Reusable Fuzzy Extractors via Digital Lockers

Ran Canetti∗ Benjamin Fuller† Omer Paneth‡ Leonid Reyzin§ Adam Smith¶

October 29, 2015

Abstract

Fuzzy extractors (Dodis et al., Eurocrypt 2004) convert repeated noisy readings of a secret into the same uniformly distributed key. To eliminate noise, they require an initial enrollment phase that takes the first noisy reading of the secret and produces a nonsecret helper string to be used in subsequent readings. Reusable fuzzy extractors (Boyen, CCS 2004) remain secure even when this initial enrollment phase is repeated multiple times with noisy versions of the same secret, producing multiple helper strings (for example, when a single person’s biometric is enrolled with multiple unrelated organizations).

We construct the first reusable fuzzy extractor that makes no assumptions about how multiple readings of the source are correlated (the only prior construction assumed a very specific, unrealistic class of correlations). The extractor works for binary strings with Hamming noise; it achieves computational security under assumptions on the security of hash functions or in the random oracle model. It is simple and efficient and tolerates near-linear error rates.

Our reusable extractor is secure for source distributions of linear min-entropy rate. The construction is also secure for sources with much lower entropy rates—lower than those supported by prior (nonreusable) constructions—assuming that the distribution has some additional structure, namely, that random subsequences of the source have sufficient minentropy. We show that such structural assumptions are necessary to support low entropy rates.

We then explore further how different structural properties of a noisy source can be used to construct fuzzy extractors when the error rates are high, providing a computationally secure and an information-theoretically secure construction for large-alphabet sources.

Keywords Fuzzy extractors, reusability, key derivation, error-correcting codes, computational entropy, digital lockers, point obfuscation.

1 Introduction

Fuzzy Extractors Cryptography relies on long-term secrets for key derivation and authentication. However, many sources with sufficient randomness to form long-term secrets provide similar but not identical values of the secret at repeated readings. Prominent examples include biometrics and other human-generated data [Dau04, ZH93, BS00, EHMS00, MG09, MRW02], physically unclonable functions (PUFs) [PRTG02], TSS∗06, [GCVDD02, SD07], and quantum information [BBR88]. Turning similar

∗ Email: canetti@cs.bu.edu. Boston University and Tel Aviv University.
† Email: bfuller@cs.bu.edu. Boston University and MIT Lincoln Laboratory.
‡ Email: paneth@cs.bu.edu. Boston University.
§ Email: reyzin@cs.bu.edu. Boston University.
¶ Email: asmith@cse.psu.edu. Pennsylvania State University; work performed while at Boston University’s Hariri Institute for Computing and RISCS Center, and Harvard University’s “Privacy Tools” project.
readings into identical values is known as information reconciliation; further converting those values into uniformly random secret strings is known as privacy amplification \cite{BBR88}. Both of these problems have interactive and non-interactive versions. In this paper, we are interested in the non-interactive case, which is useful for a single user trying to produce the same key from multiple noisy readings of a secret at different times. A fuzzy extractor is the primitive that accomplishes both information reconciliation and privacy amplification non-interactively; fuzzy extractors are defined in \cite{DORS08}.

Fuzzy extractors consist of a pair of algorithms: \texttt{Gen} (used once, at “enrollment”) takes a source value \( w \), and produces a key \( r \) and a public helper value \( p \). The second algorithm \texttt{Rep} (used subsequently) takes this helper value \( p \) and a close \( w' \) to reproduce the original key \( r \). The standard correctness guarantee is that \( r \) will be correctly reproduced by \texttt{Rep} as long as \( w' \) is no farther than \( t \) from \( w \) according to some notion of distance (specifically, we work with Hamming distance; our primary focus is on binary strings, although we also consider larger alphabets). The security guarantee is that \( r \) produced by \texttt{Gen} is indistinguishable from uniform, even given \( p \). In this work, we consider computational indistinguishability \cite{FMR13} rather than the more traditional information-theoretic notion. (Note that so-called “robust” fuzzy extractors \cite{Mau97, MW97, BDK+05, CDF+08, DKK+12} additionally protect against active attackers who modify \( p \); we do not consider them here, except to point out that our constructions can be easily made robust by the random-oracle-based transform of \cite[Theorem 1]{BDK+05}.)

**Reusability** A fuzzy extractor is reusable \cite{Boy04} if it remains secure even when a user enrolls the same or correlated values multiple times. For example, if the source is a biometric reading, the user may enroll the same biometric with different noncooperating organizations. Reusability is particularly important for biometrics which, unlike passwords, cannot be changed or created. It is also useful in other contexts, for example, to permit a user to reuse the same visual password across many services or to make a single physical token (embodying a PUF) usable for many applications.

Each enrollment process will get a slightly different enrollment reading \( w_i \), and will run \texttt{Gen}(\( w_i \)) to get a key \( r_i \) and a helper value \( p_i \). Security for each \( r_i \) should hold even when an adversary is given all the values \( p_1, \ldots, p_p \) and even the keys \( r_j \) for \( j \neq i \) (because one organization cannot be sure how other organizations will use the derived keys).

Most constructions of fuzzy extractors are not reusable \cite{Boy04, STP09, BA12, BA13}. In fact, the only known construction of reusable fuzzy extractors \cite{Boy04} requires very particular relationships between \( w_i \) values\footnote{Specifically, Boyen’s construction requires that the exclusive or \( w_i \oplus w_j \) of any two secrets not leak any information about \( w_i \).} which are unlikely to hold for a practical source.

### 1.1 Our Contribution

**A Reusable Fuzzy Extractor** We construct the first reusable fuzzy extractor whose security holds even if the multiple readings \( w_i \) used in \texttt{Gen} are arbitrarily correlated, as long as the fuzzy extractor is secure for each \( w_i \) individually. This construction is the first to provide reusability for a realistic class of correlated readings. Our construction is based on digital lockers; in the most efficient instantiation, it requires only evaluation of cryptographic hash functions and is secure in the random oracle model or under strong computational assumptions on the hash functions\footnote{The term “digital lockers” was introduced by Canetti and Dakdouk \cite{CD08}; the fact that such digital lockers can be built easily out cryptographic hash functions was shown by \cite[Section 4]{LPS04}.}

Our construction is secure if the bits of \( w \) are partially independent. Namely, we require that, for some known parameter \( k \), the substring formed by the bits at \( k \) randomly chosen positions in \( w \) is unguessable.
(i.e., has minentropy that is superlogarithmic is the security parameter). We call sources with this feature “sources with high-entropy samples.” This requirement is in contrast to most constructions of fuzzy extractors that require $w$ to have sufficient minentropy.

All sources of with sufficient minentropy have high-entropy samples (because sampling preserves the entropy rate [Vad03]). However, sources with high-entropy samples also include a large family of low entropy sources. They are easy to construct artificially: for example, we can build a source of length $n$ whose bits are $k$-wise independent by multiplying (over GF(2)) a fixed $n \times k$ matrix of rank $k$ by a random $k$-bit vector. They also arise naturally whenever $w$ exhibits a lot of redundancy. For example, when the binary string $w$ is obtained via signal processing from some underlying reading (such as an image of an iris or an audio recording of a voice), the signal itself is likely to have a lot of redundancy (for example, nearby pixels of an image are highly correlated). By requiring only high-entropy samples rather than a high entropy rate, we free the signal processing designer from the need to remove redundancy when converting the underlying reading to a string $w$ used in the fuzzy extractor. Thus, we enable the use of oversampled signals.

Our construction can tolerate $\frac{n \ln n}{k}$ errors (out of the $n$ bits of $w$) if we allow the running time of the construction to be linear in $n$ (more generally, $c^{\frac{n \ln n}{k}}$ errors for running time linear in $n^c$). Note that, since in principle $k$ needs to be only slightly superlogarithmic to ensure the high-entropy condition on the samples, our allowable error rate is only slightly sublinear.

The Advantage of Exploiting the Structure of the Distribution Following the tradition of extractor literature [CG88, NZ93], much work on fuzzy extractors has focused on providing constructions that work for any source of a given minentropy $m$. In contrast, our construction exploits more about the structure of the distribution than just its entropy. As a result, it supports not only all sources of a given (sufficiently high) minentropy, but also many sources with an entropy rate much lower than the error rate. We know of no prior constructions with this property. We now explain why, in order to achieve this property, exploiting the structure of the distribution is necessary.

A fuzzy extractor that supports $t$ errors out of a string of $n$ bits and works for all sources of minentropy $m$ must have the entropy rate $\frac{m}{n}$ at least as big as the binary entropy of the error rate, $h_2(\frac{t}{n})$ (to be exact, $m \geq nh_2(\frac{t}{n}) - \frac{1}{2} \log n - \frac{1}{2}$). The reason for this requirement is simple: if $m$ too small, then a single ball of radius $t$, which contains at least $2^{nh_2(\frac{t}{n})-\frac{1}{2} \log n-\frac{1}{2}}$ points [Ash65, Lemma 4.7.2, Equation 4.7.5, p. 115], may contain the entire distribution of $2^n$ points inside it. For this distribution, an adversary can run Rep on the center of this ball and always learn the key $r$. This argument leads to the following proposition, which holds regardless of whether the fuzzy extractor is information-theoretic or computational, and extends even to the interactive setting.

**Proposition 1.1.** If the security guarantee of a fuzzy extractor holds for any source of minentropy $m$ and the correctness guarantees holds for any $t$ errors and $m < \log |B_t|$ (where $|B_t|$ denotes the number of points in a ball of radius $t$), the fuzzy extractor must provide no security. In particular, for the binary Hamming case, $m$ must exceed $nh_2(\frac{t}{n})-\frac{1}{2} \log n-\frac{1}{2} \approx nh_2(\frac{t}{n}) > t \log_2 \frac{n}{t}$.

Thus, in order to correct $t$ errors regardless of the structure of the distribution, we would have to assume a higher total minentropy $m$ than would be necessary if we could take advantage of specific properties of the distribution. In particular, the $m$ vs. $t$ tradeoff obtained in our construction is simply impossible to

\footnote{Binary entropy $h_2(\alpha)$ for $0 < \alpha < 1$ is defined as $-\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha)$; it is greater than $\alpha \log_2 \frac{1}{\alpha}$ and, in particular, greater than $\alpha$ for interesting range $\alpha < \frac{1}{2}$.}
obtain via a construction that works for a distribution of minentropy $m$, because in our case $m$ can be much less than $t < nh_2(\frac{1}{n})$.

