Cryptanalysis of Goldreich's PRGs with MAJ–XOR Predicates

Tianning Wang[1], Haoyang Wang[1], Fukang Liu[2], and Willi Meier[3]

[1] School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
{tianning, haoyang.wang}@sjtu.edu.cn
[2] Institute of Science Tokyo, Tokyo, Japan
liufukangs@gmail.com
[3] University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland
willimeier48@gmail.com

Abstract. Goldreich's pseudorandom generators (PRGs) expand a secret seed by evaluating a fixed low-locality predicate on random subsets of its bits. We study the concrete and asymptotic security of constructions whose local predicate combines majority with XOR. We introduce Deterministic Pooled Recovery (DPR), which guesses a set of seed positions and pools the outputs whose majority inputs are forced by the guess. The selected outputs yield noiseless sparse linear equations. For the proposed $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ challenge published in ToSC 2025, with seed length $n = 1024$ and output length $m = n^2$, DPR has an estimated recovery cost of $2^{109.71}$ operations using $\omega = 2.38$ for the linear-algebra exponent. This is below the claimed 128-bit security level. We also refine the common-bias analysis by evaluating its advantage at each fixed seed weight. This viewpoint motivates Bias-Amplified Pooled Recovery (BAPR), which uses partial majority bias to construct pooled noisy linear systems. For $m = n^s$ with fixed $s > 1$, odd majority locality $a \in \omega(1) \cap O(\log n)$, and fixed XOR locality $b \geq 3$, we prove that BAPR admits a worst-case $2^{o(n)}$-time implementation that recovers any fixed seed with high probability. This rules out exponential security in this intermediate-locality regime. In addition, we identify exploitable affine structure and residue constraints in proposed alternative predicates.

Keywords: Goldreich's PRGs · random local functions · MAJ–XOR predicates · seed recovery

1 Introduction

Pseudorandom generators (PRGs) are a fundamental cryptographic primitive that expands a short random seed into a longer string indistinguishable from uniform. Goldreich proposed a family of candidate one-way functions in which each output bit applies a fixed predicate to a small random subset of the input [20]. When more output bits than seed bits are generated, this framework yields

highly parallel candidate PRGs. Their security depends delicately on the choice of predicate, its locality, and the stretch.

Goldreich's construction also has a natural interpretation as a weak pseudorandom function (PRF). Concretely, the seed can be viewed as a weak PRF key, with each random tuple of input positions serving as an input and the corresponding generator bit as the output. Applebaum and Raykov connected expander-based local functions with weak PRFs of very low circuit complexity [5]. As highlighted by Boura et al., high-stretch constructions based on $\mathrm{MAJ}_a \oplus \mathrm{XOR}_b$ with growing a form an important class of shallow weak PRF candidates [7]. Their low evaluation complexity is attractive in secure computation, including applications to oblivious PRFs, pseudorandom correlation functions for two-party protocols, and zero-knowledge proofs [7]. Concrete uses include silent OT and QuietOT instantiated with weak PRFs based on Goldreich's construction [8,13], as well as low-bandwidth mixed arithmetic in VOLE-based zero-knowledge proofs using a Goldreich-style $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ generator [1].

The MAJ–XOR family combines a linear XOR component with a nonlinear majority component and can attain high resiliency and algebraic immunity [7,22,23,18]. Closely related structures appear in the FLIP and FiLIP stream ciphers, which were designed for efficient homomorphic evaluation [25,24]. Consequently, understanding the concrete security of these predicates is relevant both to Goldreich's construction and to the protocols and symmetric primitives built around the same low-complexity structure.

The security of Goldreich's construction has consequently been studied from both asymptotic and concrete perspectives. Early work showed that locality three is vulnerable to linear tests at sufficient stretch and that no predicate of locality at most four yields a local small-bias generator with arbitrary polynomial stretch [15,26]. Subsequent work obtained longer stretch from random local one-way functions and clarified the predicate-dependent behavior of local small-bias generators [2,3]. General attacks exploit correlations with a few input variables, low rational degree, or bounded degree and locality [6,4,27]. Couteau et al. [14] initiated a systematic study of concrete security and developed a guess-and-determine attack. Yang et al. [28] substantially improved this line of attack, introduced a guess-and-decode method for sparse predicates, and also considered extensions to MAJ–XOR predicates.

More recently, several attacks have been developed specifically for MAJ–XOR predicates. Fu et al. [19] introduced the group-and-solve (GAS) attack. It obtains noisy linear systems by grouping outputs around a common subset and gives strong seed-recovery attacks in several parameter regimes. D'Antona et al. [16] introduced the common-bias attack and obtained strong average-seed distinguishers. GAS loses equations as the common subset grows, whereas the common-bias attack is especially effective for large majority locality or high stretch. However, it remains open whether $\mathrm{MAJ}_a \oplus \mathrm{XOR}_b$ PRGs can retain exponential security in the intermediate regime $a \in \omega(1) \cap O(\log n)$.

Our contributions. Our main contributions are as follows.

- We introduce *Deterministic Pooled Recovery* (DPR), which guesses seed bits and pools outputs whose threshold values are forced by the guess. These outputs yield noiseless XOR equations. For fixed locality and balanced seeds, we establish a subexponential-time recovery bound for a range of polynomial stretches. For the $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ challenge of Boura et al. [7, Table 3], the estimated cost at balanced seed weight is below the claimed security level.
- We analyze common bias at each fixed seed weight and introduce *Bias-Amplified Pooled Recovery* (BAPR). BAPR pools outputs whose threshold values are biased, rather than forced, by the guess and forms noisy XOR equations. Under the Gauss/SubGauss cost model, its finite-parameter estimates indicate lower recovery costs than prior attacks for (near-)balanced seeds.
- We address the intermediate-locality question raised by D'Antona et al. [16]. For $m = n^s$ with fixed $s > 1$, odd $a \in \omega(1) \cap O(\log n)$, and fixed XOR locality $b \geq 3$, BAPR admits a worst-case $2^{o(n)}$-time implementation. For each fixed seed, it returns a seed consistent with the output with probability $1 - o(1)$ over the random public input sets, ruling out exponential security in this regime.
- For the alternative predicates of D'Antona et al. [16], we derive affine-recovery cost estimates for $f_{d,\epsilon,a} \oplus \mathrm{XOR}_b$ and show how partial guessing exposes residue constraints in $\mathrm{MOD3}_a \oplus \mathrm{XOR}_b$.

Organization. Section 2 gives the required notation and definitions. Section 3 introduces DPR, develops its rank analysis, and applies it to the $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ challenge. Section 4 gives the fixed-weight common-bias analysis, introduces BAPR, and studies the intermediate-locality regime. Section 5 discusses affine-approximation attacks and residue constraints for the alternative predicates. Finally, Section 6 concludes the paper.

2 Preliminaries

2.1 Notation

For a positive integer n, let $[n] = \{1, \ldots, n\}$. If $x = (x_1, \ldots, x_n) \in \mathbb{F}_2^n$ and $S \subseteq [n]$, then $x_S = (x_i)_{i \in S}$ denotes the restriction of x to the coordinates indexed by S, in increasing order. If σ is an ordered tuple of distinct elements of $[n]$, then x_σ denotes the corresponding coordinates of x in the order specified by σ. The Hamming weight of x is $\mathrm{wt}(x) = |\{i : x_i = 1\}|$. For an integer r, $\binom{[n]}{r}$ denotes the set of r-subsets of $[n]$. We use the convention $\binom{r}{u} = 0$ whenever $u < 0$ or $u > r$. For a positive integer r, write $\mathbf{1}_r$ for the all-one vector of length r.

We use $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, $o(\cdot)$, and $\omega(\cdot)$ in their standard asymptotic senses. The standalone symbol ω denotes the exponent of linear algebra. Unless stated otherwise, log denotes the natural logarithm, while $\log_2$ denotes the base-two logarithm.

2.2 Weak pseudorandom functions

A PRF is a keyed family of functions that is computationally hard to distinguish from a uniformly random function. A weak PRF asks for the same indistinguishability property only on random inputs. Thus the difference between a weak PRF and a PRF lies in the query model: for a weak PRF, the adversary receives evaluations at uniformly sampled inputs, whereas for a standard PRF, the adversary may choose its inputs adaptively.

Let $f : \mathcal{K} \times \mathcal{X} \to \mathcal{Y}$ be a keyed function. For a key $\mathsf{sk} \in \mathcal{K}$, write $f_{\mathsf{sk}}(x) = f(\mathsf{sk}, x)$. The weak PRF advantage of an adversary $\mathcal{A}$ receiving Q random input–output samples is

$$\mathrm{Adv}^{\mathrm{wPRF}}_{f,\mathcal{A}} = \left|\Pr\left[\mathcal{A}\big((\xi_i, f_{\mathsf{sk}}(\xi_i))_{i=1}^{Q}\big) = 1\right] - \Pr\left[\mathcal{A}\big((\xi_i, g(\xi_i))_{i=1}^{Q}\big) = 1\right]\right|,$$

where $\mathsf{sk} \leftarrow \mathcal{K}$ and $\xi_1, \ldots, \xi_Q \leftarrow \mathcal{X}$ are sampled uniformly and independently, $g : \mathcal{X} \to \mathcal{Y}$ is a uniformly random function, and the probabilities also include the internal randomness of $\mathcal{A}$. We say that f provides κ-bit weak PRF security if every adversary that receives at most 2^κ random input–output samples and runs in time at most that required for 2^κ evaluations of f has negligible distinguishing advantage.

2.3 Goldreich's PRGs and Local Predicates

Goldreich's PRGs are specified by a fixed local predicate and a public sequence of input tuples. Fix positive integers n, m, a Boolean predicate P, and ordered tuples $\sigma_1, \ldots, \sigma_m$, each consisting of distinct elements of $[n]$ and having the arity of P. The resulting generator is

$$G_{\Gamma,P}(x) = \big(P(x_{\sigma_1}), P(x_{\sigma_2}), \ldots, P(x_{\sigma_m})\big),$$

where $\Gamma = (\sigma_1, \ldots, \sigma_m)$ is public. In the polynomial-stretch regime, we write $m = n^s$, up to rounding, where $s > 1$ is the stretch exponent.

The same construction also admits a natural weak PRF interpretation: the seed x serves as the secret key, while a random ordered tuple of input positions serves as the input. Namely, for an ordered tuple σ of the appropriate length, define

$$f_x(\sigma) = P(x_\sigma).$$

Thus, independently sampled random queries to f_x have the same distribution as the coordinates of $G_{\Gamma,P}(x)$ when the tuples in Γ are sampled independently. This gives a single-output weak PRF formulation of the same local construction. Applebaum and Raykov [5] study a closely related batched construction in which each weak PRF input encodes a hypergraph and the output contains one predicate evaluation for each of its hyperedges. They establish weak PRF security under suitable expander-based one-wayness assumptions and parameter settings.

The choice of predicate is the main security-relevant ingredient in Goldreich's PRGs. Its arity controls both evaluation efficiency and resistance to known attacks.

The main class considered in this work is the MAJ–XOR family. We describe it using threshold functions, with majority as a special case. For $z \in \{0,1\}^a$, define the threshold function

$$\mathrm{THR}_{a,t}(z) = \begin{cases} 1, & \mathrm{wt}(z) \geq t, \\ 0, & \mathrm{wt}(z) < t. \end{cases}$$

For odd a, the majority function is the balanced threshold

$$\mathrm{MAJ}_a(z) = \mathrm{THR}_{a,(a+1)/2}(z).$$

For $u \in \{0,1\}^b$, define

$$\mathrm{XOR}_b(u) = u_1 \oplus \cdots \oplus u_b.$$

A MAJ–XOR-type output is obtained by applying the threshold and XOR parts to two disjoint local input sets and XORing the results:

$$P_{a,t,b}(x_A, x_B) = \mathrm{THR}_{a,t}(x_A) \oplus \mathrm{XOR}_b(x_B), \qquad |A| = a,\ |B| = b,\ A \cap B = \varnothing.$$

The MAJ–XOR predicate is the special case $t = (a+1)/2$.

For each output i, we sample A_i uniformly from $\binom{[n]}{a}$. We then sample B_i uniformly from $\binom{[n]\setminus A_i}{b}$. The pairs (A_i, B_i) are sampled independently across outputs. The observed output bit is

$$y_i = \mathrm{THR}_{a,t}(x_{A_i}) \oplus \mathrm{XOR}_b(x_{B_i}). \tag{1}$$

These predicates have very low evaluation complexity and can be chosen to satisfy classical design criteria such as resiliency and algebraic immunity [22,23,18]. The concrete values of these properties depend on the parameters.

3 Deterministic Pooled Recovery

In this section, we introduce *Deterministic Pooled Recovery* (DPR), a seed-recovery attack on Goldreich's PRGs instantiated with MAJ–XOR predicates. We analyze the size and rank of its selected linear systems and apply the attack to concrete challenge instances.

