X^i\Big) \cdot \Big(\sum_{i=0}^{m-1} \bar{w}_i X^{-i}\Big).$$

It is knowledge sound with a knowledge error $\kappa \leq \frac{2m}{|\mathcal{C}_{\mathcal{R}_q}|}$ from Ξ^{norm} to Ξ^{lin} with parameters

$$\Xi^{\mathsf{lin}}_{m,\hat{n},r,\sqrt{r}\beta} \cup \Xi^{\mathsf{sis}}_{2\beta'} \leftarrow\!\shortmid \Xi^{\mathsf{lin}}_{m,\hat{n}',r+\ell,\beta_{\mathsf{out}}{}'}$$

where $\hat{n}' = \hat{n} + 3$, b_{ip} and $\ell \geq \log_{b_{\mathsf{ip}}}(2\beta_{\mathbf{V}}{}^2 + 1)$ is such that $b_{\mathsf{ip}} \leq 2\beta_{\mathbf{V}}/(\sqrt{m\ell}\sqrt{\hat{\mathsf{f}}\varphi})$.

Remark 1. We note that, due to the change from Frobenius norm to maximum 2-norm over the columns of $\mathbf{W}$, we rely on a straightforward modification of analysis of Π^{norm} from [KLNO24]. Instead of arguing about the Frobenius norm, we use this as an upper-bound for the maximum 2-norm over the columns of $\mathbf{W}$. In the extraction direction, we convert the norm and argue that extracted norm is at most $\sqrt{r}\beta$.

[22] Consider $\mathcal{R} = \mathbb{Z}$. For a witness $\mathbf{w}$, the norm check computes the self-inner-product over $\mathcal{R}_q$ and thus $\|\mathbf{w}\|_2^2 = \sum_{i\in[m-1]} w_i^2$. If there is no wraparound, then checking that $\|\mathbf{w}\|_2^2 = \sum_{i=0}^{m} w_i^2 < \beta$ makes sense. However with slack, there is wraparound and indeed since $\sum_{i=0}^{3} \frac{1}{4} = 4 \cdot \frac{1}{4} = 1 \ll \beta$, the witness $\mathbf{w} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, \dots, 0)$ would pass the check even though it shouldn't. Thus Π^{norm} may fail completely if slack is present.

B.2 Structured Loop

We discuss how to compose the atomic RoKs to obtain obtain asymptotically and concrete efficient succinct arguments for the principal relation Ξ^{lin}. We begin by overviewing the functionality of each atomic RoK $\Pi^{\otimes\mathsf{RP}}$, $\Pi^{b\text{-}\mathsf{decomp}}$, Π^{split}, Π^{fold}, Π^{batch} and Π^{norm}, and also discuss considerations when composing them.

Keeping track of composition costs. The basic idea for obtaining a succinct argument system for Ξ^{lin} is to compose the reductions of knowledge Π^{split} and Π^{fold} to reduce the witness dimensions m so that the resulting witness is "relative small", i.e. until the witness contains $\mathsf{poly}(\lambda)$ ring elements.

As highlighted in [KLNO24], applying only Π^{split} and Π^{fold}, would result in an argument with both correctness and soundness gaps. The *correctness gap* refers to the round-by-round growth of the norm bound β of the intermediate witnesses in the loops, which limits the number of iterations or leads to potentially superpolynomial q. To eliminate the correctness gap, the protocol $\Pi^{b\text{-}\mathsf{decomp}}$ decomposes the witness into multiple lower-norm witnesses, and the norm bound β on the intermediate witness is controlled by this.

The *soundness gap* refers to the norm β^{ext} of a witness extracted by the knowledge extractor, as compared to the norm β considered in the correctness relation. To recall, the goal of [KLNO24] was to construct a succinct argument for Ξ^{lin} where the parameters for which correctness and soundness hold agree, i.e. the extracted witness should have the same norm bound as the witness in the original relation. To eliminate the soundness gap, the norm check protocol is used in [KLNO24], i.e. the prover sends an explicit norm claim β for the (intermediate) witness, which expands the running Ξ^{lin} relation into a Ξ^{norm} relation. Then, Π^{norm} is run to reduce this Ξ^{norm} relation back to a Ξ^{lin} relation. The effect of this procedure is to insert a "checkpoint" into the composition, so that the witness extracted at this step of the composition is "reset" to $\beta^{\mathsf{ext}} = \sqrt{r}\beta$. Since the matrix norm in [KLNO24] (namely Frobenius) differs from ours, we do not reset the norm bound precisely to β, but rather to $\sqrt{r}\beta$. Fortunately, this relaxed bound does not impact the asymptotic analysis, since $\sqrt{r} \in O(1)$ for our parameter choices, and thus $O(\sqrt{r}\beta) = O(\beta)$.