Beating the bound of Proposition 1.1 is important. For example, the IrisCode [Dau04], which is the state of the art approach to handling what is believed to be the best biometric [PPJ03], produces a source where $m$ is less than $nh_2(\frac{1}{n})$ [BH09 Section 5]. PUFs with slightly nonuniform outputs suffer from similar problems [KLRW14].

We emphasize that in applications of fuzzy extractors to physical sources, any constraint on the source—whether minentropy-based or more structured—is always, by necessity, an assumption about the physical world. It is no more possible to verify that a source has high minentropy than it is to verify that it has high-entropy samples. Both statements about the source can be derived only by modeling the source—for example, by modeling the physical processes that generate irises or PUFs.

While some prior work on key agreement from noisy data also made assumptions on the structure of the source (often assuming that it consists of independent identically distributed symbols, e.g. [Mau93, MW96, MTV09, YD10, HNSS12]), we are not aware of any work that beat the bound of Proposition 1.1 with the exception of the work by Holenstein and Renner [HR05 Theorem 4]. Their construction supports a uniform length $n$ binary $w$, with a random selection of $(n - m)$ bits leaked to the adversary and $t$ random bits flipped in $w'$. They show that it is possible to support any $m > 4t(1 - \frac{1}{n})$, which is lower than $\log |B_t| \approx nh_2(\frac{1}{n})$, but still higher than $t$.

Constructions Exploiting the Structure of the Distribution for Larger Alphabets

In addition to the binary alphabet construction that supports reuse and low entropy rates, as discussed above, we explore how low entropy rates can be supported when symbols of the string $w$ comes from a large, rather than a binary, alphabet. We obtain two additional constructions, both of which allow for distributions whose total minentropy is lower than the volume of the ball of radius $t$ (in the large-alphabet Hamming space). Unfortunately, neither of them provides reusability, but both can tolerate a linear error rate (of course, over the larger alphabet, where errors may be more likely, because each symbol carries more information).

Our second construction for large alphabets works for sources with sparse high-entropy marginals: sources for which sufficiently many symbols have high entropy individually, but no independence among symbols is assumed (thus, the total entropy may be as low as the entropy of a single symbol).

Our third construction for large alphabets provides information-theoretic, rather than computational, security. It works for sparse block sources. These are sources in which a sufficient fraction of the symbols have entropy conditioned on previous symbols.

Both constructions should be viewed as evidence that assumptions on the source other than total minentropy may provide new opportunities for increasing the error tolerance of fuzzy extractors.

Our Approach

Our approach in all three constructions is different from most known constructions of fuzzy extractors, which put sufficient information in $p$ to recover the original $w$ from a nearby $w'$ during Rep (this procedure is called a secure sketch). We deliberately do not recover $w$, because known techniques for building secure sketches do not work for sources whose entropy rate is lower than its error rate. (This is because they lose at least $\log |B_t|$ bits of entropy regardless of the source. This loss is necessary when the source is uniform [DORS08 Lemma C.1] or when reusability against a sufficiently

4However, standard heuristics for estimating entropy can also be used to indicate whether a source has high-entropy samples. For a corpus of noisy signals, repeat the following a statistically significant number of times: 1) sample $k$ indices 2) run the heuristic entropy test on the corpus which each sample restricted to the $k$ indices.
rich class of correlations is desired \cite[Theorem 11]{Boy04}; computational definitions of secure sketches suffer from similar problems \cite[Corollary 3.8]{FMRT13}.) Instead, in the computational constructions, we lock up a freshly generated random $r$ using parts of $w$ in an error-tolerant way; in the information-theoretic construction, we reduce the alphabet in order to reduce the ball volume while maintaining entropy.

We note that the idea of locking up a random $r$ has appeared in a prior theoretical construction of a computational fuzzy extractor for any source. Namely, Bitansky et al. \cite{BCKP14} show how to obfuscate a proximity point program that tests if an input $w'$ is within distance $t$ of the value $w$ hidden inside the obfuscated program and, if so, outputs the secret $r$ (such a program would be output by $\text{Gen}$ as $p$ and run by $\text{Rep}$). However, such a construction is based on very strong assumptions (semantically secure graded encodings \cite{PS13}) and, in contrast to our construction, is highly impractical in terms of efficiency. Moreover, it is not known to provide reusability, because known obfuscation of proximity point programs is not known to be composable.

2 Definitions

For a random variables $X_i$ over some alphabet $Z$ we denote by $X = X_1, ..., X_n$ the tuple $(X_1, \ldots, X_n)$. For a set of indices $J$, $X_J$ is the restriction of $X$ to the indices in $J$. The set $J^c$ is the complement of $J$. The minentropy of $X$ is $H_\infty(X) = -\log(\max_x \Pr[X = x])$, and the average (conditional) minentropy of $X$ given $Y$ is $H_\infty(X|Y) = -\log(\mathbb{E}_{y \in Y} \max_x \Pr[X = x|Y = y])$ \cite[Section 2.4]{DORS08}. For a random variable $W$, let $H_0(W)$ be the logarithm of the size of the support of $W$, that is $H_0(W) = \log |\{w| \Pr[W = w] > 0\}|$. The statistical distance between random variables $X$ and $Y$ with the same domain is $\Delta(X, Y) = \frac{1}{2} \sum_x |\Pr[X = x] - \Pr[Y = x]|$. For a distinguisher $D$ we write the computational distance between $X$ and $Y$ as $\delta^D(X, Y) = |\mathbb{E}[D(X)] - \mathbb{E}[D(Y)]|$ (we extend it to a class of distinguishers $D$ by taking the maximum over all distinguishers $D \in D$). We denote by $D_s$ the class of randomized circuits which output a single bit and have size at most $s$.

For a metric space $(M, \text{dis})$, the (closed) ball of radius $t$ around $x$ is the set of all points within radius $t$, that is, $B_t(x) = \{y|\text{dis}(x, y) \leq t\}$. If the size of a ball in a metric space does not depend on $x$, we denote by $|B_t|$ the size of a ball of radius $t$. We consider the Hamming metric over vectors in $\mathbb{Z}^n$, defined via $\text{dis}(x, y) = |\{i| x_i \neq y_i\}|$. For this metric, $|B_t| = \sum_{i=0}^t \binom{n}{i} (|Z| - 1)^i$. $U_n$ denotes the uniformly distributed random variable on $\{0,1\}^n$. Unless otherwise noted logarithms are base 2. Usually, we use capitalized letters for random variables and corresponding lowercase letters for their samples.

2.1 Fuzzy Extractors

In this section we define computational fuzzy extractors. Definitions for information-theoretic fuzzy extractors can be found in the work of Dodis et al. \cite[Sections 2.5–4.1]{DORS08}. The definition of computational fuzzy extractors allows for a small probability of error.

Definition 2.1. \cite[Definition 2.5]{EMRT13} Let $W$ be a family of probability distributions over $\mathcal{M}$. A pair of randomized procedures “generate” ($\text{Gen}$) and “reproduce” ($\text{Rep}$) is an $(\mathcal{M}, W, \kappa, t)$-computational fuzzy extractor that is $(\varepsilon_{\sec}, \delta_{\sec})$-hard with error $\delta$ if $\text{Gen}$ and $\text{Rep}$ satisfy the following properties:

- The generate procedure $\text{Gen}$ on input $w \in \mathcal{M}$ outputs an extracted string $r \in \{0,1\}^\kappa$ and a helper string $p \in \{0,1\}^*$.
- The reproduction procedure $\text{Rep}$ takes an element $w' \in \mathcal{M}$ and a bit string $p \in \{0,1\}^*$ as inputs. The correctness property guarantees that if $\text{dis}(w, w') \leq t$ and $(r, p) \leftarrow \text{Gen}(w)$, then $\Pr[\text{Rep}(w', p) = r] \geq 1 - \delta$, where the probability is over the randomness of $(\text{Gen}, \text{Rep})$. 


• The security property guarantees that for any distribution $W \in W$, the string $r$ is pseudorandom conditioned on $p$, that is $\delta^{\mathcal{D}_{\text{sec}}}(\langle R, P \rangle, \langle U_\kappa, P \rangle) \leq \epsilon_{\text{sec}}$.

In the above definition, the errors are chosen before $P$: if the error pattern between $w$ and $w'$ depends on the output of $\text{Gen}$, then there is no guarantee about the probability of correctness. In Constructions 4.1 and 5.3 it is crucial that $w'$ is chosen independently of the outcome of $\text{Gen}$.

Information-theoretic fuzzy extractors are obtained by replacing computational distance by statistical distance. We do make a second definitional modification. The standard definition of information-theoretic fuzzy extractors considers $W$ consisting of all distributions of a given entropy. As described in the introduction, we construct fuzzy extractors for parameter regimes where it is impossible to provide security for all distributions with a particular minentropy. In both the computational and information-theoretic settings we consider a family of distributions $W$.

### 2.2 Reusable Fuzzy Extractors

A desirable feature of fuzzy extractors is reusability [Boy04]. Intuitively, it is the ability to support multiple independent enrollments of the same value, allowing users to reuse the same biometric or PUF, for example, with multiple noncooperating providers. More precisely, the algorithm $\text{Gen}$ may be run multiple times on correlated readings $w_1, \ldots, w_\rho$ of a given source. Each time, $\text{Gen}$ will produce a different pair of values $(r_1, p_1), \ldots, (r_\rho, p_\rho)$. Security for each extracted string $r_i$ should hold even in the presence of all the helper strings $p_1, \ldots, p_\rho$ (the reproduction procedure $\text{Rep}$ at the $i$th provider still obtains only a single $w'_i$ close to $w_i$ and uses a single helper string $p_i$). Because the multiple providers may not trust each other, a stronger security feature (which we satisfy) ensures that each $r_i$ is secure even when all $r_j$ for $j \neq i$ are also given to the adversary.

Our ability to construct reusable fuzzy extractors depends on the types of correlations allowed among $w_1, \ldots, w_\rho$. Boyen [Boy04] showed how to do so when each $w_i$ is a shift of $w_1$ by a value that is oblivious to the value of $w_1$ itself (formally, $w_i$ is a result of a transitive isometry applied to $w_1$). Boyen also showed that even for this weak class of correlations, any secure sketch must lose at least $\log |B_i|$ entropy [Boy04, Theorem 11].

We modify the definition of Boyen [Boy04, Definition 6] for the computational setting. We first present our definition and then compare to the definitions of Boyen.

