Electronic Colloquium on Computational Complexity, Report No. 154 (2024)

Improved Condensers for Chor-Goldreich Sources

Jesse Goodman[*]
The University of Texas at Austin
jpmgoodman@utexas.edu

Xin Li[†]
Johns Hopkins University
lixints@cs.jhu.edu

David Zuckerman[‡]
The University of Texas at Austin
diz@cs.utexas.edu

October 10, 2024

Abstract

One of the earliest models of weak randomness is the Chor-Goldreich (CG) source. A (t, n, k)-CG source is a sequence of random variables $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_t) \sim (\{0,1\}^n)^t$, where each $\mathbf{X}_i$ has min-entropy k conditioned on any fixing of $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}$. Chor and Goldreich proved that there is no deterministic way to extract randomness from such a source. Nevertheless, Doron, Moshkovitz, Oh, and Zuckerman showed that there is a deterministic way to condense a CG source into a string with small entropy gap. They gave applications of such a condenser to simulating randomized algorithms with small error and to certain cryptographic tasks. They studied the case where the block length n and entropy rate k/n are both constant.

We study the much more general setting where the block length can be arbitrarily large, and the entropy rate can be arbitrarily small. We construct the first explicit condenser for CG sources in this setting, and it can be instantiated in a number of different ways. When the entropy rate of the CG source is constant, our condenser requires just a constant number of blocks t to produce an output with entropy rate 0.9, say. In the low entropy regime, using $t = \text{poly}(n)$ blocks, our condenser can achieve output entropy rate 0.9 even if each block has just 1 bit of min-entropy. Moreover, these condensers have exponentially small error.

Finally, we provide strong existential and impossibility results. For our existential result, we show that a random function is a seedless condenser (with surprisingly strong parameters) for any small family of sources. As a corollary, we get new existential results for seeded condensers and condensers for CG sources. For our impossibility result, we show the latter result is nearly tight, by giving a simple proof that the output of any condenser for CG sources must inherit the entropy gap of (one block of) its input.

[*] Supported by a Simons Investigator Award (#409864, David Zuckerman).
[†] Supported by NSF CAREER Award CCF-1845349 and NSF Award CCF-2127575.
[‡] Supported by NSF Grant CCF-2312573 and a Simons Investigator Award (#409864).

ISSN 1433-8092

Contents

1 Introduction

Randomness is extremely useful in computing, yet it is difficult or expensive to obtain high-quality randomness. It is therefore important to understand what can be done with low-quality, or weak, random sources. Researchers have studied models of weak random sources for decades. One of the earliest models is the Chor-Goldreich (CG) source [CG88], which generalized the related Santha-Vazirani source [SV86].

Definition 1. *The* min-entropy *of a random variable* $\mathbf{X}$ *is given by* $H_\infty(\mathbf{X}) = \min_{x \in \mathit{support}(\mathbf{X})} \log_2(\frac{1}{\Pr[\mathbf{X}=x]})$. *We say* $\mathbf{X}$ *is an* (n,k) *source if* $\mathbf{X}$ *is over* $\{0,1\}^n$ *and has min-entropy* $H_\infty(\mathbf{X}) \geq k$.

Definition 2. *A random variable* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_t) \sim (\{0,1\}^n)^t$ *is called a* (t,n,k)*-CG source if for all* $i \in [t]$ *and all* $(x_1, \ldots, x_{i-1}) \in (\{0,1\}^n)^{i-1}$*, it holds that* $H_\infty(\mathbf{X}_i | \mathbf{X}_1 = x_1, \ldots, \mathbf{X}_{i-1} = x_{i-1}) \geq k$*. Each* $\mathbf{X}_i$ *is called a* block.

We would like to make use of a CG source knowing only the parameters t, n, and k. That is, our algorithms should work for all (t,n,k)-CG sources; an adversary can pick a (t,n,k)-CG source after seeing our algorithm.

The most natural way to use a weak source is to convert it into high quality randomness. However, generalizing the argument by Santha and Vazirani, Chor and Goldreich showed that it is impossible to deterministically extract even one nearly-uniform bit from a CG source (if $k \leq n-1$). It was therefore accepted by the community that one needed to add more randomness, either in the form of a random seed or a second CG source, to do anything useful.

That changed recently when Doron, Moshkovitz, Oh, and Zuckerman [DMOZ23] showed how to deterministically condense a CG source. Specifically, they showed how to efficiently output a string $\mathbf{Z} \sim \{0,1\}^m$ with small *entropy gap*, defined as $g := m - H_\infty(\mathbf{Z})$. (Strictly speaking, their condenser only outputs a string that is close in variation distance to a distribution with small entropy gap.)

Distributions with small entropy gap are useful in certain applications. They can be used to simulate algorithms with small error probability. They are also useful for unpredictability applications in cryptography. For example, they can be used as the input for a one-way function, and as the key to generate message authentication codes. Note that seeded extractors are not so useful in these applications, since cycling over seeds is not realistic in a cryptographic setting. For more on the utility of small entropy gap, see the work of Doron et al. [DMOZ23].

Thus, CG sources are intermediate in the following sense. A very general source, such as an (n,k)-source (which is a CG source with $t=1$), does not admit any deterministic condensing. Other, less general sources such as affine sources admit deterministic extraction. CG sources are one of the few models where we can do something extremely useful deterministically, even though we can't extract a single random bit.

Doron et al. construct their deterministic condenser by using the CG source to take a random walk on a lossless expander. They show that for any constant block length n, constant entropy rate k/n, and constant error ε, they can output a string that contains a constant fraction of the original entropy, and has a constant entropy gap.

In this paper, we study whether their results can be generalized to the case of a small number t of long blocks, as well as to subconstant entropy rate. This is natural and important for a few reasons. First, small t allows for much more general sources; indeed, $t=1$ gives the most general model of an (n,k)-source. It is interesting to find the most general model of a weak source where we can condense deterministically, and CG sources with few blocks seem like a natural candidate. Second, such CG sources often appear as intermediate objects in extractor constructions, where they are often called block sources. Third, long blocks seem even more likely to model natural defective random sources. It allows for more short-range correlations, and if there aren't too many long-range correlations then it should be a CG source with long blocks.

It appears hard to generalize the techniques of [DMOZ23] to work for long blocks. This is because known constructions of lossless expanders are not good enough. First, to obtain results for any entropy rate, Doron et al. had to use a two-level construction, where one level relied on a brute force construction of a small lossless expander. For long block lengths this is infeasible.

Second, for longer blocks, one could try higher degree lossless expanders, such as those by Guruswami, Umans, and Vadhan [GUV09]. However, the price of their extremely good lossless expansion is that the entropy gap becomes too large.

We study deterministic condensers for CG sources with few large blocks, and obtain improved results. Before describing our constructions, we briefly mention that we show the entropy gap g' in the output of any condenser for CG sources must always be at least the entropy gap $g = n - k$ of the last block $\mathbf{X}_t$ of the CG source. Thus, our goal is to ideally achieve $g' = O(g)$, while preserving almost all of the entropy.

1.1 Our results

Explicit constructions

For our main theorem, we construct the first explicit condenser for Chor-Goldreich sources that can be instantiated with any block length n, any min-entropy k, and any error ε. We present the general version of our condenser below, and then proceed to highlight two interesting instantiations.

Theorem 1 (Explicit condensers for CG sources)**.** *For any $\alpha > 0$, there is a constant $C \geq 1$ such that the following holds. For all $t, n \in \mathbb{N}$ and $\delta, \varepsilon > 0$, there is an explicit condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$ for $(t, n, k = \delta n = n - g)$-CG sources with output length $m = k' + g'$, output entropy $k' \geq (1-\alpha)kt$, output gap $g' \leq C \cdot (1/\delta)^C \cdot (g + \log(1/\varepsilon))$, and error ε.*

Thus, our explicit condenser is able to preserve 99% of the min-entropy, while achieving a gap that is only $\mathrm{poly}(1/\delta)$ times larger than the gap g of a single block. Moreover, there is no restriction on how the input parameters can be set, and we highlight two interesting settings below.

We first consider the case where the entropy rate δ is constant, as in [DMOZ23]. Here, we obtain qualitatively similar results, but ours works for arbitrarily large blocks (instead of constant-sized blocks) and has exponentially small error. Moreover, we only need the number of blocks t to be a large enough constant to output entropy rate 0.9. This constant is a polynomial in $1/\delta$.

Corollary 1. *For any constant $\delta > 0$, there exists a constant $C > 0$ such that the following holds. For any $t, n \in \mathbb{N}$, there exists an explicit condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^{k'+g'}$ for $(t, n, k := \delta n)$-CG sources, which has output entropy $k' \geq 0.99kt$, output gap $g' \leq Cn$, and error $\varepsilon = 2^{-n}$.*

Next, we dramatically improve the entropy requirement from $k = 0.01n$ to just $k = 1$, while the entropy gap grows by just a polynomial factor. As a result, we only need a polynomial number of blocks t to output entropy rate 0.9.

Corollary 2. *There exists a universal constant $C > 0$ such that the following holds. For any $t, n \in \mathbb{N}$, there exists an explicit condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^{k'+g'}$ for $(t, n, k := 1)$-CG sources, which has output entropy $k' \geq 0.99kt$, output gap $g' \leq n^C$, and error $\varepsilon = 2^{-n}$.*

In fact, looking at Theorem 1, our condenser can even handle CG sources that have min-entropy $k \ll 1$, while achieving error $\varepsilon \ll 2^{-n}$. However, it is worth pointing out that this result is only useful when the stated output gap g' is less than tg, since this is the original entropy gap in the input CG source.

Overall, as we mentioned, our condensers work for smaller entropy rates and larger blocks than those in [DMOZ23]. Moreover, our condensers achieve exponentially small error, while the constructions in [DMOZ23] have constant error. Nevertheless, their condenser does have some advantages over ours. First, their condenser works in an online manner, and ours doesn't. Second, they analyzed their condenser for almost-CG sources, and we haven't. That said, our condensers do extend to at least one notion of "almost," as we describe next.

Remark 1 (Explicit condensers for almost CG sources)**.** *Our explicit condensers can also be extended to certain notions of almost CG sources, such as* suffix-friendly *almost CG sources, as defined in [DMOZ23]. This is because such sources can be reduced to standard block sources, simply by grouping together blocks. While such a reduction will produce uneven block lengths (unlike standard CG sources), our constructions can easily be adapted to handle this more general setting.*

Existential results

We complement our explicit constructions with strong existential results. For our main existential result, we show that a random function is a seedless condenser (with surprisingly strong parameters) for any small family of sources. Throughout, we use capital letters to denote exponential versions of lower-case letters.[1]

Theorem 2 (Existential results for any small family)**.** *There exist universal constants $C, c > 0$ such that the following holds. Let $\mathcal{X}$ be a family of (n, k)-sources. For any $\ell \in [0, k]$ and $g > 0$ such that $m := k - \ell + g$ is an integer, and any $\varepsilon \in (0, 1]$, the following holds. If $|\mathcal{X}| \leq 2^{c\varepsilon K \psi}$, where*

$$\psi = \max\left\{g - \frac{1}{\lfloor L \rfloor}\log(1/\varepsilon) - C, \quad g - \frac{1}{\lfloor L \rfloor}\log(C2^g g/\varepsilon)\frac{C2^g}{g}\right\},$$

then there exists a condenser $\mathsf{Cond} : \{0,1\}^n \to \{0,1\}^m$ *for $\mathcal{X}$ with loss ℓ, gap g, and error ε.*

The above can be viewed as a condenser version of the classic result that there exist good seedless extractors for any small family of sources. In fact, it strictly generalizes it.[2] Overall, this result shows that condensers can handle much larger families of sources than extractors, while outputting much more of the original min-entropy. In particular, the classical existential result for extractors only works for families of size $2^{\Omega(\varepsilon^2 K)}$, and requires the extractor to lose $\ell = 2\log(1/\varepsilon)$ bits of min-entropy. The above result shows that condensers can handle families of size up to $2^{\Omega(g\varepsilon K)}$, provided the gap is of the form $g = O(\frac{1}{L}\log(1/\varepsilon))$. This means that allowing just $g = 1$ bit of gap can significantly increase the size of the family that can be handled, while decreasing the loss to $\ell = \log\log(1/\varepsilon) + O(1)$. Furthermore, the loss can be decreased all the way to $\ell = 0$, at the price of a slightly larger gap $g = O(\log(1/\varepsilon))$.[3]

As an immediate corollary, we get improved existential results for seeded condensers.

[1] For example, $L := 2^\ell, K := 2^k$, and so on.

[2] This is because the extractor case corresponds to the case where the error is $\varepsilon/2$ and the gap is $g = \varepsilon/2$, as this implies an error of ε and a gap of 0.

[3] In fact, note that this gap can be reduced to $g = 1\log(1/\varepsilon) + O(1)$ if we only wish to handle families of size $2^{\Omega(\varepsilon K)}$.

Corollary 3 (Existential results for seeded condensers)**.** *There is a universal constant $C \geq 1$ such that the following holds. There exists a seeded condenser* $\mathsf{sCond} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for (n,k)-sources with output length $m = k + d - \ell + g$, error ε, loss ℓ, and gap*

$$g \leq \frac{1}{\lfloor L \rfloor} \log(1/\varepsilon) + C,$$

provided that $d \geq \log(\frac{n-k}{\varepsilon}) + C$.

We note that we can improve the seed length requirement to $d \geq \log(\frac{n-k}{\varepsilon g})$, if one is willing to increase the gap to $g = \frac{2}{\lfloor L \rfloor} \log(1/\varepsilon) + C$.[4] Previously, a work of Aviv and Ta-Shma [AT19] established similar existential bounds for seeded condensers, but in the lossless regime $\ell = 0$ their result required entropy gap $g = O(\frac{\log(1/\varepsilon)}{\varepsilon})$, while we only require $g = O(\log(1/\varepsilon))$.[5]

For our last existential result, we show the existence of good condensers for Chor-Goldreich sources. Since the number of such sources is very large, we cannot apply Theorem 2 to obtain this result. Instead, we show that one can iteratively condense CG sources using seeded condensers (in the spirit of [NZ96], but with a correlated seed). Then, we plug in the seeded condensers from Corollary 3 to obtain the following, which we take some time to digest immediately after.

Corollary 4 (Existential results for CG sources)**.** *There is a constant $C \geq 1$ such that the following hold.*

- ***Two blocks:*** *There exists a condenser* $\mathsf{Cond} : (\{0,1\}^n)^2 \to \{0,1\}^m$ *for $(2, n, k = n-g)$-CG sources with output length $m = 2k - \ell + g$, error ε, loss ℓ, and gap*

  $$g' \leq g + \frac{1}{\lfloor L \rfloor} (g + \log(1/\varepsilon)) + C,$$

  provided that $k \geq \log(g/\varepsilon) + C$.

- ***More than two blocks:*** *There exists a condenser* $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$ *for $(t, n, k = n-g)$-CG sources with output length $m = kt + g'$, error ε, loss $\ell = 0$, and gap*

  $$g' \leq g + 2^{C(\log^* t)^2} \cdot (g + \log(1/\varepsilon) + C \log^* t),$$

  provided that $k \geq \log(g/\varepsilon) + C$.

  On the other hand, if $m = kt - \ell + g'$ and the loss is $\ell = 2(\log^* t)^2$, then one can obtain gap

  $$g' \leq g + C \cdot 2^{-\log^* t} \cdot (g + \log(1/\varepsilon)) + C \log^* t,$$

  provided that $k \geq \log(g/\varepsilon) + 2\log^* t + C$.

Thus, it is possible to condense Chor-Goldreich sources, even when there are just $t = 2$ blocks with logarithmic min-entropy. In the multi-block setting $t > 2$, we obtain a full tradeoff between the loss ℓ and gap g' (Section 6.3), but only highlight the extreme regimes above, for simplicity. In particular, the

[4] Moreover, Theorem 2 can be used to give a more general version of the above result, which recovers known existential results for seeded extractors, but we only present the above version for simplicity.

[5] It is worth noting that they focused on *strong* seeded condensers, while we focus on standard seeded condensers, since our result is just a corollary of our existential seedless condensers (Theorem 2), for which there is no notion of "strong." However, it should be relatively straightforward to extend our result to obtain strong seeded condensers, using standard tricks.

above shows that in the *lossless* regime $\ell = 0$, one can condense from multi-block CG sources with a modest multiplicative blow-up of $2^{O((\log^* t)^2)}$ in the gap (where $\log^*$ denotes the iterated logarithm). On the other hand, if one is willing to lose a little more min-entropy, this blow-up can be improved to an additive $O(\log^* t)$. Moreover, we note that at the expense of a significantly greater loss in min-entropy, it is possible to blow-up the gap by an additive constant (with no dependence on t), and refer the reader to Section 6.3.2 for more.

Impossibility results

Finally, we show a lower bound, which says that the gap in the CG source must propagate to the output.

Theorem 3 (Impossibility results for CG sources)**.** *Fix any $0 \le g \le m \le n \in \mathbb{N}$ and $\varepsilon \in [0, 1)$. For every function $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$, there exists a $(t, n, n-g)$-CG source $\mathbf{X}$ such that $\mathsf{Cond}(\mathbf{X})$ is ε-far from every $(m, m-g+c_\varepsilon)$-source, where $c_\varepsilon := \log(\frac{1}{1-\varepsilon})$.*[6]

This impossibility result is a strengthening of the fact that it is impossible to condense general (n, k)-sources, and was independently obtained by Chattopadhyay, Gurumukhani, and Ringach [CGR24].[7]

Organization The rest of this paper is organized as follows. We start with an overview of our techniques in Section 2. Then, after some preliminaries in Section 3, we provide a collection of (mostly new) tools and tricks around block sources in Section 4, which we'll use throughout the paper. In Section 5, we provide our main explicit condenser for Chor-Goldreich sources, and prove Theorem 1. Following this, we provide our existential results in Section 6 and our impossibility results in Section 7. Finally, we conclude with some open problems in Section 8.

2 Overview of our techniques

To begin, we give an informal overview of the techniques used in our constructions and proofs.

2.1 Explicit constructions

As discussed in the introduction, it seems difficult to extend the techniques of Doron et al. [DMOZ23] to obtain a condenser that can handle CG sources with long blocks. This is because their construction involves the use of excellent lossless expanders, which we don't know how to explicitly construct. They get around this problem by considering constant block length n, which allows them to obtain the lossless expanders via a brute-force search. But since we want to work with a larger block length, this is no longer possible. Thus, we need a new idea.

High-level plan

Our idea is to return to a classical paradigm in the construction of seedless extractors for independent sources, and show that it can be adapted to get seedless condensers for Chor-Goldreich sources. Intuitively,

[6] We remark that c_ε is an unavoidable term, since sources with 0 min-entropy are still ε-close to min-entropy c_ε.

[7] Beyond this impossibility result, there is little overlap between our two works, which will both appear at FOCS 2024. This is because we focus on explicit condensers for CG sources, whereas [CGR24] focuses on existential and impossibility results for almost CG sources, and other more general models.

this makes sense given that CG sources are a natural generalization of independent sources, and condensers are a natural generalization of extractors.

The well-known paradigm that we use involves taking a single independent source, expanding it into a table where one row is uniform (or has high entropy), and gradually collapsing that table (with the help of the other independent sources) until all that remains is that one good row [Ta-96,Rao09,BRSW12,Li13,Coh16]. Our goal is to extend this paradigm so that it still works even if the sources are not truly independent sources, but blocks coming from a CG source.

In order to make this happen, the core tool that we use is a simple observation, which says that every seeded condenser (and thus seeded extractor) still works even if its seed is "CG-correlated" with the source. In more detail, suppose that a seeded condenser was expecting to receive an (n, k)-source $\mathbf{X}$ and independent seed $\mathbf{Y} \sim \{0,1\}^d$ as input, but instead received an (n, k)-source $\mathbf{X}$ and a correlated seed $\mathbf{Y} \sim \{0,1\}^d$, which is only guaranteed have min-entropy $d - g$ on each fixing of $\mathbf{X}$. The core tool we use says that the error of the seeded condenser blows up from $\varepsilon \to \varepsilon 2^g$, while its output gap g' blows up from $g' \to g' + g$. This key observation has appeared a few times in prior work, with slightly weaker parameters [BCDT19] or in a slightly different context [BGM22]. We record it as Lemma 8.

By combining the paradigm for extracting from independent sources with the above tool, we now have a very high-level plan for condensing CG sources with long blocks. However, several challenges arise along the way, since the known tools for collapsing the table (such as non-malleable extractors and mergers)[8] cannot be ported over in a black-box manner. Instead, we must construct our own non-malleable condensers (for CG sources) from scratch, and we do so via a simple composition of seeded extractors. With this high-level plan in mind, we proceed with a more detailed description of our condenser.

A detailed description of our condenser

Given a CG source (or block source)[9] $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_t)$, we work backwards starting with $\mathbf{X}_t$, and try to extract (almost) all of the entropy out of $\mathbf{X}$ while preserving a small entropy gap. To get the condenser, we first convert the block $\mathbf{X}_t$ into a *somewhere high-entropy source*, which is a table with some number of rows such that at least one row has high entropy rate. If the entropy rate of $\mathbf{X}_t$ is relatively large (e.g., any constant), we can use well-known somewhere condensers [BKS+10, Raz05], which produce a small (constant) number of rows with exponentially small error.

Our next step is to use the other blocks to reduce the number of rows in this table, while preserving almost all of the entropy. If the blocks were independent, prior work shows that we could eventually reduce the table to a single row that is close to uniform (which gives an extractor). However, when we only have a block source, for technical reasons we'll soon explain, this is no longer possible and eventually we get one row with large entropy and small entropy gap. This gives a condenser.

The non-malleable condenser The key ingredient in achieving this is a new merger (or *non-malleable condenser*, similar to those defined in [Li12, Li15]), that we design to merge two rows in the table while using a few additional blocks. Our final condenser is obtained by repeatedly applying this merger, until the number of rows in the table reduces to one. To illustrate our ideas, let's consider the simplified case of merging two rows in the table (say $\mathbf{Y}_1, \mathbf{Y}_2$), where at least one row is uniform (but we don't know which

[8] Here, a non-malleable extractor can roughly be thought of as a seeded extractor $\mathsf{sExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ that offers an additional "robustness" guarantee. This robustness guarantee says that the output of sExt not only looks uniform, but also looks *independent* of (an output produced by) an additional call to sExt on the same source and a correlated seed. Such an object can be used to break the correlation between pairs of rows in the table, while keeping the good row looking uniform (or high-entropy).

[9] Recall that a block source is a generalization of a CG source in that the blocks need not have the same length.

one). Our basic merger works as follows. First take two other blocks (say $\mathbf{X}_1, \mathbf{X}_2$), and take two slices: $\mathbf{Z}_1$ from $\mathbf{Y}_1$ and $\mathbf{Z}_2$ from $\mathbf{Y}_2$. We use a standard seeded extractor sExt which works even when the seed only has entropy rate 0.9 [GUV09], and compute $\mathbf{W}_1 = \mathsf{sExt}(\mathbf{X}_2, \mathbf{Z}_1), \mathbf{W}_2 = \mathsf{sExt}(\mathbf{X}_2, \mathbf{Z}_2)$. In this step we make sure that the sizes of these random variables satisfy

$$|\mathbf{W}_1| \gg |\mathbf{W}_2| \gg |\mathbf{Z}_2| \gg |\mathbf{Z}_1|.$$

Next we apply the seeded extractor again and compute $\mathbf{S}_1 = \mathsf{sExt}(\mathbf{X}_1, \mathbf{W}_1)$ and $\mathbf{S}_2 = \mathsf{sExt}(\mathbf{X}_1, \mathbf{W}_2)$, where $|\mathbf{S}_1| = |\mathbf{S}_2|$. Finally, we output $\mathbf{V} = \mathbf{S}_1 \oplus \mathbf{S}_2$, where $\oplus$ denotes bit-wise XOR.

The analysis For the analysis, let us first consider the case where all the blocks are *independent*. We have two cases. If $\mathbf{Y}_1$ is the uniform row, then $\mathbf{Z}_1$ is uniform and therefore $\mathbf{W}_1$ is uniform (ignoring the error of the extractor). Since $|\mathbf{W}_1| \gg |\mathbf{W}_2| \gg |\mathbf{Z}_2|$, we can fix $\mathbf{Z}_2$ and $\mathbf{W}_2$, and conditioned on this fixing, (with high probability) $\mathbf{W}_1$ still has entropy rate say > 0.9. Notice at this point, $\mathbf{X}_1$ is still independent of $\mathbf{W}_1$. Now we can further fix $\mathbf{S}_2 = \mathsf{sExt}(\mathbf{X}_1, \mathbf{W}_2)$, which is a deterministic function of $\mathbf{X}_1$ since $\mathbf{W}_2$ is fixed. Conditioned on this fixing, $\mathbf{X}_1$ still has good entropy as long as the size of $\mathbf{S}_2$ is not too large. Hence, $\mathbf{S}_1 = \mathsf{sExt}(\mathbf{X}_1, \mathbf{W}_1)$ is close to uniform and so is $\mathbf{V} = \mathbf{S}_1 \oplus \mathbf{S}_2$.

In the other case, $\mathbf{Y}_2$ is the uniform row, and thus $\mathbf{Z}_2$ is uniform.[10] Now we first fix $\mathbf{Z}_1$ and $\mathbf{W}_1$. Notice that since $|\mathbf{Z}_2| \gg |\mathbf{Z}_1|$, conditioned on this fixing $\mathbf{Z}_2$ still has entropy rate > 0.9. Furthermore when $\mathbf{Z}_1$ is fixed, $\mathbf{W}_1$ is a deterministic function of $\mathbf{X}_2$. Thus conditioned on the further fixing of $\mathbf{W}_1$, $\mathbf{X}_2$ and $\mathbf{Z}_2$ are still independent while $\mathbf{X}_2$ still has good entropy as long as the size of $\mathbf{W}_1$ is not too large. Therefore $\mathbf{W}_2 = \mathsf{sExt}(\mathbf{X}_2, \mathbf{Z}_2)$ is close to uniform even conditioned on $\mathbf{W}_1$. Now, we can further fix $\mathbf{S}_1 = \mathsf{sExt}(\mathbf{X}_1, \mathbf{W}_1)$, which is a deterministic function of $\mathbf{X}_1$ since $\mathbf{W}_1$ is fixed. Conditioned on this fixing, $\mathbf{X}_1$ is still independent of $\mathbf{W}_2$ and still has good entropy as long as the size of $\mathbf{S}_1$ is not too large. Hence, $\mathbf{S}_2 = \mathsf{sExt}(\mathbf{X}_1, \mathbf{W}_2)$ is close to uniform and so is $\mathbf{V} = \mathbf{S}_1 \oplus \mathbf{S}_2$.

Extending the analysis to correlated blocks Now let us see what happens if the blocks are not independent, but rather form a block source. We will again use the core observation that for any seeded extractor, if the seed and the source form a block source, then the output of the extractor becomes a source that suffers roughly the same entropy gap as the seed.