In our succinct argument construction, the use of unstructured short challenges induces *slack*. To "reset" the slack, we employ $\Pi^{\otimes\mathsf{RP}}$, which splits the claim into two claims with norm β and β_{RP}, respectively. The original claim is then folded and later adjoined to the second one. In the extraction direction, the second claim, which was *not* folded and therefore does not have slack, is used to argue about the norm of the original relation, effectively upgrading a relaxed guarantee (i.e. with slack) to an exact guarantee (i.e. without slack).

The RoKs Π^{norm} and $\Pi^{\otimes\mathsf{RP}}$ add rows to $\underline{\mathbf{F}}$, which would increase the communication cost. To prevent this, we employ Π^{batch} for batching newly added rows, effectively "resetting" the number of rows.

When composing the atomic RoKs, we need to track the following:

- The changes to the parameters $\underline{n}, n, m, r, \beta$. In particular, the hypothetical norm bound β^{ext} of an extracted witness must not exceed the allowed budget β^{sis}, i.e. $\beta^{\mathsf{ext}} < \beta^{\mathsf{sis}}/2$.
- The accumulated soundness error due to composition. The total soundness error must never exceed the allowed budget $2^{-\lambda}$.
- The accumulated communication cost.

We note that the hypothetical norm bound β^{ext} of the extracted witness is not a function of the parameters of preceding RoKs in the chain of composition, but rather a function of the parameters of subsequent RoKs. In other words, as we insert more RoKs into the composition, the β^{ext} values of all previous RoKs may change.

The overall communication cost of the composition, which is a natural target for optimisation, is the sum of the communication costs of all instances of atomic RoKs involved plus the size of the final witness. We note, however, that each atomic RoK also has different prover and verifier time costs which are harder to keep track. In the following, we focus only on minimising the communication cost for simplicity.

Composition strategy for minimising communication. We next discuss a strategy for minimising the overall communication cost. To aid reasoning, we first recall the effects on parameters that each atomic RoK has. We will omit soundness costs since parameters can be easily set to make them negligible.

– $\Pi^{b\text{-decomp}}$: Shrink β but grow r and β^{ext}. Cost $(\ell_{\mathsf{d}}-1)\hat{n}r$ of $\mathcal{R}_q$ communication, where ℓ_{d} a length of the decomposition basis.
– $\Pi^{\text{split-}\otimes}$: Shrink m but grow r and β^{ext}. Cost $r((d^2-1)\cdot(\hat{n}-\overline{n})+(d-1)\cdot\overline{n})$ of $\mathcal{R}_q$ communication.
– Π^{fold}: Reset r to some fixed (e.g. the initial) value but grow β and β^{ext}. Cost 0 communication.
– Π^{norm}: Reset β^{ext} to β but grow $\underline{n}$, r and β (for the next round). Cost $\ell_{\mathsf{n}}\hat{n}+3\cdot(r+\ell_{\mathsf{n}})$ of $\mathcal{R}_q$ communication, where ℓ_{n} a length of the decomposition basis.
– $\Pi^{\otimes\mathsf{RP}}$: Produce a second relation Ξ^{lin} of $\beta_{\mathsf{RP}}>\beta$. In the extraction direction, it uses the second relation to cancel the slack, setting ϱ to 1, but increases β^{ext}. Cost $r+\overline{n}+1$ of $\mathcal{R}_q$ communication.
– Π^{join}: Set β to the maximum of norms of the two claims, i.e., the a folded relation and a projected one in our case. In the extraction direction, β^{ext} remains unaltered for both branches. Cost 0 communication.
– Π^{batch}: Shrink $\underline{n}$ to 1. Cost 0 communication.

To achieve asymptotically efficient succinct arguments for the relation Ξ^{lin}, we repeat the following sequence in a loop:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-decomp}} \to \Pi^{\text{split-}\otimes} \to \Pi^{\otimes\mathsf{RP}} \to \begin{matrix}\Pi^{\mathsf{fold}}\\ \Pi^{\mathsf{id}}\end{matrix} \to \Pi^{\mathsf{join}} \to \Pi^{\mathsf{batch}}$$

where Π^{id} is a dummy RoK which acts as the identity function.

The loop is repeated $\mu = O(\log_{m_{\mathsf{rp}}}(m/\lambda))$ times, so that $m \in O(\lambda)$ after the final iteration.

SIS hardness. To set up the baseline, we directly employ calculations from [KLNO24]. This means that the size of a single $\mathcal{R}_q$ element is asymptotically

$$\Omega\left(\frac{\lambda\cdot(\log m+\log\lambda)^2}{\log\lambda}\right)=\Omega\left(\frac{\lambda\cdot\log^2 m}{\log\lambda}\right)$$

to achieve binding of single-row vSIS commitment under vSIS hardness assumptions. We will assume this size of an $\mathcal{R}_q$ element in the analysis below.