3.1 The DPR Method

The method guesses seed bits and selects outputs whose threshold inputs contain sufficiently many guessed positions. When the guess forces their threshold values, the selected outputs yield noiseless linear equations.

Consider the $\mathrm{THR}_{a,t} \oplus \mathrm{XOR}_b$ output model from (1). Let $K \subseteq [n]$ be a set of seed positions of size k, and write $N = n - k$ for the number of remaining coordinates. We guess that all bits in K are equal to 1, namely

$$x_K = \mathbf{1}_k.$$

For an integer q, define the pooled selected index set

$$I_{K,q} = \{\, i \in [m] : |A_i \cap K| \geq q \,\}.$$

Write $e = |I_{K,q}|$ for the number of selected outputs. Unlike GAS, which selects outputs containing one fixed guessed subset [19], DPR pools every output whose threshold input contains at least q positions from the larger guessed set K.

Lemma 1 (Distribution of the selected-pool size). *For any fixed $K \subseteq [n]$ with $|K| = k$, assume that the threshold input sets $A_1, \ldots, A_m$ are chosen independently and uniformly from $\binom{[n]}{a}$. Define the selection probability*

$$p_{\mathrm{sel}} = \sum_{j=q}^{\min(a,k)} \frac{\binom{k}{j}\binom{N}{a-j}}{\binom{n}{a}}. \tag{2}$$

Then

$$e \sim \mathrm{Bin}(m, p_{\mathrm{sel}}), \qquad \mu := \mathbb{E}[e] = m p_{\mathrm{sel}}. \tag{3}$$

Here Bin *denotes the binomial distribution.*

Proof. For each i, counting according to $|A_i \cap K|$ shows that $\Pr[|A_i \cap K| \geq q] = p_{\mathrm{sel}}$. Since the sets A_i are sampled independently, the corresponding selection indicators are independent. Their sum is e, which gives the stated distribution and expectation.

Assume that $q \geq t$ and that the guess $x_K = \mathbf{1}_k$ is correct. For every $i \in I_{K,q}$, we have $|A_i \cap K| \geq q \geq t$. Hence, $\mathrm{THR}_{a,t}(x_{A_i}) = 1$. The corresponding output equation becomes

$$\mathrm{XOR}_b(x_{B_i}) = y_i \oplus 1.$$

After substituting the guessed variables in $B_i \cap K$, this is a noiseless linear equation in the remaining N seed bits.

Algorithm 1 summarizes the DPR procedure for a candidate set K. Note that it describes a single trial of the attack. In the randomized version, we repeat this trial with independently and uniformly sampled k-subsets K; alternatively, the candidate sets may be enumerated.

The attacker does not know a correct all-one set K. For a fixed seed of Hamming weight w, the probability that a uniformly chosen k-subset K satisfies $x_K = \mathbf{1}_k$, assuming $k \leq w$, is $\binom{w}{k}/\binom{n}{k}$. Thus the expected number of independent trials in the randomized search is

$$S_1(n, w, k) = \frac{\binom{n}{k}}{\binom{w}{k}}.$$

Algorithm 1 Deterministic Pooled Recovery (DPR) for $\mathrm{THR}_{a,t} \oplus \mathrm{XOR}_b$.

Require: Public sets $(A_i, B_i)_{i=1}^m$, output $y \in \{0,1\}^m$, and parameters k, q with $q \geq t$, and a rank-deficiency bound δ.
Ensure: A seed candidate $x \in \{0,1\}^n$, or failure.
Choose a candidate set $K \subseteq [n]$ with $|K| = k$.
Guess $x_K = \mathbf{1}_k$.
Set $I_{K,q} = \{i : |A_i \cap K| \geq q\}$ and $e = |I_{K,q}|$.
if $e < N$ **then**
    return failure.
end if
for each $i \in I_{K,q}$ **do**
    Form the linear equation implied by the guess

$$\bigoplus_{v \in B_i \setminus K} x_v = y_i \oplus 1 \oplus (|B_i \cap K| \bmod 2).$$

end for
Apply Gaussian elimination to the resulting system over $\mathbb{F}_2$.
if the system is inconsistent or has rank less than $N - \delta$ **then**
    return failure.
end if
Enumerate its at most 2^δ solutions and verify them against the full PRG output.
return the first candidate passing verification, or failure.

An all-zero guess instead forces the threshold output to 0 when $q \geq a - t + 1$. For majority with odd a and balanced seeds, it has the same search cost as the all-one guess.

Remark 1 (Trying both constant values). For the same candidate set K, DPR may try both $x_K = \mathbf{1}_k$ and $x_K = \mathbf{0}_k$. For a seed of weight w, the probability that one of these guesses is correct is

$$\frac{\binom{w}{k} + \binom{n-w}{k}}{\binom{n}{k}}.$$

For majority with odd a, both branches have the same forcing threshold $t = (a+1)/2$. They therefore use the same selected rows and the same coefficient matrix M_K, so Gaussian elimination can be performed once with two right-hand sides. At $w = n/2$, trying both values halves the expected number of candidate sets and reduces the search exponent by one bit. The improvement in the balanced-case cost is small and does not affect the asymptotic estimates. We therefore analyze the all-one branch below and retain its cost as a conservative estimate.

For a candidate K, the selected equations above give a linear system over these N unknowns. Let M_K be its coefficient matrix after deleting the guessed columns, and let

$$d_K = N - \mathrm{rank}(M_K)$$

be its rank deficiency.

For a correct guess $x_K = \mathbf{1}_k$ with $q \geq t$, the selected system is consistent and its affine solution space contains exactly 2^{d_K} candidates. Hence, if $d_K \leq \delta$, enumerating this space requires at most 2^δ candidate checks against the complete PRG output. This recovers the original seed whenever it is the unique candidate passing verification.

We next account for trials in which the guess on K is incorrect. Let C_{sel} denote the cost of locating the selected rows for one candidate K, let V denote the cost of verifying one seed candidate against the full PRG output, and let ω be the linear-algebra exponent. A direct scan of the m threshold input sets gives $C_{\text{sel}} = O(ma)$, while elimination over the N remaining unknowns is charged $O(N^\omega)$ operations in the concrete estimates.

Incorrect guesses. For an incorrect guess, the assumed threshold values and the substituted guessed bits induce an error in the right-hand side. Consistency imposes $e - \text{rank}(M_K)$ independent parity checks on this error.

The following estimate uses a random-syndrome model for incorrect guesses: conditional on M_K, we model the outcomes of these checks as uniformly distributed in $\mathbb{F}_2^{e-\text{rank}(M_K)}$. Under this model, a false system is consistent with probability

$$2^{-(e-\text{rank}(M_K))}.$$

Conditional on consistency, its affine solution space contains $2^{N-\text{rank}(M_K)}$ vectors. Hence the expected number of candidates generated by a false guess, conditional on e and M_K, is

$$2^{-(e-\text{rank}(M_K))}2^{N-\text{rank}(M_K)} = 2^{N-e}.$$

In particular, this estimate is independent of the rank of M_K. We discard trials with $e < N$ before candidate enumeration. For every remaining false trial, $e \geq N$, so $2^{N-e} \leq 1$. Thus, within the random-syndrome model, the expected verification cost of a false trial is at most $O(V)$. The probability that a correct guess is discarded is included in the retry estimate below.

Remark 2. This calculation is an idealized reference, not a bound for every incorrect guess under the actual distribution of (A_i, B_i). In particular, its consistency checks need not be uniform for a fixed incorrect K. Appendix A instead bounds the aggregate number of false candidates directly and derives a cost guarantee without this assumption.

Not every correct guess produces a usable system. For an integer $\delta \geq 0$, write

$$p_{\text{use}}(\delta) = \Pr[e \geq N,\ d_K \leq \delta \mid x_K = \mathbf{1}_k]$$

for the probability that a correct guess yields enough rows and rank deficiency at most δ. Unusable systems are discarded. Using an independent-trial approximation for the retry count, the estimated number of trials before obtaining a usable correct system is

$$\frac{S_1(n, w, k)}{p_{\text{use}}(\delta)}.$$

This gives the total-cost estimate

$$O\left(\frac{S_1(n,w,k)}{p_{\mathrm{use}}(\delta)}\big(C_{\mathrm{sel}}+N^{\omega}+V\big)+2^{\delta}V\right).$$

Here $V = O(m(a+b))$ for full verification. The corresponding bit-complexity exponent is

$$\max\big\{\log_2 S_1(n,w,k)-\log_2 p_{\mathrm{use}}(\delta)+\log_2\big(C_{\mathrm{sel}}+N^{\omega}+V\big),\delta+\log_2 V\big\}+O(1).$$

Within this model, verification in false trials is charged at its expected cost $O(V)$, while the successful trial contributes at most $2^{\delta}V$.

3.2 Rank Model

We next study the rank of the selected XOR system and quantify the effect of rank deficiency on the recovery cost.

The selected XOR matrix is not an arbitrary random matrix, because the rows are chosen according to the condition $|A_i \cap K| \geq q$. Conditional on A_i, the set B_i is sampled uniformly from the b-subsets of $[n]\setminus A_i$. The selection condition therefore induces a small but explicit conditioning effect on the remaining XOR rows.

This conditioning effect can be written explicitly. Suppose that a selected threshold set satisfies $|A_i \cap K| = j$. Since B_i must be disjoint from A_i, the available coordinates for B_i consist of $k-j$ guessed coordinates and $N-(a-j)$ remaining unknown coordinates. Hence

$$\Pr[|B_i \setminus K| = \ell \mid |A_i \cap K| = j] = \frac{\binom{N-(a-j)}{\ell}\binom{k-j}{b-\ell}}{\binom{n-a}{b}}, \qquad 0 \leq \ell \leq b. \tag{4}$$

After substituting the guessed variables, the selected equations therefore look like sparse XOR equations with the row-weight distribution in (4), averaged over the selected values of j. Conditional on row weight ℓ, the unknown coordinates in a row are uniform among the ℓ-subsets of the N remaining variables. Thus M_K is a mixed-row-weight random binary matrix, or equivalently the transpose of a random hypergraph incidence matrix. Rank deficiency and its relation to the hypergraph 2-core have been studied for both fixed and mixed row weights in [12,17]; more general full-rank criteria appear in [11]. When $k = o(n)$, the distribution above has $\ell = b$ with probability $1-o(1)$, so the fixed-row-weight model is the natural first approximation. The lower-weight rows, although asymptotically a vanishing fraction, should not be discarded in a proof because they can remove parity obstructions.

For the corresponding fixed-row-weight model, consider an $e \times N$ random binary matrix whose rows are sampled independently and uniformly from the weight-b vectors. Cooper–Frieze–Pegden [12, Theorem 3] show that, for fixed $b \geq 3$, the transition to maximum possible rank occurs around

$$e = \frac{N}{b}(\log N + \gamma_N).$$

Here γ_N measures the offset from the threshold. The probability of maximum rank tends to zero when $\gamma_N \to -\infty$ and to one when $\gamma_N \to +\infty$. The largest rank allowed by row parity is N for odd b and $N-1$ for even b, since the all-one vector lies in the right kernel in the latter case. Thus the natural rank scale for sparse XOR systems is $N \log N$. This result does not apply directly to our matrix M_K, whose row weights follow the conditioned mixture in (4), but it provides the correct heuristic scale and motivates the rank checks below.

The same model also quantifies the zero-column obstruction. A zero column of M_K corresponds to an unknown seed coordinate absent from every selected XOR equation, so the selected system cannot determine that bit. Conditional on $i \in I_{K,q}$, let η_j be the probability that $|A_i \cap K| = j$, and let $\bar{\jmath}$ be its expected value:

$$\eta_j = \frac{\binom{k}{j}\binom{N}{a-j}}{\sum_{h=q}^{\min(a,k)} \binom{k}{h}\binom{N}{a-h}}, \qquad \bar{\jmath} = \sum_{j=q}^{\min(a,k)} j\eta_j.$$

For a fixed unknown coordinate $v \in [n] \setminus K$, the probability that a selected XOR row contains v is

$$\rho = \frac{b}{n-a}\left(1 - \frac{a-\bar{\jmath}}{N}\right).$$

Indeed, given $|A_i \cap K| = j$, the coordinate v lies outside A_i with probability $1-(a-j)/N$, after which it is selected into B_i with probability $b/(n-a)$. If e rows are selected and Z denotes the number of zero columns of M_K, independence across output rows gives

$$\mathbb{E}[Z \mid |I_{K,q}| = e] = N(1-\rho)^e, \qquad d_K \geq Z. \tag{5}$$

Since $e \sim \mathrm{Bin}(m, p_{\mathrm{sel}})$, averaging (5) over e gives

$$\mathbb{E}[Z] = N(1 - p_{\mathrm{sel}}\rho)^m. \tag{6}$$

Consequently, $N(1-p_{\mathrm{sel}}\rho)^m \leq \mathbb{E}[d_K]$. This is a lower bound on the expected rank deficiency.