With this property in hand, our previous analysis can go through with a few modifications. Most importantly, some of the random variables in $\{\mathbf{W}_1, \mathbf{W}_2, \mathbf{S}_1, \mathbf{S}_2, \mathbf{V}\}$ will no longer be uniform, since the entropy gap of the seed will be inherited in the output when we apply a seeded extractor. In addition, we need to set the errors in the extractors appropriately so that the blow up factor 2^g in the error can be absorbed. Finally, in the analysis, when we fix certain random variables (e.g., $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{W}_1, \mathbf{W}_2$), this may affect other blocks besides the block from which the random variable is produced, because now the blocks are no longer independent. However, as long as we keep the sizes of the random variables relatively small compared to the entropy in each block, after conditioning the blocks still have enough entropy left and thus they are still (close to) a block source.

Tracking the gap Notice that if initially the "good" row in $\mathbf{Y}_1, \mathbf{Y}_2$ has some entropy gap g', then the final entropy gap g'' of our basic merger will be a constant factor larger than g', due to our conditioning argument and the requirement that the seed used in the extractor has entropy rate > 0.9. Therefore when

[10] Note that we are only analyzing the case where at least one of $\mathbf{Y}_1, \mathbf{Y}_2$ are uniform, since this is true for some pair of consecutive rows in the table, and we don't care what happens when we merge other pairs of consecutive rows.

we repeatedly apply the basic merger, the entropy gap will increase by a constant factor at each step. As a result, the final entropy gap will become larger than g, the entropy gap of each block in the CG source. To see this, consider the case where $k = \delta n$ for some constant $\delta > 0$ and we start with a somewhere high-entropy condenser as in [BKS+10, Raz05]. If we boost the entropy rate to say 0.99, then the initial entropy gap of the good row will be $\mathrm{poly}(\delta)g$, since the length of each row in the table is $\mathrm{poly}(\delta)n$. However, the table itself also has $\mathrm{poly}(1/\delta)$ rows, so by the above analysis, we eventually get an entropy gap of $C^{\log(\mathrm{poly}(1/\delta))}\,\mathrm{poly}(\delta)g = \mathrm{poly}(1/\delta)g$, since each time we use the non-malleable condenser, we halve the number of rows in the table, but blow up the gap by a constant factor C. Thus if δ was originally a constant, the final entropy gap is $\mathrm{poly}(1/\delta)g = O(g)$. Note that this is as expected since our impossibility result shows that an entropy gap of g is necessary.

This also results in another modification we must make in our constructions. Specifically, when the entropy gap becomes large in the process of repeated merging, in order to obtain a seed for the extractor with entropy rate > 0.9, it is no longer enough to just use the entropy from one block. Rather, at this point we need to use the concatenation of several blocks as the source in the seeded extractor to get sufficient entropy. In doing so, we need a slightly larger seed length over time, but this does not drastically change any of the parameters. In fact, since the seed length must grow anyway, we can (for free) force the error of each merging step to be half the error of the prior step, resulting in a geometric series of errors. As a result, the overall error of the condenser is simply the error of the first (somewhere-condensing) step, which is just $2^{-\mathrm{poly}(\delta)n}$ if we start with a $(t, n, \delta n)$-CG source.

Pre- and post-processing Finally, we have just two loose ends that we need to tie up. First, the discussion above assumed that we started with a $(t, n, \delta n)$-CG source for some constant $\delta > 0$, so that we could apply the somewhere-condensers of [BKS+10, Raz05] to create the initial table. However, what if we want to condense from CG sources with sub-constant δ? As it turns out (and is well-known), these somewhere-condensers can actually handle an input source of min-entropy $\delta n = n^{0.99}$, and thus the construction can still be applied even if we start off with a CG-source with $\delta n = n^{0.99}$. More importantly, if we start off with a $(t, n, \delta n)$-CG source with $\delta n \ll n^{0.99}$, we can always turn it into a $(t/b, nb, \delta nb)$-CG source with $\delta nb \geq (nb)^{0.99}$, via a *pre-processing step*, where we simply group the blocks into "super-blocks" containing sufficiently many blocks b each. This will slightly impact the parameters of our condenser, but not enough to be noticeable (when compared to the impact of the other steps).

Second, the discussion above gave a detailed overview of how we can condense the CG source into a string $\mathbf{Z}$ with high entropy rate, but what if this was done using a relatively small number of blocks in the CG source, and most of the entropy in the CG source still remains (i.e., in the unused blocks $\mathbf{X}^\star$)? To deal with this, we append a simple *post-processing* step to our condenser. As it turns out, since we already have obtained a (perhaps short) string with high entropy rate, it is relatively easy to condense the rest of the min-entropy out of the CG source. Indeed, since the entropy rate is so high, we can use our core tool that a seeded extractor can handle CG-correlated seeds, and suffer very little loss. Thus, a first attempt to get the rest of the min-entropy out may involve calling a seeded extractor with $\mathbf{X}^\star$ as the source and $\mathbf{Z}$ as the seed. However, it may be the case that $\mathbf{X}^\star$ is extremely long compared to $\mathbf{Z}$, which would make this approach fail. Instead, the right approach is to use classic block-source extraction framework of Nisan and Zuckerman [NZ96], or rather a slight generalization that works for condensers and a CG-correlated seed. With this approach, we can successfully condense the rest of the min-entropy out of $\mathbf{X}^\star$, even if $\mathbf{Z}$ is very tiny in comparison.

2.2 Existential results

Next, we briefly discuss the ideas that go into our existential results. As a reminder, our main result shows that there exist great seedless condensers for any small enough family $\mathcal{X}$ of sources. As a corollary (i.e., by picking the appropriate family $\mathcal{X}$), we immediately get our existential results for seeded condensers. Then, by plugging these seeded condensers into the (slight generalization of the) block-source extraction framework described in the paragraph above, we immediately get our existential results for CG sources. Thus, all that remains is to show that there exist great seedless condensers for any small family of sources.

In order to show the above, we show that a random function $f : \{0,1\}^n \to \{0,1\}^m$ is, with high probability, a great seedless condenser for a single source $\mathbf{X} \sim \{0,1\}^n$ (and apply a union bound over all $\mathbf{X} \in \mathcal{X}$). As it turns out, if one wishes to get good parameters, this is quite nontrivial to show.

The overall approach is as follows. First, we recall that a random variable $f(\mathbf{X}) \sim \{0,1\}^m$ is ε-close (in statistical distance) to min-entropy k' iff for every $S \subseteq \{0,1\}^m$, it holds that

$$\Pr[f(\mathbf{X}) \in S] \leq |S| \cdot 2^{-k'} + \varepsilon.$$

Thus, it is tempting to fix a set S, show that the above is true with high probability over f, and then union bound over all $S \subseteq \{0,1\}^m$. However, there are simply too many sets S for this to yield good parameters.

As a second approach, one may recall a classical lemma (in, e.g., [GUV09, Lemma 6.2]), which says that if you want to ensure that $f(\mathbf{X})$ is ε-close to min-entropy k', it is enough to show that there exists no *small* set $S \subseteq \{0,1\}^m$ of size $\leq \varepsilon 2^{k'}$ such that

$$\Pr[f(\mathbf{X}) \in S] \geq \varepsilon.$$

This is much better, since we have greatly reduced the number of sets $S \subseteq \{0,1\}^m$ that we ultimately need to union bound over. However, we can still do even better.

The key realization (which is inspired by existence proofs for lossless condensers) is that we can specify $S \subseteq \{0,1\}^m$ by instead specifying its *preimage* $f^{-1}(S)$. Thus, instead of counting sets from $\{0,1\}^m$ (for the union bound), we can count sets from support($\mathbf{X}$). This is much better when support($\mathbf{X}$) $\ll 2^m$, which happens when we are targetting a regime where the gap of the condenser will need to exceed the loss (e.g., the lossless regime) and $\mathbf{X}$ is flat.[11] But what if $\mathbf{X}$ is not flat? When talking about *seeded condensers*, one can often assume that $\mathbf{X}$ is flat for free. But this is not true for *seedless condensers* (for an explanation why, see Section 6.1).

In order to deal with an (n, k)-source $\mathbf{X}$ that may not be flat, we break its support into two parts X_1, X_2. We pick some threshold T and let X_1 contain the heaviest T elements in support($\mathbf{X}$), while X_2 contains the rest. Then, instead of analyzing the performance of f on the entirety of $\mathbf{X}$, we analyze it on the subdistributions of $\mathbf{X}$ over X_1 and X_2 (and make sure that the images of X_1 and X_2 do not interact too much). If we pick the threshold T correctly, then the subdistribution on X_1 will look roughly flat, while the subdistribution on X_2 has much higher entropy than $\mathbf{X}$. This is exactly what we want, because the former allows us to safely count tests via their preimages in X_1, while the latter allows us to safely count tests by picking them from $\{0,1\}^m$ (since f will be nowhere close to the lossless regime for the subdistribution on X_2, as it has much higher min-entropy than $\mathbf{X}$). All that remains is to ensure that the images of X_1 and X_2 do not interact too much, which follows without too much additional trouble.

[11] As a reminder, a flat source is uniform over its support.

2.3 Impossibility results

Finally, our impossibility result for condensing CG sources is a simple extension and generalization of the well-known impossibility result for extracting from CG sources [CG88], which uses backwards induction on the blocks. Indeed, the latter can be viewed as a special case of the former.

3 Preliminaries

Before we dive into our main proofs, we collect some preliminaries that will be used throughout the paper.

Notation We adopt the convention that capital letters denote the exponential version of lower-case letters. For example, $N := 2^n, D := 2^d$, and so on. Given a string $x \in \{0,1\}^n$, we let x_i denote the value it holds at its i^{th} index, and for a set $S \subseteq [n]$ we let x_S denote $(x_i)_{i \in S}$ (concatenated in increasing order of i). We also use $x_{<i}$ as shorthand for $x_{[1,i-1]}$, and we define $x_{\leq i}, x_{>i}$, and $x_{\geq i}$ similarly. All logs are base 2, unless otherwise noted. In particular, we write $\log() := \log_2()$ and $\ln() := \log_e()$.

3.1 Probability

We use bold letters, such as $\mathbf{X}$, to refer to random variables (which we often call *sources*). We let $\mathbf{U}_n$ denote the uniform random variable over $\{0,1\}^n$, and more generally say that a random variable $\mathbf{X}$ is *flat* if it is uniform over its support. Furthermore, if support$(\mathbf{X}) \subseteq V$, we say that $\mathbf{X}$ is supported on V and denote this by $\mathbf{X} \sim V$. Finally, for any two random variables $\mathbf{X}, \mathbf{Y}$ defined over the same space, and $y \in$ support$(\mathbf{Y})$, we let $(\mathbf{X} \mid \mathbf{Y} = y)$ denote a random variable that hits x with probability $\Pr[\mathbf{X} = x \mid \mathbf{Y} = y]$.

Statistical distance

Next, we introduce a standard way to measure the distance between two random variables.

Definition 3 (Statistical distance)**.** *The* statistical distance *between random variables* $\mathbf{X}, \mathbf{Y} \sim V$ *is defined*

$$|\mathbf{X} - \mathbf{Y}| := \max_{S \subseteq V} \Pr[\mathbf{X} \in S] - \Pr[\mathbf{Y} \in S] = \frac{1}{2} \sum_{v \in V} |\Pr[\mathbf{X} = v] - \Pr[\mathbf{Y} = v]|.$$

We say that $\mathbf{X}, \mathbf{Y}$ *are* ε*-close and write* $\mathbf{X} \approx_\varepsilon \mathbf{Y}$ *iff* $|\mathbf{X} - \mathbf{Y}| \leq \varepsilon$*. If* $\mathbf{X}, \mathbf{Y}$ *are* 0*-close then we write* $\mathbf{X} \equiv \mathbf{Y}$*. If* $\mathbf{X}, \mathbf{Y}$ *are not* ε*-close, we say they are* ε*-far and write* $\mathbf{X} \not\approx_\varepsilon \mathbf{Y}$*.*

Statistical distance is a metric, which means that it satisfies the triangle inequality.

Fact 1 (Triangle inequality)**.** *For any random variables* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \sim V$*,*

$$|\mathbf{X} - \mathbf{Z}| \leq |\mathbf{X} - \mathbf{Y}| + |\mathbf{Y} - \mathbf{Z}|.$$

Throughout this paper, we will often want to bound the statistical distance between random variables. A classic tool for this is the following.

Fact 2 (Data-processing inequality)**.** *For any random variables* $\mathbf{X}, \mathbf{Y} \sim V$ *and function* $f : V \to W$*,*

$$|\mathbf{X} - \mathbf{Y}| \geq |f(\mathbf{X}) - f(\mathbf{Y})|.$$

Another tool that is useful for bounding statistical distance is the coupling lemma:

Lemma 1 (Coupling lemma)**.** *For any two random variables* $\mathbf{X}, \mathbf{Y} \sim V$*, the following holds. For every pair of jointly distributed random variables* $(\mathbf{X}', \mathbf{Y}')$ *with* $\mathbf{X}' \equiv \mathbf{X}$ *and* $\mathbf{Y}' \equiv \mathbf{Y}$*, it holds that*

$$|\mathbf{X} - \mathbf{Y}| \leq \Pr[\mathbf{X}' \neq \mathbf{Y}'].$$

Moreover, there exists a pair of jointly distributed random variables $(\mathbf{X}^\star, \mathbf{Y}^\star)$ *with* $\mathbf{X}^\star \equiv \mathbf{X}$ *and* $\mathbf{Y}^\star \equiv \mathbf{Y}$ *such that*

$$|\mathbf{X} - \mathbf{Y}| = \Pr[\mathbf{X}^\star \neq \mathbf{Y}^\star].$$

Convex combinations

We will also frequently use the notion of convex combinations. We say $\mathbf{X}$ is a convex combination of distributions from $\mathcal{Y}$ if there exist probabilities $\{p_i\}$ summing to 1 and distributions $\mathbf{Y}_i \in \mathcal{Y}$ such that $\mathbf{X} = \sum_i p_i \mathbf{Y}_i$, meaning that $\mathbf{X}$ samples from $\mathbf{Y}$ with probability p_i. The following fact will be quite useful.

Fact 3. *Let* $\mathbf{X} \sim V$ *and* $\mathbf{A} \sim W$ *be (arbitrarily correlated) random variables, and let* $\mathcal{X}$ *be a family of random variables over* V*. Suppose that* $\Pr_{a \sim \mathbf{A}}[\mathbf{X} \notin \mathcal{X} \mid \mathbf{A} = a] \leq \varepsilon$*. Then* $\mathbf{X}$ *is* ε*-close to a convex combination of random variables from* $\mathcal{X}$*.*

Proof. For every fixed a such that $(\mathbf{X} \mid \mathbf{A} = a) \in \mathcal{X}$, define $\mathbf{Y}^a := (\mathbf{X} \mid \mathbf{A} = a)$. For all other a, define $\mathbf{Y}^a$ to be an arbitrary member of $\mathcal{X}$. Consider the convex combination $\mathbf{Y}^\star := \sum_a \Pr[\mathbf{A} = a] \cdot \mathbf{Y}^a$. Clearly, it is a convex combination of distributions from $\mathcal{X}$. It is also straightforward to verify $\mathbf{Y}^\star \approx_\varepsilon \mathbf{X}$. □

Concentration bounds

Finally, we will use the following version of the multiplicative Chernoff bound, which works even if we only know an upper bound on the expectation of the random variable of interest.

Theorem 4 (Chernoff bound)**.** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be a sequence of independent random variables, where each* $\mathbf{X}_i \sim \{0, p_i\}$ *for some* $p_i \in [0, 1]$*, and let* $\mathbf{X} := \sum_i \mathbf{X}_i$ *denote their sum. Then for any* $\delta > 0$ *and* $\mu \geq \mathbb{E}[\mathbf{X}]$*,*

$$\Pr[\mathbf{X} \geq (1 + \delta)\mu] \leq \left(\frac{e^\delta}{(1 + \delta)^{(1+\delta)}} \right)^\mu .$$

3.2 Entropy

In extractor theory, the standard way to measure the randomness content of a source is via its *min-entropy*.

Definition 4 (Min-entropy)**.** *The* min-entropy *of a random variable* $\mathbf{X} \sim \{0, 1\}^n$ *is defined*

$$H_\infty(\mathbf{X}) := \min_{x \in \mathit{support}(\mathbf{X})} \log \left(\frac{1}{\Pr[\mathbf{X} = x]} \right),$$

while its min-entropy gap *is defined as* $n - H_\infty(\mathbf{X})$*.*[12]

[12] For convenience, from here on out, whenever we say “entropy” we really mean “min-entropy.”

It is often the case that a random variable $\mathbf{X} \sim \{0,1\}^n$ does not exactly have high min-entropy, but is (statistically) close to a random variable that does. In many applications, this is just as good as $\mathbf{X}$ having high min-entropy itself, and as a result, this notion has earned its own name: *smooth min-entropy*. In order to formally introduce this definition, we first let $\mathcal{B}_\varepsilon(\mathbf{X})$ denote the set of random variables $\mathbf{Y} \sim \{0,1\}^n$ that are ε-close to $\mathbf{X}$ in statistical distance. Then, we define smooth min-entropy as follows.

Definition 5 (Smooth min-entropy)**.** *The* ε-smooth min-entropy *of a random variable* $\mathbf{X} \sim \{0,1\}^n$ *is defined as*

$$H_\infty^\varepsilon(\mathbf{X}) := \sup_{\mathbf{Y} \in \mathcal{B}_\varepsilon(\mathbf{X})} H_\infty(\mathbf{Y}) = \max_{\mathbf{Y} \in \mathcal{B}_\varepsilon(\mathbf{X})} H_\infty(\mathbf{Y}).$$

Looking at this definition, a few remarks are in order. First, we note that we were able to replace the supremum with a maximum due to standard tools from analysis.[13] (In doing so, we know there always exists some distribution $\mathbf{Y} \approx_\varepsilon \mathbf{X}$ such that $H_\infty(\mathbf{Y}) = H_\infty^\varepsilon(\mathbf{X})$.) Next, in order to highlight that smooth min-entropy is a weaker notion than standard min-entropy (and thus easier to obtain), we point out that there are other well-studied notions of entropy that imply much better guarantees on the former than the latter.[14] Finally, we mention two strange artifacts of the above definition, which distinguish it from other notions of entropy: First, note that a constant random variable has ε-smooth min-entropy $c_\varepsilon := \log(\frac{1}{1-\varepsilon})$, which is > 0 for $\varepsilon > 0$. Second, notice that when ε is large, the smooth min-entropy of $\mathbf{X}$ can actually depend on the *ambient space* on which $\mathbf{X}$ was defined! While this may seem concerning at first, one may find comfort in thinking of smooth min-entropy simply as convenient shorthand for the expression in Definition 5, instead of as a true "entropy."

Next, we record a very useful characterization of smooth min-entropy, which will be used throughout. This has appeared a few times in prior work, albeit in slightly different forms (see, e.g., [Zuc07, Lemma 2.2] or [GUV09, Lemma 6.2]).

Lemma 2 (A characterization of smooth min-entropy)**.** *For any* $\mathbf{X} \sim \{0,1\}^n$ *and* $k \leq n$*,*

$$H_\infty^\varepsilon(\mathbf{X}) \geq k \iff \forall S : \Pr[\mathbf{X} \in S] \leq |S| \cdot 2^{-k} + \varepsilon.$$

Proof. ($\Longrightarrow$) Let $\mathbf{X}' \sim \{0,1\}^n$ be a source of min-entropy at least k such that $\mathbf{X}' \approx_\varepsilon \mathbf{X}$. Then for any S,

$$\Pr[\mathbf{X} \in S] \leq \Pr[\mathbf{X}' \in S] + \varepsilon \leq |S| \cdot 2^{-k} + \varepsilon.$$

($\Longleftarrow$) Let Heavy be the set of elements that $\mathbf{X}$ assigns probability $> 2^{-k}$, and let $\mathsf{Light} := \{0,1\}^n \setminus \mathsf{Heavy}$. Notice that since $n \geq k$, we have $\Pr[\mathbf{X} \in \mathsf{Heavy}] - 2^{-k} \cdot |\mathsf{Heavy}| \leq 2^{-k} \cdot |\mathsf{Light}| - \Pr[\mathbf{X} \in \mathsf{Light}]$. In other words, there is a way to shift the excess weight that $\mathbf{X}$ assigns to Heavy onto Light without going over probability 2^{-k} on any of these elements. Let $\mathbf{X}' \sim \{0,1\}^n$ denote this new source, and note $H_\infty(\mathbf{X}') = k$. By our construction of $\mathbf{X}'$ and the hypothesis, we have

$$\begin{aligned}
|\mathbf{X} - \mathbf{X}'| &= \max_S |\Pr[\mathbf{X} \in S] - \Pr[\mathbf{X}' \in S]| \\
&= \Pr[\mathbf{X} \in \mathsf{Heavy}] - \Pr[\mathbf{X}' \in \mathsf{Heavy}] \\
&\leq |\mathsf{Heavy}| \cdot 2^{-k} + \varepsilon - |\mathsf{Heavy}| \cdot 2^{-k} \\
&\leq \varepsilon,
\end{aligned}$$

[13] In particular, one can argue that $\mathcal{B}_\varepsilon(\mathbf{X})$ is closed and bounded, and by the Heine-Borel theorem for finite-dimensional normed vector spaces, it is also compact. Then, since the min-entropy function $H_\infty()$ is continuous, $H_\infty(\mathcal{B}_\varepsilon(\mathbf{X}))$ is also compact (and therefore closed and bounded). It follows that $\sup H_\infty(\mathcal{B}_\varepsilon(\mathbf{X})) \in H_\infty(\mathcal{B}_\varepsilon(\mathbf{X}))$, allowing us to replace sup with max.

[14] Consider the Rényi entropy of a random variable, defined $H_2(\mathbf{X}) := \log\left(\frac{1}{\sum_x \Pr[\mathbf{X}=x]^2}\right)$. Comparing this to min-entropy, we have $H_\infty(\mathbf{X}) \geq \frac{1}{2}H_2(\mathbf{X})$, but if we compare this to smooth min-entropy, it is known that $H_\infty^\varepsilon(\mathbf{X}) \geq H_2(\mathbf{X}) - \log(1/\varepsilon)$ [RW04, Lemma 4.2]. A similar connection was used in [DMOZ23] in order to use the ℓ_q norm as a proxy for smooth min-entropy.

as desired. □

In fact, notice that the proof of the lemma above actually proved the following stronger result.

Lemma 3 (A characterization of smooth min-entropy)**.** *For any* $\mathbf{X} \sim \{0,1\}^n$*,* $0 \leq k \leq n$*, and* $\varepsilon \in [0,1]$*, the following holds. If we define* $\mathsf{Heavy} := \left\{x \in \{0,1\}^n : \Pr[\mathbf{X} = x] > 2^{-k}\right\}$*, then we have the equivalence*

$$H_\infty^\varepsilon(\mathbf{X}) \geq k \iff \Pr[\mathbf{X} \in \mathsf{Heavy}] \leq |\mathsf{Heavy}| \cdot 2^{-k} + \varepsilon.$$

Finally, we record one technical lemma that will be useful later on.

Claim 1. *Consider any random variables* $\mathbf{A}, \mathbf{A}' \sim A$ *and* $\mathbf{B}, \mathbf{B}' \sim B$ *such that* $(\mathbf{A}, \mathbf{B}) \approx_\varepsilon (\mathbf{A}', \mathbf{B}')$*. Then*

$$\Pr_{a \sim \mathbf{A}}[H_\infty^\gamma(\mathbf{B} \mid \mathbf{A} = a) < k] \leq \Pr_{a \sim \mathbf{A}'}[H_\infty^{\gamma/2}(\mathbf{B}' \mid \mathbf{A}' = a) < k] + 4\varepsilon/\gamma + \varepsilon.$$

Proof. Let $\mathsf{BAD} := \{a : H_\infty^\gamma(\mathbf{B} \mid \mathbf{A} = a) < k\}$, let $\mathsf{BAD}' := \{a : H_\infty^{\gamma/2}(\mathbf{B}' \mid \mathbf{A}' = a) < k\}$, and define $S := \mathsf{BAD} \setminus \mathsf{BAD}'$. Note that

$$\begin{aligned}
\Pr_{a \sim \mathbf{A}}[H_\infty^\gamma(\mathbf{B} \mid \mathbf{A} = a) < k] &= \Pr_{a \sim \mathbf{A}}[a \in \mathsf{BAD}] \\
&\leq \Pr_{a \sim \mathbf{A}}[a \in S] + \Pr_{a \sim \mathbf{A}}[a \in \mathsf{BAD}'] \\
&\leq \Pr_{a \sim \mathbf{A}}[a \in S] + \Pr_{a \sim \mathbf{A}'}[a \in \mathsf{BAD}'] + \varepsilon \\
&= \Pr_{a \sim \mathbf{A}}[a \in S] + \Pr_{a \sim \mathbf{A}'}[H_\infty^{\gamma/2}(\mathbf{B}' \mid \mathbf{A}' = a) < k] + \varepsilon,
\end{aligned}$$

and thus all that remains is to bound $\Pr_{a \sim \mathbf{A}}[a \in S]$. Towards this end, notice that for every $a \in S$, it holds that $H_\infty^\gamma(\mathbf{B} \mid \mathbf{A} = a) < k$ and $H_\infty^{\gamma/2}(\mathbf{B}' \mid \mathbf{A}' = a) \geq k$. In other words, $(\mathbf{B}' \mid \mathbf{A}' = a)$ is $(\gamma/2)$-close to the family $\mathcal{X}$ of sources with min-entropy at least k, yet $(\mathbf{B} \mid \mathbf{A} = a)$ has distance $> \gamma$ from this same family. By the triangle inequality, this means $(\mathbf{B} \mid \mathbf{A} = a)$ has distance $> \gamma/2$ from $(\mathbf{B}' \mid \mathbf{A}' = a)$.