Challenge Set $\mathcal{C}_{\mathcal{R}}$ for Π^{fold}. The main conceptual difference with [KLNO24] is that we consider a challenge set $\mathcal{C}_{\mathcal{R}}$ of low-norm elements, which is unstructured. On the one hand, as $\mathcal{C}_{\mathcal{R}}$ is not subtractive, we must accept the possibility of extracting witnesses which satisfy a relaxed variant of the Ξ^{lin} relation, i.e. with slack. On the other hand, we can choose an exponential-size challenge set, which is beneficial for efficiency. For concreteness, we can for example consider the set:

$$\mathcal{C}_{\mathcal{R}} \subseteq \{a\in\mathcal{R}_q \mid \mathfrak{cf}(a)\in\{0,1\}^{\varphi}\}$$

of cardinality $2^{\varphi}\in\Theta(2^{\lambda})$, and with ring expansion factor $\Gamma=O(\lambda)$ for the challenges (cf. Definition 5).

Round Communication Complexity. Let us start with the "correctness direction" of norms and parameters. To aid discussion, we introduce auxiliary variables for keeping track of how the parameters of the Ξ^{lin} relation change throughout a single loop of the protocol.

$$\begin{aligned}&(m_0,n_0,r_0,\beta_0)\xrightarrow{\Pi^{\mathsf{norm}}}(m_0,n_1,r_1,\beta_1)\xrightarrow{\Pi^{b\text{-decomp}}}(m_0,n_1,r_2,\beta_2)\xrightarrow{\Pi^{\mathsf{split}}}(m_1,n_1,r_3,\beta_2)\\ &\xrightarrow{\Pi^{\otimes\mathsf{RP}}}\begin{matrix}(m_1,n_1,r_3,\beta_2)\\ (m_1,\hat{n},r_4,\beta_{\mathsf{RP}})\end{matrix}\xrightarrow{\Pi^{\mathsf{fold}}}(m_1,n_2,r_4,\beta_3)\xrightarrow{\Pi^{\mathsf{join}}}(m_2,n_1,r_5,\beta_4)\xrightarrow{\Pi^{\mathsf{batch}}}(m_2,n_0,r_5,\beta_4).\end{aligned}$$

We will use the parameters for analysing the communication complexity. For simplicity, we skip superscripts in the parameters, which indicate the round of the loop, i.e. we write r_j instead of $r_j^{(i)}$. This is enough the asymptotic analysis, yet concretely the parameters are not expected to be exactly the same in each round. For simple recursion, we assume that parameters are chosen so that $\beta_0=\beta_4$.

For Π^{norm}, we set $\beta_1=O(\sqrt{m\hat{\mathfrak{f}}\varphi}\,\beta_0)$ and $r_1\in O(1)$ which is consistent with Lemma 12 as long as β_0 is sufficiently larger than $\sqrt{m\hat{\mathfrak{f}}\varphi}\in\sqrt{m}\cdot\mathsf{poly}(\lambda)$., e.g. if $\beta_0\geq m\hat{\mathfrak{f}}\varphi$. After $\Pi^{b\text{-decomp}}$, we get $\beta_2=\sqrt{m\hat{\mathfrak{f}}\varphi}\,b_{\mathsf{d}}$.

Parameters	description	instantiation
λ	security parameter	–
κ	statistical knowledge error	$2^{O(\lambda)}$
m	height of the witness matrix $\mathbf{W}$	–
$\tilde{m}$	height of the witness in the unstructured loop	$O(\lambda)$
q	argument system modulus	$\mathsf{poly}(\lambda, m)$
$\mathcal{R}_i = \mathbb{Z}[\mathfrak{f}_i]$	intermediate rings in the chain: $\mathcal{R} = \mathcal{R}_\ell \supset \mathcal{R}_{\ell-1} \supset \cdots \supset \mathcal{R}_1 \supset \mathcal{R}_0 = \mathbb{Z}$.	$[\mathcal{R}_{i+1} : \mathcal{R}_i] \leq O(1)$
e_i	irr. factors of cyclotomic polynomial of $\mathcal{R}_{q,i}$	$O(\varphi_i)$ for $\varphi_i = \varphi(\mathfrak{f}_i)$
$\mathfrak{f}$	conductor of the ring	$O(\lambda \log m / \log \lambda)$
φ	ring dimension	$\varphi(\mathfrak{f})$
$r = r_0^{(0)}$	width of the witness matrix $\mathbf{W}$	$O(1)$
$r_0^{(i)}, \ldots r_5^{(i)}$	intermediate widths of the witness in the i-th round	$O(1)$
n	height of the matrix $\mathbf{F}$	1
β^{sis}	norm of the (presumably hard) vSIS instance	$\mathsf{poly}(\lambda, m)$
μ	invocations of the protocol in the structured loop	$O(\log(m/\lambda))$
$\tilde{\mu}$	invocations of the protocol in the unstructured loop	$O(\log(\tilde{m}))$
$\overline{n}$	number of top rows of $\mathbf{H}$	1
$\underline{n}$	number of bottom rows of $\mathbf{H}$	1
$\hat{n}$	number of rows of $\mathbf{H}$	$\overline{n} + \underline{n}$
β_0	initial norm bound	$\mathsf{poly}(\lambda, m)$
r_0	initial width of the witness	$O(1)$
b_{n}	decomposition base for the coeffs. of the Laurent poly.	β_1
ℓ_{n}	length of the decomposition basis w.r.t.	$O(1)$
b_{d}	decomposition base	$O\left(\left(m\hat{\mathfrak{f}}^{3/2}\lambda^{3/2}\right)^{1/\Theta(1)}\right)$
ℓ_{d}	length of the decomposition basis w.r.t. b_{d}	$O(1)$
$d^{(i)}$	folding factor of i-th round	$O(1)$ for $i > 0$, m_{rp} for $i = 0$
m_{rp}	width of the matrix $\mathbf{J}$	$O(\log_2(1/\kappa)) = O(\lambda)$
n_{rp}	height of the matrix $\mathbf{J}$	$O(\log_2(1/\kappa)) = O(\lambda)$
δ	soundness error of the projection	$2^{-n_{\mathsf{rp}}}$
k_{rp}	ratio of the matrix $\mathbf{J}$, $m_{\mathsf{rp}}/n_{\mathsf{rp}}$	$O(1)$ s.t $k_{\mathsf{rp}} = d^{(i)} \cdot r_2^{(i)}/r_4^{(i)}$
t_i	number of rows of the challenge matrix $\mathbf{C}_{i-1}$	$t_i \cdot e_i = \Omega(\lambda/\log\lambda)$