We next use this row distribution to derive a high-probability upper bound on the rank deficiency d_K. For $0 \leq \ell \leq b$, the probability that a selected row has Hamming weight ℓ is

$$\nu_\ell := \Pr[|B_i \setminus K| = \ell \mid i \in I_{K,q}] = \sum_{j=q}^{\min(a,k)} \eta_j \frac{\binom{N-(a-j)}{\ell}\binom{k-j}{b-\ell}}{\binom{n-a}{b}},$$

and, for $0 \leq z \leq N$, define

$$\phi_z = \sum_{\ell=0}^{b} \nu_\ell \frac{\sum_{\substack{0 \leq h \leq \ell \\ h \text{ even}}} \binom{z}{h}\binom{N-z}{\ell-h}}{\binom{N}{\ell}}.$$

Thus ϕ_z is the probability that the inner product over $\mathbb{F}_2$ between one selected row and a fixed vector of Hamming weight z is zero.

Lemma 2 (Bounding the rank deficiency). *Conditional on $|I_{K,q}| = e$, set*

$$\Psi_e = \sum_{z=0}^{N} \binom{N}{z} \phi_z^e.$$

Then, for every integer $D \geq 0$,

$$\mathbb{E}[2^{d_K} \mid |I_{K,q}| = e] = \Psi_e, \qquad \Pr[d_K \geq D \mid |I_{K,q}| = e] \leq 2^{-D}\Psi_e.$$

Thus Ψ_e is the expected size of $\ker(M_K)$, conditional on there being e selected rows.

Proof. By the rank–nullity theorem, $|\ker(M_K)| = 2^{d_K}$. For a fixed vector $v \in \mathbb{F}_2^N$ of weight z, each selected row has zero inner product with v with probability ϕ_z. Conditional on $|I_{K,q}| = e$, the selected rows are independent, and hence

$$\Pr[M_K v = 0 \mid |I_{K,q}| = e] = \phi_z^e.$$

Summing over all $\binom{N}{z}$ vectors of each weight z gives

$$\mathbb{E}[2^{d_K} \mid |I_{K,q}| = e] = \sum_{z=0}^{N} \binom{N}{z} \phi_z^e = \Psi_e.$$

Applying Markov's inequality to the nonnegative random variable 2^{d_K} now yields

$$\Pr[d_K \geq D \mid |I_{K,q}| = e] = \Pr[2^{d_K} \geq 2^D \mid |I_{K,q}| = e] \leq 2^{-D}\Psi_e.$$

3.3 An Asymptotic Estimate

We now give an asymptotic estimate of how many deterministic equations DPR can collect. This identifies a range of stretches and localities in which pooling can yield subexponential recovery.

In the fixed-locality regime with $k = o(n)$, (6) shows that $\Theta(n)$ selected equations leave $\Theta(n)$ zero columns in expectation. The natural scale is instead $\Theta(n \log n)$ selected rows, consistent with the fixed-row-weight rank threshold discussed above.

Theorem 1 (Pooled selection at the sparse-matrix coverage scale). *Let a, t, b be fixed constants with $1 \leq t \leq a$, let $m = n^s$, and suppose $1 < s < t + 1$. Set $\alpha = (t + 1 - s)/t \in (0, 1)$. If*

$$k = \Theta\Big(n^{\alpha}(\log n)^{1/t}\Big),$$

and $q = t$, then the expected number of DPR equations is

$$\mu = \Theta(n \log n).$$

For balanced seeds, the expected search factor is $2^{(1+o(1))k}$*. If, for a fixed correct guess, the selected system has rank deficiency at most* $\delta = O(k)$ *with probability* $1 - o(1)$*, DPR has a variant with worst-case running time*

$$2^{O\left(n^{\alpha}(\log n)^{1/t}\right)} \operatorname{poly}(n).$$

It succeeds with probability $1 - o(1)$ *over the sampled pairs* (A_i, B_i) *and the random guesses. Since* $\alpha < 1$*, this running time is* $2^{o(n)}$*.*

Proof. For fixed a, t and $k = o(n)$, the dominant term in (2) with $q = t$ is the term $j = t$. Hence

$$\Pr[|A_i \cap K| \geq t] = \Theta\left(\left(\frac{k}{n}\right)^t\right).$$

Therefore

$$\mu = n^s \cdot \Theta\left(\left(\frac{k}{n}\right)^t\right) = \Theta\left(n^{s-t}k^t\right).$$

With $k = \Theta(n^{\alpha}(\log n)^{1/t})$, this gives $\mu = \Theta(n \log n)$. Since e has mean $\Theta(n \log n)$, a Chernoff bound [9] gives $\Pr[e < N] = o(1)$. Together with the assumed rank bound, this implies $p_{\text{use}}(\delta) = 1 - o(1)$. For a balanced seed, the expected number of guesses for an all-one K is $\binom{n}{k}/\binom{n/2}{k} = 2^{(1+o(1))k}$ when $k = o(n)$. Run at most $R = \lceil S_1 \log n \rceil$ trials, allowing every retained trial, including incorrect guesses, to enumerate all its at most 2^{δ} candidates. The probability of encountering a correct guess is at least $1 - 1/n$. The first such guess is independent of the sampled pairs (A_i, B_i) and is usable with probability $p_{\text{use}}(\delta) = 1 - o(1)$, so recovery succeeds with probability $1 - o(1)$. Pool construction and elimination take polynomial time per trial, and rank screening bounds every trial's enumeration cost by $2^{\delta} \operatorname{poly}(n)$. The total work, including all incorrect guesses, is therefore at most

$$R\, 2^{\delta} \operatorname{poly}(n) = 2^{O(k)} \operatorname{poly}(n).$$

Fixing t threshold inputs reduces $\mathrm{THR}_{a,t} \oplus \mathrm{XOR}_b$ to a degree-one affine function, so the bit-fixing criterion of Applebaum and Lovett [4] applies when $s > t + 1$. DPR extends this viewpoint to $1 < s < t + 1$ by pooling the forcing subsets contained in a larger guessed set, at the cost of a subexponential search.

For MAJ_a, we have $t = (a + 1)/2$, and hence

$$\alpha = 1 - \frac{2(s-1)}{a+1}.$$

For fixed s, this exponent approaches 1 as a grows. Since α determines the leading power of n in the search exponent, DPR gives its largest asymptotic improvement for small majority localities.

3.4 Application to the $\mathbf{MAJ_7 \oplus XOR_4}$ Challenge

To evaluate the effectiveness of DPR at concrete parameters, we apply it to a representative challenge instance with small majority locality a. The SoK of Boura et al. lists $n = 1024$, $s = 2$, and the predicate $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ as a 128-bit challenge parameter [7, Table 3].

For MAJ_7, we take $t = q = 4$. Specializing Theorem 1 gives

$$k = \Theta\Big(n^{(5-s)/4}(\log n)^{1/4}\Big).$$

For $s = 2$, this yields the subexponential search factor

$$2^{\Theta(n^{3/4}(\log n)^{1/4})}.$$

Concrete instantiation for $n = 1024$ and $s = 2$. We evaluate the attack at balanced seed weight $w = 512$. Evaluating the finite-parameter tradeoff between search cost and the size and rank of the selected system gives the suitable choice $k = 80$, and hence $N = 944$. Equations (2) and (3), together with the binomial law in Lemma 1, give the expected row count $\mu = \mathbb{E}[e]$, the probability that the actual count e falls below N, and the search exponent:

$$\mu \approx 1060.22, \qquad \Pr[e < N] \approx 1.31 \cdot 10^{-4}, \qquad \log_2 S_1(1024, 512, 80) \approx 84.83.$$

Combining the bound on $\Pr[e < N]$ with Lemma 2 gives

$$\begin{aligned} p_{\mathrm{use}}(39) &= \Pr[e \geq N] \Pr[d_K \leq 39 \mid e \geq N] \\ &\geq (1 - 1.31 \cdot 10^{-4})(1 - 1.09 \cdot 10^{-4}) > 0.99976. \end{aligned}$$

Taking $\delta = 39$ and $\omega = 2.38$, the search and final candidate enumeration have estimated costs $2^{109.71}$ and $2^{62.46}$, respectively. Hence the total work is dominated by $2^{109.71}$, below the claimed 128-bit security level. Appendix A independently establishes a recovery probability exceeding 0.63 within the same rounded work estimate, under the actual distribution of (A_i, B_i). Even with $\omega = 3$, the estimated work factor is about $2^{114.5}$ and remains below 2^{128}.

For small majority locality a, the most relevant prior seed-recovery method is the GAS attack of Fu et al. [19]. For a fixed set R of size r, GAS selects outputs with $R \subseteq A_i$, guesses $x_R = \mathbf{1}_r$, and solves the resulting noisy XOR system using Gauss or SubGauss.

Table 1 compares GAS with DPR for balanced seeds and several stretches. Here β is the probability that a selected equation is correct. For each DPR row, δ is the smallest integer for which

$$\Pr[d_K > \delta \mid e \geq N] \leq 1.31 \cdot 10^{-4},$$

and the reported value of p_{use} is a lower bound. The last column reports the corresponding seed-recovery cost estimate $\log_2 T_{\mathrm{rec}}$. Appendix A gives an additional analysis with explicit budgets and numerical candidate-count evaluations for these rows.

n	s	method	parameter	μ	β	δ	p_{use}	$\log_2 T_{\text{rec}}$
512	2.0	GAS [19]	$r = 1$	3584.0	0.6553	–	–	334.1
512	2.0	DPR	$k = 47,\ q = 4$	468.7	1	29	0.5746	73.9
512	2.5	GAS [19]	$r = 2$	952.2	0.8113	–	–	178.0
512	2.5	DPR	$k = 22,\ q = 4$	492.3	1	27	0.5468	53.2
1024	2.0	GAS [19]	$r = 1$	7168.0	0.6558	–	–	647.6
1024	2.0	DPR	$k = 80,\ q = 4$	1060.2	1	39	0.9997	109.7
1024	3.0	GAS [19]	$r = 2$	43050.0	0.8119	–	–	333.7
1024	3.0	DPR	$k = 15,\ q = 4$	1097.4	1	36	0.9965	69.5
1024	3.3	GAS [19]	$r = 3$	1684.9	0.9371	–	–	123.1
1024	3.3	DPR	$k = 10,\ q = 4$	1366.7	1	20	0.9999	56.5

Table 1. Comparison between GAS and DPR for $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ at balanced seed weight. The GAS entries are recomputed using the formulas of Fu et al. [19]. The estimates use $\omega = 2.38$.

s	k	$\mathbb{E}[e]$	N	d_K^{th}	δ	mean d_K	max d_K
2.0	80	1060.22	944	14.78	40	16.60	36
2.5	32	867.50	992	124.50	231	123.27	209
2.5	38	1755.17	986	1.01	14	1.02	6
2.5	42	2635.85	982	0.032	12	0.029	1
2.7	28	1994.81	996	0.399	13	0.42	3
2.7	32	3470.00	992	0.0012	12	0.00	0

Table 2. Rank experiment for DPR with $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$, $n = 1024$, and $q = 4$. Here d_K^{th} denotes the theoretical lower-bound estimate for $\mathbb{E}[d_K]$ defined above, while δ is the conservative threshold obtained from Lemma 2. The last two columns report the mean and maximum deficiency among 2000 matrices.

Experiment. We experimentally examine the rank behavior of the selected DPR systems. For each pair (s, k), we sampled e using Lemma 1 and generated the rows from (4), with 2000 independent matrices per pair.

Since $\mathrm{rank}(M_K) \le e$ and $d_K \ge Z$, we use the following theoretical lower-bound estimate for the rank deficiency:

$$d_K^{\text{th}} := \max\{\mathbb{E}[\max\{N - e, 0\}], \mathbb{E}[Z]\} \le \mathbb{E}[d_K].$$

The two terms are evaluated using Lemma 1 and (6), respectively. We also report a conservative threshold δ, defined as the smallest value for which Lemma 2, combined with the binomial distribution of e, gives $\Pr[d_K > \delta] \le 1.31 \cdot 10^{-4}$. This threshold is unconditional; Table 1 instead conditions on $e \ge N$, because systems with fewer rows are discarded.

At the challenge point $(s, k) = (2, 80)$, the observed mean deficiency 16.60 is close to the zero-column lower bound 14.78. This suggests that the dominant source of rank loss is the absence of some columns, rather than more complicated sparse dependencies. Moreover, the maximum observed deficiency was 36, below both the conditional threshold 39 used in the attack estimate and the unconditional threshold 40 in the table. The estimate from Lemma 2 is much

looser than the observed rank behavior, but it gives a rigorous high-probability estimate under the mixed-row model. For all rows in Table 2, the maximum observed deficiency is below the corresponding conservative threshold δ.