**Definition 2.2 (Reusable Fuzzy Extractors).** Let $W$ be a family of distributions over $\mathcal{M}$. Let $(\text{Gen}, \text{Rep})$ be a $(\mathcal{M}, W, \kappa, t)$-computational fuzzy extractor that is $(s_{\text{sec}}, s_{\text{sec}})$-hard with error $\delta$. Let $(W^1, W^2, \ldots, W^\rho)$ be $\rho$ correlated random variables such that each $W^i \in W$. Let $D$ be an adversary. Define the following game for all $j = 1, \ldots, \rho$:

- **Sampling** The challenger samples $w^i \leftarrow W^i$ and $u \leftarrow \{0, 1\}^\kappa$.

- **Generation** The challenger computes $(r^i, p^i) \leftarrow \text{Gen}(w^i)$.

- **Distinguishing** The advantage of $D$ is

$$
\text{Adv}(D) \overset{\text{def}}{=} \Pr[D(r^1, \ldots, r^j-1, r^j, r^{j+1}, \ldots, r^\rho, p^1, \ldots, p^\rho) = 1] - \Pr[D(r^1, \ldots, r^j-1, u, r^{j+1}, \ldots, r^\rho, p^1, \ldots, p^\rho) = 1].
$$

\footnote{Reusability and unlinkability are two different properties. Unlinkability prevents an adversary from telling if two enrollments correspond to the same physical source [CS08, KBK+11]. We do not consider this property in this work.}
(Gen, Rep) is \((\rho, \epsilon_{sec}, s_{sec})\)-reusable if for all \(D \in \mathcal{D}_{sec}\) and for all \(j = 1, \ldots, \rho\), the advantage is at most \(\epsilon_{sec}\).

**Comparison with the definition of Boyen**  Boyen considers two versions of reusable fuzzy extractors. In the first version (called "outsider security" [Boy04, Definition 6]), the adversary sees \(p^1, \ldots, p^\rho\) and tries to learn about the values \(w^1, \ldots, w^\rho\) or the keys \(r^1, \ldots, r^\rho\). This version is weaker than our version, because the adversary is not given any \(r^i\) values. In the second version (called "insider security" [Boy04, Definition 7]), the adversary controls some subset of the servers and can run Rep on arbitrary \(\tilde{p}^i\). This definition allows the adversary, in particular, to learn a subset of keys \(r^i\) (by performing key generation on the valid \(p^i\)), just like in our definition. However, it also handles the case when the \(p^i\) values are actively compromised. We do not consider such an active compromise attack. As explained in Section 1, protection against such an attack is called "robustness" and can be handled separately—for example, by techniques from [BDK^+05, Theorem 1].

In Boyen’s definitions, the adversary creates a perturbation function \(f^i\) after seeing \(p^1, \ldots, p^{i-1}\) (and generated keys in case of insider security) and the challenger generates \(w^i = f^i(w^1)\). The definition is parameterized by the class of allowed perturbation functions. Boyen constructs an outsider reusable fuzzy extractor for unbounded \(\rho\) when the perturbation family is a family of transitive isometries; Boyen then adds insider security using random oracles.

In contrast, instead of considering perturbation functions to generate \(w^i\), we simply consider all tuples of distributions as long as each distribution is in \(W\), because we support arbitrary correlations among them.

### 3 Tools: Digital Lockers, Point Functions, and Hash Functions

Our main construction uses digital lockers, which are computationally secure symmetric encryption schemes that retain security even when used multiple times with correlated and weak (i.e., nonuniform) keys [CKVW10]. In a digital locker, obtaining any information about the plaintext from the ciphertext is as hard as guessing the key. They have the additional feature that the wrong key can be recognized as such (with high probability). We use notation \(c = \text{lock}(\text{key}, \text{val})\) for the algorithm that performs the locking of the value \(\text{val}\) using the key \(\text{key}\), and \(\text{unlock}(\text{key}, c)\) for the algorithm that performs the unlocking (which will output \(\text{val}\) if \(\text{key}\) is correct and \(\bot\) with high probability otherwise).

The following simple and efficient construction of digital lockers was shown to provide the desired security in the random oracle model of [BR93] by Lynn, Prabhakaran, and Sahai [LPS04, Section 4]. Let \(H\) be a cryptographic hash function, modeled as a random oracle. The locking algorithm \(\text{lock}(\text{key}, \text{val})\) outputs the pair \(\text{nonce}, H(\text{nonce}, \text{key}) \oplus (\text{val}||0^s)\), where \(\text{nonce}\) is a nonce, || denotes concatenation, and \(s\) is a security parameter. As long as the entropy of \(\text{key}\) is superlogarithmic, the adversary has negligible probability of finding the correct \(\text{key}\); and if the adversary doesn’t find the correct \(\text{key}\), then the adversarial knowledge about \(\text{key}\) and \(\text{val}\) is not significantly affected by this locker. Concatenation with \(0^s\) is used to make sure that \(\text{unlock}\) can tell (with certainty \(1 - 2^{-s}\)) when the correct value is unlocked.

It is seems plausible that in the standard model (without random oracles), specific cryptographic hash functions, if used in this construction, will provide the necessary security [CD08 Section 3.2], [Dak09 Section 8.2.3]. Moreover, Bitansky and Canetti [BC10], building on the work of [CD08, CKVW10], show how to obtain composable digital lockers based on a strong version of the Decisional Diffie-Hellman assumption without random oracles.

The security of digital lockers is defined via virtual-grey-box simulatability [BC10], where the simulator
is allowed unbounded running time but only a bounded number of queries to the ideal locker. The simulator is allowed to depend on the desired quality of the simulation. Formally, let idealUnlock(key, val) be the oracle that returns val when given key, and ⊥ otherwise.

**Definition 3.1.** The pair of algorithm (lock, unlock) with security parameter λ is an ℓ-composable secure digital locker with error γ if the following hold:

- **Correctness** For all key and val, Pr[unlock(key, lock(key, val)) = val] ≥ 1 − γ. Furthermore, for any key′ ≠ key, Pr[unlock(key′, lock(key, val))] = ⊥] ≥ 1 − γ.

- **Security** For every PPT adversary A and every positive polynomial p, there exists a (possibly inefficient) simulator S and a polynomial q(λ) such that for any sufficiently large s, any polynomially-long sequence of values (vali, keyi) for i = 1, ..., ℓ, and any auxiliary input z ∈ {0, 1}∗,\[
\left| \Pr \left[ A \left( z, \{\text{lock}(key_i, val_i)\}_i \right) = 1 \right] - \Pr \left[ S \left( z, \{|key_i|, |val_i|\}_i \right) = 1 \right] \right| \leq \frac{1}{p(s)}
\]

where S is allowed q(λ) oracle queries to the oracles \{idealUnlock(keyi, vali)\}i=1, and every positive polynomial \(\rho\).

In one of the constructions for large alphabets, we use a weaker primitive: an obfuscated point function. This primitive can be viewed as a digital locker without the plaintext: it simply outputs 1 if the key is correct and 0 otherwise. Such a function can be easily constructed from the digital locker above with the empty ciphertext, or from a strong version of the Decisional Diffie-Hellman assumption [Can97]. We use notation \(c = \text{lock}(key)\) and unlock(key, c); security is defined the same way as for digital lockers with a fixed plaintext.

## 4 Main Result: Reusable Construction for Sources with High-Entropy Samples

### Sources with High-Entropy Samples

Let the source \(W = W_1, ..., W_n\) consist of strings of length \(n\) over some arbitrary alphabet \(Z\) (the case of greatest interest is that of the binary alphabet \(Z = \{0, 1\}\); however, we describe the construction more generally). For some parameters \(k, \alpha\), we say that the source \(W\) is a source with \(\alpha\)-entropy \(k\)-samples if \(H_\infty(W_{j_1}, ..., W_{j_k}) \geq \alpha\) for uniformly random \(1 \leq j_1, ..., j_k \leq n\). See Section 5 for a discussion of how sources with this property come up naturally.

### The Sample-then-Lock Construction

The construction first chooses a random \(r\) to be used as the output of the fuzzy extractor. It then samples a random subset of symbols \(v_1 = w_{j_1}, ..., w_{j_k}\) and creates a digital locker that hides \(r\) using \(v_1\). This process is repeated to produce some number \(\ell\) of digital lockers all containing \(r\), each unlockable with \(v_1, ..., v_\ell\), respectively. The use of the composable digital lockers allows us to sample multiple times, because we need to argue only about individual entropy of \(V_i\). Composability also allows reusability.

---

6We present and analyze the construction with uniformly random subsets; however, if necessary, it is possible to substantially decrease the required public randomness and the length of \(p\) by using more sophisticated samplers. See [Gol11] for an introduction to samplers.  
7For the construction to be reusable \(\rho\) times the digital locker must be composable \(\ell \cdot \rho\) times.
Construction 4.1 (Sample-then-Lock). Let $\mathcal{Z}$ be an alphabet, and let $W = W_1, \ldots, W_n$ be a source with $\alpha$-entropy $k$-samples, where each $W_j$ is over $\mathcal{Z}$. Let $\ell$ be a parameter, to be determined later. Let $\text{lock, unlock}$ be an $\ell$-composable secure digital locker with error $\gamma$ (for $k$-bit values and keys over $\mathcal{Z}^k$). Define $\text{Gen, Rep}$ as:

\[\text{Gen}\]

1. \textbf{Input:} $w = w_1, \ldots, w_n$
2. \textbf{Sample} $r \leftarrow \{0, 1\}^\kappa$.
3. \textbf{For} $i = 1, \ldots, \ell$:
   - (i) Choose uniformly random $1 \leq j_{i,1}, \ldots, j_{i,k} \leq n$.
   - (ii) Set $v_i = w_{j_{i,1}}, \ldots, w_{j_{i,k}}$.
   - (iii) Set $c_i = \text{lock}(v_i, r)$.
   - (iv) Set $p_i = c_i, (j_{i,1}, \ldots, j_{i,k})$.
4. \textbf{Output} $(r,p)$, where $p = p_1 \ldots p_\ell$.

\[\text{Rep}\]

1. \textbf{Input:} $(w' = w'_1, \ldots, w'_n, p = p_1 \ldots p_\ell)$
2. \textbf{For} $i = 1, \ldots, \ell$:
   - (i) Parse $p_i$ as $c_i, (j_{i,1}, \ldots, j_{i,k})$.
   - (ii) Set $v'_i = w'_{j_{i,1}}, \ldots, w'_{j_{i,k}}$.
   - (iii) Set $r_i = \text{unlock}(v'_i, c_i)$. If $r_i \neq \perp$ output $r_i$.
3. \textbf{Output} $\perp$.