Table 3: Parameter instantiation for the protocol.

For $\Pi^{\otimes\mathsf{RP}}$, we find $\beta_{\mathsf{RP}} \coloneqq m_{\mathsf{rp}}\beta_2$. After Π^{join}, we get $\beta_3 = \beta_2 r_3 \Gamma$, where Γ is the expansion factor of challenge set and b_{d} as in Table 3. Finally, running Π^{join} gives us $\beta_4 \coloneqq \max(\beta_3, \beta_{\mathsf{RP}})$.

To argue about the feasibility of the setting, we need to establish that there exists b_{d} such that $\ell_{\mathsf{d}} = O(1)$ defined as in Table 3 and $\beta_4 = \beta_0$. We observe that

$$\beta_4 = O\left((r_3\gamma + m_{\mathsf{rp}})\sqrt{m\hat{\mathfrak{f}}\varphi}\sqrt{m\hat{\mathfrak{f}}\varphi}\, b_{\mathsf{d}}\right)$$

is the largest norm which occurs during a loop (namely, at the end). After substituting $\beta_4 = \beta_0$, $\beta_0 = b_{\mathsf{d}}^{\Theta(1)}$ and r_1, r_2, r_3 with asymptotic expressions from Table 3, we establish the condition for b_{d} as

$$b_{\mathsf{d}} = O\left(\left(m\mathfrak{f}^2\lambda\right)^{1/\Theta(1)}\right).$$

We now turn towards the "extraction direction". To justify security, we must ensure that during the extraction of the intermediate RoKs the norm bound β^{sis} is never exceeded. We assume a witness of norm β^{ext} is obtained as the final loop output.

- For Π^{batch}, the extractor retrieves the unaltered witness or produces a vSIS break with norm $2\beta^{\mathsf{ext}}$.
- The extractor for Π^{join} does not alter the norm and extracts witnesses for two relation Ξ^{lin} of norm β^{ext}.

– The extractor for Π^{fold} extracts a witness to the relaxed relation of norm $\Xi^{\mathsf{lin\text{-}rel}}$ with norm $2\beta^{\mathsf{ext}}$ and slack 2Γ.
– The extractor for $\Pi^{\otimes\mathsf{RP}}$ extracts a witness of norm $4\sqrt{\mathsf{rad}(\mathfrak{f})m_{\mathsf{rp}}}\beta^{\mathsf{ext}}$ or vSIS-break of norm $4\sqrt{d}2\Gamma\beta^{\mathsf{ext}}$.
– The extractor for Π^{split} extracts a witness of norm $4\sqrt{d}\sqrt{\mathsf{rad}(\mathfrak{f})m_{\mathsf{rp}}}\beta^{\mathsf{ext}}$.
– The extractor for $\Pi^{b\text{-}\mathsf{decomp}}$ recomposes the witness and the norm bound becomes $\left(4\sqrt{d}\sqrt{\mathsf{rad}(\mathfrak{f})m_{\mathsf{rp}}}\beta^{\mathsf{ext}}\right)^{\ell_{\mathsf{d}}}$.
– For Π^{norm}, the extractor extracts a witness of norm β^{ext} or results in a vSIS break with norm $2\left(4\sqrt{d}\sqrt{\mathsf{rad}(\mathfrak{f})m_{\mathsf{rp}}}\beta^{\mathsf{ext}}\right)^{\ell_{\mathsf{d}}}$.