4 Bias-Amplified Pooled Recovery

For majority predicates, DPR is most effective when the majority locality a is small: forcing every selected threshold value requires $q \geq t$, which becomes increasingly costly as $t = (a+1)/2$ grows. To extend pooled selection to a broader range of localities, we introduce *Bias-Amplified Pooled Recovery* (BAPR). BAPR retains the DPR selection rule but relaxes the forcing requirement. A correct guess now biases the selected threshold values rather than determining them, producing a system of noisy XOR equations for seed recovery.

The common-bias test of [16] gives strong average-case attacks for several parameter sets, but its reported advantage is averaged over a uniformly random secret seed. We refine this analysis by evaluating the advantage separately at each fixed seed weight.

4.1 A Fixed-Weight Refinement of Common Bias

We condition on $\mathrm{wt}(x) = w$ and evaluate the distinguishing advantage separately for each Hamming weight. This distinguishes average-seed performance from guarantees that hold uniformly across seed weights.

Fix $\lambda \geq 1$ and set $L = n + \lambda$. To describe the test using a common constant, define

$$\text{c-bias}^{L}_{\text{const}}(\mathrm{THR}_{a,t}) = \max_{c\in\{0,1\}} \Pr[\mathrm{THR}_{a,t}(x_{A_1}) = \cdots = \mathrm{THR}_{a,t}(x_{A_L}) = c]\,. \quad (7)$$

The probability is taken over the seed and the random public threshold input sets in the average-case model.

When this event occurs, XORing the output with $c\mathbf{1}_L$ removes the threshold contribution and leaves a linear system from the XOR parts. Linear dependencies among these equations therefore give consistency checks on the output.

Theorem 2 (Common-bias consistency test [16]). *Consider L output equations from a Goldreich PRG instantiated with $\mathrm{THR}_{a,t} \oplus \mathrm{XOR}_b$, and let $M \in \mathbb{F}_2^{L\times n}$ be the selected XOR matrix. There exists a rank-λ matrix $H \in \mathbb{F}_2^{\lambda\times L}$ such that $HM = 0$. Fix a constant $c \in \{0,1\}$ attaining the maximum in (7). The test that accepts an output $y \in \{0,1\}^L$ exactly when $H(y \oplus c\mathbf{1}_L) = 0$ has distinguishing advantage at least*

$$\text{c-bias}^{L}_{\text{const}}(\mathrm{THR}_{a,t}) - 2^{-\lambda}.$$

To expose the seed-weight dependence, for $0 \leq h \leq n$ define

$$q_{a,t}(n,h) = \sum_{u=t}^{a} \frac{\binom{h}{u}\binom{n-h}{a-u}}{\binom{n}{a}}. \quad (8)$$

This is the probability that one threshold input evaluates to 1, conditioned on a seed of Hamming weight h. The restricted common bias above can therefore be written as

$$\text{c-bias}^{L}_{\text{const}}(\text{THR}_{a,t}) = 2^{-n} \max \left\{ \sum_{h=0}^{n} \binom{n}{h} q_{a,t}(n,h)^L, \sum_{h=0}^{n} \binom{n}{h} (1 - q_{a,t}(n,h))^L \right\}.$$

Unbalanced seeds may dominate this average even when the same event is unlikely for a fixed balanced seed.

Now fix a seed $x \in \{0,1\}^n$ with $\text{wt}(x) = w$. Equation (8) gives

$$\Pr_A[\text{THR}_{a,t}(x_A) = 1] = q_{a,t}(n,w),$$

where A is uniform in $\binom{[n]}{a}$. Without selection, define

$$P^{\text{no sel}}_{\text{good}}(w) = \max \left\{ q_{a,t}(n,w)^L, (1 - q_{a,t}(n,w))^L \right\}.$$

The corresponding fixed-weight advantage lower bound is $\text{Adv}^{\text{no sel}}_{w} = P^{\text{no sel}}_{\text{good}}(w) - 2^{-\lambda}$, which is positive only when $P^{\text{no sel}}_{\text{good}}(w) > 2^{-\lambda}$.

Table 3 refines the no-selection analysis of [16] for $\text{THR}_{64,32}$ at fixed weights. The cost is $\tau_0 = Ln^2$, as in [16, Remark 3]. Each weight range is exactly the set of weights for which $\text{Adv}^{\text{no sel}}_{w} > 0$. The cost column gives the largest cost in that range, attained at the boundary with the smaller positive advantage.

n	λ	$\log_2 \tau_0$	weight range	$\log_2(\tau_0/\text{Adv}^{\text{no sel}}_{w})$	seed fraction
256	21	24.11	$w \geq 149$ or $w \leq 103$	43.68	$2^{-7.34}$
512	39	27.11	$w \geq 302$ or $w \leq 202$	64.57	$2^{-15.09}$
894	65	29.51	$w \geq 530$ or $w \leq 350$	95.01	$2^{-25.92}$
2048	144	33.10	$w \geq 1220$ or $w \leq 797$	177.07	$2^{-58.64}$

Table 3. Fixed-weight evaluation of the common-bias test without selection for $\text{THR}_{64,32}$. The seed fraction is the probability of this range for $W \sim \text{Bin}(n, 1/2)$.

Effect of selection. D'Antona et al. [16] choose the common intersection adaptively from the public input sets. This conditions the remaining input sets on the selection process, so their independence and uniformity do not follow directly. We make a small change to obtain an exact fixed-weight calculation: fix $0 \leq r \leq a$ and choose $R \leftarrow \binom{[n]}{r}$ independently of the sampled input sets. Scan the outputs and retain the first $L = n + \lambda$ for which $R \subseteq A_i$. Each output qualifies with probability $\binom{a}{r}/\binom{n}{r}$. Conditional on qualifying, its threshold set

has the form $A_i = R \cup C_i$, where the retained C_i are independent and uniform in $\binom{[n]\setminus R}{a-r}$. Define

$$\pi_j(w,r) = \frac{\binom{w}{j}\binom{n-w}{r-j}}{\binom{n}{r}},$$

this is the probability that R contains exactly j ones, conditioned on $\mathrm{wt}(x) = w$. Given this event, the remaining $a-r$ threshold positions are sampled from the $n-r$ coordinates outside R, among which $w-j$ are ones. Hence

$$\theta_j(n,w,a,t,r) = \sum_{u=t-j}^{a-r} \frac{\binom{w-j}{u}\binom{n-r-(w-j)}{a-r-u}}{\binom{n-r}{a-r}}.$$

is the conditional probability that one selected threshold output equals 1. The fixed-weight favourable probability in this model is therefore

$$P^{\mathrm{sel}}_{\mathrm{good}}(w;r) = \max\left\{ \sum_{j=0}^{r} \pi_j(w,r)\theta_j(n,w,a,t,r)^L, \right.$$
$$\left. \sum_{j=0}^{r} \pi_j(w,r)\big(1-\theta_j(n,w,a,t,r)\big)^L \right\}.$$

Conditional on obtaining L rows, the corresponding lower-bound advantage is $\mathrm{Adv}^{\mathrm{sel}}_w(r) = P^{\mathrm{sel}}_{\mathrm{good}}(w;r) - 2^{-\lambda}$.

Table 4 reports fixed-weight results for the common-bias test with selection and $\mathrm{THR}_{64,32}$. Each weight range is exactly the set of weights for which $\mathrm{Adv}^{\mathrm{sel}}_w(r) > 0$. We use the (r,λ) values of [16, Table 1], where r is denoted by s. Accounting for row selection and the linear consistency test, the expected cost is $\tau = aL\binom{n}{r}/\binom{a}{r} + Ln^2$.

n	λ	r	$\log_2 \tau$	weight range	$\log_2(\tau/\mathrm{Adv}_w)$	seed fraction
256	21	4	24.46	$w \geq 144$ or $w \leq 108$	46.57	$2^{-4.90}$
512	39	4	28.17	$w \geq 290$ or $w \leq 214$	66.16	$2^{-9.26}$
894	65	5	35.17	$w \geq 500$ or $w \leq 379$	103.32	$2^{-12.13}$
2048	144	4	37.31	$w \geq 1163$ or $w \leq 852$	180.28	$2^{-31.08}$

Table 4. Fixed-weight evaluation of the common-bias test with selection for $\mathrm{THR}_{64,32}$, using the parameters of [16, Table 1]. Here $\mathrm{Adv}_w = \mathrm{Adv}^{\mathrm{sel}}_w(r)$. The cost is the maximum over the displayed weight range attained at its weaker boundary, and the seed fraction is the probability of that range for $W \sim \mathrm{Bin}(n, 1/2)$.

Tables 3 and 4 evaluate boundary costs at different seed weights, so these costs should not be compared directly. The lower bound is weakest near balance, while unbalanced seeds contribute most of the average advantage. This motivates using the guessed bits to form biased linear equations instead of requiring all selected threshold values to agree.

Remark 3. Appendix C of D'Antona et al. [16] extends the common-bias test to uniformly random balanced seeds. It selects outputs whose threshold inputs lie in $n' < n$ coordinates and averages over the hypergeometric seed weight on this subset. However, the method requires a large stretch to obtain $n + \lambda$ selected outputs. Its concrete costs are substantially higher than in the random-seed analysis, showing that balanced seeds considerably reduce its effectiveness.

4.2 Bias-Amplified Pooled Recovery

The GAS attack of Fu et al. [19] exploits bias in the threshold values, but each group is restricted to one fixed common subset and can be costly when increasing the bias leaves too few equations. Although their FiLIP attack uses a larger guessed set to handle per-output whitening, our focus is seed recovery for Goldreich PRGs when fixed-subset groups yield too few equations.

We retain the notation $K, k, N, I_{K,q}, e, p_{\text{sel}}, \mu$, and S_1 from Section 3.1. The formulas in (2) and (3) remain valid for $q < t$. We now use such a value of q and condition on a correct guess $x_K = \mathbf{1}_k$.

Pooled selection. For fixed seed weight w, conditioning on $|A_i \cap K| = j$ leaves $a - j$ threshold positions outside K, among which $w - k$ of the N available seed coordinates are ones. Using (8), set

$$\beta_j(w) = \begin{cases} q_{a-j,t-j}(N, w-k), & j < t, \\ 1, & j \geq t. \end{cases}$$

Conditioned on selection, the possible values of j have weights proportional to $\binom{k}{j}\binom{N}{a-j}$. Thus the effective holding probability is

$$\beta_{k,q}(w) = \frac{\sum_{j=q}^{\min(a,k)} \binom{k}{j}\binom{N}{a-j}\,\beta_j(w)}{\sum_{j=q}^{\min(a,k)} \binom{k}{j}\binom{N}{a-j}}.$$

This is the effective holding probability of the selected threshold values. When $\beta_{k,q}(w) > 1/2$, the selected outputs give biased noisy XOR equations.

For a correct guess $x_K = \mathbf{1}_k$, substituting the guessed XOR variables and predicting the selected threshold value as 1 gives, for each $i \in I_{K,q}$, the noisy equation

$$\bigoplus_{v \in B_i \setminus K} x_v = y_i \oplus 1 \oplus (|B_i \cap K| \bmod 2).$$

It is correct exactly when $\mathrm{THR}_{a,t}(x_{A_i}) = 1$, and hence with probability $\beta_{k,q}(w)$ conditioned on selection and on $\mathrm{wt}(x) = w$.

The all-zero branch instead guesses $x_K = \mathbf{0}_k$, predicts the selected threshold values as zero, and has search factor $\binom{n}{k}/\binom{n-w}{k}$. Conditioned on $|A_i \cap K| = j$, its holding probability is $1 - q_{a-j,t}(N, w)$; averaging over j uses the same selection weights as above.

The DPR method from the previous section is the special case $q \geq t$. In that case $\beta_{k,q}(w) = 1$, so the selected equations are noiseless. BAPR relaxes the selection condition, obtains many more equations, and then uses the bias $\beta_{k,q}(w) > 1/2$ algorithmically.

Rank of a random subsystem. The row distribution and rank bounds of Section 3.2 also apply when $q < t$. At the row-count thresholds used in the concrete estimates below, both solvers sample N distinct equations uniformly from the selected pool. Write R for the sampled indices, $M_{K,R}$ for the submatrix of M_K formed by those rows, and $d_{K,R} = N - \operatorname{rank}(M_{K,R})$. An inconsistent system is discarded; otherwise, its affine solution space contains exactly $2^{d_{K,R}}$ candidates, which are enumerated and checked against the full output.

For each pool size $e \geq N$, averaging over the sampled pairs (A_i, B_i) and the random choice of R, the sampled equations have the distribution of N independent selected rows. Lemma 2 therefore gives

$$\mathbb{E}[2^{d_{K,R}} \mid e] = \Psi_N.$$

Since inconsistent subsystems yield no candidates, Ψ_N bounds the expected candidate count per sampled subsystem.