**Correctness vs. Efficiency Tradeoff:** Recall we assume that $\text{dis}(w, w') \leq t$. For any given $i$, the probability that $v'_i = v_i$ is at least $(1 - \frac{t}{n})^k \approx e^{-\frac{tk}{n}}$ (using $e^x \approx 1 + x$). Therefore, the probability that no $v'_i$ matches during $\text{Rep}$, causing $\text{Rep}$ output to $\perp$, is at most

\[
\left(1 - \left(1 - \frac{t}{n}\right)^k\right)\ell \approx (1 - e^{-\frac{tk}{n}})^\ell \approx \exp(-\ell e^{-\frac{tk}{n}}).
\]

In addition, $\text{Rep}$ may be incorrect due to an error in one of the lockers, which happens with probability at most $\ell \cdot \gamma$. Thus, to make the overall error probability less than fuzzy extractor’s allowable error parameter $\delta$ we need $\exp(-\ell e^{-\frac{tk}{n}}) + \ell \cdot \gamma \leq \delta$. Let us assume that $\gamma$ is small enough so that $\ell \cdot \gamma \leq \delta/2$. Then if $tk = cn \ln n$ for some constant $c$, setting $\ell \approx n^c \log \frac{2}{\delta}$ suffices.

**Theorem 4.2.** Let $\lambda$ be a security parameter. Let $\mathcal{W}$ be a family of sources over $\mathcal{Z}^n$ with $\alpha$-entropy $k$-samples for $\alpha = \omega(\log \lambda)$. Then for any $s_{sec} = \text{poly}(\lambda)$ there exists some $\epsilon_{sec} = \text{ngl}(\lambda)$ such that Construction 4.1 is a $(\mathcal{Z}^n, \mathcal{W}, \kappa, t)$-computational fuzzy extractor that is $(\epsilon_{sec}, s_{sec})$-hard with error $\delta = (1 - (1 - \frac{t}{n})^k)\ell + \ell \gamma \approx \exp(-\ell e^{-\frac{tk}{n}}) + \ell \gamma$.

**Proof.** Correctness is argued above. We now argue security.

Our goal is to show that for all $s_{sec} = \text{poly}(\lambda)$ there exists $\epsilon_{sec} = \text{ngl}(\lambda)$ such that $\delta^{\text{Dis}}(\langle R, P \rangle, \langle U, P \rangle) \leq \epsilon_{sec}$. Fix some polynomial $s_{sec}$ and let $D$ be a distinguisher of size at most $s_{sec}$. We want to bound

\[
|\mathbb{E}[D(R, P)] - \mathbb{E}[D(U, P)]|
\]

by a negligible function.

We proceed by contradiction: suppose this difference is not negligible. That is, suppose that there is some polynomial $p(\cdot)$ such that for all $\lambda_0$ there exists some $\lambda > \lambda_0$ such that

\[
|\mathbb{E}[D(R, P)] - \mathbb{E}[D(U, P)]| > 1/p(\lambda).
\]
We note that $\lambda$ is a function of $\lambda_0$ but we omit this notation for the remainder of the proof for clarity. By the security of digital lockers (Definition 3.1), there is a polynomial $q$ and an unbounded time simulator $S$ (making at most $q(\lambda)$ queries to $\{\text{idealUnlock}(v_i, r)\}_{i=1}^\ell$) such that

$$\left| \mathbb{E}[D(R, P_1, ..., P_\ell)] - \mathbb{E}\left[ S^{\text{idealUnlock}(v_i, r)}_{i=1} \left( R, \{j_{i,1}, ..., j_{i,K}\}_{i=1}^\ell, k, \kappa \right) \right] \right| \leq \frac{1}{3p(\lambda)}. \quad (2)$$

The same is true if we replaced $R$ above by and independent uniform random variable $U$ over $\{0, 1\}^\kappa$. We now prove the following lemma, which shows that $S$ cannot distinguish between $R$ and $U$.

**Lemma 4.3.** Let $U$ denote the uniform distribution over $\{0, 1\}^\kappa$. Then

$$\left| \mathbb{E}\left[ S^{\text{idealUnlock}(v_i, r)}_{i=1} \left( R, \{j_{i,1}, ..., j_{i,K}\}_{i=1}^\ell, k, \kappa \right) \right] - \mathbb{E}\left[ S^{\text{idealUnlock}(v_i, r)}_{i=1} \left( U, \{j_{i,1}, ..., j_{i,K}\}_{i=1}^\ell, k, \kappa \right) \right] \right| \leq \frac{q(q+1)}{2^\alpha}, \quad (3)$$

where $q$ is the maximum number of queries $S$ can make.

**Proof.** Fix any $u \in \{0, 1\}^\kappa$ (the lemma will follow by averaging over all $u$). Let $r$ be the correct value of $R$. The only information about whether the value is $r$ or $u$ can obtained by $S$ through the query responses. First, modify $S$ slightly to quit immediately if it gets a response not equal to $\perp$ (such $S$ is equally successful at distinguishing between $r$ and $u$, because the first non-$\perp$ response suffices). There are $q+1$ possible values for the view of $S$ on a given input ($q$ of those views consists of some number of $\perp$ responses followed by the first non-$\perp$ response, and one view has all $q$ responses equal to $\perp$). By [DORS08 Lemma 2.2b], $H_\infty(V_i|\text{View}(S), \{j_{ik}\}) \geq H_\infty(V_j|\{j_{ik}\}) - \log(q+1) \geq \alpha - \log(q+1)$. Therefore, at each query, the probability that $S$ gets a non-$\perp$ answer (equivalently, guesses $V_i$) is at most $(q+1)2^{-\alpha}$. Since there are $q$ queries of $S$, the overall probability is at most $q(q+1)/2^\alpha$. \qed

Since $2^\alpha$ is $\text{ngl}(\lambda)$, there exists some $\lambda_1 \geq \lambda_0$ such that for all $\lambda > \lambda_1$, $q(q+1)/2^\alpha \leq 1/(3p(\lambda))$. Combining Equation 2, Equation 3 and Equation 2 in which $R$ is replaced with $U$ for all $\lambda > \lambda_1$, we obtain that

$$\delta^D((R, P), (U, P)) \leq \frac{1}{p(\lambda)}.$$

This is a contradiction of Equation 1 and completes the proof of Theorem 4.2 \qed

**Reusability** The reusability of Construction 4.1 follows from the security of digital lockers. Consider any $\rho$ number of reuses. For each fixed $i \in \{1, ..., \rho\}$, we can treat the keys $r^1, ..., r^{i-1}, r^{i+1}, ..., r^\rho$ and the sampled positions as auxiliary input to the digital locker adversary. The result follows by simulatability of this adversary, using the same argument as the proof of Theorem 4.2 above. Note that this argument now requires the digital locker to be $\rho \cdot \ell$-composable.

**Theorem 4.4.** Fix $\rho$ and let all the variables be as in Theorem 4.2 except that $(\text{lock}, \text{unlock})$ is an $\ell \cdot \rho$-composable secure digital locker (for $\kappa$-bit values and keys over $\mathbb{Z}_2^k$). Then for all $s_{\sec} = \text{poly}(n)$ there exists some $\epsilon_{\sec} = \text{ngl}(u)$ such that $(\text{Gen}, \text{Rep})$ is $(\rho, \epsilon_{\sec}, s_{\sec})$-reusable fuzzy extractor.
Related Constructions  The work of Škorić and Tuyls [ST09] can be viewed as a fuzzy extractor that places the entire string into a single digital locker (in their paper, they use the language of hash functions). Their Rec procedure symbol searches for a nearby value that unlocks the digital locker, limiting Rec to a polynomial number of error patterns. We use a subset of symbols to lock and take multiple samples, greatly increasing the error tolerance.

The idea of sampling subsets of bits of \( w \) and hashing them together was used by Davida and Frankel [DF99] in order to speed up the search for a matching biometric in the context of passive identification, where a user is identified from a large set of previously enrolled users.

5  Additional Constructions for the Case of Large Alphabets

In this section we provide additional constructions of fuzzy extractors that exploit the structure of the distribution \( w \) (instead of working for all distributions of a particular min-entropy). As stated in the introduction, both constructions work for low entropy rates when \( w \) comes from a large source alphabet \( Z \).

5.1  Construction for Sources with Sparse High-Entropy Marginals

In this section, we consider an alternative construction that is suited to sources over large alphabets. Intuitively, we use single symbols of \( w \) to lock bits of a secret that we then transform into \( r \); we use error-correcting codes to handle bits of the secret that cannot be retrieved due to errors in \( w' \). Our main technical tool is obfuscated point functions (a weaker primitive than digital lockers).

This construction requires enough symbols individually to contain sufficient entropy, but does not require independence of symbols, or even “fresh” entropy from them. Unlike the previous construction, it tolerates a linear fraction of errors (but over a larger alphabet, where errors may be more likely.). However, it cannot work for small alphabets, and is not reusable.

Sources with Sparse High-Entropy Marginals  This construction works for distributions \( W = W_1, ..., W_n \) over \( Z^n \) in which enough symbols \( W_j \) are unpredictable even after adaptive queries to equality oracles for other symbols. This quality of a distribution is captured in the following definition.

**Definition 5.1.** Let \( \text{idealUnlock}(\text{key}) \) be an oracle that returns 1 when given key and 0 otherwise. A source \( W = W_1, ..., W_n \) has \( \beta \)-sparse \( \alpha \)-entropy \( q \)-marginals if there exists a set \( J \subset \{1, ..., n\} \) of size at least \( n - \beta \) such that for any unbounded adversary \( S \),

\[
\forall j \in J, \bar{H}_\infty(W_j|\text{View}(S(\cdot))) \geq \alpha.
\]

where \( S \) is allowed \( q \) queries to the oracles \( \{\text{idealUnlock}(W_i)\}_{i=1}^n \).

We show some examples of such sources in Appendix A.4. In particular, any source \( W \) where for all \( j \), \( H_\infty(W_j) \geq \alpha = \omega(\log \lambda) \) (but all symbols may arbitrarily correlated) is a source with sparse high-entropy marginals (Proposition A.10).

The Error-Correct-and-Obfuscate Construction  This construction is inspired by the construction of Canetti and Dakdouk [CD08]. Instead of having large parts of the string \( w \) unlock \( r \), we have individual symbols unlock bits of the output.
Before presenting the construction we provide some definitions from error correcting codes. We use error-correct codes over \( \{0, 1\}^n \) which correct up to \( t \) bit flips from 0 to 1 but no bit flips from 1 to 0 (this is the Hamming analog of the \( Z \)-channel [TABB02]).