We set $\beta^{\mathsf{ext}} = \sqrt{r_0}\beta_0$ due to the extraction guarantee of Π^{norm}. If an extractor runs into a vSIS break, it can immediately output it (after extending it from the intermediate $\overline{\mathbf{F}}$ to the initial one). The largest norm which occurs in this sequence is the norm occurring in the vSIS break in Π^{norm}, which is

$$\beta^{\mathsf{sis}} := 2\left(4\sqrt{d}\sqrt{\mathsf{rad}(\mathfrak{f})m_{\mathsf{rp}}}\sqrt{r_0}\beta_0\right)^{\ell_{\mathsf{d}}} = \mathsf{poly}(m,\lambda),$$

and therefore we conclude that is suffices to set $\beta^{\mathsf{sis}} < q = \mathsf{poly}(m,\lambda)$. All paramaters discussed above are summarised in Table 3.

Finally, we are ready to estimate the proof size. We track the communication cost *from prover to verifier* for atomic RoKs using the variables defined in Table 3. The verifier's random challenges are irrelevant after Fiat–Shamir compilation.[23]

– Π^{norm} involves sending $3(r_0+\ell_{\mathsf{n}})+\ell_{\mathsf{n}}\hat{n} = O(1)$ $\mathcal{R}_q$ elements,
– $\Pi^{b\text{-}\mathsf{decomp}}$ involves sending $\hat{n}(\ell_{\mathsf{n}}-1)r_0 = O(1)$ $\mathcal{R}_q$ elements,
– $\Pi^{\mathsf{split}\text{-}\otimes}$ involves sending $\overline{n}(d-1)r_3 + (\hat{n}-\overline{n})(d^2-1)r_3 = O(1)$ $\mathcal{R}_q$ elements,
– $\Pi^{\otimes\mathsf{RP}}$ involves sending $O(r_2) = O(1)$ $\mathcal{R}_q$ elements,
– Π^{fold} incurs no communication cost.
– Π^{join} involves sending $O(r_4) = O(1)$ ring elements,
– Π^{batch} incurs no communication cost.

We conclude that the total communication cost for a single round is $O(1)$ $\mathcal{R}_q$ elements, expressed in bits as

$$O\left(\frac{\lambda}{\log\lambda}\cdot\log^2 m\right).$$

Total Communication Complexity. We assume that $r_0^{(0)} = 1$, i.e. that we start from a single-column relation. Further, due to the way we define Π^{fold} $r_4^{(i)} = 1$. We observe that for the i-th round, $r_5^{(i)} = 2$ and further $r_0^{(i)} = 2$ for $i > 0$.

Since $\mu = O(\log m)$, we conclude that the total size of the proof for the structured loop, excluding the folded witness, is expessed in $O(\mu)$ $\mathcal{R}_q$ elements, expressed in bits as

$$O\left(\frac{\lambda\cdot\log^3 m}{\log\lambda}\right).$$

After μ rounds, the height of the witness is $d_{\mu-1} = \Theta(\lambda)$ and width is $r = O(1)$. We cannot continue structured folding as $\underline{\mathbf{F}}$ is unstructured. The size (in bits) of such a witness is

$$O\left(\frac{\lambda^2\cdot\log^2 m}{\log\lambda}\right).$$

To achieve overall communication of $O(\lambda\log(m)^3/\log(\lambda))$, the witness itself cannot be sent in plain and has to be further reduced via *unstructured* folding. We denote the dimensions of the witness at the end of structured folding by $\tilde{m}$ and $\tilde{r}$.

[23] More precisely, they are $O(r) = O(1)$ ring elements in all protocols except for the random projection. The projections themselves are at most $O(\lambda^3)$ bits uncompressed, but can be derived from a PRG to with $O(\lambda)$ seed length. This increases the soundness of by a negligible term.

Verifier Efficiency. In the structured loop, due to the row-tensor structure of $\mathbf{F}$ and small dimension of random projection matrix, the verifier's runtime for each of the protocols Π^{norm}, $\Pi^{b\text{-}\mathsf{decomp}}$, Π^{batch}, $\Pi^{\mathsf{split}\text{-}\otimes}$, Π^{fold}, and $\Pi^{\otimes\mathsf{RP}}$ is succinct, requiring $O_\lambda(\log m)$ operations per protocol.

B.3 Unstructured Loop

While switching to the unstructured folding, formally, we change the relation from $\Xi^{\mathsf{lin}\text{-}\otimes}$ to $\Xi^{\mathsf{lin}\text{-}\bar{\otimes}}$ using the $\Pi^{\mathsf{sf}\text{-}\mathsf{forget}}$ protocol.