Recovery cost. The symbols N, e, m, and ω retain their meanings from the previous section; let $\beta > 1/2$ denote the holding probability of a noisy equation and let e denote the base of the natural logarithm.

To solve the noisy equations, we use the Gauss and SubGauss solvers of Fu et al. [19] under their solver-cost assumptions. Both solvers repeatedly choose a subsystem and provisionally treat its noisy equations as correct. They then solve the subsystem by Gaussian elimination, enumerate any remaining free variables, and verify the resulting candidates against the full PRG output. The Gauss solver uses subsystems of size $\min\{e, N\}$, whereas SubGauss uses subsystems of size $\min\{\lfloor e\beta \rfloor, N\}$ and tries several random subsystems from the same pool. The Gauss estimate is

$$T_{\mathrm{G}}(N, e, \beta; m) = \beta^{-\min(e,N)} \begin{cases} m + N^{\omega}, & e \geq N, \\ m + 2^{N-e} e^{\omega}, & e < N. \end{cases}$$

The SubGauss estimate is

$$T_{\mathrm{SG}}(N, e, \beta; m) = \frac{\mathrm{e}}{\mathrm{e} - 1} \begin{cases} m + \beta^{-N} N^{\omega}, & e \geq N/\beta, \\ m + 2^{N - e\beta} \beta^{-e\beta} (e\beta)^{\omega}, & e < N/\beta. \end{cases}$$

For a fixed (k, q), set $\beta = \beta_{k,q}(w)$. We discard pools with $e < N$ for Gauss or $e < \lceil N/\beta \rceil$ for SubGauss. Both solvers therefore use N-row subsystems, whose

Algorithm 2 Bias-Amplified Pooled Recovery (BAPR) with Gauss or SubGauss for $\mathrm{THR}_{a,t} \oplus \mathrm{XOR}_b$.

Require: Public sets $(A_i, B_i)_{i=1}^m$, output $y \in \{0,1\}^m$, parameters k, q, a holding estimate $\beta > 1/2$, and a choice of Gauss or SubGauss.
Ensure: A seed candidate $x \in \{0,1\}^n$, or failure.
Choose or enumerate a candidate set $K \subseteq [n]$ with $|K| = k$.
Guess $x_K = \mathbf{1}_k$.
Select the outputs indexed by $I_{K,q}$ and compute the row count e.
Set $e_0 \leftarrow N$ for Gauss or $e_0 \leftarrow \lceil N/\beta \rceil$ for SubGauss.
if $e < e_0$ **then**
    return failure.
end if
for each $i \in I_{K,q}$ **do**
    Form the noisy XOR equation

$$\bigoplus_{v \in B_i \setminus K} x_v = y_i \oplus 1 \oplus (|B_i \cap K| \bmod 2).$$

end for
for each subsystem trial prescribed by the chosen solver **do**
    Sample R uniformly among the N-subsets of $I_{K,q}$.
    Apply Gaussian elimination to this subsystem and compute $d_{K,R} = N - \mathrm{rank}(M_{K,R})$.
    if the subsystem is consistent **then**
        Enumerate its $2^{d_{K,R}}$ affine solutions and verify the completed seed candidates against the full PRG output.
        return immediately if a candidate passes verification.
    end if
end for
return failure.

rank correction is Ψ_N. The resulting heuristic all-one recovery-cost estimate is

$$T_{\mathrm{BAPR}}(w;k,q) = S_1(n,w,k) \min\left\{ \frac{\Psi_N T_{\mathrm{G}}(N,N,\beta;m)}{\Pr[e \geq N]}, \frac{\Psi_N T_{\mathrm{SG}}(N,\lceil N/\beta \rceil,\beta;m)}{\Pr[e \geq \lceil N/\beta \rceil]} \right\}.$$

The denominators account for discarded pools. Optimizing the estimate over valid parameters gives

$$T_{\mathrm{BAPR}}(w) = \min_{\substack{1 \leq k \leq w,\ 0 \leq q \leq \min(a,k),\ q<t \\ \beta_{k,q}(w) > 1/2}} T_{\mathrm{BAPR}}(w;k,q).$$

Algorithm 2 summarizes the BAPR procedure used for these estimates.

4.3 Concrete Examples

A challenge predicate. We first apply the model to a predicate appearing in proposed challenge instances, namely $\mathrm{THR}_{64,32} \oplus \mathrm{XOR}_{10}$ [7]. Its threshold is $t = 32$. Deterministic pooling would therefore require $q \geq 32$, which is too expensive at the stretches considered here. BAPR instead uses $q < 32$, producing biased noisy XOR equations.

D'Antona et al. [16, Table 11] report distinguishing-cost estimates for uniformly random balanced seeds. We compare these estimates with the BAPR recovery costs at balanced seed weight in Table 5. For the CB rows, n' is the number of coordinates supporting the selected threshold inputs, while λ is the number of independent consistency checks obtained from $n + \lambda$ selected outputs.

n	s	method	parameter	μ	β	$\log_2(\text{cost})$
512	4.5	CB [16]	$n' = 375,\ \lambda = 71$	–	–	112.84
512	4.5	BAPR	$k = 19,\ q = 14$	647.7	0.9756	62.1
894	3.4	CB [16]	$n' = 702,\ \lambda = 159$	–	–	194.31
894	3.4	BAPR	$k = 45,\ q = 15$	921.7	0.9778	99.1
2048	3.0	CB [16]	$n' = 1625,\ \lambda = 388$	–	–	423.60
2048	3.0	BAPR	$k = 118,\ q = 16$	2184.1	0.9824	200.2

Table 5. Attack-cost estimates for $\mathrm{THR}_{64,32} \oplus \mathrm{XOR}_{10}$ with balanced seeds. The CB results are from [16, Table 11]. The BAPR estimates use $\omega = 2.38$.

In particular, the BAPR estimates do not come from averaging over weak, unbalanced seeds. They apply to the balanced case itself, and therefore indicate that the corresponding parameters cannot obtain their claimed security merely by restricting attention to typical balanced seeds.

More instances. We further evaluate BAPR on the parameters proposed by Fu et al. [19, Table 2]. For even n, we take $w = n/2$; for odd n, we use the nearest balanced weight. Table 6 summarizes the comparison.

The much smaller $\mathrm{CB}_{\mathrm{avg}}$ costs may appear stronger, but they average over the random seed and do not apply uniformly across seed weights. Their lower effectiveness at balance indicates that the average advantage is driven by more unbalanced, weak seeds. For a fixed weight w, we evaluate the common-bias cost as $\log_2(\tau_0 / \mathrm{Adv}_w^{\text{no sel}})$, where $\tau_0 = (n + \lambda)n^2$. Table 7 quantifies its coverage across seed weights. The columns under coverage at $\mathrm{CB}_{\mathrm{avg}}$ list the weights whose cost does not exceed the reported average-seed cost.

At the $\mathrm{CB}_{\mathrm{avg}}$ costs reported in [16, Table 4], the covered seed fraction ranges from approximately 0.036% to 11.8%. Allowing any cost below the respective value of κ increases the maximum coverage to approximately 14.3%–53.0%, but still leaves a central range of seed weights uncovered by this bound. At the balanced or nearest-balanced weights in Table 6, it gives no positive advantage lower bound, whereas the heuristic BAPR recovery estimates remain below the

κ	n	m	(a,b)	λ	$\mathrm{CB}_{\mathrm{avg}}$	$\mathrm{CB}_{\mathrm{bal}}$	BAPR	(k,q)
80	91	281	$(47,7)$	12	$2^{25.94}$	none	$2^{69.26}$	$(25,15)$
128	141	485	$(47,7)$	16	$2^{32.62}$	none	$2^{93.90}$	$(40,16)$
256	285	1170	$(47,7)$	30	$2^{49.16}$	none	$2^{161.23}$	$(72,15)$
80	105	1075	$(57,8)$	12	$2^{26.33}$	none	$2^{64.12}$	$(23,16)$
128	155	1929	$(57,8)$	16	$2^{32.20}$	none	$2^{81.33}$	$(32,16)$
256	303	5274	$(57,8)$	28	$2^{47.17}$	none	$2^{130.20}$	$(58,16)$
80	146	21316	$(75,10)$	13	$2^{28.61}$	none	$2^{60.72}$	$(25,19)$
128	205	42025	$(75,10)$	17	$2^{34.08}$	none	$2^{76.30}$	$(37,21)$
256	365	133225	$(75,10)$	28	$2^{47.23}$	none	$2^{108.75}$	$(55,20)$

Table 6. Attacks on the challenge parameters of Fu et al. [19, Table 2]. The $\mathrm{CB}_{\mathrm{avg}}$ column reproduces the average-seed costs reported in [16, Table 4], whereas $\mathrm{CB}_{\mathrm{bal}}$ evaluates the same advantage bound at $w = \lfloor n/2 \rfloor$; *none* means that this bound is non-positive. The BAPR estimates use $\omega = 2.38$.

(n,κ)	λ	coverage at $\mathrm{CB}_{\mathrm{avg}}$		$\lambda^\star$	cost below κ		
		weight range	seed fraction		weight range	seed fraction	cost
$(91,80)$	12	$\leq 37,\ \geq 54$	$2^{-3.43}$	55	$\leq 42,\ \geq 49$	$2^{-0.92}$	74.03
$(141,128)$	16	$\leq 56,\ \geq 85$	$2^{-5.79}$	80	$\leq 64,\ \geq 77$	$2^{-1.68}$	101.25
$(285,256)$	30	$\leq 112,\ \geq 173$	$2^{-11.43}$	211	$\leq 130,\ \geq 155$	$2^{-2.69}$	235.70
$(105,80)$	12	$\leq 44,\ \geq 61$	$2^{-3.08}$	41	$\leq 48,\ \geq 57$	$2^{-1.20}$	59.83
$(155,128)$	16	$\leq 63,\ \geq 92$	$2^{-5.37}$	84	$\leq 71,\ \geq 84$	$2^{-1.58}$	105.09
$(303,256)$	28	$\leq 121,\ \geq 182$	$2^{-10.84}$	207	$\leq 139,\ \geq 164$	$2^{-2.57}$	231.53
$(146,80)$	13	$\leq 62,\ \geq 84$	$2^{-3.61}$	49	$\leq 67,\ \geq 79$	$2^{-1.46}$	68.98
$(205,128)$	17	$\leq 85,\ \geq 120$	$2^{-5.85}$	87	$\leq 94,\ \geq 111$	$2^{-1.92}$	108.96
$(365,256)$	28	$\leq 149,\ \geq 216$	$2^{-10.87}$	205	$\leq 168,\ \geq 197$	$2^{-2.81}$	230.23

Table 7. Fixed-weight coverage of the common-bias test for the instances in Table 6. The value $\lambda^\star$ maximizes the seed fraction for which the cost remains below κ. The final column gives the fixed-weight cost at the boundary closest to $n/2$. The security target κ and all cost exponents are given in bits.

corresponding values of κ. Under the stated solver-cost model, BAPR therefore provides attack estimates in a central seed-weight region not covered by the common-bias advantage bound.

4.4 The Intermediate-Locality Regime

D'Antona et al. [16] leave open whether $\mathrm{MAJ}_a \oplus \mathrm{XOR}_b$ can retain exponential security when $a \in \omega(1) \cap O(\log n)$. We next analyze BAPR asymptotically to determine whether bias-amplified pooling remains effective in this intermediate-locality regime.

Theorem 3 (Subexponential seed recovery with BAPR). *Let $a = a(n)$ be odd with $a \in \omega(1) \cap O(\log n)$, set $t = (a+1)/2$, and let $m = n^s$ for a fixed*

$s > 1$. Let the XOR locality $b \geq 3$ be fixed, and choose

$$q = \left\lceil \sqrt{a \log a} \right\rceil, \qquad k = \left\lceil 2\frac{q}{a} n \left(n^{1-s} \log n \right)^{1/q} \right\rceil.$$

BAPR admits an implementation with worst-case running time $2^{o(n)}$ and seed-recovery probability $1-o(1)$ over independently sampled pairs (A_i, B_i), uniformly over all seeds.