Now, define $p := \Pr[\mathbf{A} \in S]$, and partition S into subsets X_1, X_2 such that $\Pr[\mathbf{A} = a] \geq \Pr[\mathbf{A}' = a]$ for all $a \in X_1$, and $\Pr[\mathbf{A} = a] < \Pr[\mathbf{A}' = a]$ for all $a \in X_2$. Since X_2, X_2 is a partition, it must hold that either $\Pr[\mathbf{A} \in X_1] \geq p/2$ or $\Pr[\mathbf{A} \in X_2] \geq p/2$. Suppose the former is true, and recall that for all $a \in X_1 \subseteq S$, it holds that $(\mathbf{B} \mid \mathbf{A} = a) \not\approx_{\gamma/2} (\mathbf{B}' \mid \mathbf{A}' = a)$. By definition of statistical distance, this means that for every $a \in X_1$ there is a set Q_a such that $\Pr[(\mathbf{B} \mid \mathbf{A} = a) \in Q_a] - \Pr[(\mathbf{B}' \mid \mathbf{A}' = a) \in Q_a] \geq \gamma/2$. Thus

$$\begin{aligned}
|(\mathbf{A}, \mathbf{B}) - (\mathbf{A}', \mathbf{B}')| &\geq \sum_{a \in X_1, b \in Q_a} \Pr[(\mathbf{A}, \mathbf{B}) = (a, b)] - \Pr[(\mathbf{A}', \mathbf{B}') = (a, b)] \\
&= \sum_{a \in X_1, b \in Q_a} \Pr[\mathbf{A} = a] \cdot \Pr[\mathbf{B} = b \mid \mathbf{A} = a] - \Pr[\mathbf{A}' = a] \cdot \Pr[\mathbf{B}' = b \mid \mathbf{A}' = a] \\
&\geq \sum_{a \in X_1, b \in Q_a} \Pr[\mathbf{A} = a] \cdot (\Pr[\mathbf{B} = b \mid \mathbf{A} = a] - \Pr[\mathbf{B}' = b \mid \mathbf{A}' = a]) \\
&= \sum_{a \in X_1} \Pr[\mathbf{A} = a] \sum_{b \in Q_a} (\Pr[\mathbf{B} = b \mid \mathbf{A} = a] - \Pr[\mathbf{B}' = b \mid \mathbf{A}' = a]) \\
&\geq \frac{\gamma}{2} \sum_{a \in X_1} \Pr[\mathbf{A} = a] \geq p\gamma/4.
\end{aligned}$$

Consider now the case that $\Pr[\mathbf{A} \in X_2] \geq p_2$, and recall that for all $a \in X_2 \subseteq S$ it holds that $(\mathbf{B} \mid \mathbf{A} = a) \not\approx_{\gamma/2} (\mathbf{B}' \mid \mathbf{A}' = a)$. By definition of statistical distance, for every $a \in X_2$ there is a set Q_a such that $\Pr[(\mathbf{B}' \mid \mathbf{A}' = a) \in Q_a] - \Pr[(\mathbf{B} \mid \mathbf{A} = a) \in Q_a] \geq \gamma/2$. By definition of X_2, we know that $\Pr[\mathbf{A}' = a] > \Pr[\mathbf{A} = a]$ for all $a \in X_2$, which implies that $\Pr[\mathbf{A}' \in X_2] \geq p/2$. Thus

$$\begin{aligned} |(\mathbf{A}, \mathbf{B}) - (\mathbf{A}', \mathbf{B}')| &\geq \sum_{a \in X_2, b \in Q_a} \Pr[(\mathbf{A}', \mathbf{B}') = (a, b)] - \Pr[(\mathbf{A}, \mathbf{B}) = (a, b)] \\ &= \sum_{a \in X_2, b \in Q_a} \Pr[\mathbf{A}' = a] \cdot \Pr[\mathbf{B}' = b \mid \mathbf{A}' = a] - \Pr[\mathbf{A} = a] \cdot \Pr[\mathbf{B} = b \mid \mathbf{A} = a] \\ &\geq \sum_{a \in X_2} \Pr[\mathbf{A}' = a] \sum_{b \in Q_a} (\Pr[\mathbf{B}' = b \mid \mathbf{A}' = a] - \Pr[\mathbf{B} = b \mid \mathbf{A} = a]) \\ &\geq \frac{\gamma}{2} \sum_{a \in X_2} \Pr[\mathbf{A}' = a] \geq p\gamma/4. \end{aligned}$$

Thus we see that no matter what, $|(\mathbf{A}, \mathbf{B}) - (\mathbf{A}', \mathbf{B}')| \geq p\gamma/4$. And since this statistical distance is at most ε by the hypothesis, we get that $p \leq 4\varepsilon/\gamma$. Since p was defined to be $\Pr_{a \sim \mathbf{A}}[a \in S]$, the result follows. □

3.3 Condensers

At last, we are ready to present a formal definition for the main objects of study in this paper.

Definition 6 (Condenser). *Let $\mathcal{X}$ be a family of (n, k)-sources. A function* $\mathsf{Cond} : \{0,1\}^n \to \{0,1\}^m$ *is called a* condenser *for $\mathcal{X}$ with error ε, loss $\ell \in [0, k]$, and gap g, if $m = k - \ell + g$ and for every $\mathbf{X} \in \mathcal{X}$,*

$$H_\infty^\varepsilon(\mathsf{Cond}(\mathbf{X})) \geq k - \ell.$$

We call $k' := k - \ell$ the output entropy *of the condenser.*

Note that after specifying the family $\mathcal{X}$ and the error ε of the condenser, there are many equivalent ways to describe the remaining parameters. In particular, one may choose to specify its loss and gap, or its output entropy and gap, or its loss and output length, and so on (and the other parameters can be inferred).[15] Our choice will often depend on whichever feels the most appropriate in context. One important note, however, is that the output entropy simply describes *a lower bound on* the actual (smooth) min-entropy of the output, while the output gap describes an upper bound on the actual gap. Indeed, the parameters of the condenser should not change as you plug in different sources from $\mathcal{X}$!

Now, while the above definition seems to describe "deterministic" or "seedless" condensers, it is easy to see that it also captures seeded condensers, simply by setting $\mathcal{X}$ to consist of all sources of the form $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is an (n, k)-source and $\mathbf{Y}$ is an independent (d, d)-source. Still, it is helpful to introduce a separate (perhaps redundant) definition, which makes it easier to refer to their parameters.

Definition 7 (Seeded condenser). *A function* $\mathsf{sCond} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is an* $(n, k) \times (d, d) \to_\varepsilon (m, k')$ seeded condenser *if for any (n, k)-source $\mathbf{X}$, it holds that $H_\infty^\varepsilon(\mathsf{sCond}(\mathbf{X}, \mathbf{U}_d)) \geq k'$.*

Next, note that condensers (as put forth in Definition 6) strictly generalize extractors, which correspond to the case where $g = 0$. As a result, the same is true of seeded condensers and seeded extractors. Still, it will be handy to record a separate definition of these objects, for ease of reference.

[15] When there is a notion of "gap" in the input, we often refer to the gap of the condenser as the "output gap" to avoid confusion.

Definition 8 (Seeded extractor)**.** *A function* $\mathsf{sExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* (k, ε)-seeded extractor *if for any* (n, k)*-source* $\mathbf{X}$*, it holds that* $\mathsf{sExt}(\mathbf{X}, \mathbf{U}_d) \approx_\varepsilon \mathbf{U}_m$*.*

Note that such a seeded extractor is automatically a $(n, k) \times (d, d) \to_\varepsilon (m, m)$-seeded condenser.

Finally, we record a useful "trick," which can be thought of as a trivial condenser. In the world of extractors, it is well-known that you can shorten the output length "for free," simply by taking a prefix (i.e., this operation won't harm the other parameters of the extractor). In the world of condensers, this may harm the overall output entropy *rate* k'/m, but it cannot harm the absolute *gap*.

Fact 4. *If* $\mathbf{X} \sim \{0,1\}^n$ *has min-entropy gap* $\leq g$*, its prefix* $\mathbf{X}_{[p]}$ *of length* p *has min-entropy gap* $\leq g$*.*

Proof. Let x be the most likely element hit by $\mathbf{X}_{[p]}$, and suppose it is hit with probability $2^{-\ell}$. Conditioned on $\mathbf{X}_{[p]} = x$, there is some element $y \in \{0,1\}^{n-p}$ hit by $\mathbf{X}_{[p+1,n]}$ with probability at least $2^{-(n-p)}$. This means $\mathbf{X}$ hits (x, y) with probability at least $2^{-\ell-(n-p)}$. Since $\mathbf{X}$ has min-entropy gap $\leq g$, this means that $p - \ell \leq g$, and the result follows. □

4 Basics of block sources

In this section, we'll introduce some definitions, facts, and tools related to CG sources and block sources. Many of the tools we develop here are new, and they find good use throughout the rest of the paper.

4.1 Definitions

First, recall that an (n, k)-source is simply a random variable $\mathbf{X} \sim \{0,1\}^n$ with min-entropy at least k. Chor-Goldreich sources generalize (n, k)-sources in the following way:

Definition 9 (CG sources)**.** *A source* $\mathbf{X} \sim (\{0,1\}^n)^t$ *is called a* (t, n, k)*-Chor-Goldreich source if*

$$H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x) \geq k$$

for all $i \in [t]$ *and* $x \in (\{0,1\}^n)^{i-1}$*.*

Note that a $(1, n, k)$-Chor-Goldreich source is exactly an (n, k)-source. A $(t, 1, k)$-Chor-Goldreich source, on the other hand, is known as a *Santha-Vazirani source* [SV86]. Next, the following allows us to assume that every Chor-Goldreich source has some nice structure.

Fact 5. *If* $\mathbf{X} \sim (\{0,1\}^n)^t$ *is a* (t, n, k)*-Chor-Goldreich source, then it is a convex combination of* (t, n, k)*-Chor-Goldreich sources* $\mathbf{X}' \sim (\{0,1\}^n)^t$ *such that for any* $i \in [t]$ *and* $x \in (\{0,1\}^n)^{i-1}$*,*

$$(\mathbf{X}_i \mid \mathbf{X}_{<i} = x)$$

is a flat (n, k)*-source.*

Proof. It is well-known that any (n, k)-source (with k an integer) is a convex combination of flat (n, k)-sources [Vad12, Lemma 6.10]. Iteratively apply this to blocks $\mathbf{X}_1, \ldots, \mathbf{X}_t$, using the fact that under any conditioning on $\mathbf{X}_{<i}$, the block $\mathbf{X}_i$ is still an (n, k)-source (by definition of Chor-Goldreich source). □

In our constructions, we'll often need to work with a generalization of CG sources, where the block lengths are uneven. These are called *block sources*.

Definition 10 (Block sources)**.** *A source* $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_t)$ *is called an* $((n_1, k_1), (n_2, k_2), \dots, (n_t, k_t))$*-block source if each* $\mathbf{X}_i$ *is over* n_i *bits, and*

$$H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x) \geq k_i$$

for all $i \in [t]$ *and* $x \in \{0,1\}^{n_1} \times \{0,1\}^{n_2} \times \cdots \times \{0,1\}^{n_{i-1}}$*.*

Note that a (t, n, k)-CG source is just a block source with $n_1 = \cdots = n_t = n$ and $k_1 = \cdots = k_t = k$.

To streamline our proofs, it will be convenient to take this generalization two steps further. We use the following definition, which generalizes block sources by only requiring each block $\mathbf{X}_i$ to be *close* to having a min-entropy guarantee, and only requiring this closeness to hold for *most* fixings of the prefix $\mathbf{X}_{<i}$.

Definition 11 (Almost block sources)**.** *A source* $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_t)$ *is called an* $((\eta_1, \gamma_1), \dots, (\eta_t, \gamma_t))$*-almost* $((n_1, k_1), \dots, (n_t, k_t))$*-block source if each* $\mathbf{X}_i$ *is over* n_i *bits, and for every* $i \in [t]$ *it holds that*

$$\Pr_{x \sim \mathbf{X}_{<i}}[(\mathbf{X}_i \mid \mathbf{X}_{<i} = x) \text{ is } \gamma_i\text{-close to an } (n_i, k_i)\text{-source}] \geq 1 - \eta_i.$$

This notion is a generalization of the first type of almost block sources studied in [DMOZ23, Definition 1.3] (which correspond to the special case where $\gamma_1 = \cdots = \gamma_t = \gamma$ and $\eta_1 = \cdots = \eta_t = 0$), and a specialization of the third type of almost block sources studied in [DMOZ23, Definition 8.3] (with $\lambda = 0$).

4.2 Almost block sources are close to block sources

As it turns out, it is not too difficult to show that an almost block source is close to a true block source.

Lemma 4. *If* $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_t)$ *is an* $((\eta_1, \gamma_1), \dots, (\eta_t, \gamma_t))$*-almost* $((n_1, k_1), \dots, (n_t, k_t))$*-block source, then* $\mathbf{X}$ *is* ε*-close to an* $((n_1, k_1), \dots, (n_t, k_t))$*-block source* $\mathbf{X}^\star$*, where* $\varepsilon = \sum_{i \in [t]}(\eta_i + \gamma_i)$*.*

The key tool is the following, which can be viewed as a tightness result for a key lemma (on amplifying statistical distance) of Chattopadhyay, Goodman, and Zuckerman [CGZ22, Lemma 1, ECCC version]. More formally, it can be viewed as a "local-to-global" closeness result for sequences of correlated random variables. It also generalizes the classic fact that a sequence of *independent* random variables, each close to uniform, is itself (relatively) close to uniform (e.g., [Rao09, Proposition 2.11]).

Lemma 5. *Let* $\mathbf{X} \sim V_1 \times \cdots \times V_t$ *and* $\mathbf{Y} \sim V_1 \times \cdots \times V_t$ *each be a sequence of (not necessarily independent) random variables. Suppose that for every* $i \in [t]$ *and* $v \in$ *support*$(\mathbf{X}_{<i}) \cap$ *support*$(\mathbf{Y}_{<i})$*,*

$$|(\mathbf{X}_i \mid \mathbf{X}_{<i} = v) - (\mathbf{Y}_i \mid \mathbf{Y}_{<i} = v)| \leq \varepsilon_i.$$

Then

$$|\mathbf{X} - \mathbf{Y}| \leq \sum_{i \in [t]} \varepsilon_i.$$

We first prove the key tool above.

Proof. We proceed via a coupling argument. Namely, we will define jointly distributed random variables $\mathbf{X}', \mathbf{Y}' \sim V_1 \times \cdots \times V_t$ such that $\mathbf{X}' \equiv \mathbf{X}, \mathbf{Y}' \equiv \mathbf{Y}$, and so that it is easy to get a good upper bound on $\Pr[\mathbf{X}' \neq \mathbf{Y}']$. The result will then follow by the first part of the coupling lemma (Lemma 1).

In order to actually construct $\mathbf{X}', \mathbf{Y}'$, we will use the second part of the coupling lemma (Lemma 1). In more detail, we define these random variables iteratively (from $i = 1, 2, \dots, t$), as follows. For every $i \in \{1, 2, \dots, t\}$, we will define a new pair of jointly distributed random variables $(\mathbf{X}'_i, \mathbf{Y}'_i)$ such that for every $(u, v) \in \text{support}(\mathbf{X}'_{<i}, \mathbf{Y}'_{<i})$, all of the following bullet points hold:

- $\big(\mathbf{X}'_i \mid (\mathbf{X}'_{<i}, \mathbf{Y}'_{<i}) = (u,v)\big) \equiv (\mathbf{X}_i \mid \mathbf{X}_{<i} = u)$.
- $\big(\mathbf{Y}'_i \mid (\mathbf{X}'_{<i}, \mathbf{Y}'_{<i}) = (u,v)\big) \equiv (\mathbf{Y}_i \mid \mathbf{Y}_{<i} = v)$.
- $\Pr[\mathbf{X}'_i \neq \mathbf{Y}'_i \mid (\mathbf{X}'_{<i}, \mathbf{Y}'_{<i}) = (u,v)] = |(\mathbf{X}_i \mid \mathbf{X}_{<i} = u) - (\mathbf{Y}_i \mid \mathbf{Y}_{<i} = v)|$.

We now show (via induction on i) that such $(\mathbf{X}'_i, \mathbf{Y}'_i)$ exist, and that $\mathbf{X}'_{\leq i} \equiv \mathbf{X}_{\leq i}$ and $\mathbf{Y}'_{\leq i} \equiv \mathbf{Y}_{\leq i}$.

When $i = 1$, we know that such random variables $(\mathbf{X}'_1, \mathbf{Y}'_1)$ exist via the second part of the coupling lemma (Lemma 1).[16] Furthermore, the bullet points tell us that $\mathbf{X}'_{\leq 1} \equiv \mathbf{X}_{\leq 1}$ and $\mathbf{Y}'_{\leq 1} \equiv \mathbf{Y}_{\leq 1}$.

When $i > 1$, we assume (via the induction hypothesis) that $\mathbf{X}'_{<i} \equiv \mathbf{X}_{<i}$ and $\mathbf{Y}'_{<i} \equiv \mathbf{Y}_{<i}$. As a result, $(u,v) \in \text{support}(\mathbf{X}'_{<i}, \mathbf{Y}'_{<i})$ implies that both $u \in \text{support}(\mathbf{X}_{<i})$ and $v \in \text{support}(\mathbf{Y}_{<i})$, and so $(\mathbf{X}_i \mid \mathbf{X}_{<i} = u)$ and $(\mathbf{Y}_i \mid \mathbf{Y}_{<i} = v)$ are well-defined. Thus, the second part of the coupling lemma (Lemma 1) once again tells us that there exist random variables $(\mathbf{X}'_i, \mathbf{Y}'_i)$ satisfying all three bullets. Furthermore, we assert that $\mathbf{X}'_{\leq i} \equiv \mathbf{X}_{\leq i}$ and $\mathbf{Y}'_{\leq i} \equiv \mathbf{Y}_{\leq i}$. To see why the former holds, note that for all $x \in \text{support}(\mathbf{X}'_{\leq i})$,

$$\begin{aligned}\Pr[\mathbf{X}'_{\leq i} = x] &= \Pr[\mathbf{X}'_{<i} = x_{<i}] \cdot \Pr[\mathbf{X}'_i = x_i \mid \mathbf{X}'_{<i} = x_{<i}] \\ &= \Pr[\mathbf{X}_{<i} = x_{<i}] \cdot \Pr[\mathbf{X}'_i = x_i \mid \mathbf{X}'_{<i} = x_{<i}],\end{aligned}$$

since the induction hypothesis tells us that $\mathbf{X}'_{<i} \equiv \mathbf{X}_{<i}$. Then, by the law of total probability,

$$\begin{aligned}\Pr[\mathbf{X}'_i = x_i \mid \mathbf{X}'_{<i} = x_{<i}] &= \sum_{y \in \text{support}(\mathbf{Y}'_{<i} \mid \mathbf{X}'_{<i} = x_{<i})} \Pr[\mathbf{X}'_i = x_i \wedge \mathbf{Y}'_{<i} = y \mid \mathbf{X}'_{<i} = x_{<i}] \\ &= \sum_{y \in \text{support}(\mathbf{Y}'_{<i} \mid \mathbf{X}'_{<i} = x_{<i})} \Pr[\mathbf{Y}'_{<i} = y \mid \mathbf{X}'_{<i} = x_{<i}] \cdot \Pr[\mathbf{X}'_i = x_i \mid \mathbf{X}_{<i'} = x_{<i}, \mathbf{Y}'_{<i} = y] \\ &= \Pr[\mathbf{X}_i = x_i \mid \mathbf{X}_{<i} = x_{<i}] \cdot \sum_{y \in \text{support}(\mathbf{Y}'_{<i} \mid \mathbf{X}'_{<i} = x_{<i})} \Pr[\mathbf{Y}'_{<i} = y \mid \mathbf{X}'_{<i} = x_{<i}] \\ &= \Pr[\mathbf{X}_i = x_i \mid \mathbf{X}_{<i} = x_{<i}],\end{aligned}$$

where the penultimate equality follows from the first bullet above. Thus

$$\begin{aligned}\Pr[\mathbf{X}'_{\leq i} = x] &= \Pr[\mathbf{X}_{<i} = x_{<i}] \cdot \Pr[\mathbf{X}'_i = x_i \mid \mathbf{X}'_{<i} = x_{<i}] \\ &= \Pr[\mathbf{X}_{<i} = x_{<i}] \cdot \Pr[\mathbf{X}_i = x_i \mid \mathbf{X}_{<i} = x_{<i}] \\ &= \Pr[\mathbf{X}_{\leq i} = x].\end{aligned}$$

As a result, we have that $\mathbf{X}'_{\leq i} \equiv \mathbf{X}_{\leq i}$, and an identical argument shows that $\mathbf{Y}'_{\leq i} \equiv \mathbf{Y}_{\leq i}$.

Finally, we now have joint random variables $\mathbf{X}', \mathbf{Y}'$ such that $\mathbf{X}' \equiv \mathbf{X}$ and $\mathbf{Y}' \equiv \mathbf{Y}$. Thus, by the first part of the coupling lemma (Lemma 1), we know that

$$|\mathbf{X} - \mathbf{Y}| \leq \Pr[\mathbf{X}' \neq \mathbf{Y}'],$$

[16] Formally, note that when $i = 1$, the phrase "for every $(u,v) \in \text{support}(\mathbf{X}'_{<i}, \mathbf{Y}'_{<i})$" is removed, and there is no conditioning.

and so all that remains it to upper bound this probability. To do so, note that

$$\begin{aligned}
\Pr[\mathbf{X}' \neq \mathbf{Y}'] &= \sum_{i\in[t]} \Pr[\mathbf{X}'_i \neq \mathbf{Y}'_i \wedge \mathbf{X}'_{<i} = \mathbf{Y}'_{<i}] \\
&= \sum_{i\in[t]} \sum_{(v,v)\in \text{support}(\mathbf{X}'_{<i},\mathbf{Y}'_{<i})} \Pr[\mathbf{X}'_i \neq \mathbf{Y}'_i \wedge (\mathbf{X}'_{<i},\mathbf{Y}'_{<i}) = (v,v)] \\
&= \sum_{i\in[t]} \sum_{(v,v)\in \text{support}(\mathbf{X}'_{<i},\mathbf{Y}'_{<i})} \Pr[(\mathbf{X}'_{<i},\mathbf{Y}'_{<i}) = (v,v)] \cdot \Pr[\mathbf{X}'_i \neq \mathbf{Y}'_i \mid (\mathbf{X}'_{<i},\mathbf{Y}'_{<i}) = (v,v)] \\
&= \sum_{i\in[t]} \sum_{(v,v)\in \text{support}(\mathbf{X}'_{<i},\mathbf{Y}'_{<i})} \Pr[(\mathbf{X}'_{<i},\mathbf{Y}'_{<i}) = (v,v)] \cdot |(\mathbf{X}_i \mid \mathbf{X}_{<i} = v) - (\mathbf{Y}_i \mid \mathbf{Y}_{<i} = v)| \\
&\leq \sum_{i\in[t]} \varepsilon_i \cdot \sum_{(v,v)\in \text{support}(\mathbf{X}'_{<i},\mathbf{Y}'_{<i})} \Pr[(\mathbf{X}'_{<i},\mathbf{Y}'_{<i}) = (v,v)] \\
&\leq \sum_{i\in[t]} \varepsilon_i,
\end{aligned}$$

where the last equality follows from the third bullet point above, and the penultimate inequality follows from the lemma hypothesis, since

$$(v,v) \in \text{support}(\mathbf{X}'_{<i},\mathbf{Y}'_{<i}) \implies v \in \text{support}(\mathbf{X}'_{<i}) \cap \text{support}(\mathbf{Y}'_{<i}) \implies v \in \text{support}(\mathbf{X}_{<i}) \cap \text{support}(\mathbf{Y}_{<i}).$$

Thus

$$|\mathbf{X} - \mathbf{Y}| \leq \Pr[\mathbf{X}' \neq \mathbf{Y}'] \leq \sum_{i\in[t]} \varepsilon_i,$$

as desired. □

With this tool in hand, it is now easy to show that almost block sources are close to true block sources.

Proof of Lemma 4. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_t)$ be an $((\eta_1,\gamma_1),\ldots,(\eta_t,\gamma_t))$-almost $((n_1,k_1),\ldots,(n_t,k_t))$-block source. We first show how to "zero out" the η_i terms, and then the γ_i terms.

Zeroing out the η_i terms in X. We start by defining, for every $i \in [t]$, the set

$$\mathsf{Good}_i := \{x \in \text{support}(\mathbf{X}_{<i}) : (\mathbf{X}_i \mid \mathbf{X}_{<i} = x) \text{ is } \gamma_i\text{-close to an } (n_i,k_i)\text{-source}\}.$$

Then, we define a source $\mathbf{X}' = (\mathbf{X}'_1, \ldots, \mathbf{X}'_t)$ such that for every $i \in \{1,2,\ldots,t\}$ and $x \in \text{support}(\mathbf{X}'_{<i})$,

$$(\mathbf{X}'_i \mid \mathbf{X}'_{<i} = x) \equiv \begin{cases} (\mathbf{X}_i \mid \mathbf{X}_{<i} = x) & \text{if } x \in \mathsf{Good}_i, \\ \mathbf{U}_{n_i} & \text{otherwise.} \end{cases}$$

It is immediate that $\mathbf{X}'$ is an $((0,\gamma_1),\ldots,(0,\gamma_t))$-almost $((n_1,k_1),\ldots,(n_t,k_t))$-block source.

Furthermore, observe that for any $x \in \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_t}$ with $x_{<i} \in \mathsf{Good}_i$ for all $i \in [t]$,

$$\Pr[\mathbf{X}' = x] = \Pr[\mathbf{X} = x].$$

Thus, for any set $S \subseteq \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_t}$, we have that

$$\begin{aligned}\Pr[\mathbf{X} \in S] &\leq \Pr\left[\mathbf{X} \in S \wedge \mathbf{X}_{<i} \in \mathsf{Good}_i, \forall i \in [t]\right] + \Pr\left[\exists i \in [t] : \mathbf{X}_{<i} \notin \mathsf{Good}_i\right] \\ &\leq \Pr\left[\mathbf{X}' \in S \wedge \mathbf{X}'_{<i} \in \mathsf{Good}_i, \forall i \in [t]\right] + \sum_{i \in [t]} \Pr[\mathbf{X}_{<i} \notin \mathsf{Good}_i] \\ &\leq \Pr[\mathbf{X}' \in S] + \sum_{i \in [t]} \eta_i.\end{aligned}$$

In other words, $\mathbf{X}'$ is $\left(\sum_{i \in [t]} \eta_i\right)$-close to $\mathbf{X}$.

Zeroing out the γ_i terms in $\mathbf{X}'$. Since $\mathbf{X}'$ is an $((0,\gamma_1),\dots,(0,\gamma_t))$-almost $((n_1,k_1),\dots,(n_t,k_t))$-block source, we know that for every $i \in [t]$ and $x \in \mathrm{support}(\mathbf{X}'_{<i})$, it holds that $\left(\mathbf{X}'_i \mid \mathbf{X}'_{<i} = x\right)$ is γ_i-close to an (n_i,k_i)-source, which we will call $\mathbf{Z}_i^{(x)}$. Using this, we define a new source $\mathbf{X}^\star = (\mathbf{X}^\star_1,\dots,\mathbf{X}^\star_t)$ such that for every $i \in \{1,2,\dots,t\}$ and $x \in \mathrm{support}(\mathbf{X}^\star_{<i})$,

$$(\mathbf{X}^\star_i \mid \mathbf{X}^\star_{<i} = x) \equiv \begin{cases}\mathbf{Z}_i^{(x)} & \text{if } x \in \mathrm{support}(\mathbf{X}'_{<i}), \\ \mathbf{U}_{n_i} & \text{otherwise.}\end{cases}$$

It is immediate that $\mathbf{X}^\star$ is an $((0,0),\dots,(0,0))$-almost $((n_1,k_1),\dots,(n_t,k_t))$-block source; or in other words, an $((n_1,k_1),\dots,(n_t,k_t))$-block source.