To achieve asymptotically efficient succinct arguments for the unstructured loop, the composition strategy is instantiated with following loop phase:

$$\Pi^{\mathsf{norm}} \to \Pi^{b\text{-}\mathsf{decomp}} \to \Pi^{\mathsf{split}\text{-}\bar{\otimes}} \to \Pi^{\mathsf{RP}} \to \Pi^{\mathsf{fold}} \to \Pi^{\mathsf{batch}}$$

To aid discussion, we again introduce auxiliary variables for keeping track of how the parameters change throughout a single loop of the protocol.

$$\begin{aligned}&(m_0, n_0, r_0, \beta_0) \xrightarrow{\Pi^{\mathsf{norm}}} (m_0, n_1, r_1, \beta_1) \xrightarrow{\Pi^{b\text{-}\mathsf{decomp}}} (m_0, n_1, r_2, \beta_2) \xrightarrow{\Pi^{\mathsf{split}\text{-}\bar{\otimes}}} \\ &(m_1, n_2, r_3, \beta_2) \xrightarrow{\Pi^{\mathsf{RP}}} (m_1, n_3, r_3, \beta_2) \xrightarrow{\Pi^{\mathsf{fold}}} (m_1, n_3, r_4, \beta_3) \xrightarrow{\Pi^{\mathsf{batch}}} (m_1, n_0, r_4, \beta_3)\end{aligned}$$

Let us start with the "correctness direction" of norms and parameters. To argue about the feasibility, setting the parameters almost exactly as in the structured case is sufficient.

We now turn towards the "extraction direction". To justify security, we must ensure that during extraction of the intermediate RoKs works the norm bound β^{sis} is never exceeded. We assume a witness of norm β_0 is obtained as the final loop output.

- In Π^{batch}, the extractor retrieves the unaltered witness or produces a vSIS break with norm $2\beta^{\mathsf{ext}}$.
- The extractor for Π^{fold} extracts a witness to the relaxed relation of norm $\Xi^{\mathsf{lin}\text{-}\mathsf{rel}}$ with norm $2\beta^{\mathsf{ext}}$ and slack Γ.
- The extractor for Π^{RP} extracts a witness of norm $2m_1\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\beta^{\mathsf{ext}}$ or vSIS-break of norm $4\Gamma\beta^{\mathsf{ext}}$.
- The extractor for $\Pi^{\mathsf{split}\text{-}\bar{\otimes}}$ extracts a witness of norm $2\sqrt{d}m_1\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\beta^{\mathsf{ext}}$.
- The extractor for $\Pi^{b\text{-}\mathsf{decomp}}$ recomposes the witness and the norm becomes $\left(2\sqrt{d}m_1\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\beta^{\mathsf{ext}}\right)^{\ell_{\mathsf{d}}}$.
- Similarly, for Π^{norm}, the extractor extracts a witness of norm β^{ext} or results in a vSIS break with norm $2\left(2\sqrt{d}m_1\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\beta^{\mathsf{ext}}\right)^{\ell_{\mathsf{d}}}$.

We set $\beta^{\mathsf{ext}} = \sqrt{r_0}\beta_0$ due to the extraction guarantee of Π^{norm}. The largest norm which occurs in this sequence is the norm occuring in the vSIS break in Π^{norm}, which is

$$\beta^{\mathsf{sis}} \geq 2\left(2\sqrt{d}m_1\varphi\sqrt{n_{\mathsf{rp}}\mathsf{rad}(\mathfrak{f})}\sqrt{r_0}\beta_0\right)^{\ell_{\mathsf{d}}} = \mathsf{poly}(m, \lambda),$$

and therefore, we conclude that is suffices to set $\beta^{\mathsf{sis}} < q = \mathsf{poly}(m, \lambda)$.

Communication cost of Π^{RP}. We assume the tower structure of $\mathcal{R}_q$:

$$\mathcal{R} = \mathcal{R}_\ell \supset \mathcal{R}_{\ell-1} \supset \cdots \supset \mathcal{R}_1 \supset \mathcal{R}_0 = \mathbb{Z}.$$

and we write

$$\mathcal{Q}_i \cong (\mathbb{F}_{q^{e_i}})^{\varphi_i/e_i}.$$

We set $e_i = O(\varphi_i)$. and consider each extension to be d-smooth for some constant $d \in O(1)$, i.e.,

$$d_i \coloneqq [\mathcal{R}_{i+1} : \mathcal{R}_i] \leq d,$$

with a slight abuse of notation.

Our goal is to maintain a minimal communication, while keeping the knowledge error negligible. We upper-bound the knowledge error of Π^{RP} by $\kappa \leq \delta + m_0^2/(\sum_{i\in[\ell]} q^{t_i \cdot e_i})$. Therefore, for negligible error, it is enough to set $t_i \cdot e_i = \Omega(\lambda/\log\lambda)$.