Proof. Write $N = n - k$. To simplify notation, set $\zeta_n = (n^{1-s} \log n)^{1/q}$. Since $q = \Theta(\sqrt{a \log a}) = o(a)$ and $q = O(\sqrt{\log n \log\log n}) = o(\log n)$, we have $q \to \infty$ and eventually $q < t$. Moreover,

$$\log(q\zeta_n) = \log q - \frac{(s-1)\log n - \log\log n}{q} \longrightarrow -\infty,$$

because $\log n/q \gg \log q$. As $\zeta_n = n^{-o(1)}$, it follows that $k/q \sim 2n\zeta_n/a \to \infty$, $k = n^{1-o(1)} = o(n)$, and $ak/n \sim 2q\zeta_n = o(1)$. Thus $q \leq k$ for large n. Fix a k-subset K independently of the sampled pairs (A_i, B_i). For $J = |A_i \cap K|$, the event $J = q$ gives

$$\begin{aligned} p_{\text{sel}} \geq \Pr[J = q] &= (1 + o(1)) \binom{a}{q} \left(\frac{k}{n}\right)^q \\ &\geq (1 - o(1)) 2^q n^{1-s} \log n. \end{aligned}$$

Here the approximation uses $a = O(\log n)$, $k = n^{1-o(1)}$, and $ak/n = o(1)$; the last inequality uses $\binom{a}{q} \geq (a/q)^q$ and the choice of k. Thus $\mu = n^s p_{\text{sel}} \geq (1 - o(1)) 2^q n \log n$, so $\mu/(N \log N) \to \infty$.

For the rank bound, let e_b count selected rows of weight b outside K. They form a submatrix of M_K. Conditional on a selected A_i, $\Pr[B_i \cap K \neq \varnothing] \leq bk/(n-a) = o(1)$, so $\mathbb{E}[e_b] = (1 - o(1))\mu$. We take the target number of weight-b rows to be

$$L_N := \left\lceil \frac{N}{b} (\log N + q) \right\rceil.$$

As $L_N = \Theta(n \log n)$ and $\mu/L_N = \Omega(2^q) \to \infty$, a Chernoff bound gives $e_b \geq L_N$ with probability $1 - o(1)$. Conditional on selection and row weight b, their supports are independent and uniform among the b-subsets of $[n] \setminus K$. The first L_N such rows have a repeated support with probability at most

$$\frac{\binom{L_N}{2}}{\binom{N}{b}} = O(N^{2-b} \log^2 N) = o(1).$$

Without repetition, the supports are sampled uniformly without replacement. Since $q \to \infty$ and $q = o(\log N)$, Cooper–Frieze–Pegden [12, Theorem 3] gives rank N for odd b and $N - 1$ for even b, with probability $1 - o(1)$. Hence

$$N - \operatorname{rank}(M_K) \leq 1$$

with probability $1 - o(1)$.

For a worst-case time bound, we replace Gauss and SubGauss retries with error enumeration. The Cooper–Frieze–Pegden rank bound above guarantees a basis of at least $N - 1$ original rows with high probability. Since this basis is selected according to the coefficients, the single-trial probability β^N for uniformly sampled N-row subsystems does not directly apply. We therefore bound the number of errors in the chosen basis conditional on the coefficients and enumerate all error patterns up to a sublinear weight.

The implementation enumerates every k-subset K and both guesses $x_K = c\mathbf{1}_k$, $c \in \{0, 1\}$. It discards a branch if $\operatorname{rank}(M_K) < N - 1$; otherwise it chooses a basis of original rows in a fixed order, indexed by R. This basis depends only on the selected indices and XOR supports. Under a correct guess, we bound the error probability of each selected equation by ε_n and enumerate error patterns of weight at most h, where

$$\varepsilon_n = \exp\left(-\frac{2q^2}{9a}\right), \qquad h = \lceil N\sqrt{\varepsilon_n}\,\rceil .$$

For $i \in R$, let $\ell_i = y_i \oplus c \oplus c(|B_i \cap K| \bmod 2)$. The algorithm enumerates $z \in \{0, 1\}^{|R|}$ with $\operatorname{wt}(z) \le h$, solves $M_{K,R}u = \ell_R \oplus z$, and verifies its at most two completed seeds against the full output. It returns the first one that passes.

To bound errors for this coefficient-dependent basis, fix a seed x with $w \ge n/2$ and a correct all-one set K independent of the sampled pairs. For a selected row, condition on its XOR support B and on $J = |A \cap K| = j \ge q$. Put $r = |B \setminus K| \le b$ and $v = \operatorname{wt}(x_{B\setminus K}) \le r$. The number U of ones in $A \setminus K$ has distribution

$$U \mid (B, J = j) \sim \operatorname{Hypergeom}(N - r, w - k - v, a - j).$$

Indeed, $A \setminus K$ is uniform among the $(a-j)$-subsets outside $K \cup B$. For $q \le j < t$, $w \ge n/2$ gives

$$\begin{aligned}\mathbb{E}[U \mid B, J = j] - (t - j) &\ge \frac{j-1}{2} - \frac{(a-j)(k+b)}{2(N-b)} \\ &\ge \frac{q}{3}\end{aligned}$$

for large n, since $ak/n = o(1)$ and $ab/n = o(1)$. The hypergeometric lower-tail bound [10] gives

$$\Pr[\mathrm{MAJ}_a(x_A) = 0 \mid B, J = j] \le \exp\left(-\frac{2q^2}{9(a-j)}\right) \le \varepsilon_n.$$

For $j \ge t$, the threshold value is forced to one. Averaging over $j \ge q$, still conditional on B, therefore yields

$$\Pr[\mathrm{MAJ}_a(x_A) = 0 \mid B,\ |A \cap K| \ge q] \le \varepsilon_n = o(1). \tag{9}$$

Let $\mathcal{F}$ record all XOR supports B_i and selection indicators $\mathbf{1}_{\{|A_i \cap K| \ge q\}}$. The row basis R depends only on $\mathcal{F}$. Independence across outputs extends (9) to

conditioning on all of $\mathcal{F}$. For discarded pools set $R = \varnothing$. Let H count errors in R. Linearity of expectation and Markov's inequality give

$$\mathbb{E}[H \mid \mathcal{F}] \le |R|\varepsilon_n \le N\varepsilon_n, \qquad \Pr[H > h \mid \mathcal{F}] \le \frac{N\varepsilon_n}{h} \le \sqrt{\varepsilon_n} = o(1).$$

If the pool has rank at least $N-1$ and $H \le h$, enumeration includes the true error pattern and hence the true seed among its candidates. By a union bound, this event has probability $1 - o(1)$.

For each fixed x with $w \ge n/2$, choose a correct all-one set $K_*(x)$ depending only on x; it exists since $k = o(n)$. The algorithm visits this branch unless it has already returned a verified seed. The same argument applies when $w < n/2$ to an all-zero set, by complement symmetry of odd majority. The bounds are uniform over x.

The rank check leaves at most two solutions per error pattern on every branch, including incorrect guesses. Since $k, h = o(n)$, the binomial entropy bound gives $\binom{n}{k} = 2^{o(n)}$ and $\sum_{j=0}^{h}\binom{N}{j} = 2^{o(n)}$. All remaining work per candidate is polynomial, so the worst-case time is

$$T_{\mathrm{BAPR}}^{\mathrm{enum}} \le \binom{n}{k}\,\mathrm{poly}(n,m,a)\sum_{j=0}^{h}\binom{N}{j} = 2^{o(n)}.$$

Connection with concrete parameters. The instances in Table 6 illustrate the parameter scaling in Theorem 3. Across the table, the ratio $a/\log_2 n$ lies between 5.76 and 10.43, placing these majority localities on the logarithmic scale considered in the theorem. For $a = 47, 57, 75$, the asymptotic choice $q = \lceil\sqrt{a \log a}\rceil$ gives $14, 16, 18$, respectively, while the optimized values in the table are 15–16, 16, and 19–21. This agreement indicates that the asymptotic rule captures the selection-threshold scale at these concrete parameters.

5 Further Observations on Alternative Predicates

The preceding sections focus on THR(MAJ)–XOR predicates. We now consider two alternative predicate families proposed in [16].

5.1 An Affine Approximation Attack on $f_{d,\epsilon,a}$

For $0 \le d \le a/2$, $\epsilon \in \{0,1\}$, and $z \in \{0,1\}^a$, let $u = \mathrm{wt}(z)$. The interpolation predicate $f_{d,\epsilon,a}$ is defined by

$$f_{d,\epsilon,a}(z) = \begin{cases} 0, & u < d, \\ u + \epsilon \pmod 2, & d \le u \le a - d, \\ 1, & u > a - d. \end{cases}$$

Thus, in the middle range $d \le u \le a - d$, the predicate is the XOR of its input bits with the constant ϵ, and hence is affine over $\mathbb{F}_2$. This feature reduces common

bias, but it also creates a different weakness. Whenever $d \le \mathrm{wt}(x_{A_i}) \le a - d$, an output satisfies

$$\begin{aligned} y_i &= f_{d,\epsilon,a}(x_{A_i}) \oplus \mathrm{XOR}_b(x_{B_i}) \\ &= \epsilon \oplus \bigoplus_{v \in A_i \cup B_i} x_v. \end{aligned}$$

The second equality follows from $A_i \cap B_i = \varnothing$. Thus every such output gives a noiseless affine equation in the seed bits. The attack below uses the same affine equation for all outputs and succeeds when the chosen equations fall in Hamming-weight regions on which this approximation is correct.

For a fixed seed of weight w, let $p_{\mathrm{aff}}(n, w)$ be the probability that the affine equation above is correct for a random output. It is given by

$$p_{\mathrm{aff}}(n, w) = \sum_{\substack{0 \le u \le a \\ f_{d,\epsilon,a}(u) = u + \epsilon \pmod 2}} \frac{\binom{w}{u}\binom{n-w}{a-u}}{\binom{n}{a}},$$

where $f_{d,\epsilon,a}(u)$ denotes the value of the symmetric predicate on an input of weight u. The sum includes the middle interval and any extreme weights at which the affine expression happens to agree with the constant part of the predicate.

The coefficient vector of an affine equation is the incidence vector of $A_i \cup B_i$, and hence has weight $a + b$. The maximum possible number of independent equations is

$$N^\star = \begin{cases} n, & a + b \text{ odd}, \\ n - 1, & a + b \text{ even}. \end{cases}$$

In the even case, every coefficient row has even weight, so the all-one vector belongs to the kernel. We estimate the attack under the model that a maximum-rank set of rows is available and that its affine equations are correct independently with probability $p_{\mathrm{aff}}(n, w)$. One solve then succeeds with probability $p_{\mathrm{aff}}(n, w)^{N^\star}$. Including the cost of solving and checking a candidate against the available outputs gives

$$T_{\mathrm{aff}}(w) \approx p_{\mathrm{aff}}(n, w)^{-N^\star} \left(m + (N^\star)^\omega\right).$$

When a and b are both odd, the complete predicate is invariant under complementing the seed. In this case $a+b$ is even and the attack recovers a representative of the equivalent pair $\{x, \mathbf{1} \oplus x\}$.

Proposition 1 (Affine-approximation criterion). *Suppose that n is even, the seed is balanced, and the stretch is polynomial. Under the affine recovery model above, the attack has $T_{\mathrm{aff}} = 2^{o(n)}$ whenever*

$$\frac{a/2 - d}{\sqrt{a}} \longrightarrow \infty.$$

In particular, for every constant $c > 0$, it is sufficient that $d \le (1/2 - c)a$ and a tends to infinity. The attack is polynomial under the stronger condition

$$\frac{(a/2 - d)^2}{a} \ge \left(\frac{1}{2} + c\right) \log n.$$

Algorithm 3 Affine recovery for $f_{d,\epsilon,a} \oplus \mathrm{XOR}_b$.

Require: Public sets $(A_i, B_i)_{i=1}^m$, output $y \in \{0,1\}^m$, and a trial bound $R_{\max}$.
Ensure: A seed representative $x \in \{0,1\}^n$, or failure.

```
for j = 1, ..., R_max do
    Select N* rows whose affine coefficient vectors are independent.
    Form ⊕_{v∈A_i∪B_i} x_v = y_i ⊕ ε for each selected row.
    Solve the resulting affine system over F_2.
    Check its candidate solutions against the PRG output.
    if a candidate passes verification then
        return the candidate.
    end if
end for
return failure.
```

n	s	(d, ϵ, a, b)	d/a	$p_{\mathrm{aff}}(n, n/2)$	$\log_2 T_{\mathrm{aff}}$
256	2.0	$(17, 0, 101, 10)$	0.168	$1 - 2^{-65.8}$	19.2
512	4.5	$(37, 0, 127, 23)$	0.291	$1 - 2^{-27.6}$	40.5
894	3.4	$(29, 0, 71, 17)$	0.408	0.9696	73.2
1024	2.0	$(17, 0, 35, 10)$	0.486	0.6995	551.9
1536	3.0	$(25, 0, 51, 15)$	0.490	0.6668	929.3
2000	4.0	$(33, 0, 67, 21)$	0.493	0.6463	1302.7
2048	3.0	$(25, 0, 51, 15)$	0.490	0.6661	1232.9

Table 8. Affine recovery estimate for $f_{d,\epsilon,a} \oplus \mathrm{XOR}_b$ at balanced seed weight. The cost includes candidate verification and uses $\omega = 2.38$.