Furthermore, observe that for any $i \in [t]$ and $x \in \mathrm{support}(\mathbf{X}'_{<i}) \cap \mathrm{support}(\mathbf{X}^\star_{<i})$,

$$\left|(\mathbf{X}'_i \mid \mathbf{X}'_{<i} = x) - (\mathbf{X}^\star_i \mid \mathbf{X}^\star_{<i} = x)\right| \leq \gamma_i.$$

As a result, Lemma 5 immediately tells us that $\mathbf{X}^\star$ is $\left(\sum_{i \in [t]} \gamma_i\right)$-close to $\mathbf{X}'$.

Wrapping up By a standard application of the triangle inequality, we get that $\mathbf{X}^\star$ is ε-close to $\mathbf{X}$, where $\varepsilon = \sum_{i \in [t]}(\eta_i + \gamma_i)$. Since $\mathbf{X}^\star$ is an $((n_1,k_1),\dots,(n_t,k_t))$-block source, this completes the proof. □

4.3 Keeping a block source fresh while fixing correlated randomness

In extractor theory, the situation often arises that you have a collection of independent random variables $\mathbf{X}_1,\dots,\mathbf{X}_t$, and additional random variables $\mathbf{X}'_1,\dots,\mathbf{X}'_t$ where each $\mathbf{X}'_i$ is a deterministic function $\mathbf{X}_i$. The latter variables often get in the way of the analysis, and the goal is usually to condition ("fix") them to constant values, while keeping the entropy and independence in $\mathbf{X}_1,\dots,\mathbf{X}_t$. The classic tool used for this is the chain rule for min-entropy.

Lemma 6 (Min-entropy chain rule [MW97])**.** *For any random variables $\mathbf{X} \sim X$ and $\mathbf{Y} \sim Y$,*

$$\Pr_{y \sim \mathbf{Y}}\left[H_\infty(\mathbf{X} \mid \mathbf{Y} = y) \geq H_\infty(\mathbf{X}) - \log(|Y|) - \log(1/\varepsilon)\right] \geq 1 - \varepsilon.$$

Indeed, as long as the entropy in each $\mathbf{X}_i$ is larger than the length (support size) of each $\mathbf{X}'_i$, the above lemma can be used to fix $\mathbf{X}'_1,\dots,\mathbf{X}'_t$ without losing the independence of $\mathbf{X}_1,\dots,\mathbf{X}_t$ or too much entropy. But what if $\mathbf{X}_1,\dots,\mathbf{X}_t,\mathbf{X}'_1,\dots,\mathbf{X}'_t$ have correlations among them? As we will see, this situation will frequently arise in our analysis of CG sources. In this section, we establish a formal way to deal with this. We prove the following, which shows how to keep a block source "fresh" (looking like a block source) while fixing a series of correlated random variables.

Lemma 7. *Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_t)$ be an $((n_1, k_1), (n_2, k_2), \dots, (n_t, k_t))$-block source, and let $\mathbf{X}' = (\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_t)$ be another sequence of (possibly correlated) random variables satisfying the following.*

- *$\mathbf{X}'_i$ is supported on a set of size at most $2^{n'_i}$, for every $i \in [\tau]$.*
- *The random variables $\left(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}'_{\geq i} = x'\right)$ and $(\mathbf{X}'_{<i} \mid \mathbf{X}_{<i} = x, \mathbf{X}'_{\geq i} = x')$ are independent, for every $i \in [t]$, $x \in \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_{i-1}}$, $x' \in \{0,1\}^{n'_i} \times \cdots \times \{0,1\}^{n'_t}$.*

Then

$$\Pr_{x' \sim \mathbf{X}'}\left[(\mathbf{X} \mid \mathbf{X}' = x') \text{ is not } t\sqrt{\varepsilon}\text{-close to an } ((n_1, \ell_1), (n_2, \ell_2), \dots, (n_t, \ell_t))\text{-block source}\right] \leq t\sqrt{\varepsilon},$$

where each $\ell_i := k_i - \sum_{j \geq i} n'_j - \log(1/\varepsilon)$.

When we construct our condenser, it is crucial that the entropy loss on $\mathbf{X}_i$ only comes from $\mathbf{X}'_j, j \geq i$.

Proof. Pick any index $i \in [t]$, and define $\ell_i := k_i - \sum_{j \geq i} n'_j - \log(1/\varepsilon)$. Note that for every fixed x, $(\mathbf{X}_i \mid \mathbf{X}_{<i} = x)$ has min-entropy at least k_i (since it is a block source), and thus the min-entropy chain rule (Lemma 6) tells us that

$$\Pr_{x' \sim \mathbf{X}'_{\geq i}}[H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}'_{\geq i} = x') < \ell_i] \leq \varepsilon.$$

By the independence guaranteed in the second bullet of the lemma, we know that for any fixed $x, x^\star$, the distributions $(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}' = x^\star)$ and $\left(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}'_{\geq i} = x^\star_{\geq i}\right)$ are identical. Thus we know that

$$\Pr_{x^\star \sim \mathbf{X}'}[H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}' = x^\star) < \ell_i] \leq \varepsilon$$

for every fixed x. As a result, we have that

$$\Pr_{\substack{x^\star \sim \mathbf{X}' \\ x \sim \mathbf{X}_{<i}}}[H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}' = x^\star) < \ell_i] \leq \varepsilon.$$

Using an averaging argument, this gives

$$\Pr_{x^\star \sim \mathbf{X}'}\left[\Pr_{x \sim \mathbf{X}_{<i}}[H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}' = x^\star) < \ell_i] \geq \sqrt{\varepsilon}\right] \leq \sqrt{\varepsilon},$$

and a union bound tells us

$$\Pr_{x^\star \sim \mathbf{X}'}\left[\exists i \in [t] : \Pr_{x \sim \mathbf{X}_{<i}}[H_\infty(\mathbf{X}_i \mid \mathbf{X}_{<i} = x, \mathbf{X}' = x^\star) < \ell_i] \geq \sqrt{\varepsilon}\right] \leq t\sqrt{\varepsilon},$$

In other words, we get that $\mathbf{X}$ becomes an $\sqrt{\varepsilon}$-almost $((n_1, \ell_1), (n_2, \ell_2), \dots, (n_t, \ell_t))$-block source except with probability at most $t\sqrt{\varepsilon}$ over fixing $\mathbf{X}' = x^\star$. Applying Lemma 4 completes the proof. □

4.4 Seeded condensers automatically work for two-block sources

A core tool we use is the fact that seeded condensers can be used on block sources, while suffering just a small loss in parameters. This observation has been made in prior work, with slightly weaker parameters [BCDT19, Lemma 28], or using a slightly different language [BGM22, Proof of Theorem 4.4].

Lemma 8. *Let* $\mathsf{sCond} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a seeded* $(n,k) \to_\varepsilon (m,k')$ *condenser. Then for any* $((n,k),(d,d-g))$*-block source* $(\mathbf{X},\mathbf{Y})$*, it holds that* $H_\infty^{2^g\varepsilon}(\mathsf{sCond}(\mathbf{X},\mathbf{Y})) \geq k' - g$*.*

In other words, the output loses g bits of entropy, and the error blows up by a factor of 2^g.

Proof. Let $(\mathbf{X},\mathbf{Y}) \sim \{0,1\}^n \times \{0,1\}^d$ be an $((n,k),(d,d-g))$ block source, and let $\mathbf{Y}^* \sim \{0,1\}^d$ be an independent uniform random variable. Notice that for any fixed x, S we have

$$\Pr[\mathsf{sCond}(x,(\mathbf{Y} \mid \mathbf{X}=x)) \in S] \leq 2^g \cdot \Pr[\mathsf{sCond}(x,\mathbf{Y}^*) \in S],$$

since if we define $S_x := \{y : \mathsf{sCond}(x,y) \in S\}$ then $\Pr[\mathsf{sCond}(x,\mathbf{Y}^*) \in S] = 2^{-d}|S_x|$ and $\Pr[\mathsf{sCond}(x,(\mathbf{Y} \mid \mathbf{X}=x)) \in S] \leq 2^{-(d-g)}|S_x|$. Thus we have

$$\begin{aligned}\Pr[\mathsf{sCond}(\mathbf{X},\mathbf{Y}) \in S] &= \sum_x \Pr[\mathbf{X}=x] \cdot \Pr[\mathsf{sCond}(x,(\mathbf{Y} \mid \mathbf{X}=x)) \in S] \\ &\leq 2^g \sum_x \Pr[\mathbf{X}=x] \cdot \Pr[\mathsf{sCond}(x,\mathbf{Y}^*) \in S] \\ &= 2^g \Pr[\mathsf{sCond}(\mathbf{X},\mathbf{Y}^*) \in S].\end{aligned}$$

Since sCond is a seeded $(n,k) \to_\varepsilon (m,k')$ condenser, the above expression is at most

$$\begin{aligned}&\leq 2^g \cdot (|S| \cdot 2^{-k'} + \varepsilon) \\ &= |S| \cdot 2^{-k'+g} + 2^g\varepsilon.\end{aligned}$$

The result now follows by the standard characterization of smooth min-entropy (Lemma 2). □

4.5 Iterative condensing of multi-block sources

Finally, the following generalizes well-known block-source extraction and condensing results, such as in [NZ96, DMOZ23]. For example, if instantiated with an $((n_1,k_1),\ldots,(n_{t-1},k_{t-1}),(n_t,n_t))$-block-source and seeded condensers with gap 0 (i.e., seeded extractors), then you get well-known results about extracting from block sources with a small seed (which is constant for constant error). We will use this framework in both our explicit and existential constructions, in order to handle sources with a very large number of blocks.

Lemma 9. *Consider a sequence of functions* $\mathsf{sCond}_1, \mathsf{sCond}_2, \ldots, \mathsf{sCond}_{t-1}$*, where each* $\mathsf{sCond}_i : \{0,1\}^{n_i} \times \{0,1\}^{m_{i+1}} \to \{0,1\}^{m_i}$ *is a seeded* $(n_i,k_i) \to_{\varepsilon_i} (m_i, m_i - g_i)$ *condenser. Furthermore, consider any pair of nonnegative real numbers* (n_t,k_t) *such that* $m_t = n_t$*, and define* $g_t := n_t - k_t$ *and* $\varepsilon_t := 0$*.*

Now, define a function $\mathsf{Cond}' : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \times \cdots \times \{0,1\}^{n_t} \to \{0,1\}^{m_1} \times \{0,1\}^{m_2} \times \cdots \times \{0,1\}^{m_t}$ *as* $\mathsf{Cond}'(x_1,x_2,\ldots,x_t) := (y_1,y_2,\ldots,y_t)$*, where* $y_t := x_t$*, and for all other* $i \in [t-1]$*,*

$$y_i := \mathsf{sCond}_i(x_i, y_{i+1}).$$

Then the function $\mathsf{Cond} : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \times \cdots \times \{0,1\}^{n_t} \to \{0,1\}^{m_1}$ *defined as* $\mathsf{Cond}(x_1,x_2,\ldots,x_t) := y_1$ *is a condenser for* $((n_1,k_1),(n_2,k_2),\ldots,(n_t,k_t))$*-block sources with output gap* $g := \sum_{i\in[t]} g_i$ *and error* $\varepsilon := \sum_{i\in[t]} \varepsilon_i \cdot 2^{\sum_{j\in(i,t]} g_j}$*.*

Proof. Let $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_t)$ be an arbitrary $((n_1,k_1),(n_2,k_2),\ldots,(n_t,k_t))$-block source, and define $\mathbf{Y} = (\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_t) := \mathsf{Cond}'(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_t)$ as in the lemma statement. We will prove a stronger claim than in the lemma, and show that for every $a \in [t]$ and $x \in \mathrm{support}(\mathbf{X}_{<a})$,

$$(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x) \text{ is } \varepsilon'_a\text{-close to an } (m_a, m_a - g'_a)\text{-source,}$$

where $\varepsilon'_a := \sum_{i\in[a,t]} \varepsilon_i \cdot 2^{\sum_{j\in(i,t]} g_j}$, and $g'_a := \sum_{i\in[a,t]} g_i$.[17]

The proof will proceed via backwards induction on a. We start by noting the claim is easy when $a = t$. Indeed, recall that $\mathbf{Y}_t := \mathbf{X}_t$, and that $\mathbf{X}$ is an $((n_1, k_1), (n_2, k_2), \dots, (n_t, k_t))$-block source. This means that for every fixing of $\mathbf{X}_{<t}$, it holds that $\mathbf{X}_t$ (and thus $\mathbf{Y}_t$) is an (n_t, k_t)-source. In other words, every $(\mathbf{Y}_t \mid \mathbf{X}_{<t} = x)$ is ε'_t-close to an $(m_t, m_t - g'_t)$-source, since $\varepsilon'_t = 0$ and $(m_t, m_t - g'_t) = (n_t, k_t)$.

Next, consider any $1 \le a < t$ and $x \in \text{support}(\mathbf{X}_{<a})$, and assume the claim holds for $a + 1$. Recall that $\mathbf{Y}_a := \mathsf{sCond}_a(\mathbf{X}_a, \mathbf{Y}_{a+1})$, and thus

$$(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x) = (\mathsf{sCond}_a(\mathbf{X}_a, \mathbf{Y}_{a+1}) \mid \mathbf{X}_{<a} = x).$$

Now, since $\mathbf{X}$ is a block source, we know that $(\mathbf{X}_a \mid \mathbf{X}_{<a} = x)$ is an (n_a, k_a)-source. Furthermore, the induction hypothesis tells us that for every $x' \in \text{support}(\mathbf{X}_a \mid \mathbf{X}_{<a} = x)$, it holds that $(\mathbf{Y}_{a+1} \mid \mathbf{X}_{<a} = x, \mathbf{X}'_a = x')$ is ε'_{a+1}-close to an $(m_{a+1}, m_{a+1} - g'_{a+1})$-source. This means that the source

$$((\mathbf{X}_a, \mathbf{Y}_{a+1}) \mid \mathbf{X}_{<a} = x)$$

is an $\big((0,0), (0, \varepsilon'_{a+1})\big)$-almost $\big((n_a, k_a), (m_{a+1}, m_{a+1} - g'_{a+1})\big)$-block source. And by Lemma 4, this means it is ε'_{a+1}-close to some $\big(\big((n_a, k_a), (m_{a+1}, m_{a+1} - g'_{a+1})\big)$-block source $(\mathbf{X}^\star_a, \mathbf{Y}^\star_{a+1})$. Thus, by a standard application of the data-processing inequality (Fact 2), we have that

$$\begin{aligned}(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x) &= (\mathsf{sCond}_a(\mathbf{X}_a, \mathbf{Y}_{a+1}) \mid \mathbf{X}_{<a} = x)\\ &\approx_{\varepsilon'_{a+1}} \mathsf{sCond}_a(\mathbf{X}^\star_a, \mathbf{Y}^\star_{a+1}).\end{aligned}$$

Now, using the fact that seeded condensers automatically work for block sources (Lemma 8), we get that $\mathsf{sCond}_a(\mathbf{X}^\star_a, \mathbf{Y}^\star_{a+1})$ is $\left(2^{g'_{a+1}}\varepsilon_a\right)$-close to some $(m_a, m_a - g_a - g'_{a+1})$-source. Thus, the triangle inequality tells us $(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x)$ is $\left(\varepsilon'_{a+1} + 2^{g'_{a+1}}\varepsilon_a\right)$-close to an $(m_a, m_a - g_a - g'_{a+1})$-source. And by definition,

$$\varepsilon'_{a+1} + 2^{g'_{a+1}}\varepsilon_a = \left(\sum_{i\in[a+1,t]} \varepsilon_i \cdot 2^{\sum_{j\in(i,t]} g_j}\right) + \left(2^{\sum_{j\in[a+1,t]} g_j}\varepsilon_a\right) = \sum_{i\in[a,t]} \varepsilon_i \cdot 2^{\sum_{j\in(i,t]} g_j} = \varepsilon'_a,$$

and

$$g_a + g'_{a+1} = g_a + \sum_{i\in[a+1,t]} g_i = \sum_{i\in[a,t]} g_i = g'_a.$$

Thus, we get that $(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x)$ is ε'_a-close to an $(m_a, m_a - g'_a)$-source, as desired.

To conclude, we now know that for all $a \in [t]$ and $x \in \text{support}(\mathbf{X}_{<a})$,

$$(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x) \text{ is } \varepsilon'\text{-close to an } (m_a, m_a - g')\text{-source,}$$

where $\varepsilon' = \sum_{i\in[a,t]} \varepsilon_i \cdot 2^{\sum_{j\in(i,t]} g_j}$ and $g'_a = \sum_{i\in[a,t]} g_i$. This completes the proof, since the lemma statement corresponds to the special case where $a = 1$. □

[17] We also note that when $a = 1$, the expression $(\mathbf{Y}_a \mid \mathbf{X}_{<a} = x)$ should be interpreted as just $\mathbf{Y}_1$.

5 Explicit constructions

We are now ready to build our condenser for Chor-Goldreich sources, and ultimately prove Theorem 1.

5.1 Somewhere-condensers from non-malleable condensers

Our condenser will be built by expanding the last block of the CG source into a *somewhere-random* source, and iteratively purifying it until we are left with just a single row that has high entropy. To make things formal, we'll need some definitions.

Definition 12 (Somewhere-ℓ-sources)**.** *A source* $\mathbf{Y} \sim (\{0,1\}^m)^D$ *is called a* somewhere-ℓ-source *if there exists some* $i \in [D]$ *such that* $\mathbf{Y}_i$ *has min-entropy at least* ℓ.

Definition 13 (Somewhere-condensers for CG sources)**.** *A function* $\mathsf{sCond} : (\{0,1\}^n)^t \to (\{0,1\}^w)^D$ *is a somewhere-ℓ-condenser for* (t,n,k)*-CG sources with error* ε *if for any* (t,n,k)*-CG source* $\mathbf{X} \sim (\{0,1\}^n)^t$, $\mathsf{sCond}(\mathbf{X})$ *is* ε*-close to a convex combination of somewhere-ℓ-sources.*

Definition 14 (Non-malleable condensers for block sources)**.** *A function* $\mathsf{nmCond} : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^w \times [2] \to \{0,1\}^m$ *is a non-malleable condenser (with advice) for* $((n,k),(n,k),(w,\ell))$*-block sources with error* ε *and output entropy* r *if the following holds. For any* $\mathbf{X}, \mathbf{Y} \sim \{0,1\}^n$ *and* $\mathbf{Z}_1, \mathbf{Z}_2 \sim \{0,1\}^w$ *such that at least one of the sequences* $(\mathbf{X},\mathbf{Y},\mathbf{Z}_1)$ *and* $(\mathbf{X},\mathbf{Y},\mathbf{Z}_2)$ *is an* $((n,k),(n,k),(w,\ell))$*-block source,*

$$\mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{Z}_1,1) \oplus \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{Z}_2,2)$$

is ε*-close to an* (m,r)*-source.*

In our first key lemma, we show how a non-malleable condenser can be used to improve the quality of a somewhere-condenser. We prove the following, which we will eventually apply iteratively.

Lemma 10 (Purifying a somewhere-condenser)**.** *Suppose you have the following objects.*

- $\mathsf{sCond} : (\{0,1\}^n)^t \to (\{0,1\}^w)^{2^d}$ *a somewhere-ℓ-condenser for* (t,n,k)*-CG sources with error* ε_1.
- $\mathsf{nmCond} : \{0,1\}^{nb} \times \{0,1\}^{nb} \times \{0,1\}^w \times [2] \to \{0,1\}^m$ *a non-malleable condenser (with advice) for* $((nb, kb-d-\log(1/\varepsilon_2)), (nb, kb-d-\log(1/\varepsilon_2)), (w,\ell))$*-block sources, which has error* ε_2 *and output entropy* r.

Consider the function $\mathsf{sCond}^\star : (\{0,1\}^n)^b \times (\{0,1\}^n)^b \times (\{0,1\}^n)^t \to (\{0,1\}^m)^{2^{d-1}}$ *whose* i^{th} *output is*

$$\mathsf{sCond}^\star_i(X,Y,Z) := \mathsf{nmCond}(X,Y,\mathsf{sCond}(Z)_{2i-1},1) \oplus \mathsf{nmCond}(X,Y,\mathsf{sCond}(Z)_{2i},2).$$

Then, $\mathsf{sCond}^\star$ *is a somewhere-r-condenser for* $(2b+t,n,k)$*-CG sources with error* $\varepsilon = \varepsilon_1 + 4\sqrt{\varepsilon_2} + \varepsilon_2$.

The core technical claim we use is the following.

Claim 2. *Let* $\mathbf{X},\mathbf{Y} \sim \{0,1\}^n$ *and* $\mathbf{Z} := (\mathbf{Z}_1,\dots,\mathbf{Z}_D) \sim (\{0,1\}^w)^D$ *be random variables such that:*

- $(\mathbf{X},\mathbf{Y})$ *is an* $((n,k),(n,k))$*-block source.*
- $\forall x,y \in \{0,1\}^n$, $(\mathbf{Z} \mid \mathbf{X}=x, \mathbf{Y}=y)$ *is* ε_1*-close to a convex combination of somewhere-ℓ-sources.*

Then $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ *is* $(\varepsilon_1 + 4\sqrt{\varepsilon_2})$*-close to a convex combination of sources of the form* $(\mathbf{X}',\mathbf{Y}',\mathbf{Z}')$*, where:*

- *$\exists i \in [D]$ s.t. $(\mathbf{X}', \mathbf{Y}', \mathbf{Z}'_i)$ is an $((n, k-d-\log(1/\varepsilon_2)), (n, k-d-\log(1/\varepsilon_2)), (w, \ell))$-block source.*

Proof. By the lemma hypothesis, we know that for every fixed x, y, $(\mathbf{Z} \mid \mathbf{X} = x, \mathbf{Y} = y)$ is ε_1-close to a convex combination of somewhere-ℓ-sources. This means that for every fixed x, y, there is some convex combination of the form $\mathbf{R}^{x,y} := \sum_{i \in [D]} p_i^{x,y} \cdot \mathbf{R}^{x,y,i}$ such that

$$(x, y, (\mathbf{Z} \mid \mathbf{X} = x, \mathbf{Y} = y)) \approx_{\varepsilon_1} (x, y, \mathbf{R}^{x,y}), \tag{1}$$

where we may assume that each $\mathbf{R}^{x,y,i} \sim (\{0,1\}^w)^D$ is not only a somewhere-ℓ-source, but in fact has its entropy in its i^{th} row. That is, $\mathbf{R}_i^{x,y,i} \sim \{0,1\}^w$ has min-entropy at least ℓ. Moreover, since Equation (1) is true for all fixed x, y, it follows that

$$(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \approx_{\varepsilon_1} (\mathbf{X}, \mathbf{Y}, \mathbf{R}^{\mathbf{X},\mathbf{Y}}).$$

We henceforth focus on $(\mathbf{X}, \mathbf{Y}, \mathbf{R}^{\mathbf{X},\mathbf{Y}})$. Towards this end, recall that $\mathbf{R}^{\mathbf{X},\mathbf{Y}}$ is a convex combination of the form $\mathbf{R}^{\mathbf{X},\mathbf{Y}} = \sum_{i \in [D]} p_i^{\mathbf{X},\mathbf{Y}} \cdot \mathbf{R}^{\mathbf{X},\mathbf{Y},i}$, where for every fixed x, y it holds that $\mathbf{R}^{x,y,i} \sim \{0,1\}^w$ has min-entropy at least ℓ. Notice that because of this structure, we can equivalently sample $(\mathbf{X}, \mathbf{Y}, \mathbf{R}^{\mathbf{X},\mathbf{Y}})$ as follows. First, define a new random variable $\mathbf{A} \sim [D]$ that depends on $\mathbf{X}, \mathbf{Y}$ in the following way: for every fixed x, y, define

$$\Pr[\mathbf{A} = i \mid \mathbf{X} = x, \mathbf{Y} = y] := p_i^{x,y}.$$

Then, for every fixed x, y define a new random variable $\mathbf{T}^{x,y} \sim (\{0,1\}^w)^D$ independent of $\mathbf{X}, \mathbf{Y}$[18] such that

$$(\mathbf{T}^{x,y} \mid \mathbf{A} = i) \equiv \mathbf{R}^{x,y,i}.$$

This means that the random variable $(\mathbf{T}^{x,y} \mid \mathbf{A} = i) \sim (\{0,1\}^w)^D$ has entropy at least ℓ in its i^{th} row: in other words, $(\mathbf{T}_i^{x,y} \mid \mathbf{A} = i) \sim \{0,1\}^w$ has entropy at least ℓ for all fixed x, y, i. Given these definitions, it is straightforward to verify that

$$(\mathbf{X}, \mathbf{Y}, \mathbf{R}^{\mathbf{X},\mathbf{Y}}) \equiv (\mathbf{X}, \mathbf{Y}, \mathbf{T}^{\mathbf{X},\mathbf{Y}}).$$

This is useful, because the latter three random variables are defined in the same space as another random variable $\mathbf{A} \sim [D]$, which has the property that $(\mathbf{T}_i^{\mathbf{X},\mathbf{Y}} \mid \mathbf{A} = i) \sim \{0,1\}^w$ has min-entropy at least ℓ for all i. Moreover, recall that $(\mathbf{X}, \mathbf{Y})$ is an $((n, k), (n, k))$-block source. Thus we can apply our lemma on fixing randomness against block sources (Lemma 7) to get

$$\Pr_{i \sim \mathbf{A}}[(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} = i) \text{ is not } 2\sqrt{\varepsilon_2}\text{-close to an } ((n, k'), (n, k'))\text{-block source}] \leq 2\sqrt{\varepsilon_2},$$

where $k' = k - d - \log(1/\varepsilon_2)$. Thus we get that upon fixing $\mathbf{A} = i$, both of the following hold (except with probability at most $2\sqrt{\varepsilon_2}$):

- $(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} = i)$ is $2\sqrt{\varepsilon_2}$-close to an $((n, k'), (n, k'))$-block source, and
- $(\mathbf{T}_i^{\mathbf{X},\mathbf{Y}} \mid \mathbf{A} = i) \sim \{0,1\}^w$ has min-entropy at least ℓ. In fact, for all fixed x, y, it remains true that $(\mathbf{T}_i^{\mathbf{X},\mathbf{Y}} \mid \mathbf{A} = i, \mathbf{X} = x, \mathbf{Y} = y) \sim \{0,1\}^w$ has min-entropy at least ℓ.

Applying a standard fact about convex combinations (Fact 3), we therefore get that $(\mathbf{X}, \mathbf{Y}, \mathbf{T}^{\mathbf{X},\mathbf{Y}}) \sim \{0,1\}^n \times \{0,1\}^n \times (\{0,1\}^w)^D$ is $2\sqrt{\varepsilon_2}$-close to a convex combination of distributions of the form $(\mathbf{X}^\star, \mathbf{Y}^\star, \mathbf{Z}^\star) \sim \{0,1\}^n \times \{0,1\}^n \times (\{0,1\}^w)^D$ satisfying:

[18] By this, we mean $\mathbf{T}^{x,y}$ is independent of $\mathbf{X}, \mathbf{Y}$. Later, we will use $\mathbf{T}^{\mathbf{X},\mathbf{Y}}$, which is of course not independent of $\mathbf{X}, \mathbf{Y}$. However, the independence assumption tells us that $(\mathbf{T}^{\mathbf{X},\mathbf{Y}} \mid \mathbf{X} = x, \mathbf{Y} = y) \equiv (\mathbf{T}^{x,y} \mid \mathbf{X} = x, \mathbf{Y} = y) \equiv \mathbf{T}^{x,y}$.