We analyse the communication cost incurred by the prover in the Π^{RP} protocol. First, the prover sends the vector $\mathbf{t}_0$ consisting of m_0 $\mathbb{Z}_q$ elements. In each subsequent iteration ($i > 0$) of the (inner) loop, the prover sends t_i elements of $\mathcal{Q}_i$ elements. In asympototic terms, it amounts to sending $O(\lambda/\log\lambda)$ $\mathbb{Z}_q$ element. After converging to a single row (or $O(1)$ many rows) after $\log\lambda$ iterations, the prover sends $O(1)$ ring elements. As such, the communication is upper-bounded by $O(m_0 + \lambda) = O(\lambda)$ $\mathbb{Z}_q$ elements and $O(1)$ $\mathcal{R}_q$ elements, which could be simplified to $O(1)$ ring elements.

Round Communication Complexity. Finally, we are ready to estimate the proof size. Again, we will only count the *prover messages*. We track the communication cost for atomic RoKs not included in the structured loop:

- $\Pi^{\mathsf{split}\text{-}\otimes}$ involves sending $(d-1)\overline{n} + (d^2-1)\underline{n} = O(1)$ $\mathcal{R}_q$ elements,
- Π^{RP} involves sending $O(1)$ $\mathcal{R}_q$ elements,

Analogous to our analysis for the structured loop, we conclude that the total communication cost for a single round is $O(1)$ $\mathcal{R}_q$ elements, which leads to bit-communication at most

$$O\left(\frac{\lambda}{\log\lambda} \cdot \log^2 m\right) = O(\lambda \log\lambda)$$

assuming $m \in O(\lambda)$.

Total Communication Complexity. We assume that $r_0^{(0)} = 2$, i.e. that we start from a single relation. Further, observe that $r_4^{(i-1)} = r_0^{(i)} = 1$ for $i > 0$.

Since $\tilde{\mu} = O(\log\tilde{m})$, we conclude that the total size of the proof for the structured loop, excluding the folded witness, is expessed in $O(\tilde{\mu})$ $\mathcal{R}_q$ elements, expressed in bits as

$$O\left(\frac{\lambda \cdot \log^3 m}{\log\lambda}\right) = O\left(\lambda \cdot \log^2 \lambda\right)$$

assuming $m \in O(\lambda)$.

After $\tilde{\mu}$ rounds, the height of the witness is $O(1)$ and its width is $r = O(1)$. The size (in bits) of such a witness is

$$O\left(\frac{\lambda \cdot \log^2 m}{\log\lambda}\right) = O\left(\lambda \cdot \log\lambda\right)$$

assuming $m \in O(\lambda)$. The witness is sent directly to the verifier and the relation is verified in plain.

Verifier Efficiency. In the unstructured loop, the verifier's runtime is not sublinear in $\tilde{m}$ and verifier runs $O_\lambda(\tilde{m})$ ring operations. Nevertheless, as $\tilde{m} = O(\lambda)$ the verifier remains succinct in the unstructured loop as well.

B.4 Combined Protocol

We conclude that the total communication cost is

$$O\left(\frac{\lambda \cdot \log^3 m}{\log\lambda}\right).$$

B.5 Parameters Selection

We propose a concrete parameter selection for the protocol, optimised to minimise the proof size while adhering to the asymptotic composition described earlier. A slight adjustment was made by moving $\Pi^{b\text{-decomp}}$ to the start of the loop, as suggested in [KLNO24], and incorporating an additional use of Π^{batch}.

For the random projections, we employed Lemma 5 to estimate the number of rows and columns in the projection matrix $\mathbf{J}$, allowing us to set more aggressive norm bounds. Additionally, alongside the canonical ℓ_2-norm, we also tracked the ℓ_∞-norm throughout all RoKs. This approach results in slightly more favourable parameters in the "correctness" direction. For random projection bounds, we use more fine-grained parameters selection compared to Theorems 1 and 2. We rely on Lemma 5, which provides a more precise estimate of the projected norm compared to the one used in Lemma 4. This allows us to reduce the proof size by $\approx 20\%$

To demonstrate the competitive proof size, we compare our protocol with the one from [KLNO24], showing that our approach achieves a smaller proof size. Notably, our improvements particularly enhance scalability with respect to the security parameter. We present the scalability results for common values of the security parameters λ and κ.

The detailed parameter selection for each batch is presented in Table 2. For further details, the parameter estimation process is documented in the following notebook, which includes the code and the output of the parameter selection process. The notebook is available at the following link:

anonymous.4open.science/r/rok-and-roll/estimates.ipynb.