Proof. For a balanced seed, $U = \mathrm{wt}(x_{A_i})$ is hypergeometric with mean $a/2$. The affine approximation is always correct when $d \le U \le a - d$. The local weight is concentrated around $a/2$, and Hoeffding's inequality for sampling without replacement [21] gives

$$1 - p_{\mathrm{aff}}(n, n/2) \le 2 \exp\left(-\frac{2(a/2 - d)^2}{a} \right).$$

The first condition makes the right-hand side $o(1)$. Since $-\log(1 - z) = O(z)$ as $z \to 0$, the repetition term contributes $o(n)$ to $\log_2 T_{\mathrm{aff}}$, while $m + (N^\star)^\omega$ is polynomial. Under the stronger condition the right-hand side is $O(n^{-1-2c})$, so the expected number of trials is bounded and the total cost is polynomial.

Thus the relevant quantity is not only the ratio d/a, but also the width of the affine interval measured relative to the natural fluctuation $\sqrt{a}$. Values of d close to $a/2$ leave only a narrow affine region and do not satisfy this criterion.

Table 8 applies this estimate to the $f_{d,\epsilon,a}$-type parameters appearing in the challenge table of [16]. All listed values of n are even, so the estimate is evaluated at the balanced weight $w = n/2$, and uses $\omega = 2.38$.

The first three rows have a sufficiently wide affine interval to yield a low-cost recovery attack. In the last four rows, d/a is close to $1/2$, so a random local

input falls in the affine region too infrequently for this attack to be effective. This is separate from common bias. A predicate may have low common bias while remaining vulnerable because it agrees with an affine function on most local inputs.

5.2 Residue Constraints in $\mathbf{MOD3}_a \oplus \mathbf{XOR}_b$

For $t \in \{0,1,2\}$, D'Antona et al. [16] define the residue-class predicate $t\mathrm{MOD3}_a$ on $\{0,1\}^a$ by

$$t\mathrm{MOD3}_a(z) = 1 \quad \Longleftrightarrow \quad \mathrm{wt}(z) \equiv t \pmod 3.$$

They write $\mathrm{MOD3}_a$ for the case $t = 0$, which is the predicate considered here. The corresponding PRG output has the form

$$y_i = \mathrm{MOD3}_a(x_{A_i}) \oplus \mathrm{XOR}_b(x_{B_i}).$$

The security discussion in [16] relates this predicate to alternating-moduli systems. If the public sets A_i are ignored, however, the MOD3 term has an LPN-like interpretation. For a uniformly random local input $z \in \{0,1\}^a$,

$$\Pr[\mathrm{MOD3}_a(z) = 1] = 2^{-a} \sum_{\substack{0 \le u \le a \\ u \equiv 0 \pmod 3}} \binom{a}{u} = \frac{1}{3}\left(1 + 2^{1-a}\cos\frac{a\pi}{3}\right).$$

The last equality follows by evaluating the binomial sum with the cube roots of unity. In particular, this probability differs from $1/3$ by at most $2^{1-a}/3$. Predicting this term as zero therefore produces noisy XOR equations with error probability close to $1/3$. However, this interpretation loses information because the apparent errors are evaluations on public subsets of the same secret seed.

Constraints after a partial guess. Fix $K \subseteq [n]$ with $|K| = k$ and consider the outputs satisfying $B_i \subseteq K$. The number of such outputs has distribution

$$\mathrm{Bin}\left(m, \frac{\binom{k}{b}}{\binom{n}{b}}\right).$$

After enumerating x_K, a correct guess determines the XOR part of every such output and gives the constraint

$$\mathrm{MOD3}_a(x_{A_i}) = y_i \oplus \mathrm{XOR}_b(x_{B_i}).$$

Thus the residual bit is not discarded as noise. It restricts $\mathrm{wt}(x_{A_i})$ modulo 3.

If A_i also intersects K, the guessed values reduce this to a constraint on the remaining positions of A_i. Let $j = |A_i \cap K|$ and $u = \mathrm{wt}(x_{A_i \cap K})$. For a

uniformly random assignment to the remaining seed bits, the probability that $\mathrm{MOD3}_a(x_{A_i}) = 1$ is

$$2^{-(a-j)} \sum_{\substack{0 \le v \le a-j \\ v \equiv -u \pmod 3}} \binom{a-j}{v} = \frac{1}{3} + O\Big(2^{-(a-j)}\Big).$$

Consequently, a residual value of one retains approximately one third of the possible completions, while a residual value of zero retains approximately two thirds. For rows whose sets $A_i \setminus K$ are pairwise disjoint, these retention factors multiply.

Comparison with LPN. In standard LPN with noise probability 1/3, learning an error bit places no restriction on the remaining secret variables. Here the corresponding bit is an explicit residue constraint on the seed. Hence modeling the MOD3 terms as LPN noise discards exploitable information. Turning many overlapping residue constraints into a scalable recovery algorithm remains open. We leave as an open question whether these structured MOD3 constraints can be exploited to obtain a practical recovery attack beyond the LPN abstraction.

6 Conclusion

We studied pooled-selection attacks on Goldreich's PRGs with MAJ–XOR predicates. DPR turns a correct partial seed guess into noiseless sparse equations. We analyzed the size and rank of the selected systems and supported the rank analysis with experiments. For the proposed $\mathrm{MAJ}_7 \oplus \mathrm{XOR}_4$ challenge with $n = 1024$ and $m = n^2$, DPR has an estimated recovery cost of $2^{109.71}$ operations when $\omega = 2.38$, below the claimed 128-bit security level.

Our fixed-weight common-bias analysis motivated BAPR, which uses partial majority bias to form noisy equations. Its concrete estimates show the benefit of pooling at (near-)balanced seed weights. For $m = n^s$ with fixed $s > 1$, odd majority locality $a \in \omega(1) \cap O(\log n)$, and fixed XOR locality $b \ge 3$, BAPR also admits a worst-case $2^{o(n)}$-time implementation. For every fixed seed, it returns a seed consistent with the output with probability $1 - o(1)$ over the sampled pairs (A_i, B_i) and the attack's guesses, without relying on the noisy-solver cost models used for the concrete estimates.

We also identified exploitable affine structure and residue constraints in proposed alternative predicates. Efficiently solving the overlapping residue constraints remains open. Overall, our results show that security depends not only on the predicate itself, but also on how its output equations interact after partial seed guessing.

References

1. Agarwal, A., Baum, C., Braun, L., Scholl, P.: Low-bandwidth mixed arithmetic in VOLE-based ZK from low-degree PRGs. In: Fehr, S., Fouque, P. (eds.) Advances

in Cryptology - EUROCRYPT 2025 - 44th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Madrid, Spain, May 4-8, 2025, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 15604, pp. 396–426. Springer (2025). https://doi.org/10.1007/978-3-031-91134-7_14

2. Applebaum, B.: Pseudorandom generators with long stretch and low locality from random local one-way functions. In: Proceedings of the 44th Annual ACM Symposium on Theory of Computing. pp. 805–816. ACM (2012). https://doi.org/10.1145/2213977.2214050
3. Applebaum, B., Bogdanov, A., Rosen, A.: A dichotomy for local small-bias generators. In: Cramer, R. (ed.) Theory of Cryptography - 9th Theory of Cryptography Conference, TCC 2012, Taormina, Sicily, Italy, March 19-21, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7194, pp. 600–617. Springer (2012). https://doi.org/10.1007/978-3-642-28914-9_34
4. Applebaum, B., Lovett, S.: Algebraic attacks against random local functions and their countermeasures. SIAM J. Comput. **47**(1), 52–79 (2018). https://doi.org/10.1137/16M1085942
5. Applebaum, B., Raykov, P.: Fast pseudorandom functions based on expander graphs. In: Hirt, M., Smith, A.D. (eds.) Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I. Lecture Notes in Computer Science, vol. 9985, pp. 27–56 (2016). https://doi.org/10.1007/978-3-662-53641-4_2
6. Bogdanov, A., Qiao, Y.: On the security of Goldreich's one-way function. Comput. Complex. **21**(1), 83–127 (2012). https://doi.org/10.1007/S00037-011-0034-0
7. Boura, C., Couteau, G., Perrin, L., Rotella, Y.: SoK: On shallow weak PRFs A common symmetric building block for MPC protocols. IACR Trans. Symmetric Cryptol. **2025**(3), 289–336 (2025). https://doi.org/10.46586/TOSC.V2025.I3.289-336
8. Bui, D., Couteau, G., Meyer, P., Passelègue, A., Riahinia, M.: Fast public-key silent OT and more from constrained naor-reingold. In: Joye, M., Leander, G. (eds.) Advances in Cryptology - EUROCRYPT 2024 - 43rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zurich, Switzerland, May 26-30, 2024, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 14656, pp. 88–118. Springer (2024). https://doi.org/10.1007/978-3-031-58751-1_4
9. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. The Annals of Mathematical Statistics **23**(4), 493–507 (1952). https://doi.org/10.1214/aoms/1177729330
10. Chvátal, V.: The tail of the hypergeometric distribution. Discrete Mathematics **25**(3), 285–287 (1979). https://doi.org/10.1016/0012-365X(79)90084-0
11. Coja-Oghlan, A., Gao, P., Hahn-Klimroth, M., Lee, J., Müller, N., Rolvien, M.: The full rank condition for sparse random matrices. Combinatorics, Probability and Computing **33**(5), 643–707 (2024). https://doi.org/10.1017/S096354832400021X
12. Cooper, C., Frieze, A.M., Pegden, W.: On the rank of a random binary matrix. Electron. J. Comb. **26**(4), 4 (2019). https://doi.org/10.37236/8092
13. Couteau, G., Devadas, L., Devadas, S., Koch, A., Servan-Schreiber, S.: Quietot: Lightweight oblivious transfer with a public-key setup. In: Chung, K., Sasaki, Y. (eds.) Advances in Cryptology - ASIACRYPT 2024 - 30th International Conference on the Theory and Application of Cryptology and Information Security, Kolkata, India, December 9-13, 2024, Proceedings, Part II. Lecture Notes in Computer Science, vol. 15485, pp. 197–231. Springer (2024). https://doi.org/10.1007/978-981-96-0888-1_7

14. Couteau, G., Dupin, A., Méaux, P., Rossi, M., Rotella, Y.: On the concrete security of Goldreich's pseudorandom generator. In: Peyrin, T., Galbraith, S.D. (eds.) Advances in Cryptology - ASIACRYPT 2018 - 24th International Conference on the Theory and Application of Cryptology and Information Security, Brisbane, QLD, Australia, December 2-6, 2018, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11273, pp. 96–124. Springer (2018). https://doi.org/10.1007/978-3-030-03329-3_4
15. Cryan, M., Miltersen, P.B.: On pseudorandom generators in NC^0. In: Mathematical Foundations of Computer Science 2001. Lecture Notes in Computer Science, vol. 2136, pp. 272–284. Springer (2001). https://doi.org/10.1007/3-540-44683-4_24
16. D'Antona, A.G., Méaux, P., Ünal, A.: On the common bias of majorities: Polytime attacks on THR-XOR PRGs. IACR Cryptol. ePrint Arch. **2026**, 880 (2026), https://eprint.iacr.org/2026/880
17. Darling, R.W.R., Penrose, M.D., Wade, A.R., Zabell, S.L.: Rank deficiency in sparse random GF(2) matrices. Electronic Journal of Probability **19**(83), 1–36 (2014). https://doi.org/10.1214/EJP.v19-2458
18. Dupin, A., Méaux, P., Rossi, M.: On the algebraic immunity–resiliency trade-off, implications for Goldreich's pseudorandom generator. Designs, Codes and Cryptography **91**(9), 3035–3079 (2023). https://doi.org/10.1007/s10623-023-01220-w
19. Fu, X., Li, M., Lyu, S., Liu, C.: Attacks on Goldreich's pseudorandom generators by grouping and solving. In: Daemen, J., Thomé, E. (eds.) Advances in Cryptology - EUROCRYPT 2026 - 45th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Rome, Italy, May 10-14, 2026, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 16546, pp. 3–32. Springer (2026). https://doi.org/10.1007/978-3-032-25333-0_1
20. Goldreich, O.: Candidate one-way functions based on expander graphs. Electronic Colloquium on Computational Complexity (ECCC) **7**(90) (2000), https://eccc.weizmann.ac.il/eccc-reports/2000/TR00-090/index.html
21. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58**(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830
22. Méaux, P.: On the fast algebraic immunity of threshold functions. Cryptography and Communications **13**(5), 741–762 (2021). https://doi.org/10.1007/s12095-021-00505-y
23. Méaux, P.: On the algebraic immunity of direct sum constructions. Discrete Applied Mathematics **320**, 223–234 (2022). https://doi.org/10.1016/j.dam.2022.05.021
24. Méaux, P., Carlet, C., Journault, A., Standaert, F.: Improved filter permutators for efficient FHE: better instances and implementations. In: Hao, F., Ruj, S., Gupta, S.S. (eds.) Progress in Cryptology - INDOCRYPT 2019 - 20th International Conference on Cryptology in India, Hyderabad, India, December 15-18, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11898, pp. 68–91. Springer (2019). https://doi.org/10.1007/978-3-030-35423-7_4
25. Méaux, P., Journault, A., Standaert, F., Carlet, C.: Towards stream ciphers for efficient FHE with low-noise ciphertexts. In: Fischlin, M., Coron, J. (eds.) Advances in Cryptology - EUROCRYPT 2016 - 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part I. Lecture Notes in Computer Science, vol. 9665, pp. 311–343. Springer (2016). https://doi.org/10.1007/978-3-662-49890-3_13

26. Mossel, E., Shpilka, A., Trevisan, L.: On ϵ-biased generators in NC^0. Random Structures & Algorithms **29**(1), 56–81 (2006). https://doi.org/10.1002/rsa.20112
27. Ünal, A.: Worst-case subexponential attacks on prgs of constant degree or constant locality. In: Hazay, C., Stam, M. (eds.) Advances in Cryptology - EUROCRYPT 2023 - 42nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Lyon, France, April 23-27, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 14004, pp. 25–54. Springer (2023). https://doi.org/10.1007/978-3-031-30545-0_2
28. Yang, J., Guo, Q., Johansson, T., Lentmaier, M.: Revisiting the concrete security of Goldreich's pseudorandom generator. IEEE Trans. Inf. Theory **68**(2), 1329–1354 (2022). https://doi.org/10.1109/TIT.2021.3128315

A Bounding Incorrect-Guess Costs in DPR

This appendix bounds the work caused by incorrect guesses through their average candidate count. We relate this count to recovery cost and success, bound it, and evaluate the result at concrete parameters.