**Definition 5.2.** Let \( e, c \in \{0, 1\}^n \) be vectors. Let \( x = \text{Err}(c, e) \) be defined as follows

\[
x_i = \begin{cases} 
1 & c_i = 1 \lor e_i = 1 \\
0 & \text{otherwise}
\end{cases}
\]

**Definition 5.3.** A set \( C \) (over \( \{0, 1\}^n \)) is a \( (t, \delta_{\text{code}}) \)-\( Z \) code if there exists an efficient procedure Decode such that

\[ \forall e \in \{0, 1\}^n | \text{Wgt}(e) \leq t, \Pr_{c \in C} [\text{Decode}(\text{Err}(c, e)) \neq c] \leq \delta_{\text{code}}. \]

**Construction 5.4** (Lock-and-Error-Correct). Let \( Z \) be an alphabet and let \( W = W_1, ..., W_n \) be a distribution over \( Z^n \). Let \( C \subset \{0, 1\}^n \) be \( (t, \delta_{\text{code}}) \)-\( Z \) code. Let \( \text{lock}, \text{unlock} \) be an \( n \)-composable secure obfuscated point function with error \( \gamma \) (for keys over \( Z \)). Define \( \text{Gen, Rep} \) as:

\begin{align*}
\text{Gen} & \quad \text{Input: } w = w_1, ..., w_n \\
& \quad 2. \text{Sample } c \leftarrow C. \\
& \quad 3. \text{For } j = 1, ..., n: \\
& \quad \quad (i) \text{ If } c_j = 0: p_j = \text{lock}(w_j). \\
& \quad \quad (ii) \text{ Else: } r_j \leftarrow Z. \\
& \quad \quad \quad \text{Let } p_j = \text{lock}(r_j). \\
& \quad 4. \text{Output } (c, p), \text{ where } p = p_1 ... p_n.
\end{align*}

\begin{align*}
\text{Rep} & \quad \text{Input: } (w', p) \\
& \quad 2. \text{For } j = 1, ..., n: \\
& \quad \quad (i) \text{ If } \text{unlock}(w'_j, p_j) = 1: \text{ set } c'_j = 0. \\
& \quad \quad (ii) \text{ Else: set } c'_j = 1. \\
& \quad 3. \text{Set } c = \text{Decode}(c'). \\
& \quad 4. \text{Output } c.
\end{align*}

As presented, Construction 5.4 is not yet a computational fuzzy extractor. The codewords \( c \) are not uniformly distributed and it is possible to learn some bits of \( c \) (for the symbols of \( W \) without much entropy). However, we can show that \( c \) looks like it has entropy to a computationally bounded adversary who knows \( p \). Applying a randomness extractor with outputs over \( \{0, 1\}^k \) (technically, an average-case computational randomness extractor) to \( c \), and adding the extractor seed to \( p \), will give us the desired fuzzy extractor. See Appendix A.1 for the formal details.

Construction 5.4 is secure if no distinguisher can tell whether it is working with \( r_j \) or \( w_j \). By the security of point obfuscation, anything learnable from the obfuscation is learnable from oracle access to the function. Therefore, our construction is secure as long as enough symbols are unpredictable even after adaptive queries to equality oracles for individual symbols, which is exactly the property satisfied by sources with sparse high-entropy marginals.

The following theorem formalizes this intuition (proof in Appendix A.1).

---

\[8\] Any code that corrects \( t \) Hamming errors also corrects \( t \) \( 0 \rightarrow 1 \) errors, but more efficient codes exist for this type of error [TABB02]. Codes with \( 2^{\Theta(n)} \) codewords and \( t = \Theta(n) \) over the binary alphabet exist for Hamming errors and suffice for our purposes (first constructed by Justensen [Jus72]). These codes also yield a constant error tolerance for \( 0 \rightarrow 1 \) bit flips. The class of errors we support in our source (\( t \) Hamming errors over a large alphabet) and the class of errors for which we need codes (\( t 0 \rightarrow 1 \) errors) are different.
Theorem 5.5. Let $\lambda$ be a security parameter. Let $Z$ be an alphabet. Let $W$ be a family of sources with $\beta$-sparse $\alpha = \omega(\log \lambda)$-entropy $q$-marginals over $Z^n$, for any $q = \text{poly}(n)$. Furthermore, let $C$ be a $(t, \delta_{\text{code}})$-$Z$ code over $Z^n$. Then for any $s_{\text{sec}} = \text{poly}(n)$ there exists some $\varepsilon_{\text{sec}} = \text{ngl}(n)$ such that Construction 5.4, followed by a $\kappa$-bit randomness extractor (whose required input entropy is $\leq H_0(C) - \beta$), is a $(Z^n, W, \kappa, t)$-computational fuzzy extractor that is $(\varepsilon_{\text{sec}}, s_{\text{sec}})$-hard with error $\delta_{\text{code}} + n(1/|Z| + \gamma)$.

Entropy vs. Error Rate The minimum entropy necessary to satisfy Definition 5.1 is $\omega(\log \lambda)$ (for example, when all symbols are completely dependent but are all individually unguessable). The construction corrects a constant fraction of errors. When $n = \lambda^{1/c}$ then the entropy is smaller than the number of errors $m = \omega(\log \lambda) < \Theta(n) = \lambda^{1/c}$.

Improvements If most codewords have Hamming weight close to $1/2$, we can decrease the error tolerance needed from the code from $t$ to about $t/2$, because roughly half of the mismatches between $w$ and $w'$ occur where $c_j = 1$.

If the output length of the computational extractor $\kappa$ is not sufficient for a cryptographic key the construction can be run multiple times with the same input and independent randomness to expand the key.

Lack of Reusability Even though Construction 5.4 uses composable obfuscated point functions, it is not reusable. Definition 5.1 allows sources with some “weak” symbols that can be completely learned by an adversary observing $p$. If a source is enrolled multiple times this partial information may add up over time to reveal the original value $w_1$. In contrast, Construction 4.1 leaks no partial information for the supported sources, allowing reusability.

5.2 Information-Theoretic Construction for Sparse Block Sources

The construction in this section has information-theoretic security, in contrast to only computational security of the previous two constructions. It uses symbol-by-symbol condensers to reduce the alphabet size while preserving most of the entropy, and then applies a standard fuzzy extractor to the resulting string.

This construction requires less entropy from each symbol than the previous construction; however, it places more stringent independence requirements on the symbols. It tolerates a linear number of errors.

Sparse Block Sources This construction works for sources $W = W_1, \ldots, W_n$ over $Z^n$ in which enough symbols $W_j$ contribute fresh entropy conditioned on previous symbols. We call this such sources sparse block sources, weakening the notion of block sources (introduced by Chor and Goldreich [CG88]), which require every symbol to contribute fresh entropy.

Definition 5.6. A distribution $W = W_1, \ldots, W_n$ is an $(\alpha, \beta)$-sparse block source if there exists a set of indices $J$ where $|J| \geq n - \beta$ such that the following holds:

$$\forall j \in J, \forall w_1, \ldots, w_{j-1} \in W_1, \ldots, W_{j-1}, H_\infty(W_j | W_1 = w_1, \ldots, W_{j-1} = w_{j-1}) \geq \alpha.$$
The Condense-then-Fuzzy-Extract Construction

The construction first condenses entropy from each symbol of the source and then applies a fuzzy extractor to the condensed symbols. We'll denote the fuzzy extractor on the smaller alphabet as \((\text{Gen}', \text{Rep}')\). A condenser is like a randomness extractor but the output is allowed to be slightly entropy deficient. Condensers are known with smaller entropy loss than possible for randomness extractors (e.g. [DPW14]).

**Definition 5.7.** A function \(\text{cond} : Z \times \{0,1\}^d \rightarrow \mathcal{Y}\) is a \((m, \tilde{m}, \epsilon)\)-randomness condenser if whenever \(H_\infty(W) \geq m\), then there exists a distribution \(Y\) with \(H_\infty(Y) \geq \tilde{m}\) and \((\text{cond}(W, seed), seed) \approx_\epsilon (Y, seed)\).

The main idea of the construction is that errors are “corrected” on the large alphabet (before condensing) while the entropy loss for the error correction is incurred on a smaller alphabet (after condensing).

**Construction 5.8.** Let \(Z\) be an alphabet and let \(W = W_1, ..., W_n\) be a distribution over \(Z^n\). We describe Gen, Rep as follows:

Gen

1. **Input:** \(w = w_1, ..., w_n\)
2. For \(j = 1, ..., n:\)
   (i) Sample \(seed_i \leftarrow \{0,1\}^d\).
   (ii) Set \(v_i = \text{cond}(w_i, seed_i)\).
3. Set \((r, p') \leftarrow \text{Gen}'(v_1, ..., v_n)\).
4. Set \(p = (p', seed_1, ..., seed_n)\).
5. **Output** \((r, p)\).

Rep

1. **Input:** \((w', p = (p', seed_1, ..., seed_n))\)
2. For \(j = 1, ..., n:\)
   (i) Set \(v'_i = \text{cond}(w'_i, seed_i)\).
3. **Output** \(r = \text{Rep}'(v', p')\).

The following theorem shows the security of this construction (proof in Appendix B).

**Theorem 5.9.** Let \(\mathcal{W}\) be a family of \((\alpha = \Omega(1), \beta \leq n(1 - \Theta(1)))\)-sparse block sources over \(Z^n\) and let \(\text{cond} : Z \times \{0,1\}^d \rightarrow \mathcal{Y}\) be a \((\alpha, \tilde{\alpha}, \epsilon_{\text{cond}})\)-randomness conductor. Define \(\mathcal{V}\) as the family of all distributions with minentropy at least \(\tilde{\alpha}(n - \beta)\) and let \((\text{Gen}', \text{Rep}')\) be \((3^n, \mathcal{V}, \kappa, t, \epsilon_{\text{fext}})\)-fuzzy extractor with error \(\delta\). Then \((\text{Gen}, \text{Rep})\) is a \((\mathcal{Z}^n, \mathcal{W}, \kappa, t, n\epsilon_{\text{cond}} + \epsilon_{\text{fext}})\)-fuzzy extractor with error \(\delta\).