C Omitted Proofs

C.1 Proof of Lemma 5

Proof. Let β'_∞ be arbitrary; we choose its value later. First, observe that $\|\mathbf{x}\|_\infty > \beta'_\infty$ implies that some component has $|x_i| > \beta'_\infty$. Suppose w.l.o.g. that this is x_0. Second, note that we can consider the first component of $\mathbf{Ax}$ separately, i.e., $\mathbf{A}_{\bullet,0}x_0 + \mathbf{A} \cdot (0, x_1, \ldots)^\intercal$. By independence of the two summands, it suffices to show that

$$\sup_{\mathbf{y}\in\mathbb{R}^n} \mathsf{Pr}[\|\mathbf{z} \cdot x_0 + \mathbf{y} \mod q\|_2 \leq \beta] \leq \delta_\infty$$

where $\mathbf{z} \leftarrow \chi^n_{\mathsf{ghl}}$. Recall that $|-x_0 \mod q| = |x_0 \mod q| = |x_0|$, where the last equality follows since $\mathbf{x}$ is reduced modulo q.[24]

We now strictly lower-bound $|z_i x_0 + y_i \mod q|$ by $\beta'_\infty/2$ with probability $\frac{1}{2}$:

- If $|y_i| > \frac{\beta'_\infty}{2}$, then for $z_i = 0$ we get $|z_i x_0 + y_i \mod q| > \frac{\beta'_\infty}{2}$.
- If $|y_i| \leq \frac{\beta'_\infty}{2}$, then for $z_i = \pm 1$ we get $|\pm x_0 + y_i \mod q| > \frac{\beta'_\infty}{2}$, because otherwise $\pm x_0 + y \mod q \in [\pm\frac{\beta'_\infty}{2}]$ and $|y_i| \leq \frac{\beta'_\infty}{2}$, then $\pm x_0 \in [\pm\beta'_\infty]$ which contradicts $|x| > \beta'_\infty$.

Moreover, there exists a (tiny) $\epsilon > 0$, such that $|z_i x_0 + y_i \mod q| \geq \beta'_\infty/2 \cdot (1+\epsilon)$ still holds with probability $\frac{1}{2}$, since the left hand side has finite support. Furthermore, the inequality

$$|z_i \cdot x_0 + y_i \mod q|_2 \geq |z_i| \cdot \frac{\beta'_\infty}{2} \cdot (1+\epsilon)$$

holds in distribution (where we use that $|z_i|$ is distributed as $\mathsf{Bin}(1, \frac{1}{2})$). It then also follows that

$$\|\mathbf{z} \cdot x_0 + \mathbf{y} \mod q\|_2 \geq \|\mathbf{z}\|_2 \cdot \frac{\beta'_\infty}{2} \cdot (1+\epsilon)^n$$

holds in distribution.

[24] For odd q, $|-x_0 \mod q| = |x_0|$ is clear (when $x_0 = x_0 \mod q$ is reduced modulo q). For even q, we have $[-\frac{q}{2}, \frac{q}{2} - 1]$ as representatives. For $x_0 \neq -\frac{q}{2}$, we clearly get $|-x_0 \mod q| = |x_0|$. For $x_0 = -\frac{q}{2}$, we have $-x_0 \mod q = -x_0$, so obviously $|-x_0 \mod q| = |x_0|$ also holds.

The random variable $Z_n \coloneqq \|\mathbf{z}\|_2^2 = \|\mathbf{z}\|_0 = \sum_{i\in[n]} z_i$ is distributed like $\mathsf{Bin}(n, \frac{1}{2})$. Now, consider the claimed expression and assume $\|\mathbf{x}\|_\infty > \beta'_\infty$ and $\beta'_\infty \coloneqq \frac{2}{\alpha_\infty}\beta$. We find

$$\begin{aligned}
\mathsf{Pr}[\|\mathbf{A}\mathbf{x} \mod q\|_2 \leq \beta] &\leq \sup_{\mathbf{y}\in\mathbb{R}^n} \mathsf{Pr}[\|\mathbf{z}\cdot x_0 + \mathbf{y} \mod q\|_2 \leq \beta] \\
&\leq \mathsf{Pr}[\|\mathbf{z}\|_2 \cdot \frac{\beta'_\infty}{2} \cdot (1+\epsilon)^n \leq \beta] \\
&\leq \mathsf{Pr}[\|\mathbf{z}\|_2 \cdot \frac{\beta'_\infty}{2} < \beta] \\
&\leq \mathsf{Pr}[\sqrt{Z_n}\frac{\beta'_\infty}{2} < \beta] \\
&= \mathsf{Pr}[\sqrt{Z_n} < \frac{2\beta}{\beta'_\infty}] \\
&\leq \mathsf{Pr}[Z_n < \alpha_\infty^2] \\
&\leq \delta_\infty
\end{aligned}$$

where we use that $\frac{2\beta}{\beta'_\infty} \leq \alpha_\infty \iff \beta'_\infty \geq \frac{2}{\alpha_\infty\beta}$. This concludes the proof for the ∞-norm lower bound on $\mathbf{x}$. For the 2-norm, the claim follows from $\|\mathbf{x}\|_\infty \geq \frac{1}{\sqrt{n}}\|\mathbf{x}\|_2$. □