- $(\mathbf{X}^\star, \mathbf{Y}^\star) \sim \{0,1\}^n \times \{0,1\}^n$ is $2\sqrt{\varepsilon_2}$-close to an $((n,k'),(n,k'))$-block source, and
- $\mathbf{Z}^\star \sim (\{0,1\}^w)^D$ admits some $i \in [D]$ such that $H_\infty(\mathbf{Z}_i^\star \mid \mathbf{X}^\star = x, \mathbf{Y}^\star = y) \geq \ell$ for all x, y.

Finally, let $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star})$ be the $((n,k'),(n,k'))$-block source that $(\mathbf{X}^\star, \mathbf{Y}^\star)$ is $2\sqrt{\varepsilon_2}$-close to, and define a new random variable $\mathbf{Z}^{\star\star}$ as follows:

$$(\mathbf{Z}^{\star\star} \mid \mathbf{X}^{\star\star} = x, \mathbf{Y}^{\star\star} = y) \equiv \begin{cases} (\mathbf{Z}^\star \mid \mathbf{X}^\star = x, \mathbf{Y}^\star = y) & \text{if } (x,y) \in \text{support}(\mathbf{X}^\star, \mathbf{Y}^\star), \\ \mathbf{U} & \text{otherwise.} \end{cases}$$

It is straightforward to verify the following about $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star}, \mathbf{Z}^{\star\star}) \sim \{0,1\}^n \times \{0,1\}^n \times (\{0,1\}^w)^D$.

- $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star})$ is an $((n,k'),(n,k'))$-block source.
- There exists some $i \in [D]$ such that $H_\infty(\mathbf{Z}_i^{\star\star} \mid \mathbf{X}^{\star\star} = x, \mathbf{Y}^{\star\star} = y) \geq \ell$ for all x, y.
- $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star}, \mathbf{Z}^{\star\star}) \approx_{2\sqrt{\varepsilon_2}} (\mathbf{X}^\star, \mathbf{Y}^\star, \mathbf{Z}^{\star\star})$.

Note that the first two conditions in fact imply that there exists some $i \in [D]$ such that $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star}, \mathbf{Z}_i^{\star\star})$ is an $((n,k'),(n,k'),(w,\ell))$-block source, where recall that $k' = k - d - \log(1/\varepsilon_2)$. Thus $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star}, \mathbf{Z}^{\star\star})$ has the exact structure we were originally looking for. To summarize, recall that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \approx_{\varepsilon_1} (\mathbf{X}, \mathbf{Y}, \mathbf{R}^{\mathbf{X},\mathbf{Y}}) \equiv (\mathbf{X}, \mathbf{Y}, \mathbf{T}^{\mathbf{X},\mathbf{Y}})$, and the latter is $2\sqrt{\varepsilon_2}$-close to a convex combination of distributions $(\mathbf{X}^\star, \mathbf{Y}^\star, \mathbf{Z}^\star)$ of the form specified above, and each of these is $2\sqrt{\varepsilon_2}$-close to a distribution $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star}, \mathbf{Z}^{\star\star})$ of the desired structure. Applying the triangle inequality (Fact 1), we immediately get that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is $(\varepsilon_1 + 4\sqrt{\varepsilon_2})$-close to a convex combination of distributions $(\mathbf{X}^{\star\star}, \mathbf{Y}^{\star\star}, \mathbf{Z}^{\star\star})$ of the desired form. □

Given the above claim, it is now straightforward to show that a non-malleable condenser can be used to purify a somewhere-condenser.

Proof of Lemma 10. Let $\mathbf{B} \sim (\{0,1\}^n)^{2b+t}$ be a $(2b+t, n, k)$-CG source. Observe that we can parse it as an $((nb,kb),(nb,kb),(nt,kt))$-block source $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \sim (\{0,1\}^n)^b \times (\{0,1\}^n)^b \times (\{0,1\}^n)^t$, with the additional property that for every fixed x, y, $(\mathbf{Z} \mid \mathbf{X} = x, \mathbf{Y} = y)$ is a (t, n, k)-CG source. The goal is to show $\mathsf{sCond}^\star(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is ε-close to a somewhere-r-source $\mathbf{A} \sim (\{0,1\}^m)^{2^{d-1}}$. Recalling the definition of $\mathsf{sCond}^\star$, this means we must show that the random variable

$$\mathbf{T} := \begin{pmatrix} \mathsf{nmCond}(\mathbf{X}, \mathbf{Y}, \mathsf{sCond}_1(\mathbf{Z}), 1) \oplus \mathsf{nmCond}(\mathbf{X}, \mathbf{Y}, \mathsf{sCond}_2(\mathbf{Z}), 2), \\ \mathsf{nmCond}(\mathbf{X}, \mathbf{Y}, \mathsf{sCond}_3(\mathbf{Z}), 1) \oplus \mathsf{nmCond}(\mathbf{X}, \mathbf{Y}, \mathsf{sCond}_4(\mathbf{Z}), 2), \\ \vdots \\ \mathsf{nmCond}(\mathbf{X}, \mathbf{Y}, \mathsf{sCond}_{D-1}(\mathbf{Z}), 1) \oplus \mathsf{nmCond}(\mathbf{X}, \mathbf{Y}, \mathsf{sCond}_D(\mathbf{Z}), 2) \end{pmatrix}$$

is ε-close to a convex combination of somewhere-r-sources. Towards this end, define for each $i \in [D]$ a random variable $\mathbf{W}_i := \mathsf{sCond}_i(\mathbf{Z}) \sim \{0,1\}^w$, and let $\mathbf{W} := (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_D)$. We can rewrite $\mathbf{T}$ as

$$\mathbf{T} := \Big(\begin{array}{c} \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{W}_1,1) \oplus \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{W}_2,2), \\ \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{W}_3,1) \oplus \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{W}_4,2), \\ \vdots \\ \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{W}_{D-1},1) \oplus \mathsf{nmCond}(\mathbf{X},\mathbf{Y},\mathbf{W}_D,2) \end{array} \Big)$$

Now, recall that $\mathbf{Z}$ is a (t,n,k)-CG source, even conditioned on any fixing of $\mathbf{X}=x, \mathbf{Y}=y$. Furthermore, recall that sCond is a somewhere-ℓ-condenser for (t,n,k)-CG sources with error ε_1. We can therefore say the following about the random variables $\mathbf{X},\mathbf{Y},\mathbf{W}$:

- $(\mathbf{X},\mathbf{Y})$ is an $((nb,kb),(nb,kb))$-block source.
- $\forall x,y \in \{0,1\}^{nb}$, $(\mathbf{W} \mid \mathbf{X}=x, \mathbf{Y}=y)$ is ε_1-close to a convex combination of somewhere-ℓ-sources.

Applying our core technical claim (Claim 2), we know that $(\mathbf{X},\mathbf{Y},\mathbf{W})$ is $(\varepsilon_1 + 4\sqrt{\varepsilon_2})$-close to a convex combination of sources of the form $(\mathbf{X}',\mathbf{Y}',\mathbf{W}')$ where:

- $\exists i \in [D]$ such that $(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_i)$ is an $((nb, kb-d-\log(1/\varepsilon_2)), (nb, kb-d-\log(1/\varepsilon_2)), (w,\ell))$-block source.

By a straightforward application of the data-processing inequality (Fact 2), this means that $\mathbf{T}$ is $(\varepsilon_1 + 4\sqrt{\varepsilon_2})$-close to a convex combination of distributions of the form

$$\mathbf{T}' = \Big(\begin{array}{c} \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_1,1) \oplus \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_2,2), \\ \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_3,1) \oplus \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_4,2), \\ \vdots \\ \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_{D-1},1) \oplus \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_D,2) \end{array} \Big),$$

where $(\mathbf{X}',\mathbf{Y}',\mathbf{W}')$ have the guarantee that there exists some $i \in [D]$ such that $(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_i)$ is an $((nb, kb-d-\log(1/\varepsilon_2)), (nb, kb-d-\log(1/\varepsilon_2)), (w,\ell))$-block source. Call this the "good" index i. Now, for all $i \in [D/2]$, define

$$\mathbf{R}'_i := \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_{2i-1}) \oplus \mathsf{nmCond}(\mathbf{X}',\mathbf{Y}',\mathbf{W}'_{2i}))$$

so that we may write

$$\mathbf{T}' = (\mathbf{R}'_1, \mathbf{R}'_2, \ldots, \mathbf{R}'_{D/2}).$$

If $i^\star$ denotes the "good" index, then the definition of non-malleable condensers (Definition 14) tells us that $\mathbf{R}'_{i^\star}$ is ε_2-close to an (m,r)-source $\mathbf{R}''_{i^\star} \sim \{0,1\}^m$. Furthermore, we can define a random variable $\mathbf{R}''_{-i^\star} \sim (\{0,1\}^m)^{D/2-1}$ such that for every fixed $r \in \{0,1\}^m$,

$$(\mathbf{R}''_{-i^\star} \mid \mathbf{R}''_{i^\star} = r) \equiv \begin{cases} (\mathbf{R}'_{-i^\star} \mid \mathbf{R}'_{i^\star} = r) & \text{if } r \in \text{support}(\mathbf{R}'_{i^\star}), \\ \mathbf{U} & \text{otherwise.} \end{cases}$$

Then, if we define $\mathbf{T}'' := (\mathbf{R}''_{i^\star}, \mathbf{R}''_{-i^\star}) \sim (\{0,1\}^m)^{D/2}$, it is straightforward to verify that $\mathbf{T}'' \approx_{\varepsilon_2} \mathbf{T}'$, and moreover $\mathbf{T}''_{i^\star}$ is an (m,r)-source. In other words, $\mathbf{T}''$ is a somewhere-r-source, and thus $\mathbf{T}'$ is ε_2-close to a somewhere-r-source. Recall that at the beginning, we showed that $\mathbf{T}$ is $(\varepsilon_1 + 4\sqrt{\varepsilon_2})$-close to a convex combination of such sources $\mathbf{T}'$, and we now know that each such source $\mathbf{T}'$ is ε_2-close to a somewhere-r-source $\mathbf{T}''$. As a result, it immediately follows that $\mathbf{T}$ is $(\varepsilon_1 + 4\sqrt{\varepsilon_2} + \varepsilon_2)$-close to a convex combination of somewhere-r-sources, as desired. □

5.2 Non-malleable condensers from seeded extractors

While it is known how to explicitly construct somewhere-condensers, it is not known how to construct non-malleable condensers for block sources. In this section, we show how to use basic seeded extractors to construct them. Later, we'll instantiate the recipe below in order to obtain our non-malleable condensers.

Lemma 11 (Non-malleable condensers from seeded extractors)**.** *Suppose you have the following objects.*

- $\mathsf{sExt}_1 : \{0,1\}^n \times \{0,1\}^{p_1} \to \{0,1\}^{d_1}$ *a* (k_0, ε_1)*-seeded extractor.*
- $\mathsf{sExt}'_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m}$ *a* (k_0, ε'_1)*-seeded extractor.*
- $\mathsf{sExt}_2 : \{0,1\}^n \times \{0,1\}^{p_2} \to \{0,1\}^{d_2}$ *a* (k_0, ε_2)*-seeded extractor.*
- $\mathsf{sExt}'_2 : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{m}$ *a* (k_0, ε'_2)*-seeded extractor.*

Consider the function $\mathsf{nmCond} : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^w \times [2] \to \{0,1\}^m$ *defined as*

$$\mathsf{nmCond}(X, Y, Z, b) := \mathsf{sExt}'_b(X, \mathsf{sExt}_b(Y, Z_{[p_b]}))$$

Then nmCond *is a non-malleable condenser (with advice) for* $((n,k),(n,k),(w,w-g))$*-block sources with output entropy* $m - (g + 2d_1 + p_2 + \log(1/\varepsilon_1) + \log(1/\varepsilon_2))$ *and error* $2^{g+p_2+3}\varepsilon_1^{1/4} + 2^{g+4}\varepsilon_2^{1/4}$*, as long as:*

- $k \geq k_0 + m + 2d_1 + d_2 + p_2 + \log(1/\varepsilon_1) + \log(1/\varepsilon_2)$
- $\varepsilon_1 = \varepsilon'_1$ *and* $\varepsilon'_2 = \varepsilon_2 \cdot 2^{-2d_1}$

Proof. Consider any $\mathbf{X}, \mathbf{Y} \sim \{0,1\}^n$ and $\mathbf{Z}^1, \mathbf{Z}^2 \sim \{0,1\}^w$ such that either $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^1)$ or $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^2)$ is an $((n,k),(n,k),(w,w-g))$-block source. Unwrapping the definition of nmCond, the goal is to show that

$$\mathsf{sExt}'_1(\mathbf{X}, \mathsf{sExt}_1(\mathbf{Y}, \mathbf{Z}^1_{[p_1]})) \oplus \mathsf{sExt}'_2(\mathbf{X}, \mathsf{sExt}_2(\mathbf{Y}, \mathbf{Z}^2_{[p_2]}))$$

is ε-close to an (m,r)-source. We must show this to be true in two cases: the case where $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^1)$ is the block source, and the case where $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^2)$ is the block source. We proceed with each case separately. But for both cases, it will be useful to define the following auxiliary random variables:

$$\begin{aligned} \underline{\mathbf{Z}}^1 &:= \mathbf{Z}^1_{[p_1]} & \underline{\mathbf{Z}}^2 &:= \mathbf{Z}^2_{[p_2]} \\ \mathbf{W}_1 &:= \mathsf{sExt}_1(\mathbf{Y}, \underline{\mathbf{Z}}^1) & \mathbf{W}_2 &:= \mathsf{sExt}_2(\mathbf{Y}, \underline{\mathbf{Z}}^2) \\ \mathbf{S}_1 &:= \mathsf{sExt}'_1(\mathbf{X}, \mathbf{W}_1) & \mathbf{S}_2 &:= \mathsf{sExt}'_2(\mathbf{X}, \mathbf{W}_2) \end{aligned}$$

With this notation, the goal is simply to show that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is ε-close to an (m,r)-source. Let's get started.

Case 1. In this case, we assume that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^1)$ is the block source. In order to show that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is ε-close an (m, r)-source, the idea is to find a sequence of fixings that will force $\mathbf{S}_2$ to become a constant, but under which $\mathbf{S}_1$ can be shown to have high min-entropy. In order to fix $\mathbf{S}_2$, we will actually fix the entire sequence of random variables $(\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2)$, and argue that the sequence $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$ maintains its structure.

To make things more formal, let's start by better understanding the structure of $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$. Recall that in this case, we assumed that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^1)$ is an $((n, k), (n, k), (w, w-g))$-block source. This means that for every fixing of $\mathbf{X}, \mathbf{Y}$, $\mathbf{Z}^1$ still has min-entropy at least $w - g$. And if $\mathbf{Z}^1$ has min-entropy at least $w - g$, it is not hard to show that its prefix $\underline{\mathbf{Z}}^1 = \mathbf{Z}^1_{[p_1]}$ of length p_1 has entropy at least $p_1 - g$ (Fact 4). This tells us that $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$ is a $((n, k), (n, k), (p_1, p_1 - g))$-block source.

Next, let's better understand the structure of $(\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2)$, and how it relates to $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$. Towards this end, first note that $\mathbf{S}_2, \mathbf{W}_2$ and $\underline{\mathbf{Z}}^2$ are supported on sets of size $2^m, 2^{d_2}$ and 2^{p_2}, respectively. Then, observe the following independence relationships between $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$ and $(\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2)$:

- Upon fixing $\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^2$, the random variables $\mathbf{S}_2, \mathbf{W}_2$ become a constant. As a result, we know that $(\underline{\mathbf{Z}}^1 \mid \mathbf{X} = x, \mathbf{Y} = y, \underline{\mathbf{Z}}^2 = z_2)$ and $(\mathbf{S}_2, \mathbf{W}_2 \mid \mathbf{X} = x, \mathbf{Y} = y, \underline{\mathbf{Z}}^2 = z_2)$ are independent, $\forall x, y, z_2$.
- Upon fixing $\mathbf{X}, \mathbf{W}_2, \underline{\mathbf{Z}}^2$, the random variable $\mathbf{S}_2$ becomes a constant. As a result, we know that $(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2)$ and $(\mathbf{S}_2 \mid \mathbf{X} = x, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2)$ are independent, $\forall x, w_2, z_2$.

Because of these independence relationships between $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$ and $(\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2)$, it turns out that we can safely fix the latter sequence without severely affecting the structure of the former. In particular, by combining the above observations with our lemma on fixing randomness against block sources (Lemma 7), we immediately get the following, for any $\gamma > 0$.

$$\Pr_{(s_2, w_2, z_2) \sim (\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2)}\Big[(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1 \mid \mathbf{S}_2 = s_2, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2) \text{ is not } 3\sqrt{\gamma}\text{-close to an } ((n, k'), (n, k'), (p_1, \ell'))\text{-block source}\Big] \leq 3\sqrt{\gamma}, \tag{2}$$

where $k' = k - (m + d_2 + p_2 + \log(1/\gamma))$ and $\ell' = p_1 - (g + p_2 + \log(1/\gamma))$. The reason why this bound will be useful is because it says that with high probability over fixing $\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2$, it follows that $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$ is still a block source, and of course $\mathbf{S}_2$ is also a constant. Thus, if we can just show that $\mathbf{S}_1 = \mathsf{sExt}'_1(\mathbf{X}, \mathsf{sExt}_1(\mathbf{Y}, \underline{\mathbf{Z}}^1))$ has high entropy whenever $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1)$ is a block source, we will be done, since we just needed $\mathbf{S}_1 \oplus \mathbf{S}_2$ to have high entropy, and this is true if $\mathbf{S}_1$ has high entropy and $\mathbf{S}_2$ is constant.

More formally, consider an arbitrary $((n, k'), (n, k'), (p_1, \ell'))$-block source $(\mathbf{A}, \mathbf{B}, \mathbf{C})$. Let's analyze what $\mathsf{sExt}'_1(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C}))$ looks like. By definition of block source, we know that for every a, it holds that $(\mathbf{B}, \mathbf{C} \mid \mathbf{A} = a)$ is an $((n, k'), (p_1, \ell'))$-block source. Furthermore, recall that $\mathsf{sExt}_1 : \{0,1\}^n \times \{0,1\}^{p_1} \to \{0,1\}^{d_1}$ is a (k_0, ε_1)-seeded extractor, and thus it is trivially a seeded $(n, k_0) \to_{\varepsilon_1} (d_1, d_1)$ condenser. Since every seeded condenser also works for block sources (Lemma 8), it follows that $\mathsf{sExt}_1(\mathbf{B}, \mathbf{C} \mid \mathbf{A} = a)$ is $(2^{p_1 - \ell'}\varepsilon_1)$-close to a source $\mathbf{Q}_a \sim \{0,1\}^{d_1}$ with min-entropy at least $d_1 - (p_1 - \ell')$, provided that $k' \geq k_0$. Since it holds that $(a, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C} \mid \mathbf{A} = a)) \approx_{2^{p_1 - \ell'}\varepsilon_1} (a, \mathbf{Q}_a)$ for every fixed a, it follows that the random variables $(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C}))$ and $(\mathbf{A}, \mathbf{Q}_{\mathbf{A}})$ enjoy the same statistical distance bound. And by a straightforward application of the data-processing inequality (Fact 2), it also follows that

$$\mathsf{sExt}'_1(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C})) \approx_{2^{p_1 - \ell'}\varepsilon_1} \mathsf{sExt}'_1(\mathbf{A}, \mathbf{Q}_{\mathbf{A}}).$$

Moreover, observe that $(\mathbf{A}, \mathbf{Q_A})$ is in fact a $((n, k'), (d_1, d_1 - (p_1 - \ell')))$-block source. Repeating an identical analysis to what was done above, we can thus conclude that

$$\mathsf{sExt}'_1(\mathbf{A}, \mathbf{Q_A}) \approx_{2^{p_1-\ell'}\varepsilon'_1} \mathbf{R}^\star,$$

where $\mathbf{R}^\star \sim \{0,1\}^m$ is some source with min-entropy at least $m - (p_1 - \ell')$, provided that $k' \geq k_0$.

To summarize, we get that for any $((n, k'), (n, k'), (p_1, \ell'))$-block source $(\mathbf{A}, \mathbf{B}, \mathbf{C})$,

$$\mathsf{sExt}'_1(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C})) \approx_{2^{p_1-\ell'}(\varepsilon_1+\varepsilon'_1)} \mathbf{R}^\star,$$

where $\mathbf{R}^\star$ is some source with min-entropy at least $m - (p_1 - \ell')$, provided that $k' \geq k_0$. Moreover, it is straightforward to see that if $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is actually only ξ-close to a block source $(\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{C}^\star)$ of the above type, then the data-processing inequality (Fact 2) tells us that $\mathsf{sExt}'_1(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C}))$ is ξ-close to $\mathsf{sExt}'_1(\mathbf{A}^\star, \mathsf{sExt}_1(\mathbf{B}^\star, \mathbf{C}^\star))$, which we showed above to be $2^{p_1-\ell'}(\varepsilon_1 + \varepsilon'_1)$-close to $\mathbf{R}^\star$.

Thus, we get that for any $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ that is ξ-close to an $((n, k'), (n, k'), (p_1, \ell'))$-block source,

$$\mathsf{sExt}'_1(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C})) \approx_{2^{p_1-\ell'}(\varepsilon_1+\varepsilon'_1)+\xi} \mathbf{R}^\star,$$

where $\mathbf{R}^\star$ is an $(m, m - (p_1 - \ell'))$-source, provided that $k' \geq k_0$.

Put differently, if $\mathsf{sExt}'_1(\mathbf{A}, \mathsf{sExt}_1(\mathbf{B}, \mathbf{C}))$ is *not* $(2^{p_1-\ell'}(\varepsilon_1 + \varepsilon'_1) + \xi)$-close to any such source $\mathbf{R}^\star$, then we know that $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is not ξ-close to an $((n, k'), (n, k'), (p_1, \ell'))$-block source. Thus, if we define $g' := p_1 - \ell'$, $\xi' := 2^{p_1-\ell'}(\varepsilon_1 + \varepsilon'_1) + \xi$, and $\xi = 3\sqrt{\gamma}$, we can combine the above with Equation (2) to obtain

$$\begin{aligned}
&\Pr_{(s_2,w_2,z_2)\sim(\mathbf{S}_2,\mathbf{W}_2,\underline{\mathbf{Z}}^2)}\Big[(\mathbf{S}_1 \oplus \mathbf{S}_2 \mid \mathbf{S}_2 = s_2, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2) \text{ is not } \xi'\text{-close to an } (m, m - g')\text{-source}\Big] \\
&= \Pr_{(s_2,w_2,z_2)}\Big[(\mathbf{S}_1 \oplus s_2 \mid \mathbf{S}_2 = s_2, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2) \text{ is not } \xi'\text{-close to an } (m, m - g')\text{-source}\Big] \\
&= \Pr_{(s_2,w_2,z_2)}\Big[(\mathbf{S}_1 \mid \mathbf{S}_2 = s_2, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2) \text{ is not } \xi'\text{-close to an } (m, m - g')\text{-source}\Big] \\
&= \Pr_{(s_2,w_2,z_2)}\Big[\Big(\mathsf{sExt}'_1(\mathbf{X}, \mathsf{sExt}_1(\mathbf{Y}, \underline{\mathbf{Z}}^1))\Big) \mid \mathbf{S}_2 = s_2, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2) \text{ is not} \\
&\qquad\qquad\qquad\qquad \xi'\text{-close to an } (m, m - g')\text{-source}\Big] \\
&\leq \Pr_{(s_2,w_2,z_2)}\Big[(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^1 \mid \mathbf{S}_2 = s_2, \mathbf{W}_2 = w_2, \underline{\mathbf{Z}}^2 = z_2) \text{ is not} \\
&\qquad\qquad\qquad\qquad \xi\text{-close to an } ((n, k'), (n, k'), (p_1, \ell'))\text{-block source}\Big] \\
&\leq \xi.
\end{aligned}$$

To summarize, we've shown that there exists a random variable $\mathbf{V} := (\mathbf{S}_2, \mathbf{W}_2, \underline{\mathbf{Z}}^2)$ such that

$$\Pr_{v\sim\mathbf{V}}\big[(\mathbf{S}_1 \oplus \mathbf{S}_2 \mid \mathbf{V} = v) \text{ is not } \xi'\text{-close to an } (m, m - g')\text{-source}\big] \leq \xi.$$

By a standard fact about convex combinations (Fact 3), it immediately follows that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is ξ-close to a convex combination of sources that *are* ξ'-close to an $(m, m - g')$-source. As such, it holds that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is

$(\xi+\xi')$-close to a convex combination of $(m, m-g')$-sources. Since a convex combination of $(m, m-g')$-sources is, itself, an $(m, m-g')$-source, we conclude that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is $(\xi+\xi')$-close to an $(m, m-g')$-source. Finally, recall that

$$\xi + \xi' = 6\sqrt{\gamma} + 2^{p_1-\ell'}(\varepsilon_1 + \varepsilon_1') = 6\sqrt{\gamma} + 2^{g+p_2+\log(1/\gamma)}(\varepsilon_1 + \varepsilon_1'),$$

for any $\gamma > 0$. If we set $\varepsilon_1 = \varepsilon_1'$ and $\gamma = (2\varepsilon_1)^{1/2}$, this is at most $2^{g+p_2+3}\varepsilon_1^{1/4}$. Furthermore, recall that

$$g' = p_1 - \ell' = g + p_2 + \log(1/\gamma) \leq g + p_2 + \log(1/\varepsilon_1)$$

Recall that to make everything work, we needed $k' \geq k_0$, and plugging in our definition of k' from before, this requirement becomes (no worse than)

$$k \geq k_0 + m + d_2 + p_2 + \log(1/\gamma) = k_0 + m + d_2 + p_2 + \log(1/\varepsilon_1).$$

Thus, we get that the output of the non-malleable condenser is $2^{g+p_2+3}\varepsilon_1^{1/4}$-close to an $(m, m-(g+p_2+\log(1/\varepsilon_1)))$-source, as long as $k \geq k_0 + m + d_2 + p_2 + \log(1/\varepsilon_1)$ and $\varepsilon_1 = \varepsilon_1'$.

Case 2. We now proceed to the second case, where we assume that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^2)$ is the block source. In order to show that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is ε-close to an (m, r)-source, we now seek a sequence of fixings that will force $\mathbf{S}_1$ to be constant, but under which $\mathbf{S}_2$ can be shown to have high min-entropy.