**Overcoming Proposition 1.1** Proposition 1.1 shows that no fuzzy extractor can be secure for all sources of a given minentropy \(m < \log |B_t|\). Construction 5.8 supports sparse block sources whose overall entropy is less than \(\log |B_t|\). The structure of a sparse block source implies that \(H_\infty(W) \geq \alpha(n - \beta) = \Theta(n)\). We assume that \(H_\infty(W) = \Theta(n)\). Using standard fuzzy extractors (for \(\text{Gen}', \text{Rep}')\) it is possible to correct \(t = \Theta(n)\) errors, yielding \(\log |B_t| > \Theta(n)\) when \(|Z| = \omega(1)|.\)

**Acknowledgements**

The authors are grateful to Nishanth Chandran, Nir Bitansky, Sharon Goldberg, Gene Itkis, Bhavana Kanukurthi, and Mayank Varia for helpful discussions, creative ideas, and important references.

---

9We actually need \((\text{Gen}', \text{Rep}')\) to be an average case fuzzy extractor (see [DORS08] Definition 4 and the accompanying discussion). Most known constructions of fuzzy extractors are average-case fuzzy extractors. For simplicity we refer to \(\text{Gen}', \text{Rep}'\) as simply a fuzzy extractor.
Ran Canetti is supported by the NSF MACS project, an NSF Algorithmic foundations grant 1218461, the Check Point Institute for Information Security, and ISF grant 1523/14. Omer Paneth is additionally supported by the Simons award for graduate students in theoretical computer science. The work of Benjamin Fuller is sponsored in part by US NSF grants 1012910 and 1012798 and the United States Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Leonid Reyzin is supported in part by US NSF grants 0831281, 1012910, 1012798, and 1422965. Adam Smith is supported in part by NSF awards 0747294 and 0941553.

References


A Analysis of Construction 5.4

A.1 Computational Fuzzy Conductors and Computational Extractors

In this section we introduce tools necessary to convert Construction 5.4 to a computational fuzzy extractor. We first define an object weaker than a computational fuzzy extractor: it outputs a key with computational entropy (instead of a pseudorandom key). We call this object a computational fuzzy conductor. It is the computational analogue of a fuzzy conductor (introduced by Kanukurthi and Reyzin [KR09]). Before defining this object, we define conditional computational “HILL” ([HILL99]) entropy.

Definition A.1. [HLR07, Definition 3] Let \((W, S)\) be a pair of random variables. \(W\) has HILL entropy at least \(m\) conditioned on \(S\), denoted \(H_{\text{HILL}}^{\text{sec}}(W|S) \geq m\) if there exists a joint distribution \((X, S)\), such that \(\tilde{H}_{\infty}(X|S) \geq m\) and \(\delta_{\text{sec}}((W, S), (X, S)) \leq \epsilon_{\text{sec}}\).

Definition A.2. A pair of randomized procedures “generate” \((\text{Gen})\) and “reproduce” \((\text{Rep})\) is an \((M, W, \tilde{m}, t)\)-computational fuzzy conductor that is \((\epsilon_{\text{sec}}, s_{\text{sec}})\)-hard with error \(\delta\) if \(\text{Gen}\) and \(\text{Rep}\) satisfy Definition 2.1, except the last condition is replaced with the following weaker condition:

- for any distribution \(W \in W\), the string \(r\) has high HILL entropy conditioned on \(P\). That is \(H_{\text{HILL}}^{\text{sec}}(R|P) \geq \tilde{m}\).

Computational fuzzy conductors can be converted to computational fuzzy extractors (Definition 2.1) using standard techniques, as follows. The transformation uses a computational extractor. A computational extractor is the adaption of a randomness extractor to the computational setting. Any information-theoretic randomness extractor is also a computational extractor; however, unlike information-theoretic extractors, computational extractors can expand their output arbitrarily via pseudorandom generators once a long-enough output is obtained. We adapt the definition of Krawczyk [Kra10] to the average case:

Definition A.3. A function \(\text{cext} : \{0, 1\}^n \times \{0, 1\}^d \rightarrow \{0, 1\}^\kappa\) a \((m, \epsilon_{\text{sec}}, s_{\text{sec}})\)-average-case computational extractor if for all pairs of random variables \(X, Y\) (with \(X\) over \(\{0, 1\}^n\)) such that \(\tilde{H}_{\infty}(X|Y) \geq m\), we have \(\delta_{\text{sec}}((\text{cext}(X; U_d), U_d, Y), U_\kappa \times U_d \times Y) \leq \epsilon_{\text{sec}}\).

Combining a computational fuzzy conductor and a computational extractor yields a computational fuzzy extractor:

Lemma A.4. Let \((\text{Gen}', \text{Rep}')\) be a \((M, W, \tilde{m}, t)\)-computational fuzzy conductor that is \((\epsilon_{\text{cond}}, s_{\text{cond}})\)-hard with error \(\delta\) and outputs in \(\{0, 1\}^n\). Let \(\text{cext} : \{0, 1\}^n \times \{0, 1\}^{\kappa} \rightarrow \{0, 1\}^{\kappa}\) be a \((\tilde{m}, \epsilon_{\text{ext}}, s_{\text{ext}})\)-average case computational extractor. Define \((\text{Gen}, \text{Rep})\) as:
• \text{Gen}(w; seed) (where seed \in \{0,1\}^d): run \ (r', p') = \text{Gen}'(w) and output \ r = \text{cext}(r'; seed), p = (p', seed).

• \text{Rep}(w', (p', seed)): run \ r' = \text{Rep}'(w'; p') and output \ r = \text{cext}(r'; seed).

Then \ (\text{Gen}, \text{Rep}) is a \ ((\mathcal{M}, \mathcal{W}, \kappa, t))\text{-computational fuzzy extractor} that is \ (\epsilon_{\text{cond}}, \epsilon_{\text{ext}}')\text{-hard with error} \ \delta \ where \ s' = \min\{s_{\text{cond}} - |\text{cext}| - d, s_{\text{ext}}\}.

\textbf{Proof.} It suffices to show if there is some distinguisher \ D' of size \ s' then \ \delta^{D'}(\text{cext}(X; U_d), U_d, P')) > \epsilon_{\text{cond}} + \epsilon_{\text{ext}}

then there is an distinguisher \ D of size \ s_{\text{cond}} such that for all \ Y \ with \ \tilde{H}_\infty(Y|P') \geq \tilde{m},

\delta^{D}((X, P'), (Y, P')) \geq \epsilon_{\text{cond}}.

Let \ D' be such a distinguisher. That is,

\delta^{D'}(\text{cext}(X; U_d) \times U_d \times P', U_\kappa \times U_d \times P') > \epsilon_{\text{ext}} + \epsilon_{\text{cond}}.

Then define \ D as follows. On input \ (y, p') sample \ seed \leftarrow U_d, compute \ r \leftarrow \text{cext}(y; seed) and output \ D(r, seed, p'). Note that \ |D| \approx s' + |\text{cext}| + d = s_{\text{cond}}. Then we have the following:

\delta^D((X, P'), (Y, P')) = \delta^{D'}((\text{cext}(X, U_d), U_d, P'), \text{cext}(Y, U_d), U_d, P')

\geq \delta^{D'}((\text{cext}(X, U_d), U_d, P'), (U_\kappa \times U_d \times P'))

\geq \delta^{D'}((U_\kappa \times U_d \times P'), (\text{cext}(Y, U_d), U_d, P'))

> \epsilon_{\text{cond}} + \epsilon_{\text{ext}} - \epsilon_{\text{ext}} = \epsilon_{\text{cond}}.

Where the last line follows by noting that \ D' is of size at most \ s_{\text{ext}}. Thus \ D \ distinguishes \ X \ from all \ Y \ with sufficient conditional minentropy. This is a contradiction. \qed

\subsection{A.2 Security of Construction 5.4}

It suffices to prove that Construction 5.4 is a \ ((\mathcal{Z}^n, \mathcal{W}, \tilde{m} = H_0(C) - \beta, t))\text{-computational fuzzy conductor}, i.e., that \ C \ has HILL entropy \ H_0(C) - \beta \ conditioned on \ P. The final extraction step will convert it to a computational fuzzy extractor (see Lemma A.4).

The security proof of Construction 5.4 is similar to the security proof of Construction 4.1. However, it is made more complicated by the fact that the definition of sources with sparse high-entropy marginals (Definition 5.1) allows for certain weak symbols that can easily be guessed. This means we must limit our indistinguishable distribution to symbols that are difficult to guess. Security is proved via the following lemma:

\textbf{Lemma A.5.} \textit{Let all variables be as in Theorem 5.3. For every} \ s_{\text{sec}} = \text{poly}(n) \ \text{there exists some} \ \epsilon_{\text{sec}} = \text{ngl}(n) \ \text{such that} \ H^0_{\text{HILL}_{\epsilon_{\text{sec}}, s_{\text{sec}}}}(C|P) \geq H_0(C) - \beta. \ \text{(The second step is by} \ [\text{DORS08}, \text{Lemma} 2.2b])}. It is left to show \ \delta^{D_{\text{sec}}}(\mathcal{C}, \mathcal{P}) \leq \text{ngl}(n). The outline for the rest of the proof is as follows:
Let $D$ be a distinguisher between $(C, P)$ and $(C', P)$. By the security of obfuscated point functions,

$$|E[D(C, P_1, ..., P_n)] - E[S^{\text{idealUnlock}()}_{i=1}^n (C, n \cdot |Z|)]|$$

is small.

Show that even an unbounded $S$ making a polynomial number of queries to the stored points cannot distinguish between $C$ and $C'$. That is,

$$|E[S^{\text{idealUnlock}()}_{i=1}^n (C, n \cdot |Z|)] - E[S^{\text{idealUnlock}()}_{i=1}^n (C', n \cdot |Z|)]|$$

is small.

By the security of obfuscated point functions,

$$|E[S^{\text{idealUnlock}()}_{i=1}^n (C', n \cdot |Z|)] - E[D(C', P_1, ..., P_n)]|$$

is small.

Proof of Lemma A.5. The overall approach and the proof of the first and third bullet as in Theorem 4.2. We only prove the second bullet. Define the distribution $X$ as follows:

$$X_j = \begin{cases} W_j & C_j = 0 \\ R_j & C_j = 1. \end{cases}$$

Lemma A.6. $\Delta(S^{\text{idealUnlock}(X_i)}_{i=1}^n (C, n \cdot |Z|), S^{\text{idealUnlock}(X_i)}_{i=1}^n (C', n \cdot |Z|)) \leq (n - \beta)2^{-(\alpha+1)}$.

Proof. It suffices to show that for any two codewords that agree on $J$, the statistical distance is at most $(n - \beta)2^{-(\alpha+1)}$.