A.1 Recovery with Bounded Work

Fix a seed $x \in \{0,1\}^n$ of weight w, and assume $1 \le k \le w$, $k < n$, $1 \le t \le q \le \min(a,k)$, and $m \ge N = n - k$. Let $\Gamma = ((A_i, B_i))_{i=1}^m$ be sampled as in Section 2; these sets are public. In particular, B_i is uniform among the b-subsets disjoint from A_i. All probabilities below include this random choice of Γ, while x remains fixed. Successive guesses use the same Γ. Write $p = p_{\mathrm{sel}}$ for (2).

For a uniformly sampled k-subset K, independent of Γ, let $C_0(\Gamma, K)$ count seeds z with $z_K = \mathbf{1}_k$ that satisfy the selected equations when $x_K \neq \mathbf{1}_k$ and $e \ge N$; otherwise set $C_0 = 0$. The expected number of false candidates per trial, averaged over Γ and K, is

$$\mathcal{M} = \mathbb{E}_{\Gamma,K}[C_0(\Gamma, K)].$$

Write F for an upper bound on $\mathcal{M}$. Row and rank cutoffs can only reduce C_0, so the same bound applies after these discards. It gives the following recovery guarantee.

Proposition 2 (DPR with bounded search and verification). *Let $\pi = 1/S_1(n,w,k)$. For a rank-deficiency cutoff $\delta \ge 0$ and an upper row limit $N \le e_{\max} \le m$, write*

$$p_{\mathrm{use}}(\delta; e_{\max}) = \Pr[N \le e \le e_{\max},\ d_K \le \delta \mid x_K = \mathbf{1}_k].$$

Taking $e_{\max} = m$ recovers $p_{\mathrm{use}}(\delta)$ from Section 3.

Run at most $R \ge 1$ independent uniform guesses K, discard systems outside the row and rank limits, and allow at most $R + 2^\delta$ candidate checks in total. The probability of returning a seed passing verification is at least

$$p_{\mathrm{use}}(\delta; e_{\max})\left[1 - (1-\pi)^R\right] - F.$$

If $\tau(e_{\max}) = C_{\mathrm{sel}} + C_{\mathrm{lin}}$ bounds the per-trial work before candidate enumeration, including equation construction and elimination, and V bounds the work to generate and verify a candidate, the total work is bounded by

$$O\big(R\,\tau(e_{\max}) + (R + 2^{\delta})V\big)\,.$$

In particular, $R = \lceil S_1(n, w, k) \rceil$ gives success probability at least $p_{\mathrm{use}}(\delta; e_{\max})(1 - \exp(-1)) - F$.

Proof. Draw all R guesses in advance, independently of Γ. Let X count all candidates in retained incorrect trials, before applying the global verification budget or early stopping. Correct and discarded trials contribute zero. Linearity of expectation gives the first inequality below, and Markov's inequality gives the second:

$$\mathbb{E}[X] \le RF, \qquad \Pr[X > R] \le \frac{\mathbb{E}[X]}{R} \le F.$$

These bounds do not require independence between trials sharing Γ.

Let J be the first correct guess, if it occurs within the trial budget. Its existence has probability $1-(1-\pi)^R$, independently of Γ. Conditional on existence, K_J is uniform among correct sets and independent of Γ. Thus the probability that this first correct guess is usable equals

$$p_{\mathrm{use}}(\delta; e_{\max})\big[1 - (1-\pi)^R\big].$$

On this event and $X \le R$, preceding trials consume at most R candidate checks. The usable correct system contains the true seed and has at most 2^{δ} solutions. A seed therefore passes verification within the global budget, unless one has already been returned. Subtracting $\Pr[X > R]$ proves the success bound. The work bound follows from the two explicit budgets. The usable-guess event and the budget-overflow event need not be independent. Finally, $R\pi \ge 1$ when $R = \lceil S_1(n, w, k) \rceil$, so $(1-\pi)^R \le \exp(-1)$.

A.2 Bounding the Average Candidate Count

We group candidates by their Hamming distance h from the fixed seed. The next lemma combines their average count A_h, a one-row satisfaction bound ϑ_h, and the random number of selected rows through G.

Lemma 3 (An incorrect-guess candidate bound). *Put $M = n - a$, and define*

$$\mathcal{K}_b(M, d) = \sum_{j=0}^{b} (-1)^j \binom{d}{j} \binom{M-d}{b-j}.$$

Among the b-subsets of the M positions available to B_i, $\mathcal{K}_b(M, d)$ equals the number having an even-sized intersection with a given d-subset minus the number

having an odd-sized intersection with that subset. For $1 \le h \le n$*, let*

$$A_h = \frac{1}{\binom{n}{k}} \sum_{f=1}^{k} \binom{n-w}{f} \binom{w}{k-f} \binom{N}{h-f},$$

$$\vartheta_h = \frac{1}{2} \left(1 + \frac{\max\limits_{\max(0,h-M) \le c \le \min(a,h)} |\mathcal{K}_b(M, h-c)|}{\binom{M}{b}} \right),$$

where the maximum is over integers. For $0 \le \theta \le 1$*, put*

$$G(\theta) = \sum_{e=N}^{m} \binom{m}{e} p^e (1-p)^{m-e} \theta^e.$$

Then

$$\mathcal{M} \le \sum_{h=1}^{n} A_h G(\vartheta_h).$$

Proof. Fix K and a candidate z with $z_K = \mathbf{1}_k$. Write $D = \{v : x_v \neq z_v\}$ and $h = |D|$. A selected equation is satisfied precisely when

$$\bigoplus_{v \in B} (x_v \oplus z_v) = 1 \oplus \mathrm{THR}_{a,t}(x_A).$$

Conditional on A, put $c = |A \cap D|$. Among the M positions available to B, exactly $h - c$ belong to D. The difference between the probabilities of even and odd intersection with D is $\mathcal{K}_b(M, h-c)/\binom{M}{b}$. For either required parity, the equation therefore holds with probability at most

$$\frac{1}{2} \left(1 + \frac{|\mathcal{K}_b(M, h-c)|}{\binom{M}{b}} \right) \le \vartheta_h.$$

This inequality holds for every A, so it remains valid after conditioning on $|A \cap K| \ge q$. Conditional on their number e, the selected rows are independent with the selected-row distribution. Consequently, all selected equations hold with probability at most ϑ_h^e. Averaging over the binomial row count and discarding $e < N$ gives $G(\vartheta_h)$.

If K contains $f \ge 1$ zero positions of x, the guess is incorrect and every candidate differs from x in exactly these f guessed positions. There are $\binom{N}{h-f}$ such candidates at total distance h, and $\binom{n-w}{f}\binom{w}{k-f}$ sets K of this type. Summing candidate indicators and averaging over K proves the bound. The threshold value and XOR support retain their original dependence through A; no distributional assumption on error syndromes is needed.

Certifying the finite sum. For $0 < p < 1$, dropping the restriction $e \ge N$ and applying the binomial theorem gives

$$G(\theta) \le (1 - p + p\theta)^m.$$

If

$$r(\theta) = \frac{m-N}{N+1}\frac{p\theta}{1-p} < 1,$$

the ratios of successive summands in $G(\theta)$ are bounded by $r(\theta)$, and hence

$$G(\theta) \le \frac{\binom{m}{N}p^N(1-p)^{m-N}\theta^N}{1-r(\theta)}.$$

Taking the smaller applicable bound yields a sum of nonnegative rational expressions and integer powers. Directed rounding certifies its value without evaluating binomial tail functions. If $p = 0$, then $G = 0$; if $p = 1$, then $G(\theta) = \theta^m$.

A refined candidate count. The distance-only bound in Lemma 3 can be loose because it maximizes over the threshold input set A. A more detailed count retains this information. For an incorrect guess containing f true zero positions, put

$$\pi_f = \frac{\binom{n-w}{f}\binom{w}{k-f}}{\binom{n}{k}}, \qquad W_f = w - k + f, \qquad Z_f = n - w - f.$$

Here π_f is the probability of this type of guess, while W_f and Z_f count the true ones and zeros outside K. Classify a candidate by the number u of these ones it changes to zero and the number v of these zeros it changes to one. For a fixed K and candidate z of this type, define the selected-row probability

$$\theta_{f,u,v} = \Pr\left[\bigoplus_{j\in B}(x_j \oplus z_j) = 1 \oplus \mathrm{THR}_{a,t}(x_A) \;\middle|\; |A \cap K| \ge q\right].$$

Symmetry makes this probability depend only on (f, u, v). It is an exact rational number, obtained by counting the possible A and then the disjoint B, keeping the threshold value and required XOR parity together. Independence across output rows gives the identity

$$\mathcal{M} = \sum_{f=1}^{\min(k,n-w)} \pi_f \sum_{u=0}^{W_f}\sum_{v=0}^{Z_f}\binom{W_f}{u}\binom{Z_f}{v}G(\theta_{f,u,v}).$$

A.3 Numerical Applications

The challenge parameters. Consider $n = 1024$, $w = 512$, $a = 7$, $b = 4$, $q = t = 4$, $k = 80$, and $m = 1024^2$. Exact binomial coefficients and 70-decimal-digit outward rounding certify the bound in Lemma 3 as

$$\mathcal{M} < 5.217 \cdot 10^{-5}.$$

For the upper row limit $e_{\max} = 2N = 1888$, which bounds per-trial work, Markov's inequality applied to $(7/4)^e$ gives

$$\eta := \Pr[e > 2N] \le \left(1 + \frac{3}{4}p\right)^m \left(\frac{4}{7}\right)^{2N} < 2^{-370}.$$

With $\delta = 39$, the row and rank bounds in Section 3.4 give

$$p_{\text{use}}(39; 2N) \geq (1 - 1.31 \cdot 10^{-4})(1 - 1.09 \cdot 10^{-4}) - \eta > 0.99975.$$

Set $R = \lceil S_1(1024, 512, 80) \rceil$. Proposition 2 then gives success probability above 0.6319 with $F = 5.217 \cdot 10^{-5}$. Charging $C_{\text{sel}} = ma$, $C_{\text{lin}} = N^{\omega}$, and $V = m(a+b)$, the estimated work is $2^{109.71}$ at $\omega = 2.38$. Increasing the trial budget to $R = \lceil 10 S_1(1024, 512, 80) \rceil$ gives success probability above 0.99965 at an estimated cost of about $2^{113.04}$, still below the 128-bit target.

Other parameter sets. Numerically evaluating the refined identity for the four non-challenge DPR rows of Table 1, in their order of appearance, gives approximately

$$0.036614, \qquad 0.035871, \qquad 7.0641 \cdot 10^{-4}, \qquad 7.1 \cdot 10^{-24}.$$

These are numerical evaluations of the exact identity, not directed-rounding certificates. The DPR cost estimates in the table can be obtained with an explicit work budget: take $R = \lceil S_1 / p_{\text{use}} \rceil$, cap the total number of candidate checks at $R + 2^{\delta}$, and use $e_{\max} = 2N$. The displayed p_{use} bounds omit the negligible loss from this upper row cutoff. Together with Proposition 2, these evaluations yield constant-success estimates under the stated budgets.