To start, recall that $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}^2)$ is an $((n,k),(n,k),(w,w-g))$-block source. Using an identical argument to the one appearing at the beginning of the previous case, we know this implies that $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^2)$ is an $((n,k),(n,k),(p_2,p_2-g))$-block source. Next, we'd like to argue that $(\mathbf{X}, \mathbf{W}_2)$ is close to a block source $(\mathbf{X}', \mathbf{W}_2')$. For technical reasons that we will soon see, we actually need a slightly more involved result. In particular, we need to show there is a sequence $(\mathbf{X}', \mathbf{W}_2', \mathbf{S}_1', \mathbf{W}_1')$ such that:

- $(\mathbf{X}', \mathbf{W}_2', \mathbf{S}_1', \mathbf{W}_1')$ is close to $(\mathbf{X}, \mathbf{W}_2, \mathbf{S}_1, \mathbf{W}_1)$,
- $(\mathbf{X}', \mathbf{W}_2')$ is a block source, and
- $\mathbf{S}_1'$ is constant upon any fixing of $\mathbf{X}', \mathbf{W}_1'$.

We start by constructing $(\mathbf{X}', \mathbf{W}_2')$. To do so, first recall that $(\mathbf{X}, \mathbf{Y}, \underline{\mathbf{Z}}^2)$ is a $((n,k),(n,k),(p_2,p_2-g))$-block source. This means that for every fixed x, $(\mathbf{Y}, \underline{\mathbf{Z}}^2 \mid \mathbf{X} = x)$ is an $((n,k),(p_2,p_2-g))$-block source. Now, recall that $\mathsf{sExt} : \{0,1\}^n \times \{0,1\}^{p_2} \to \{0,1\}^{d_2}$ is a (k_0, ε_2)-seeded extractor, and is therefore also a seeded $(n, k_0) \to_{\varepsilon_2} (d_2, d_2)$ condenser. Since every seeded condenser also works for block sources (Lemma 8), it follows that $\mathsf{sExt}_2(\mathbf{Y}, \underline{\mathbf{Z}}^2 \mid \mathbf{X} = x)$ is $2^g\varepsilon_2$-close to a source $\mathbf{Q}_x \sim \{0,1\}^{d_2}$ with min-entropy at least $d_2 - g$, provided that $k \geq k_0$. Since it holds that

$$(x, \mathsf{sExt}_2(\mathbf{Y}, \underline{\mathbf{Z}}^2 \mid \mathbf{X} = x)) \approx_{2^g\varepsilon_2} (x, \mathbf{Q}_x)$$

for every fixed x, it follows that $(\mathbf{X}, \mathbf{W}_2) = (\mathbf{X}, \mathsf{sExt}_2(\mathbf{Y}, \underline{\mathbf{Z}}^2)) \approx_{2^g\varepsilon_2} (\mathbf{X}, \mathbf{Q}_\mathbf{X})$. Moreover, observe that $(\mathbf{X}, \mathbf{Q}_\mathbf{X})$ is an $((n,k),(d_2,d_2-g))$-block source. We define $(\mathbf{X}', \mathbf{W}_2') = (\mathbf{X}, \mathbf{Q}_\mathbf{X})$.

Next, let us proceed with constructing $\mathbf{S}_1', \mathbf{W}_1'$. This is not too difficult. First, we define $\mathbf{W}_1'$ by asserting that for every fixed x, w_2,

$$(\mathbf{W}_1' \mid \mathbf{X}' = x, \mathbf{W}_2' = w_2) \equiv \begin{cases} (\mathbf{W}_1 \mid \mathbf{X} = x, \mathbf{W}_2 = w_2) & \text{if } (x, w_2) \in \text{support}(\mathbf{X}, \mathbf{W}_2) \\ \mathbf{U} & \text{otherwise.} \end{cases}$$

Then, we define $\mathbf{S}_1' := \mathsf{sExt}_1'(\mathbf{X}', \mathbf{W}_1')$. This trivially satisfies the condition that $\mathbf{S}_1'$ is constant upon any fixing of $\mathbf{X}', \mathbf{W}_1'$. Moreover, recall from above that $(\mathbf{X}', \mathbf{W}_2')$ is an $((n,k),(d_2,d_2-g))$-block source. Thus all that remains is to show that $(\mathbf{X}', \mathbf{W}_2', \mathbf{S}_1', \mathbf{W}_1')$ is close to $(\mathbf{X}, \mathbf{W}_2, \mathbf{S}_1, \mathbf{W}_1)$. To see why this is true, first observe that by construction, it holds that for any $(x, w_2) \in$ support$(\mathbf{X}, \mathbf{W}_2)$,

$$\big((\mathbf{X}', \mathbf{W}_2', \mathbf{S}_1', \mathbf{W}_1') \mid \mathbf{X}' = x, \mathbf{W}_2' = w_2\big) \equiv ((\mathbf{X}, \mathbf{W}_2, \mathbf{S}_1, \mathbf{W}_1 \mid \mathbf{X} = x, \mathbf{W}_2 = w_2)).$$

Combining this with the fact that $(\mathbf{X}', \mathbf{W}_2')$ is $2^g\varepsilon_2$-close to $(\mathbf{X}, \mathbf{W}_2)$ by construction, it is straightforward to verify that

$$(\mathbf{X}', \mathbf{W}_2', \mathbf{S}_1', \mathbf{W}_1') \approx_{2^g\varepsilon_2} (\mathbf{X}, \mathbf{W}_2, \mathbf{S}_1, \mathbf{W}_1).$$

Thus, we have successfully constructed a sequence $(\mathbf{X}', \mathbf{W}_2', \mathbf{S}_1', \mathbf{W}_1')$ with all of the properties originally desired. Now, let's see how to use it.

Recall that we originally wanted to show that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is ε-close to an (m,r)-source, and planned to do so by performing some fixings that force $\mathbf{S}_1$ to be constant. The fixings that we will perform are exactly on the random variables $(\mathbf{S}_1, \mathbf{W}_1)$. To analyze the probability that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is ε-close to an (m,r)-source under these fixings, we will use the above-constructed sequence for help. In more detail, let ξ and g' be parameters that we will set later. Then, note that

$$\begin{aligned}
&\Pr_{(s_1,w_1)\sim(\mathbf{S}_1,\mathbf{W}_1)}\Big[(\mathbf{S}_1 \oplus \mathbf{S}_2 \mid \mathbf{S}_1 = s_1, \mathbf{W}_1 = w_1) \text{ is not } \xi\text{-close to an } (m, m-g')\text{-source}\Big]\\
=&\Pr_{(s_1,w_1)\sim(\mathbf{S}_1,\mathbf{W}_1)}\Big[(s_1 \oplus \mathbf{S}_2 \mid \mathbf{S}_1 = s_1, \mathbf{W}_1 = w_1) \text{ is not } \xi\text{-close to an } (m, m-g')\text{-source}\Big]\\
=&\Pr_{(s_1,w_1)\sim(\mathbf{S}_1,\mathbf{W}_1)}\Big[(\mathbf{S}_2 \mid \mathbf{S}_1 = s_1, \mathbf{W}_1 = w_1) \text{ is not } \xi\text{-close to an } (m, m-g')\text{-source}\Big]\\
=&\Pr_{(s_1,w_1)\sim(\mathbf{S}_1,\mathbf{W}_1)}\Big[\big(\mathsf{sExt}_2'(\mathbf{X}, \mathbf{W}_2) \mid \mathbf{S}_1 = s_1, \mathbf{W}_1 = w_1\big) \text{ is not } \xi\text{-close to an } (m, m-g')\text{-source}\Big].
\end{aligned}$$

Now, since we know that $(\mathbf{S}_1, \mathbf{W}_1, \mathbf{X}, \mathbf{W}_2)$ is $2^g\varepsilon_2$-close to $(\mathbf{S}_1', \mathbf{W}_1', \mathbf{X}', \mathbf{W}_2')$, we can apply Claim 1 to upper bound the above by

$$\begin{aligned}
\le \Pr_{(s_1,w_1)\sim(\mathbf{S}_1',\mathbf{W}_1')}\Big[&(\mathsf{sExt}_2'(\mathbf{X}', \mathbf{W}_2') \mid \mathbf{S}_1' = s_1, \mathbf{W}_1' = w_1) \text{ is not}\\
&\xi/2\text{-close to an } (m, m-g')\text{-source}\Big] + 4 \cdot 2^g\varepsilon_2/\xi + 2^g\varepsilon_2.
\end{aligned}\tag{3}$$

In order to continue bounding this probability, we can now apply our fixing lemma (Lemma 7) as follows. First, note that we are dealing with random variables $(\mathbf{X}', \mathbf{W}_2')$ and $(\mathbf{S}_1', \mathbf{W}_1')$, where $(\mathbf{X}', \mathbf{W}_2')$ is an $((n,k),(d_2,d_2-g))$-block source, and $\mathbf{S}_1', \mathbf{W}_1'$ are supported on sets of size 2^m and 2^{d_1}, respectively. Furthermore, note that $(\mathbf{W}_2' \mid \mathbf{X}' = x, \mathbf{W}_1' = w_1)$ and $(\mathbf{S}_1' \mid \mathbf{X}' = x, \mathbf{W}_1' = w_1)$ are independent, for all fixed x, w_1. Indeed, this is simply because we constructed $\mathbf{S}_1'$ to be constant upon any fixing of $\mathbf{X}', \mathbf{W}_1'$. Plugging these observations into Lemma 7, we immediately get that

$$\Pr_{(s_1,w_1)\sim(\mathbf{S}_1',\mathbf{W}_1')}\Big[(\mathbf{X}', \mathbf{W}_2' \mid \mathbf{S}_1' = s_1, \mathbf{W}_1' = w_1) \text{ is not } 2\sqrt{\nu}\text{-close to an } ((n,k''),(d_2,\ell''))\text{-block source}\Big] \le 2\sqrt{\nu},$$

where $k'' = k - (m + d_1 + \log(1/\nu))$ and $\ell'' = d_2 - (g + d_1 + \log(1/\nu))$. Now, consider any $((n,k''),(d_2,\ell''))$-block source $(\mathbf{A}, \mathbf{B})$, and think about what happens when you plug it into $\mathsf{sExt}_2' : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^m$, which also works as a seeded $(n, k_0) \to_{\varepsilon_2'} (m,m)$ condenser. Since every seeded condenser also works for block sources (Lemma 8), it follows that $\mathsf{sExt}_2'(\mathbf{A}, \mathbf{B})$ is $(2^{d_2-\ell''} \cdot \varepsilon_2')$-close to a source with min-entropy at least $m - (d_2 - \ell'')$, provided that $k'' \ge k_0$. Moreover, as we saw before, if $(\mathbf{A}, \mathbf{B})$ is η-close to an $((n, k''), (d_2, \ell''))$-block source, then $\mathsf{sExt}'_2(\mathbf{A}, \mathbf{B})$ is still guaranteed to be $(2^{d_2-\ell''} \cdot \varepsilon'_2 + \eta)$-close to a source with min-entropy at least $m - (d_2 - \ell'')$. Put differently, if $\mathsf{sExt}'_2(\mathbf{A}, \mathbf{B})$ were *not* this close to such a high entropy source, then we know that $(\mathbf{A}, \mathbf{B})$ is also not η-close to an $((n, k''), (d_2, \ell''))$-block source.

By the discussion above, we know that if we set $\xi/2 := (2^{d_2-\ell''}\varepsilon'_2 + \eta)$ and $\eta = 2\sqrt{\nu}$ and $g' = d_2 - \ell''$, we can upper bound Equation (3) by

$$2\sqrt{\nu} + 4 \cdot 2^g \varepsilon_2/\xi + 2^g \varepsilon_2 \leq 2\sqrt{\nu} + 5 \cdot 2^g \varepsilon_2/\xi.$$

In summary, we get that

$$\Pr_{(s_1,w_1)\sim(\mathbf{S}_1,\mathbf{W}_1)} \Big[(\mathbf{S}_1 \oplus \mathbf{S}_2 \mid \mathbf{S}_1 = s_1, \mathbf{W}_1 = w_1) \text{ is not } \xi\text{-close to an } (m, m - g')\text{-source} \Big] \leq 2\sqrt{\nu} + 5 \cdot 2^g \varepsilon_2/\xi.$$

By a standard fact about convex combinations (Fact 3), it immediately follows that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is $(2\sqrt{\nu} + 5 \cdot 2^g \varepsilon_2/\xi)$-close to a convex combination of sources that are ξ-close to an $(m, m - g')$-source. As such, it holds that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is $(2\sqrt{\nu} + 5 \cdot 2^g \varepsilon_2/\xi + \xi)$-close to a convex combination of $(m, m - g')$-sources. And since a convex combination of $(m, m - g')$-sources is, itself, an $(m, m - g')$-source, we conclude that $\mathbf{S}_1 \oplus \mathbf{S}_2$ is $(2\sqrt{\nu} + 5 \cdot 2^g \varepsilon_2/\xi + \xi)$-close to an $(m, m - g')$-source. Finally, recall that

$$\begin{aligned} 2\sqrt{\nu} + 5 \cdot 2^g \varepsilon_2/\xi + \xi &= 6\sqrt{\nu} + \frac{5 \cdot 2^g \varepsilon_2}{2^{d_2-\ell''}\varepsilon'_2 + 4\sqrt{\nu}} + 2^{d_2-\ell''}\varepsilon'_2 \\ &= 6\sqrt{\nu} + \frac{5 \cdot 2^g \varepsilon_2}{2^{g+d_1+\log(1/\nu)}\varepsilon'_2 + 4\sqrt{\nu}} + 2^{g+d_1+\log(1/\nu)}\varepsilon'_2, \end{aligned}$$

where $\nu > 0$ can be taken as anything. Taking it to be $\nu := \sqrt{\varepsilon'_2}$ allows us to upper bound the above by

$$\leq 6(\varepsilon'_2)^{1/4} + \frac{6\varepsilon_2}{2^{d_1}\sqrt{\varepsilon'_2}} + 2^{g+d_1}\sqrt{\varepsilon'_2}.$$

Then, taking $\varepsilon'_2 = \varepsilon_2 \cdot 2^{-2d_1}$ allows us to upper bound the above by $\leq 2^{g+4} \cdot \varepsilon_2^{1/4}$.

Furthermore, recall that

$$g' = d_2 - \ell'' = g + d_1 + \log(1/\nu) = g + 2d_1 + \log(1/\varepsilon_2)/2.$$

Recall that to make everything work, we needed $k'' \geq k_0$, and plugging in our definition of k'' from before, this requirement becomes

$$k \geq k_0 + m + d_1 + \log(1/\nu) = k_0 + m + 2d_1 + \log(1/\varepsilon_2)/2.$$

Thus, we get that the output of the non-malleable condenser is $2^{g+4}\varepsilon_2^{1/4}$-close to an $(m, m - (g + 2d_1 + \log(1/\varepsilon_2)/2))$-source, as long as $k \geq k_0 + m + 2d_1 + \log(1/\varepsilon_2)/2$ and $\varepsilon'_2 = \varepsilon_2 \cdot 2^{-2d_1}$. □

5.3 The main explicit condenser

Using the tools developed above, we can now construct our main explicit condenser for CG sources.

Theorem 5 (The main explicit condenser for CG sources - Theorem 1, restated)**.** *For any $\alpha > 0$, there exists a constant $C \geq 1$ such that the following holds. For all $t, n \in \mathbb{N}$ and $\delta, \varepsilon > 0$, there exists an explicit condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^{k'+g'}$ for $(t, n, k = \delta n = n - g)$-CG sources which has output entropy $k' \geq (1-\alpha)kt$, output gap $g' \leq C \cdot (1/\delta)^C \cdot (g + \log(1/\varepsilon))$, and error ε.*

The proof proceeds via three steps. First, in Section 5.3.1, we explicitly construct a non-malleable condenser for CG sources (using our framework from Section 5.2). Then, in Section 5.3.2, we present the main part of our condenser, which uses our new non-malleable condenser in the "purification" framework from Section 5.1 in order to condense CG sources to rate 0.99. Finally, in Section 5.3.3, we show how to get the remaining entropy out of the source, while maintaining a very small gap, by showing that the classical iterative condensing framework of Nisan and Zuckerman [NZ96] can be extended to handle a correlated seed.

5.3.1 Building a non-malleable condenser

We proceed to build our non-malleable condenser for CG sources. We prove the following.

Theorem 6 (Explicit non-malleable condensers)**.** *For every constant $\alpha > 0$, there exist constants $C \geq 1$ and $\gamma > 0$ such that the following holds. There exists an explicit non-malleable condenser (with advice) $\mathsf{nmCond} : \{0,1\}^n \times \{0,1\}^n \times \{0,1\}^d \times [2] \to \{0,1\}^m$ for $((n,k),(n,k),(d,(1-\gamma)d))$-block sources with error ε, output length $m = \lfloor (\frac{1}{2} - \alpha)k - C\log(n/\varepsilon) - d \rfloor$, and output gap $g' \leq C\log(n/\varepsilon) + d$, provided that $d \geq C\log(n/\varepsilon)$.*

Our construction will follow by simply plugging in known seeded extractors into our recipe from Section 5.2. We will use the following classical extractors of Guruswami, Umans, and Vadhan.

Theorem 7 (Explicit seeded extractors [GUV09])**.** *For every constant $\alpha > 0$, there is a constant $C > 0$ such that the following holds. There exists an explicit (k,ε)-seeded extractor $\mathsf{sExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with output length $m \geq (1-\alpha)k$, as long as $d \geq C\log(n/\varepsilon)$.*

With this tool in hand, we are ready to construct our non-malleable condensers.

Proof of Theorem 6. We simply plug Theorem 7 into Lemma 11, and pick parameters appropriately.

In more detail, we need to find extractors

- $\mathsf{sExt}_1 : \{0,1\}^n \times \{0,1\}^{p_1} \to \{0,1\}^{d_1}$ a (k_0, ε_1)-seeded extractor,
- $\mathsf{sExt}_1' : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m}$ a (k_0, ε_1')-seeded extractor,
- $\mathsf{sExt}_2 : \{0,1\}^n \times \{0,1\}^{p_2} \to \{0,1\}^{d_2}$ a (k_0, ε_2)-seeded extractor,
- $\mathsf{sExt}_2' : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{m}$ a (k_0, ε_2')-seeded extractor,

with parameters $p_1, d_1, p_2, d_2, d_1, m, d_2, k_0, \varepsilon_1, \varepsilon_1', \varepsilon_2, \varepsilon_2'$ that result in the non-malleable condensers we claim (using Lemma 11). To make this easy, we start by focusing on achieving error ε. In order to do so, we define $g := \gamma d$ (for some constant γ to be fixed later), and note that Lemma 11 says that we can just pick ε_1 such that $2^{g+p_2+3}\varepsilon_1^{1/4} \leq \varepsilon/2$ and $2^{g+4}\varepsilon_2^{1/4} \leq \varepsilon/2$. Moreover, it always requires that we have $\varepsilon_1 = \varepsilon_1'$ and $\varepsilon_2' = \varepsilon_2 \cdot 2^{-2d_1}$. Thus we pick errors $\varepsilon_1 = \varepsilon^4 2^{-4(g+p_2+4)}$, $\varepsilon_1' = \varepsilon_1$, $\varepsilon_2 = \varepsilon^4 \cdot 2^{-(g+5)4}$, and $\varepsilon_2' = \varepsilon_2^{-2d_1}$. This satisfies the error requirement.

Now, in order to explicitly construct these extractors, we invoke Theorem 7 so that we can handle the smallest possible seed length. Thus, we pick

- $p_2 = C\log(n/\varepsilon_2) = 4C \cdot (\log(n/\varepsilon) + g + 5) = O(\log(n/\varepsilon) + g)$,
- $p_1 = C\log(n/\varepsilon_1) = 4C \cdot (\log(n/\varepsilon) + g + p_2 + 4) = O(\log(n/\varepsilon) + g)$,

- $d_1 = C\log(n/\varepsilon_1') = 4C \cdot (\log(n/\varepsilon) + g + p_2 + 4) = O(\log(n/\varepsilon) + g)$,
- $d_2 = C\log(n/\varepsilon_2') = 4C \cdot (\log(n/\varepsilon) + g + 5 + d_1/2) = O(\log(n/\varepsilon) + g)$.

To make this work, Lemma 11 also says that we need $k_0 \leq k - (m + 2d_1 + d_2 + p_2 + \log(1/\varepsilon_1) + \log(1/\varepsilon_2))$, and recall that the right hand side is at least $k - m - O(\log(n/\varepsilon) + g)$. Thus we pick k_0 to be this value, and all that remains is to check that we didn't ask one of the extractors to output more bits than $(1-\alpha)k_0$. For this, we simply need that $m \leq (1-\alpha)k_0 = (1-\alpha)(k - m - O(\log(n/\varepsilon) + g))$, or rather that $m \leq (1/2 - \alpha)k - O(\log(n/\varepsilon) + g)$. Furthermore, recall that p_1, p_2 are prefixes of d, so we need $d \geq p_1, p_2 = O(\log(n/\varepsilon) + g)$. Now that all the conditions are satisfied, we get from Lemma 11 that the entropy gap is $O(g + \log(n/\varepsilon))$. To conclude, recall that $g = \gamma d$, and set γ to a sufficiently small constant. □

5.3.2 Condensing to rate 0.99

Now that we have our non-malleable condensers, we are ready to construct the core component of our main explicit condenser for CG sources. In this section, we prove the following.

Lemma 12 (Condensing to rate 0.99)**.** *For any constants $\alpha, C_0 > 0$, there exist constants $C_1, C_2, C_3 \geq C_0$ such that the following holds. There exists an explicit condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$ for $(t, n, \delta n)$-CG sources with output length $m \in [0.05\delta n\tau^\star, \delta n\tau^\star]$, output entropy $k' \geq (1-\alpha)m$, and error ε, provided*

$$t \geq \tau^\star := C_1 \cdot \left((1/\delta)^{C_2} + (1/\delta)^{C_3}\log(1/\varepsilon)/n\right).$$

As discussed, the key idea is to instantiate our purification framework from Section 5.1 with a baseline somewhere-condenser and a non-malleable condenser. For our non-malleable condenser, we'll use the new one constructed above. For the baseline somewhere-condenser, we'll use a classical construction due to Barak, Kindler, Shaltiel, Sudakov, and Wigderson [BKS$^+$10] and Raz [Raz05] (see also [Zuc07, Theorem 3.2]).

Theorem 8 (Explicit somewhere-condensers [BKS$^+$10, Raz05])**.** *For every constant $\beta > 0$, there exist constants $C_1, C_2, C_3 \geq 1$ such that the following holds. For any $\delta = \delta(n) > 0$, there exists an explicit somewhere-k'-condenser $\mathsf{sCond} : \{0,1\}^n \to (\{0,1\}^m)^D$ for $(n, \delta n)$-sources with output length $m = \lfloor \delta^{C_1} n \rfloor$, output entropy $k' \geq (1-\beta)m$, error $\varepsilon = 2^{-\delta^{C_2} n}$, and $D = \lceil (1/\delta)^{C_3} \rceil$ rows.*

We are now ready to condense CG sources to rate 0.99, and prove the core lemma of this paper.

Proof of Lemma 12. Let $\mathbf{X} \sim (\{0,1\}^n)^t$ be a $(t, n, k := \delta n)$-CG source. The idea is to expand the last block of $\mathbf{X}$ into a somewhere-random (SR) source (using Theorem 8), and then proceed in iterations. In each iteration, we will halve the number of rows in the SR source, using our non-malleable condenser (Theorem 6) and our purification lemma (Lemma 10).

At a high level, in order for this to work, the *row length* of the SR source must line up with the *seed length* requirement of the non-malleable condenser, and the *entropy rate* of the (good row of the) SR source must line up with the *seed (entropy) rate* requirement of the non-malleable condenser (which is roughly 0.99). Furthermore, after we have halved the number of rows in the SR source with one application of the purification lemma, we need to make sure that the new SR source has a row length and row entropy rate that is good enough for another application of the purification lemma. To make sure this happens, the output entropy rate of the first non-malleable condenser calls must be at least 0.99. But since the output gap of the non-malleable condenser is always a constant factor larger than the gap of its seed (see Theorem 6), we must make sure that its output length is also a constant factor larger than the length of its seed. And to make this happen, we must make sure that each of the two input sources to the non-malleable condenser has enough min-entropy. This is possible by concatenating several blocks of the CG source into a single "super-block." Finally, we will continue to iterate until there is just a single row left in the SR source.

Thus, the game plan is as follows. First, we fix an arbitrary constant $\alpha > 0$, which will represent the allowed missing entropy rate in the final output of the condenser. Then, we let $\varepsilon > 0$ denote another parameter, which will represent the target final error of the condenser.[19] We also set up intermediate error values $\varepsilon_0, \varepsilon_1, \varepsilon_2, \ldots$, which will represent the allowed error in each iteration of the procedure outlined above. We make sure that these are in a decaying geometric series, so that they will sum up to our overall target error ε. Finally, we determine the number of blocks that must be concatenated at each iteration (before passing them into the non-malleable condenser to collapse the SR source) in order satisfy all the requirements mentioned above. At the end, we sum up the total number of blocks we needed to fully collapse the SR source, and define τ to be exactly this value, in order to finish the proof.

More formally now, fix a parameter $\beta > 0$ to either α (from the current lemma statement) or γ (from Theorem 6, when its first parameter is fixed to 0.01) - whichever is smaller.[20] Then, let $b_0 \in \mathbb{N}$ be a "block parameter" that we will fix later, and let $\mathsf{sCond}_0 : \{0,1\}^{nb_0} \to (\{0,1\}^{m_0})^D$ be an explicit somewhere-k'_0-condenser for $(nb_0, \delta nb_0)$-sources with output length $m_0 = \lfloor \delta^{C_1} nb_0 \rfloor$, output entropy $k'_0 \geq (1-\beta)m_0$, error $\varepsilon_0 = 2^{-\delta^{C_2} nb_0}$, and $D = \lceil (1/\delta)^{C_3} \rceil_2$ rows, where $\lceil x \rceil_2$ denotes the rounding of x up to the closest power of 2. Such an explicit somewhere-condenser exists due to Theorem 8.[21]

Next, suppose there exists a sequence of explicit functions $\mathsf{nmCond}_1, \mathsf{nmCond}_2, \ldots, \mathsf{nmCond}_d$, where each nmCond_i is an explicit non-malleable condenser (with advice) for $((nb_i, \kappa_i), (nb_i, \kappa_i), (m_{i-1}, k'_{i-1}))$-block sources with error ε_i, output length $m_i = \lfloor 0.49\kappa_i - C_4 \log(nb_i/\varepsilon_i) - m_{i-1} \rfloor$, and output entropy $k'_i \geq m_i - C_4 \log(nb_i/\varepsilon_i) - m_{i-1}$, where:

- $\kappa_i := kb_i - d - \log(1/\varepsilon_i)$,
- C_4 is the constant C from Theorem 6 (when the first constant in that theorem is set to 0.01), and
- all other parameters (appearing above) not yet set will be set later.