Lemma A.7. Let $c^*$ be true value encoded in $X$ and let $c'$ a codeword in $C'$. Then,

$$\Delta(S^{\text{idealUnlock}(X_i)}_{i=1}^n (c^*, n \cdot |Z|), S^{\text{idealUnlock}(X_i)}_{i=1}^n (c', n \cdot |Z|)) \leq (n - \beta)2^{-(\alpha+1)}.$$

Proof. Recall that for all $j \in J$, $H_\infty(W_j|\text{View}(S)) \geq \alpha$. The only information about the correct value of $c^*_j$ is contained in the query responses. When all responses are 0 the view of $S$ is identical when presented with $c^*$ or $c'$. We now show that for any value of $c^*$ all queries on $j \in J$ return 0 with probability $1 - 2^{-\alpha+1}$. Suppose not. That is, suppose the probability of at least one nonzero response on index $j$ is $> 2^{-(\alpha+1)}$. Since $w, w'$ are independent of $r_j$, the probability of this happening when $c^*_j = 1$ is at most $q/|Z| \text{ or equivalently } 2^{-\log|Z|+\log q}$. Thus, it must occur with probability:

$$2^{-\alpha+1} < \Pr[\text{non zero response location } j]$$

$$= \Pr[c^*_j = 1] \Pr[\text{non zero response location } j \land c^*_j = 1]$$

$$+ \Pr[c^*_j = 0] \Pr[\text{non zero response location } j \land c^*_j = 0]$$

$$\leq 1 \times 2^{-\log|Z|+\log q} + 1 \times \Pr[\text{non zero response location } j \land c^*_j = 0]$$

We now show that for $\alpha \leq \log |Z| - \log q$:
Claim A.8. If \( W \) is a source with \( \beta \)-sparse \( \alpha \)-entropy \( q \)-marginals over \( Z \), then \( \alpha \leq \log |Z| - \log q \).

Proof. Let \( J \subset \{ 1, ..., n \} \) the set of good indices. It suffices to show that there exists an \( S \) making \( q \) queries such that for some \( j \in J \), \( \tilde{H}_\infty(W_j | S^\text{idealLock}(X_i))^q_{i=1} \leq \log |Z| - \log q \). Let \( j \in J \) be some arbitrary element of \( J \) and denote by \( w_{j,1}, ..., w_{j,q} \) the \( q \) most likely outcomes of \( W_j \) (breaking ties arbitrarily). Then \( \sum_{i=1}^q \Pr[W_j = w_{j,i}] \geq q/|Z| \). Suppose not. This means that there is some \( w_{j,i} \) with probability \( \Pr[W_j = w_{j,i}] < 1/|Z| \). Since there are \( |Z| - q \) remaining possible values of \( W_j \) for their total probability to be at least \( 1 - q/|Z| \) at least of these values has probability at least \( 1/|Z| \). This contradicts the statement \( w_{j,1}, ..., w_{j,q} \) are the most likely values. Consider \( S \) that queries the \( j \)th oracle on \( w_{j,1}, ..., w_{j,q} \). Denote by \( \text{Bad} \) the random variable when \( W_j \in \{ w_{j,1}, ..., w_{j,q} \} \) After these queries the remaining minentropy is at most:

\[
\tilde{H}_\infty(W_j | S^\text{ideal}(-)) = - \log \left( \Pr[\text{Bad} = 1] \times 1 + \Pr[\text{Bad} = 0] \times \max_w \Pr[W_j = w | \text{Bad} = 0] \right)
\leq - \log \left( \Pr[\text{Bad} = 1] \times 1 \right) = - \log \left( q/\lfloor Z \rfloor \right) = \log |Z| - \log q
\]

This completes the proof of Claim A.8.

Rearranging terms in Equation [4] we have:

\[
\Pr[\text{non zero response location } j \land c_j = 0] > 2^{-\alpha + 1} - 2^{-(\log |Z| - \log q)} = 2^{-\alpha}
\]

When there is a 1 response and \( c_j = 0 \) this means that there is no remaining minentropy. If this occurs with over \( 2^{-\alpha} \) probability this violates the condition on \( W \) (Definition 5.1). By the union bound over the indices \( j \in J \) the total probability of a 1 in \( J \) is at most \( (n - \beta)2^{-\alpha + 1} \). Recall that \( c^*, c' \) match on all indices outside of \( J \). Thus, for all \( c^*, c' \) the statistical distance is at most \( (n - \beta)2^{-\alpha + 1} \). This concludes the proof of Lemma A.7.

Lemma A.6 follows by averaging over all points in \( C' \).

A.3 Correctness of Construction 5.4

We now argue correctness of Construction 5.4. We first assume ideal functionality of the obfuscated point functions. Consider a coordinate \( j \) for which \( c_j = 1 \). Since \( w' \) is chosen independently of the points \( r_j \), and \( r_j \) is uniform, \( \Pr[r_j = w'_j] = 1/|Z| \). Thus, the probability of at least one 1 \( \rightarrow 0 \) bit flip (the random choice \( r_i \) is the same as \( w'_j \)) is \( \leq n(1/|Z|) \). Since there are most \( t \) locations for which \( w_j \neq w'_j \) there are at most \( t \) \( 0 \rightarrow 1 \) bit flips in \( c \), which the code will correct with probability \( 1 - \delta_\text{code} \), because \( c \) was chosen uniformly. Finally, since each obfuscated point function is correct with probability \( 1 - \gamma \), Construction 5.4 is correct with error at most \( \delta_\text{code} + n(1/|Z|) + \gamma \).

A.4 Characterizing sources with sparse high-entropy marginals

Definition 5.1 is an inherently adaptive definition and a little unwieldy. In this section, we partially characterize sources that satisfy Definition 5.1. The majority of the difficulty in characterizing Definition 5.1 is that different symbols may be dependent, so an equality query on symbol \( i \) may reshape the distribution.
of symbol $j$. In the examples that follow we denote the adversary by $S$ as the simulator in Definition 3.1.

We first show some sources that have sparse high-entropy marginals (Section A.4.1) and then show sources with high overall entropy that do not have sparse high-entropy marginals (Section A.4.2).

### A.4.1 Positive Examples

We begin with the case of independent symbols.

**Proposition A.9.** Let $W = W_1, ..., W_n$ be a source in which all symbols $W_j$ are mutually independent. Let $\alpha$ be a parameter. Let $J \subset \{1, ..., n\}$ be a set of indices such that for all $j \in J$, $H_\infty(W_j) \geq \alpha$. Then for any $q$, $W$ is a source with $(n - |J|)$-sparse $(\alpha - \log(q + 1))$-entropy $q$-marginals. In particular, when $\alpha = \omega(\log n)$ and $q = \text{poly}(n)$, then $W$ is a source with $(n - |J|)$-sparse $\omega(\log n)$-entropy $q$-marginals.

*Proof.* It suffices to show that for all $j \in J$, $H_\infty(W_j|\text{View}(S(\cdot))) = \alpha - \log(q + 1)$ where $S$ is allowed $q$ queries to the oracles $\{\text{idealUnlock}(W_i)\}_{i=1}^n$. We can ignore queries for all symbols but the $j$th, as the symbols are independent. Furthermore, without loss of generality, we can assume that no duplicate queries are asked, and that the adversary is deterministic ($S$ can calculate the best coins). Let $A_1, A_2, ..., A_q$ be the random variables representing the oracle answers for an adversary $S$ making $q$ queries about the $i$th symbol. Each $A_k$ is just a bit, and at most one of them is equal to 1 (because duplicate queries are disallowed). Thus, the total number of possible responses is $q + 1$. Thus, we have the following,

$$
H_\infty(W_j|\text{View}(S(\cdot))) = H_\infty(W_j|A_1, ..., A_q) = H_\infty(W_j) - |A_1, ..., A_q| = \alpha - \log(q + 1),
$$

where the second line follows from the first by [DORS08 Lemma 2.2].

In their work on computational fuzzy extractors, Fuller, Meng, and Reyzin [FMR13] show a construction for symbol-fixing sources, where each symbol is either uniform or a fixed symbol (symbol-fixing sources were introduced by Kamp and Zuckerman [KZ07]). Proposition A.9 shows that Definition 5.1 captures, in particular, this class of distributions. However, Definition 5.1 captures more distributions. We now consider more complicated distributions where symbols are not independent.

**Proposition A.10.** Let $f : \{0, 1\}^e \to \mathbb{Z}^n$ be a function. Furthermore, let $f_j$ denote the restriction of $f$’s output to its $j$th coordinate. If for all $j$, $f_j$ is injective then $W = f(U_e)$ is a source with 0-sparse $(e - \log(q + 1))$-entropy $q$-marginals.

*Proof.* Since $f$ is injective on each symbol, $H_\infty(W_j|\text{View}(S)) = H_\infty(U_e|\text{View}(S))$. Consider a query $q_k$ on symbol $j$. There are two possibilities: either $q_k$ is not in the image of $f_j$, or $q_k$ can be considered a query on the preimage $f_j^{-1}(q_k)$. Then (by assuming $S$ knows $f$) we can eliminate queries which correspond to the same value of $U_e$. Then the possible responses are strings with Hamming weight at most 1 (like in the proof of Claim A.9, and by DORS08 Lemma 2.2] we have for all $j$, $H_\infty(W_j|\text{View}(S)) \geq H_\infty(W_j) - \log(q + 1)$.

Note the total entropy of a source in Proposition A.10 is $e$, so there is a family of distributions with total entropy $\omega(\log n)$ for which Construction 5.4 is secure. For these distributions, all the coordinates are as dependent as possible: one determines all others. We can prove a slightly weaker claim when the correlation between the coordinates $W_j$ is arbitrary:
Finally this means that $i$ (where the third line follows by considering that to get $\infty$), this is true for $(\alpha - 1 - \log(q + 1))$-entropy $q$-marginals.

Proof. Intuitively, the claim is true because the oracle is not likely to return 1 on any query. Formally, we proceed by induction on oracle queries, using the same notation as in the proof of Proposition A.9. Our inductive hypothesis is that $\Pr[A_1 \neq 0 \lor \cdots \lor A_{i-1} \neq 0] \leq (i-1)2^{1-\alpha}$. If the inductive hypothesis holds, then, for each $j$, 

$$H_\infty(W_j|A_1 = \cdots = A_{i-1} = 0) \geq \alpha - 1.$$  

(5)

This is true for $i = 1$ by the condition of the theorem. It is true for $i > 1$ because, as a consequence of the definition of $H_\infty$, for any random variable $X$ and event $E$, $H_\infty(X|E) \geq H_\infty(X) + \log \Pr[E]$; and $(i-1)2^{1-\alpha} \leq 2q2^{-\alpha} \leq 1/2$.