If such a sequence of explicit functions $\mathsf{nmCond}_1, \ldots, \mathsf{nmCond}_d$ actually exists, our purification lemma (Lemma 10) immediately tells us that we can use them (along with sCond_0) to iteratively create a sequence of explicit somewhere-condensers $\mathsf{sCond}_1, \ldots, \mathsf{sCond}_d$, where sCond_d is in fact an explicit condenser for (τ, n, k)-CG sources with error

$$\varepsilon_0 + \sum_{i=1}^{d} (4\sqrt{\varepsilon_i} + \varepsilon_i),$$

output length m_d, output entropy k'_d, and $\tau = b_0 + 2\sum_{i=1}^{d} b_i$.

Thus, all that remains is to set the error parameters $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_d$ and block parameters $b_0, b_1, \ldots, b_d$ so that (1) the explicit non-malleable condensers (described above) actually exist, (2) the overall error is at most ε, and (3) the output length $m := m_d$ is in the range $m_d \in [\delta n\tau/4, \delta n\tau]$, (4) the output entropy $k' := k_d$ satisfies $k'_d \geq (1-\beta)m_d$, and (5) the threshold value τ matches its value in the lemma statement.

[19] Note that ε can depend on all other parameters arbitrarily, and thus need not be a constant.

[20] Furthermore, if both $\alpha, \gamma > 1/2$, set $\beta := 1/2$.

[21] Theorem 8 technically doesn't guarantee that the number of rows will be a power of 2, but we can easily make this happen by appending the appropriate number of dummy rows (each set to the all zeroes string) to the output of the somewhere-condenser.

Let's start by satisfying the error requirement, listed as item (2) above. Towards this end, recall that we actually already set $\varepsilon_0 := 2^{-\delta^{C_2}nb_0}$ above, so we can only control the parameter ε_0 via the unset parameter b_0. On the other hand, we have not yet set the other error parameters. We do so now, and set $\varepsilon_i := (\frac{\varepsilon}{10\cdot 2^i})^2$ for every $i \in [d]$, making the overall error of the condenser at most $\varepsilon_0 + \varepsilon/2$. Thus, in order to ensure that the overall error is at most ε, we just need that $\varepsilon_0 = 2^{-\delta^{C_2}nb_0} \leq \varepsilon/2$, or rather that $b_0 \geq \frac{\log(2/\varepsilon)}{\delta^{C_2}n}$. Thus, the only unset parameters remaining are the block parameters $b_0, b_1, \ldots, b_d$, and as long as we ultimately set b_0 so that it satisfies the above inequality, then the error requirement will be satisfied.

We now turn towards satisfying requirement (1) from above, which states that the explicit non-malleable condensers $\mathsf{nmCond}_1, \ldots, \mathsf{nmCond}_d$ actually exist. In order for this to happen, we just need to make sure that each non-malleable condenser is given a long enough seed, and that this seed has a high enough entropy rate (as dictated by Theorem 6). Towards this end, notice that for each $i \in [d]$, the non-malleable condenser nmCond_i defined above is given an (m_{i-1}, k'_{i-1})-source as a seed, where

$$m_0 = \lfloor \delta^{C_1} n b_0 \rfloor, \tag{4}$$
$$k'_0 = (1-\beta)m_0, \tag{5}$$

and for every $i \in [2, d]$,

$$m_{i-1} = \lfloor 0.49\kappa_{i-1} - C_4 \log(nb_{i-1}/\varepsilon_{i-1}) - m_{i-2} \rfloor, \tag{6}$$
$$k'_{i-1} \geq m_{i-1} - C_4 \log(nb_{i-1}/\varepsilon_{i-1}) - m_{i-2}, \tag{7}$$

where recall that we defined

$$\begin{aligned} \kappa_{i-1} &:= kb_{i-1} - d - \log(1/\varepsilon_{i-1}) \\ &= kb_{i-1} - \log\lceil(1/\delta)^{C_3}\rceil_2 - \log(1/\varepsilon_{i-1}) \\ &\geq kb_{i-1} - C_3\log(1/\delta) - 1 - \log(1/\varepsilon_{i-1}). \end{aligned}$$

Now, Theorem 6 tells us that in order for the non-malleable condensers to exist, we just need the following:

- **Sufficient seed length:** $m_{i-1} \geq C_4 \log(nb_i/\varepsilon_i)$ for all $i \in [d]$.
- **Sufficient seed entropy:** $k'_{i-1} \geq (1-\beta)m_{i-1}$ for all $i \in [d]$.

Notice that when $i = 1$, the sufficient seed entropy condition is already satisfied. And when $i > 1$, the sufficient seed entropy condition becomes $m_{i-1} \geq \frac{C_4}{\beta}\log(nb_{i-1}/\varepsilon_{i-1}) + \frac{1}{\beta}m_{i-2}$, due to the known lower bound on k'_{i-1} given earlier. Thus, we just need to set block parameters so that the following are satisfied:

$$\begin{aligned} m_0 &\geq C_4 \log(nb_1/\varepsilon_1), \\ m_{i-1} &\geq C_4\log(nb_i/\varepsilon_i) + C_4\log(nb_{i-1}/\varepsilon_{i-1})/\beta + m_{i-2}/\beta, \text{for all } i \in [2, d]. \end{aligned}$$

In order to make sure the above inequalities are satisfied, let us make them easier to digest. To do so, recall that we previously set the intermediate error parameters so that $1 \geq \varepsilon_1 \geq \cdots \geq \varepsilon_d$, and we will later set block parameters so that $2 \leq b_1 \leq \cdots \leq b_d$.[22] Next, note that $m_{i-2} \leq kb_{i-2}$ for all $i \geq 2$. Using these

[22] This will allow us to use convenient estimates, such as $\log(b_{i-1}) \geq 1$ and $\log(\frac{b_i b_{i-1}}{\varepsilon_i \varepsilon_{i-1}}) \leq 2\log(b_i/\varepsilon_i)$ for all $i \geq 2$.

observations, it is straightforward to plug in the actual values for m_{i-1} (from Equations (4) and (6)) so that the conditions above (that we need to satisfy) are satisfied if both of the following hold:

$$\begin{aligned}
&\delta^{C_1} n b_0 \geq 2C_4 \log(nb_1/\varepsilon_1),\\
&kb_{i-1} \geq 3C_3 \log(1/\delta) + \frac{18C_4}{\beta}\log(\frac{nb_i}{\varepsilon_i}) + \frac{6}{\beta}kb_{i-2}, \text{ for all } i \in [2,d].
\end{aligned}$$

Now, recall that we previously defined the error parameters as $\varepsilon_i = (\frac{\varepsilon}{10\cdot 2^i})^2$, for all $i \in [d]$. Thus we have $\varepsilon_1 \geq (\varepsilon/20)^2$, and since we previously defined $d = \log\lceil(1/\delta)^{C_3}\rceil$, we get $\varepsilon_i \geq (\varepsilon\delta^{C_3}/20)^2$ for all $i \in [2,d]$. Thus the two conditions above are satisfied if both of the following hold:

$$\begin{aligned}
&\delta^{C_1} n b_0 \geq 24C_4 \log(nb_1/\varepsilon),\\
&kb_{i-1} \geq \frac{39C_3C_4}{\beta}\log(1/\delta) + \frac{216C_4}{\beta}\log(nb_i/\varepsilon) + \frac{6}{\beta}kb_{i-2}, \text{ for all } i \in [2,d].
\end{aligned}$$

We can rewrite these in terms of block requirements as follows (recalling that $k = \delta n$):

$$\begin{aligned}
&b_0 \geq \frac{24C_4}{\delta^{C_1} n}\log(nb_1/\varepsilon),\\
&b_{i-1} \geq \frac{39C_3C_4}{\beta\delta n}\log(1/\delta) + \frac{216C_4}{\beta\delta n}\log(nb_i/\varepsilon) + \frac{6}{\beta}b_{i-2}, \text{ for all } i \in [2,d].
\end{aligned}$$

Now, if we define the constant $C_5 := 256C_1C_2C_3C_4/\beta$, then the above conditions are satisfied if both

$$b_0 \geq C_5 \log(nb_1/\varepsilon)/(\delta^{C_5} n), \tag{8}$$

$$b_{i-1} \geq C_5 \log(nb_i/\varepsilon)/(\delta^{C_5} n) + C_5 b_{i-2}, \text{ for all } i \in [2,d]. \tag{9}$$

Furthermore, recall that in order for the overall condenser to have error ε, we needed $b_0 \geq \log(2/\varepsilon)/(\delta^{C_2} n)$, and this is indeed implied by the first condition above. Thus, we have arrived at sufficient conditions on the block parameters $b_0, \ldots, b_d$ for the explicit non-malleable condensers $\mathsf{nmCond}_1, \ldots, \mathsf{nmCond}_d$ to actually exist, and for the overall error of the final condenser to be at most ε. In fact, using an almost identical argument to the one given above, it is also straightforward to show that the overall output length m_d is in the range $m_d \in [0.4kb_d, 0.49kb_d]$, and the overall output entropy is $k'_d \geq (1-\beta)m_d$, as long as

$$b_d \geq C_6 \log(nb_d/\varepsilon)/(\delta^{C_6} n) + C_6 b_{d-1} \tag{10}$$

for some constant $C_6 \geq 1$.[23] Thus, we now wish to set block parameters so they satisfy Equations (8) to (10).

Towards this end, we let $A \in \mathbb{N}$ be a sufficiently large constant, and set block parameters as follows:

$$\begin{aligned}
b_0 &:= \left\lceil \frac{\log(1/\varepsilon)}{\delta^A n} \right\rceil,\\
b_i &:= A \cdot b_{i-1}, \text{ for all } i \in [d-1],\\
b_d &:= \left\lceil A^{C_3 \log(1/\delta)+1} \right\rceil \cdot b_0
\end{aligned}$$

[23] Recall that we actually originally requested that $m_d \geq k\tau/4$, instead of $m \geq 0.4kb_d$. However, we will soon show that this follows from our setting of τ.

It is straightforward to verify that for all sufficiently large A (as a function of the constants C_5, C_6), all of Equations (8) to (10) hold. Moreover, we make sure to pick $A \geq C_0$.

All that remains is to check the total number of blocks used, and to ensure that the overall output length m_d is sufficiently large. Towards this end, the total number of blocks used is

$$\begin{aligned}\tau &:= b_0 + 2\sum_{i=1}^{d} b_i \\ &\leq 8A \cdot \Big((1/\delta)^{C_3 \log A} + (1/\delta)^{C_3 \log A + A} \log(1/\varepsilon)/n\Big) \\ &=: \tau^\star,\end{aligned}$$

while the overall output length is in the range $m_d \in [0.4kb_d, 0.49kb_d]$, which means

$$\begin{aligned}m_d &\geq 0.4kb_d \\ &\geq 0.2Ak \cdot \Big((1/\delta)^{C_3 \log A} + (1/\delta)^{C_3 \log A + A} \log(1/\varepsilon)/n\Big) \\ &\geq 0.025k\tau^\star.\end{aligned}$$

Of course, the fact that $m_d \leq 0.49kb_d$ also implies that $m_d \leq k\tau^\star$ (since we set $\tau^\star \gg b_d$ above). Thus, to conclude, as long as our CG-source originally started off with

$$t \geq \tau^\star := 8A \cdot \Big((1/\delta)^{C_3 \log A} + (1/\delta)^{C_3 \log A + A} \log(1/\varepsilon)/n\Big)$$

blocks, we can obtain $m_d \in [0.01k\tau^\star, k\tau^\star]$ output bits that are ε-close to min-entropy $k'_d \geq (1-\beta)m_d$. □

5.3.3 Condensing the rest of the entropy out

In this final step, we show how to get the rest of the entropy out of the CG source, while maintaining the gap, via *iterative condensing*. We prove the following, which will later be combined with our core lemma (Lemma 12) in order to yield our main theorem (Theorem 5).

Lemma 13 (Condensing the rest of the entropy out)**.** *For every constant $\alpha > 0$, there is a constant $C > 0$ such that there exists an explicit condenser* $\mathsf{Cond} : \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_t} \to \{0,1\}^m$ *for $((n_1, k_1), \ldots, (n_t, k_t))$-block sources with output length $m \geq (1-\alpha)k_1$, output gap $g' = g := n_t - k_t$, and error ε, provided*

$$k_{i+1} \geq C(\log(n_i/\varepsilon) + (t-i) + g)$$

for all $i \in [t-1]$.

Proof. We simply plug the GUV extractor (Theorem 7) into our iterative condensing framework (Lemma 9), recalling that an extractor is simply a condenser with output gap 0.

In more detail, if we define $m_t := n_t$, then by Theorem 7, the following holds. There exists a sequence of explicit functions $\mathsf{sCond}_1, \mathsf{sCond}_2, \ldots, \mathsf{sCond}_{t-1}$, where each $\mathsf{sCond}_i : \{0,1\}^{n_i} \times \{0,1\}^{m_{i+1}} \to \{0,1\}^{m_i}$ is a seeded $(n_i, k_i) \to_{\varepsilon_i} (m_i, m_i)$ condenser with output length $m_i \geq (1-\alpha)k_i$, as long as

$$(1-\alpha)k_{i+1} \geq C_{\mathsf{GUV}} \log(n_i/\varepsilon_i)$$

for every $i \in [t-1]$ (where C_{GUV} is a constant depending only on α).[24] Since we may assume that $\alpha < 1/2$,[25] this requirement is satisfied when $k_{i+1} \geq C \log(n_i/\varepsilon_i)$, where we have used $C := 2C_{\mathsf{GUV}}$. And if we set $\varepsilon_i := \varepsilon \cdot 2^{-g-(t-i)}$ for all $i \in [t-1]$, then the requirement is satisfied when

$$k_{i+1} \geq C(\log(n_i/\varepsilon) + (t-i) + g)$$

for all $i \in [t-1]$. Now, by our iterative condensing framework (Lemma 9), this sequence of explicit functions $\mathsf{sCond}_1, \mathsf{sCond}_2, \dots, \mathsf{sCond}_{t-1}$ can be composed to create an explicit condenser for $((n_1,k_1),\dots,(n_t,k_t))$-block sources with output length $m_1 \geq (1-\alpha)k_1$, output gap $g' := \sum_{i\in[t-1]}(m_i - m_i) + g = g$, and error

$$\sum_{i\in[t-1]} \varepsilon_i \cdot 2^g = \varepsilon \sum_{i\in[t-1]} 2^{-(t-i)} \leq \varepsilon,$$

as desired. □

While the above lemma is quite general, the following corollary will be more useful for our purposes.

Corollary 5 (Condensing a geometric block source with a high-rate final block)**.** *For any constants $\alpha_0 > 0$ and $C_0 \geq 1$, there exist constants $\beta > 0$ and $C \geq 1$ such that the following holds. There exists an explicit condenser for $((n_1,k_1),\dots,(n_t,k_t))$-block sources with output length $m \geq (1-\alpha_0)k_1$, output gap $g' = g := n_t - k_t$, and error ε, provided that all of the following hold:*

- $k_1 \geq 4k_2 \geq 4^2 k_3 \geq \cdots \geq 4^{t-1}k_t$.
- $n_1 \leq (C_0 n_2)^2 \leq (C_0 n_3)^{2^2} \leq \cdots \leq (C_0 n_t)^{2^{t-1}}$.
- $k_t \geq (1-\beta)n_t$.
- $n_t \geq C\log(1/\varepsilon) + C$.

Proof. Let $C^\star$ be the second constant from Lemma 13, when the first constant is set to α. It suffices to show

$$k_{i+1} \geq C^\star(\log(n_i/\varepsilon) + (t-i) + g)$$

for all $i \in [t-1]$. This is straightforward via a backward induction on i (using the bullet points). □

Putting everything together

At last, with all of our ingredients in place, we are ready to prove our main theorem.

Proof of Theorem 5. Recall that we wish to construct an explicit condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^{k'+g'}$ for $(t,n,k = \delta n = n-g)$-CG sources, which has output entropy $k' \geq (1-\alpha)kt$, output gap $g' \leq (1/\delta)^C \cdot (g + \log(1/\varepsilon))$, and error ε. Towards this end, let $\mathbf{X} = (\mathbf{X}_1,\dots,\mathbf{X}_t)$ be a (t,n,k)-CG source. The main idea is to use Lemma 12 to condense the last few blocks in $\mathbf{X}$ to a block with rate 0.99, and then to use this high-rate block to get the rest of the entropy out of the source, using Corollary 5.

More formally, set the constants α_0, C_0 in Corollary 5 to $\alpha/2$ and $100/\alpha$ (respectively), and let $\beta^\star, C^\star$ denote the constants β, C (in that theorem) that come out. Then, set the constants α, C_0 in Lemma 12 to $\beta^\star$

[24] Note that when $i = t-1$, the requirement is actually $m_{i+1} = n_{i+1} \geq C_{\mathsf{GUV}} \log(n_i/\varepsilon_i)$, which is weaker than what is written.

[25] This is because the lemma statement only claims a lower bound on the output length m.

and $2C^\star$ (respectively), and let C_1, C_2, C_3 be the constants that come out (corresponding to the same-named constants in that lemma statement). Finally, define a size parameter s as

$$s := \left\lceil C_1 \cdot \left((1/\delta)^{C_2} + (1/\delta)^{C_3} \log(2/\varepsilon)/n\right)\right\rceil ,$$

and let w be the largest integer such that $s \cdot \sum_{i=1}^{w} \lceil 4/\alpha \rceil^{w-i} \leq t$. Note that if $w < 2$, then the claimed gap in the theorem statement is trivial, in that it can be achieved simply by applying the identity function.

Now, henceforth assuming that $w \geq 2$, define (for every $i \in [w]$)

$$s_i := \begin{cases} s \cdot \lceil 4/\alpha \rceil^{w-i} & \text{if } i > 1, \\ t - s \cdot \sum_{i=2}^{w} \lceil 4/\alpha \rceil^{w-i} & \text{if } i = 1. \end{cases}$$

Note that $s_1 \in [s \cdot \lceil 4/\alpha \rceil^{w-1}, s \cdot \lceil 4/\alpha \rceil^{w+1}]$, and define a new source $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_w)$, where $\mathbf{Y}_1$ consists of the first s_1 blocks of $\mathbf{X}$, $\mathbf{Y}_2$ consists of the next s_2 blocks of $\mathbf{X}$, and so on. Note that $\mathbf{Y}$ is an $((ns_1, ks_1), \ldots, (ns_w, ks_w))$-block source.

Now, by Lemma 12, there exists an explicit function Cond_1 such that $\mathbf{Z} := \mathsf{Cond}_1(\mathbf{Y}_w)$ is $(\varepsilon/2)$-close to a source with length $m_w \in [0.025\delta ns, \delta ns]$ and min-entropy $k'_w \geq (1-\beta^\star)m_w$, and moreover, this is true for every fixing of the random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_{w-1}$. Thus, $\mathbf{Y}^\star := (\mathbf{Y}_1, \ldots, \mathbf{Y}_{w-1}, \mathbf{Z})$ is a $((0,0), \ldots, (0,0), (0, \varepsilon/2))$-almost $((ns_1, ks_1), \ldots, (ns_{w-1}, ks_{w-1}), (m_w, (1-\beta^\star)m_w))$-block source. Thus, by Lemma 4, $\mathbf{Y}^\star$ is $(\varepsilon/2)$-close to an $((ns_1, ks_1), \ldots, (ns_{w-1}, ks_{w-1}), (m_w, (1-\beta^\star)m_w))$-block source, $\mathbf{Y}^{\star\star}$. Now, it is straightforward to verify (given our setting of parameters) that $\mathbf{Y}^{\star\star}$ satisfies the requirements of Corollary 5, and thus there is an explicit function Cond_2 such that $\mathsf{Cond}_2(\mathbf{Y}^{\star\star})$ is $\varepsilon/2$-close to a source of length $m \geq (1-\alpha/2)ks_1$ and gap $g' \leq \beta^\star m_w$, and thus the data-processing inequality tells us that $\mathsf{Cond}_2(\mathbf{Y}^\star)$ is ε-close to a source of length $m \geq (1-\alpha/2)ks_1$ and gap $g' \leq \beta^\star m_w$. Furthermore, note that by our setting of s_i, we have $m \geq (1-\alpha)kt$, and gap

$$g' \leq \beta^\star m_w \leq m_w \leq \delta ns \leq \delta n \cdot 2C_1 \left((1/\delta)^{C_2} + (1/\delta)^{C_3} \log(2/\varepsilon)/n\right),$$

which is at most

$$C \cdot (1/\delta)^C \cdot (n + \log(1/\varepsilon)) \tag{11}$$

for some constant $C \geq 1$. Finally, we may assume that the original gap was $g > \beta^\star n$, since otherwise we could easily obtain an output gap of the form $g' \leq C \cdot (1/\delta)^C \cdot (g + \log(1/\varepsilon))$, simply by replacing Cond_1 with the identity function. Thus, we can upper bound Equation (11) by

$$\frac{C}{\beta^\star} \cdot (1/\delta)^C \cdot (g + \log(1/\varepsilon)),$$

which is again at most $C' \cdot (1/\delta)^{C'} \cdot (g + \log(1/\varepsilon))$ for a slightly larger constant C', as desired. □

6 Existential results

In this section, we present and prove all our existential results. We start by showing that a random function is a good seedless condenser for any small family (Theorem 2). Then, we instantiate this result to get improved parameters for non-explicit seeded condensers (Corollary 3). Finally, we plug the latter existential result into the iterative condensing framework in to get our existential results for CG and block sources (Corollary 4).

6.1 A random function is a seedless condenser (for any small family)

In order to show that a random function is a good seedless condenser for any small family, we show that a random function is (with high probability) a good condenser for a single source. We prove the following, which can be viewed as the condenser version of the classic observation that a random function is a good extractor [Vad12, Proposition 6.12]. (In fact, we will see that it generalizes it.)

Theorem 9 (A random function is a condenser for a single source)**.** *There exist universal constants $C, c > 0$ such that the following holds. Let $\mathbf{X}$ be an arbitrary (n, k)-source. For any $\ell \in [0, k]$ and $g > 0$ such that $m := k - \ell + g$ is an integer, and any $\varepsilon \in (0, 1]$, the following holds. If $f : \{0,1\}^n \to \{0,1\}^m$ is a uniformly random function, then*

$$\Pr_f \left[H_\infty^\varepsilon(f(\mathbf{X})) < k - \ell\right] < C \cdot 2^{-c\varepsilon K\psi},$$

where

$$\psi := \max\left\{g - \frac{1}{\lfloor L \rfloor}\log(1/\varepsilon) - C, \quad g - \frac{1}{\lfloor L \rfloor}\log(C2^g g/\varepsilon) \cdot \frac{C2^g}{g}\right\}.$$

Note that ψ evaluates to the first argument when the gap exceeds a sufficiently large constant, and the second argument for all other $g > 0$ (where 2^g becomes a constant). In all applications, one should set the gap g so that $\psi = \Omega(g)$ or $\psi = 1$.

Before we continue, we take a moment to make some remarks about the above theorem. First, we emphasize that it works for *any* (n, k)-source, not just flat ones. This is crucial to showing the existence of good seedless condensers for small families, since (unlike in the seeded setting) you cannot assume such families only contain flat sources.[26] We also note that the above strictly generalizes the classic result that a random function is a good extractor (i.e., condenser with $g = 0$) with probability $1 - 2^{-\Omega(\varepsilon^2 K)}$. This is because we can instantiate our theorem with gap $g = \varepsilon/2$ and error $\varepsilon/2$, since a source with gap g is g-close to a source with gap 0. Moreover, our generalization reveals that the well-known required loss of $\ell = 2\log(1/\varepsilon)$ for extractors generalizes to roughly $\ell = 2\log(1/g)$, meaning that the loss is primarily due to the gap, not the error. Furthermore, the success probability generalizes to $1 - 2^{-\Omega(g\varepsilon K)}$. Overall, this means that even if you are in the regime $g < 1$ (which is close to the extractor regime of $g = \varepsilon/2$), you can benefit by applying the condenser result instead of the extractor result.

Next, we record the following corollary, which is immediate via the probabilistic method.[27]

Corollary 6 (A random function is a condenser for any small family)**.** *There exist universal constants $C, c > 0$ such that the following holds. Let $\mathcal{X}$ be a family of (n, k)-sources. For any $\ell \in [0, k]$ and $g > 0$ such that $m := k - \ell + g$ is an integer, and any $\varepsilon \in (0, 1]$, the following holds. If*

$$|\mathcal{X}| \leq c \cdot 2^{c\varepsilon K\psi},$$

where ψ is as defined in Theorem 9, then there exists a condenser $\mathsf{Cond} : \{0,1\}^n \to \{0,1\}^m$ *for $\mathcal{X}$ with loss ℓ, gap g, and error ε.*

[26] This is because such existential results proceed by counting the number of sources in the family $\mathcal{X}$, and arguing that there are not too many. And while it is true that every (n, k)-source is a convex combination of flat sources, it is not true that it is a convex combination of flat sources *in that family*, which is the collection whose size was actually estimated. The family $\mathcal{X}'$ of flat sources that arises by decomposing each $\mathbf{X} \in \mathcal{X}$ into a convex combination of flat sources may have size much larger $|\mathcal{X}|$.

[27] In particular, apply Theorem 9 to each $\mathbf{X} \in \mathcal{X}$ and use a union bound.

We now proceed to prove Theorem 9. First, we prove it in the extractor (small gap) regime, via Theorem 10. Then, we prove it in the much more challenging condenser (large gap) regime, via Theorem 11. Combined, these two theorems immediately yield Theorem 9. We briefly note that from here onwards, we often simplify notation and use $[N]$ to represent $\{0,1\}^n$, and $[M]$ to represent $\{0,1\}^m$. Furthermore, we always use μ to represent the density of a set S, *not* the mean of a random variable (though they will often coincide). The set to which μ corresponds will always be clear from context.

6.1.1 The extractor regime: small gap, large loss

We start by proving our existential result for the extractor (small gap) regime.

Theorem 10 (Theorem 9, Part I)**.** *Let* $\mathbf{X}$ *be an arbitrary* (n,k)*-source. For any* $\ell \in [0,k]$ *and* $g > 0$ *such that* $m := k - \ell + g$ *is an integer, and any* $\varepsilon \in (0,1]$*, the following holds. If* $f : \{0,1\}^n \to \{0,1\}^m$ *is a uniformly random function, then*

$$\Pr_f\left[H_\infty^\varepsilon(f(\mathbf{X})) < k - \ell\right] < 2^{-\frac{\varepsilon K}{2}(g - \frac{1}{L}\frac{3G}{g}\log(\frac{2Gg}{\varepsilon}))}.$$

Note that this result is most useful in the extractor regime, i.e., when the gap is a constant or even in the range $g \in (0,1]$. (Recall that the exact extractor regime is when $g = \varepsilon/2$.) In this regime, the above bound is of the form $2^{-\frac{\varepsilon K}{2}(g - O(\frac{1}{L}\frac{1}{g}\log(g/\varepsilon)))}$. Now, in order to prove Theorem 10, we use the following proposition, which is just a restatement of (one direction of) Lemma 2.