We now show that $\Pr[A_1 \neq 0 \lor \cdots \lor A_i \neq 0] \leq i2^{1-\alpha}$, assuming that $\Pr[A_1 \neq 0 \lor \cdots \lor A_{i-1} \neq 0] \leq (i-1)2^{1-\alpha}$.

$$\Pr[A_1 \neq 0 \lor \cdots \lor A_{i-1} \neq 0 \lor A_i \neq 0] = \Pr[A_1 \neq 0 \lor \cdots \lor A_{i-1} \neq 0] + \Pr[A_1 = \cdots = A_{i-1} = 0 \land A_i = 1]$$

$$\leq (i-1)2^{1-\alpha} + \Pr[A_1 = 1 | A_1 = \cdots = A_{i-1} = 0]$$

$$\leq (i-1)2^{1-\alpha} + \max_j 2^{-H_\infty(W_j|A_1=\cdots=A_{i-1}=0)}$$

$$\leq (i-1)2^{1-\alpha} + 2^{1-\alpha}$$

$$= i2^{1-\alpha}$$

(where the third line follows by considering that to get $A_i = 1$, the adversary needs to guess some $W_j$, and the fourth line follows by (5)). Thus, using $i = q+1$ in (5), we know $H_\infty(W_j|A_1 = \cdots = A_q = 0) \geq \alpha - 1$. Finally this means that

$$\tilde{H}_\infty(W_j|A_1, \ldots, A_q) \geq -\log \left(2^{-H_\infty(W_j|A_1=\cdots=A_q=0)} \Pr[A_1 = \cdots = A_q = 0] + 1 \cdot \Pr[A_1 \neq 0 \lor \cdots \lor A_q \neq 0]\right)$$

$$\geq -\log \left(2^{-H_\infty(W_j|A_1=\cdots=A_q=0)} + q2^{1-\alpha}\right)$$

$$\geq -\log \left((q+1)2^{1-\alpha}\right) = \alpha - 1 - \log(q + 1).$$

\hfill \Box

A.4.2 Negative Examples

Propositions A.9 and A.10 rest on there being no easy “entry” point to the distribution. This is not always the case. Indeed it is possible for some symbols to have very high entropy but lose all of it after equality queries.

**Proposition A.11.** Let $W = W_1, \ldots, W_n$ be a source. Suppose that for all $j$, $H_\infty(W_j) \geq \alpha$, and that $q \leq 2^\alpha/4$ (this holds asymptotically, in particular, if $q$ is polynomial and $\alpha$ is super-logarithmic). Then $W$ is a source with $0$-sparse $(\alpha - 1 - \log(q + 1))$-entropy $q$-marginals.

**Proof.**

A.4.2 Negative Examples

Propositions A.9 and A.10 rest on there being no easy “entry” point to the distribution. This is not always the case. Indeed it is possible for some symbols to have very high entropy but lose all of it after equality queries.

**Proposition A.12.** Let $p = (\text{poly}(\lambda))$ and let $f_1, \ldots, f_n$ be injective functions where $f_j : \{0,1\}^{j \times \log p} \rightarrow \mathcal{Z}^{\lceil \log p \rceil}$. Then define the distribution $U_n$ and consider $W_1 = f_1(U_1, \ldots, \log p)$, $W_2 = f_2(U_1, \ldots, 2 \log p), \ldots, W_n = f_n(U)$. There is an adversary making $p \times n$ queries such that $\tilde{H}_\infty(W|\text{View}(S(\cdot))) = 0$.  

\footnote{Here we assume that $|\mathcal{Z}| \geq n \times \log p$, that is the source has a small number of symbols.}
Proof. Let \( x \) be the true value for \( U_{P \times n} \). We present an adversary \( S \) that completely determines \( x \). \( S \) computes \( y_1 = f_1(x_1), ..., y_n = f_2(x_n) \). Then \( S \) queries on \( (y_1), ..., (y_n) \) to the first oracle, exactly one answer returns 1. Let this value be \( y_i \) and its preimage \( x_i \). Then \( S \) computes \( y_1 = f_2(x_1, x_i), ..., y_n = f_2(x_n, x_i) \) and queries \( y_1, ..., y_n \). Again, exactly one of these queries returns 1. This process is repeated until all of \( x \) is recovered (and thus \( w \)).

The previous example relies on an adversary’s ability to determine a symbol from the previous symbols. We formalize this notion next. We define the entropy jump of a source as the remaining entropy of a symbol when previous symbols are known:

**Definition A.13.** Let \( W = W_1, ..., W_n \) be a source under ordering \( i_1, ..., i_n \). The jump of a symbol \( i_j \) is

\[
\text{Jump}(i_j) = \max_{w_{i_1}, ..., w_{i_{j-1}}} H_0(W_{i_j} | W_{i_1} = w_{i_1}, ..., W_{i_{j-1}} = w_{i_{j-1}}).
\]

An adversary who can learn symbols in succession can eventually recover the entire secret. In order for a source to have sparse high-entropy marginals, the adversary must get “stuck” early enough in this recovery process. This translates to having a super-logarithmic jump early enough.

**Proposition A.14.** Let \( W \) be a distribution and let \( q \) be a parameter, if there exists an ordering \( i_1, ..., i_n \) such that for all \( j \leq n - \beta + 1 \), \( \text{Jump}(i_j) = \log q/(n - \beta + 1) \), then \( W \) is not a source with \( \beta \)-sparse high-entropy \( q \)-marginals.

**Proof.** For convenience relabel the ordering that violates the condition as \( 1, ..., n \). We describe an unbounded adversary \( S \) that determines \( W_1, ..., W_{n-\beta+1} \). As before \( S \) queries the \( q/n \) possible values for \( W_1 \) and determines \( W_1 \). Then \( S \) queries the (at most) \( q/(n - \beta + 1) \) possible values for \( W_2 | W_1 \). This process is repeated until \( W_{n-\beta+1} \) is learned.

Presenting a sufficient condition for security is more difficult as \( S \) may interleave queries to different symbols. It seems like the optimum strategy for \( S \) is to focus on a single symbol at a time, but it is unclear how to formalize this intuition.

## B Analysis of Construction 5.8

**Proof.** Let \( W \in \mathcal{W} \). It suffices to argue correctness and security. We first argue correctness.

**Correctness:** When \( w_i = w_i' \), then \( \text{cond}(w_i, seed_i) = \text{cond}(w_i', seed_i) \) and thus \( v_i = v_i' \). Thus, for all \( w, w' \) where \( \text{dis}(w, w') \leq t \), then \( \text{dis}(v, v') \leq t \). Then by correctness of \((\text{Gen}', \text{Rep}')\), \( \text{Pr}[r, p] \leftarrow \text{Gen}'(v) \land r' \leftarrow \text{Rep}'(v', p) \land r' = r \geq 1 - \delta \).

**Security:** We now argue security. Denote by \( seed \) the random variable consisting of all \( n \) seeds and \( V \) the entire string of generated \( V_1, ..., V_n \). To show that

\[
R|P, seed \approx_{n \epsilon_{\text{cond}} + \epsilon_{\text{ext}}} U|P, seed,
\]

it suffices to show that \( H_\infty(V|seed) \) is \( n \epsilon_{\text{cond}} \) close to a distribution with average minentropy \( \bar{\alpha}(n - \beta) \). The lemma then follows by the security of \((\text{Gen}', \text{Rep}')\).

We now argue that there exists a distribution \( Y \) where \( H_\infty(Y|seed) \geq \bar{\alpha}(n - \beta) \) and \( (Y, seed_1, ..., seed_n) \approx (Y, seed_1, ..., seed_n) \). First note since \( W \) is \((\alpha, \beta)\) sparse block source that there exists a set of indices \( J \) where \( |J| \geq n - \beta \) such that the following holds:

\[
\forall j \in J, \forall w_1, ..., w_{j-1} \in W_1, ..., W_{j-1}, H_\infty(W_j | W_1 = w_1, ..., W_{j-1} = w_{j-1}) \geq \alpha.
\]

\[\text{Note, again, that} (\text{Gen}', \text{Rep}') \text{ must be an average-case fuzzy extractor. Most known constructions are average-case and we omit this notation.}\]
Then consider the first element of $j_1 \in J$, $\forall w_1, \ldots, w_{j_1-1} \in W_1, \ldots, W_{j_1-1}$,

$$H_\infty(W_{j_1}|W_1 = w_1, \ldots, W_{j_1-1} = w_{j_1-1}) \geq \alpha.$$ 

Thus, there exists a distribution $Y_{j_1}$ with $\tilde{H}_\infty(Y_{j_1}|seed_{j_1}) \geq \tilde{\alpha}$ such that

$$(\text{cond}(W_{j_1}, seed_{j_1}), seed_{j_1}, W_1, \ldots, W_{j_1-1}) \approx_{\text{cond}} (Y_{j_1}, seed_{j_1}, W_1, \ldots, W_{j_1-1})$$

and since $(seed_1, \ldots, seed_{j_1})$ are independent of these values

$$(\text{cond}(W_{j_1}, seed_{j_1}), W_{j_1-1}, \ldots, W_1, seed_{j_1}, \ldots, seed_1) \approx_{\text{cond}} (Y_{j_1}, W_{j_1-1}, \ldots, W_1, seed_{j_1}, \ldots, seed_1).$$

Consider the random variable $Z_{j_1} = (Y_{j_1}, \text{cond}(W_{j_1-1}, seed_{j_1-1}), \ldots, \text{cond}(W_1, seed_1))$ and note that

$$\tilde{H}_\infty(Z_{j_1}|seed_1, \ldots, seed_{j_1}) \geq \alpha'.$$

Applying a deterministic function does not increase statistical distance and thus,

$$(\text{cond}(W_{j_1}, seed_{j_1}), \text{cond}(W_{j_1-1}, seed_{j_1-1}), \ldots, \text{cond}(W_1, seed_1), seed_{j_1}, \ldots, seed_1) \approx_{n\text{cond}} (Z_{j_1}, seed_{j_1}, \ldots, seed_1)$$

By a hybrid argument there exists a distribution $Z$ with $\tilde{H}_\infty(Z|seed) \geq \tilde{\alpha}(n - \beta)$ where

$$(\text{cond}(W_n, seed_n), \ldots, \text{cond}(W_1, seed_1), seed_n, \ldots, seed_1) \approx_{n\text{cond}} (Z, seed_n, \ldots, seed_1).$$

This completes the proof of Theorem 5.9.