Proposition 1 (Necessary condition for condensing failure)**.** *For any fixed function* $f : \{0,1\}^n \to \{0,1\}^m$*, any* (n,k)*-source* $\mathbf{X}$*, and any* $k' \in [0,m], \varepsilon > 0$*, and* $g := m - k'$*,*

$$H_\infty^\varepsilon(f(\mathbf{X})) < k' \implies \exists S \subseteq [M] \text{ of density } \mu := |S|/M \text{ such that } \Pr[f(\mathbf{X}) \in S] > \mu G + \varepsilon.$$

In particular, we'll use the following corollary.

Corollary 7. *If* $H_\infty^\varepsilon(f(\mathbf{X})) < k'$*, then for any threshold value* $\tau \in [M]$*, one of the following must hold:*

- $\exists S \subseteq [M]$ *of size* $|S| < \tau$ *and density* $\mu := |S|/M$ *such that* $\Pr[f(\mathbf{X}) \in S] > \mu G + \varepsilon$*.*
- $\exists S \subseteq [M]$ *of size* $|S| = \tau$ *and density* $\mu := |S|/M$ *such that* $\Pr[f(\mathbf{X}) \in S] > \mu G$*.*

Proof. By Proposition 1, there exists some set $S \subseteq [M]$ of density μ such that $\Pr[f(\mathbf{X}) \in S] > \mu G + \varepsilon$. If $|S| < \tau$, then the first bullet holds. If $|S| \geq \tau$, then let $S^\star$ denote the τ elements in S hit by $f(\mathbf{X})$ with the highest probability (breaking ties arbitrarily), and let $\mu^\star$ denote the density of $S^\star$. Then

$$\Pr[f(\mathbf{X}) \in S^\star] \geq \frac{|S^\star|}{|S|} \cdot \Pr[f(\mathbf{X}) \in S] = \frac{\mu^\star}{\mu}\Pr[f(\mathbf{X}) \in S] > \mu^\star(G + \varepsilon/\mu) \geq \mu^\star G,$$

and the second bullet holds, as desired. □

Now, the idea is to eventually pick some threshold τ so that for a random function, both bullets happen with low (and close to the same) probability. We start with the first bullet.

Claim 3. *Let* $f : \{0,1\}^n \to \{0,1\}^m$ *be a uniformly random function, let* $\mathbf{X}$ *be an* (n,k)*-source, and let* $S \subseteq [M]$ *be a set of density* $\mu := |S|/M$*. Then for any* $\varepsilon > 0$ *and* $G \geq 0$*,*

$$\Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S] \geq \mu G + \varepsilon\right] \leq G^{-\varepsilon K}.$$

Proof. We may assume that $\mu > 0$, since the claim trivially holds if $\mu = 0$ (as this implies S is empty).

Now, for each $x \in \{0,1\}^n$, define the random variable

$$\mathbf{Z}_x := 1[f(x) \in S] \cdot \Pr[\mathbf{X} = x] \cdot K.$$

Note that its randomness comes from f, and it is supported on the interval $[0,1]$, since $H_\infty(\mathbf{X}) \geq k$. Furthermore, if we define $\mathbf{Z} := \sum_x \mathbf{Z}_x$, it is easy to verify that $\mathbf{Z} = \Pr_{\mathbf{X}}[f(\mathbf{X}) \in S] \cdot K$, and we also have $\mathbb{E}[\mathbf{Z}] = K|S|/M = \mu K$. Combining these observations with the Chernoff bound (Theorem 4), we have

$$\begin{aligned}
\Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S] \geq \mu G + \varepsilon\right] &= \Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S] \cdot K \geq \mu G K + \varepsilon K\right] \\
&= \Pr\left[\mathbf{Z} \geq \mathbb{E}[\mathbf{Z}] \cdot G + \varepsilon K\right] \\
&= \Pr\left[\mathbf{Z} \geq (G + \varepsilon/\mu)\mathbb{E}[\mathbf{Z}]\right] \\
&\leq \left(\frac{e^{G-1+\varepsilon/\mu}}{(G+\varepsilon/\mu)^{G+\varepsilon/\mu}}\right)^{\mu K} \qquad (12) \\
&= \exp\left(-\varepsilon K\left((1+\alpha)(\ln G + \ln(1+1/\alpha) - 1) + \alpha/G\right)\right)
\end{aligned}$$

for $\alpha := \mu G/\varepsilon$.[28] Finally, using routine calculus, it is straightforward to verify that the function

$$\phi(\alpha, g) := (1+\alpha)(\ln G + \ln(1+1/\alpha) - 1) + \alpha/G$$

is $\geq g \ln 2$ for all $g \geq 0, \alpha > 0$. The result follows. □

Next, we bound the probability that bullet two in Corollary 7 occurs for a uniformly random function. Using the same parameters and objects as defined in Claim 3, we have the following.

Claim 4.

$$\Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S] > \mu G\right] \leq \exp\big(-\mu G K(\ln G - 1 + 1/G)\big).$$

Proof. The claim is immediate via the proof of Claim 3 up to Equation (12), setting $\varepsilon = 0$. □

Using the above claims, we can show that the necessary conditions for condensing failure (Corollary 7) happen with low probability, allowing us to prove that a random function is a good condenser (Theorem 10).

Proof of Theorem 10. Let $k' := k - \ell$, $g := m - k'$, and suppose that $H_\infty^\varepsilon(f(\mathbf{X})) < k'$. By Corollary 7, we know that for any threshold value $\tau \in [M]$ (to be set momentarily), one of the following must hold:

- $\exists S \subseteq [M]$ of size $|S| < \tau$ and density $\mu := |S|/M$ such that $\Pr[f(\mathbf{X}) \in S] > \mu G + \varepsilon$.
- $\exists S \subseteq [M]$ of size $|S| = \tau$ and density $\mu := |S|/M$ such that $\Pr[f(\mathbf{X}) \in S] > \mu G$.

By combining this with Claim 3 and Claim 4, we get the following.

$$\begin{aligned}
\Pr_f\left[H_\infty^\varepsilon(f(\mathbf{X})) < k'\right] &\leq \Pr_f\left[\exists S \subseteq [M], |S| < \tau : \Pr[f(\mathbf{X}) \in S] > \mu G + \varepsilon\right] \\
&\quad + \Pr_f\left[\exists S \subseteq [M], |S| = \tau : \Pr[f(\mathbf{X}) \in S] > \mu G\right] \\
&\leq \binom{M}{<\tau} 2^{-g\varepsilon K} + \binom{M}{\tau} \exp\left(-\tau L(\ln G - 1 + 1/G)\right).
\end{aligned}$$

[28] Here and henceforth, we may assume that $\varepsilon > 0$, since the claim trivially holds if $\varepsilon = 0$.

Now, consider the quantity $\phi := (\ln G - 1 + 1/G)\log e$. If $g\varepsilon K'/\phi < 1$, then we set $\tau := \lceil g\varepsilon K'/\phi \rceil$. Otherwise, we set $\tau := \lfloor g\varepsilon K'/\phi \rfloor$. Notice that in the first case, the probability that produced the first term in the above sum would have actually realized to 0. And in the second case, observe that $g\varepsilon K \geq \tau L\phi$. Thus the above expression can be bounded by

$$\leq \binom{M}{\leq \tau} 2^{-\tau L\phi} \leq 2^{-\tau L(\phi - \frac{1}{L}\log(eM/\tau))} \leq 2^{-\frac{g\varepsilon K}{2\phi}(\phi - \frac{1}{L}\log(\frac{2eG\phi}{g\varepsilon}))}. \tag{13}$$

Finally, it is straightforward to verify that

$$\frac{2}{g\ln 2} \leq \frac{g}{\phi} \leq \frac{2G}{g\ln 2},$$

and using this observation, we can upper bound Equation (13) by

$$2^{-\frac{\varepsilon K}{2}(g - \frac{1}{L}\frac{1}{g}\frac{2G}{\ln 2})\log(\frac{g\cdot eG\ln 2}{\varepsilon})}$$

as desired. □

6.1.2 The condenser regime: large gap, small loss

Next, we turn to prove our existential result for the much more challenging condenser regime.

Theorem 11 (Theorem 9, Part II)**.** *Let $\mathbf{X}$ be an arbitrary (n,k)-source. For any $\ell \in [0,k]$ and $g > 0$ such that $m := k - \ell + g$ is an integer, and any $\varepsilon \in (0,1]$, the following holds. If $f : \{0,1\}^n \to \{0,1\}^m$ is a uniformly random function, then*

$$\Pr_f\left[H_\infty^\varepsilon(f(\mathbf{X})) < k - \ell\right] \leq 4 \cdot 2^{-\frac{\varepsilon K}{6}(g - \frac{1}{\lfloor L \rfloor}\log(1/\varepsilon) - 16)}.$$

As in the previous section, we consider several conditions which are necessary for condenser failure, and show that each happens with small probability. Similar to before, these conditions have to do with whether certain sets $S \subseteq [M]$ are assigned too much probability. This time, however, we're in a regime where we want to be able to handle very small loss, by paying some price in the gap, and thus the output length. As a result, it will be too expensive to check tests $S \subseteq [M]$ by choosing them from the set $[M]$, which may now be very large. Instead, we'll have to implicitly specify them using their preimages.

The above plan would work well if $\mathbf{X}$ had a small support size, which would be the case if $\mathbf{X}$ were flat. However, we don't want to make any such assumption, and therefore need a new idea. Our idea is to split the support of $\mathbf{X}$ into two sets: one which is small (and therefore easy to choose sets from), and one which is big (but is guaranteed to have better "local" entropy). Then, we ultimately check whether $f(\mathbf{X})$ fails the appropriate tests $S \subseteq [M]$ by specifying them through their preimages in these sets.

We now proceed to present the formal conditions we're looking for that indicate condenser failure.

Proposition 2 (Necessary conditions for condensing failure)**.** *Let $f : \{0,1\}^n \to \{0,1\}^m$ be a fixed function, and let $\mathbf{X}$ be an (n,k)-source whose support is partitioned into sets X_1, X_2. Fix any $\ell \in [0,k]$ and $\varepsilon > 0$, and define $k' := k - \ell$ and $g := m - k'$. If $H_\infty^\varepsilon(f(\mathbf{X})) < k'$, there must exist some set $S \subseteq [M]$ of density μ such that at least one of the following holds:*

1. *X_1* **has bad smooth min-entropy:** $\Pr[f(\mathbf{X}) \in S \wedge \mathbf{X} \in X_1] > \mu G + \varepsilon/3$.

2. X_2 **has bad smooth min-entropy:** $\Pr[f(\mathbf{X}) \in S \wedge \mathbf{X} \in X_2] > \mu G/L + \varepsilon/3$.

3. X_1, X_2 **have bad "joint" smooth min-entropy:** *Both of the following hold:*
   - $\Pr[f(\mathbf{X}) = v \wedge \mathbf{X} \in X_1] > \frac{L-1}{L} \cdot \frac{1}{K'}$ *for all* $v \in S$.
   - $\Pr[f(\mathbf{X}) \in S \wedge \mathbf{X} \in X_2] > \varepsilon/3$.

Proof. By definition of smooth min-entropy, we know that if $H_\infty^\varepsilon(f(\mathbf{X})) < k'$, then there is some set $S \subseteq \{0,1\}^m$ of density μ such that $\Pr[f(\mathbf{X}) \in S] > \mu G + \varepsilon$, by Proposition 1. Partition S into sets S_1, S_2 such that S_1 contains all elements $v \in \{0,1\}^m$ satisfying

$$\Pr[f(\mathbf{X}) = v \wedge \mathbf{X} \in X_1] > \frac{L-1}{L} \cdot \frac{1}{K'}.$$

Suppose that neither the first nor third case in the proposition hold. Then

$$\begin{aligned}
\Pr[f(\mathbf{X}) \in S_1] &= \Pr[f(\mathbf{X}) \in S_1 \wedge \mathbf{X} \in X_1] + \Pr[f(\mathbf{X}) \in S_1 \wedge \mathbf{X} \in X_2] \\
&\leq \frac{|S_1|}{M} G + \varepsilon/3 + \varepsilon/3 \\
&= \frac{|S_1|}{M} G + 2\varepsilon/3.
\end{aligned}$$

Furthermore, if the second case also does not hold, then

$$\begin{aligned}
\Pr[f(\mathbf{X}) \in S_2] &= \Pr[f(\mathbf{X}) \in S_2 \wedge \mathbf{X} \in X_1] + \Pr[f(\mathbf{X}) \in S_2 \wedge \mathbf{X} \in X_2] \\
&\leq \frac{L-1}{L} \cdot \frac{|S_2|}{K'} + \frac{|S_2|}{M} G/L + \varepsilon/3 \\
&= \frac{|S_2|}{M} G + \varepsilon/3.
\end{aligned}$$

But this implies that $\Pr[f(\mathbf{X}) \in S] \leq \mu G + \varepsilon$, contradicting our original assumption. □

We now show that each of these three events happens with low probability, starting with the second one.

Case 2: The subdistribution on X_2 has bad smooth min-entropy.

We prove the following, which bounds the probability that the second bullet in Proposition 2 can occur.

Lemma 14 (A random function condenses the subdistribution on X_2)**.** *Let* $\mathbf{X} \sim \{0,1\}^n$ *be a source, and let* $X \subseteq$ *support*$(\mathbf{X})$ *be a set with* $\max_{x \in X} \Pr[\mathbf{X} = x] \leq 1/K$*. For any* $\ell \in [0,k]$ *and* $g \geq 0$ *such that* $m := k - \ell + g$ *is an integer, and any* $\varepsilon \in (0,1]$*, the following holds. If* $f : \{0,1\}^n \to \{0,1\}^m$ *is a uniformly random function, then*

$$\Pr_f\left[\exists S \subseteq [M] : \Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] > \mu G + \varepsilon\right] \leq 2^{-\frac{\varepsilon K}{2}(g - \frac{1}{L}\log(2eG/\varepsilon) - \log e)}.$$

Just as in the proof of Theorem 10, we will upper bound the above probability by splitting the event in two, as prescribed by Corollary 7. To help us with this, we need subdistribution versions of Claim 3 and Claim 4, which we prove next.

Claim 5 (Claim 3, subdistribution version)**.** *Let $f : \{0,1\}^n \to \{0,1\}^m$ be a uniformly random function, let $\mathbf{X} \sim \{0,1\}^n$ be a source, and let $X \subseteq$ support$(\mathbf{X})$ be a set with $\max_{x \in X} \Pr[\mathbf{X} = x] \leq 1/K$. Then for any set $S \subseteq [M]$ of density $\mu := |S|/M$, and any $\varepsilon > 0$ and $G \geq 0$,*

$$\Pr_f \left[\Pr_{\mathbf{X}} [f(\mathbf{X}) \in S \textit{ and } \mathbf{X} \in X] > \mu G + \varepsilon \right] \leq G^{-\varepsilon K}.$$

Proof. We may assume that $\mu > 0$, since the claim trivially holds if $\mu = 0$ (as this implies S is empty).

Now, for each $x \in X$, define the random variable

$$\mathbf{Z}_x := 1[f(x) \in S] \cdot \Pr[\mathbf{X} = x] \cdot K.$$

Note that its randomness comes from f, and it is supported on the interval $[0,1]$, since $\max_{x \in X} \leq 1/K$. Furthermore, if we define $\mathbf{Z} := \sum_{x \in X} \mathbf{Z}_x$, it is easy to verify that $\mathbf{Z} = \Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] \cdot K$ and $\mathbb{E}[\mathbf{Z}] = (K|S|/M) \Pr[\mathbf{X} \in X] = \mu K \Pr[\mathbf{X} \in X] \leq \mu K$. Using these observations, we have

$$\begin{aligned}
\Pr_f \left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] \geq \mu G + \varepsilon \right] &= \Pr_f \left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] \cdot K \geq \mu G K + \varepsilon K \right] \\
&= \Pr_f[\mathbf{Z} \geq (G + \varepsilon/\mu)\mu K] \\
&\leq \left(\frac{e^{G - 1 + \varepsilon/\mu}}{(G + \varepsilon/\mu)^{G + \varepsilon/\mu}} \right)^{\mu K}, \qquad (14)
\end{aligned}$$

where the last inequality follows from the fact that the Chernoff bound (Theorem 4) can be used with just an upper bound μK on the expectation $\mathbb{E}[\mathbf{Z}]$. The remainder of the proof is now identical to the proof of Claim 3, following Equation (12). □

Next, using the same parameters and objects as described in the claim above, we prove the following.

Claim 6 (Claim 4, subdistribution version)**.**

$$\Pr_f \left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \textit{ and } \mathbf{X} \in X] > \mu G \right] \leq \exp\left(-\mu G K (\ln G - 1 + 1/G) \right).$$

Proof. The claim is immediate via the proof of Claim 5 up to Equation (14), setting $\varepsilon = 0$. □

With these claims in hand, it is now easy to prove Lemma 14.

Proof of Lemma 14. Just as in the proof to Theorem 10 (substituting in Claim 5 for Claim 3 and Claim 6 for Claim 4), we have

$$\Pr_f \left[\exists S \subseteq [M] : \Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] > \mu G + \varepsilon \right] \leq 2^{-\frac{g \varepsilon K}{2\phi}(\phi - \frac{1}{L} \log(\frac{2eG\phi}{g\varepsilon}))}, \qquad (15)$$

where $\phi := (\ln G - 1 + 1/G) \log e$. It is straightforward to verify that for all $g \geq 0$, we have

$$g - \log e \leq \phi \leq g.$$

Using this observation, we can upper bound Equation (15) by

$$2^{-\frac{\varepsilon K}{2}(g - \frac{1}{L} \log(\frac{2eG}{\varepsilon}) - \log e)}$$

as desired. □

Case 1: The subdistribution on X_1 has bad smooth min-entropy.

Next, we upper bound the probability that the first bullet in Proposition 2 can occur.

Lemma 15 (A random function condenses the subdistribution on X_1)**.** *Let $\mathbf{X}$ be an (n,k)-source, and let $X \subseteq$ support$(\mathbf{X})$ be an arbitrary set. For any $\ell \in [0,k]$ and $g \geq 0$ such that $m := k - \ell + g$ is an integer, and any $\varepsilon \in (0,1]$, the following holds. If $f : \{0,1\}^n \to \{0,1\}^m$ is a uniformly random function, then*

$$\Pr_f\left[\exists S \subseteq [M] : \Pr[f(\mathbf{X}) \in S \textit{ and } \mathbf{X} \in X] > \mu G + \varepsilon\right] \leq 2^{-\frac{\varepsilon K}{2}(g - \frac{1}{L}\log(\frac{|X|}{\varepsilon K}) - 5.886)}.$$

As before, we will upper bound this event by splitting it in two. This time, however, we will *not* ultimately specify the sets S by picking them from $[M]$. Instead, we will specify them implicitly, via their preimages. To do this, it will be useful to define a notion of "superlevel sets." Given an (n,k)-source $\mathbf{X}$ and element $v \in \{0,1\}^n$, we let SL_v denote its superlevel set, defined as follows:

$$\mathsf{SL}_v := \{x \in \{0,1\}^n : \Pr[\mathbf{X} = x] \geq \Pr[\mathbf{X} = v]\}.$$

Given this definition, we are ready to prove the preimage versions of the key claims we have been using.

Claim 7 (Claim 3, preimage version)**.** *Let $\mathbf{X}$ be an (n,k)-source. For any $\ell \in [0,k]$ and $g \geq 0$ such that $m := k - \ell + g$ is an integer, any $\varepsilon \in (0,1]$, and any $S \subseteq \{0,1\}^n$ with $\mu := |S|/M$, the following holds. If $f : \{0,1\}^n \to \{0,1\}^m$ is a uniformly random function, then*

$$\Pr_f\left[|f(S)| = |S| \textit{ and } \Pr_{\mathbf{X}}[\exists v \in S : f(\mathbf{X}) = f(v) \textit{ and } \mathbf{X} \in \mathsf{SL}_v] > \mu G + \varepsilon\right] \leq \left(\frac{e\mu}{\varepsilon}\right)^{\varepsilon K}$$

Proof. Let $f : \{0,1\}^n \to \{0,1\}^m$ be a uniformly random function. For a fixed function $h : S \to \{0,1\}^m$, let $f_h : \{0,1\}^n \to \{0,1\}^m$ be a function such that $f_h(x) = h(x)$ for all $x \in S$, and $f_h(x)$ is an independent, uniformly random value from $\{0,1\}^m$ for all other x. By the law of total probability, there exists a worst-case fixing $h^\star$ that is injective on S such that

$$\begin{aligned}&\Pr_f\left[|f(S)| = |S| \text{ and } \Pr_{\mathbf{X}}[\exists v \in S : f(\mathbf{X}) = f(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > \mu G + \varepsilon\right]\\ \leq &\Pr_{f_{h^\star}}\left[\Pr_{\mathbf{X}}[\exists v \in S : f_{h^\star}(\mathbf{X}) = f_{h^\star}(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > \mu G + \varepsilon\right].\end{aligned}$$

For ease of notation, we will henceforth use f' to denote $f_{h^\star}$.

Now, for all $v \in S$ and $x \in \mathsf{SL}_v \setminus S$, define the random variable

$$\mathbf{Z}_{x,v} := 1[f'(x) = f'(v)] \cdot \Pr[\mathbf{X} = x] \cdot K.$$

Then, for all $x \in (\cup_{v \in S}\mathsf{SL}_v) \setminus S$, define

$$\mathbf{Z}_x := \sum_{v \in S : x \in \mathsf{SL}_v} \mathbf{Z}_{x,v},$$

and finally let $\mathbf{Z} := \sum_{x \in (\cup_{v\in S}\mathsf{SL}_v)\setminus S} \mathbf{Z}_x$. Let us now make some observations about these random variables.

First, note that the randomness in these random variables comes exclusively from f', and each random variable $\mathbf{Z}_x$ is supported on $[0,1]$, since $H_\infty(\mathbf{X}) \geq k$ and since $1[f'(x) = f'(v)]$ can only equal 1 for at most one value $v \in S$ (since f' is injective on S). Furthermore, observe that

$$\mathbf{Z} = \sum_{v \in S, x \in \mathsf{SL}_v \setminus S} \mathbf{Z}_{x,v} = K \cdot \Pr_{\mathbf{X}}[\exists v \in S : f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v \setminus S].$$

Looking back at the probability we must analyze, it would be much more convenient if the expression above had the condition $\mathbf{X} \in \mathsf{SL}_v$ instead of $\mathbf{X} \in \mathsf{SL}_v \setminus S$. Luckily, it is easy to verify that

$$\begin{aligned}
&\Pr_{\mathbf{X}}[\exists v \in S : f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] \\
=&\Pr_{\mathbf{X}}[\exists v \in S : f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v \setminus S] + \Pr_{\mathbf{X}}[\mathbf{X} \in S] \\
=&\frac{1}{K} \cdot \mathbf{Z} + \Pr_{\mathbf{X}}[\mathbf{X} \in S] \\
\leq&\frac{1}{K}(\mathbf{Z} + |S|) \\
\leq&\frac{1}{K}(\mathbf{Z} + \mu G K),
\end{aligned}$$

where the penultimate inequality is because $\mathbf{X}$ has min-entropy at least k, and the final inequality is because $G \geq M/K$. Next, we can upper bound the expected value of $\mathbf{Z}$ as follows:

$$\mathbb{E}[\mathbf{Z}] = \sum_{v \in S, x \in \mathsf{SL}_v \setminus S} \mathbb{E}[\mathbf{Z}_{x,v}] = \frac{K}{M} \sum_{v \in S, x \in \mathsf{SL}_v \setminus S} \Pr[\mathbf{X} = x] = \frac{K}{M} \sum_{v \in S} \Pr[\mathbf{X} \in \mathsf{SL}_v \setminus S] \leq \frac{K}{M} \cdot |S| = \mu K.$$

Finally, since each $\mathbf{Z}_x$ is independent, and $\mathbf{Z} = \sum_x \mathbf{Z}_x$, we are ready to apply a Chernoff bound to upper bound our desired probability. In particular, we have

$$\begin{aligned}
&\Pr_{f'}\left[\Pr_{\mathbf{X}}[\exists v \in S : f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > \mu G + \varepsilon\right] \\
&= \Pr_{f'}\left[\Pr_{\mathbf{X}}[\exists v \in S : f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] \cdot K > \mu G K + \varepsilon K\right] \\
&\leq \Pr_{f'}\left[\mathbf{Z} + \mu G K > \mu G K + \varepsilon K\right] \\
&= \Pr_{f'}\left[\mathbf{Z} > (\varepsilon/\mu)\mu K\right].
\end{aligned}$$

Since we showed above that $\mu K \geq \mathbb{E}[\mathbf{Z}]$, the Chernoff bound (Theorem 4) tells us that the above is

$$\leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu K} \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu K}$$

for $\delta := \varepsilon/\mu - 1$.[29] Plugging this value of δ into the expression above yields

$$\leq (e\mu/\varepsilon)^{\varepsilon K},$$

as desired. □

[29] Note that we may assume $\delta > 0$, since otherwise $\mu/\varepsilon \geq 1$ and the bound in the claim trivially holds.

Next, using the same parameters as in the claim above (with $k' := k - \ell$), we prove the following.

Claim 8 (Claim 4, preimage version)**.**

$$\Pr_f\left[|f(S)| = |S|, \textit{ and for all } v \in S,\ \Pr_{\mathbf{X}}[f(\mathbf{X}) = f(v) \textit{ and } \mathbf{X} \in \mathsf{SL}_v] > 1/K'\right] \leq \left(\frac{4e}{G}\right)^{\mu GK}.$$

Proof. As in the proof of Claim 7, it suffices to show the claimed upper bound on the quantity

$$\Pr_{f'}\left[\forall v \in S : \Pr_{\mathbf{X}}[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > 1/K'\right], \tag{16}$$

where f' is some function that is injective (and fixed to constants) on S, and uniformly random on all other inputs. Now, let us once again proceed with defining random variables so that we can upper bound this quantity via a Chernoff bound. We must be a little more careful this time.

Towards this end, for all $v \in S$ and $x \in \mathsf{SL}_v \setminus S$, we once again want to define a random variable $\mathbf{Z}_{x,v}$. But this time, we base the definition on just how likely x is to be hit. In particular, let $X^\star$ denote the $2K$ most probable elements in support$(\mathbf{X})$, breaking ties arbitrarily. Then, define

$$\mathbf{Z}_{x,v} := \begin{cases} 1[f'(x) = f'(v)] & \text{if } x \in X^\star, \\ 1[f'(x) = f'(v)] \cdot \Pr[\mathbf{X} = x] \cdot 2K & \text{otherwise.} \end{cases}$$

Now, as before, for all $x \in (\cup_{v \in S} \mathsf{SL}_v) \setminus S$, define

$$\mathbf{Z}_x := \sum_{v \in S : x \in \mathsf{SL}_v} \mathbf{Z}_{x,v},$$

and let $\mathbf{Z} := \sum_{x \in (\cup_{v \in S} \mathsf{SL}_v) \setminus S} \mathbf{Z}_x$. Let us now make some observations about these random variables.

First, the randomness in these random variables comes exclusively from f'. Next, we claim that each random variable $\mathbf{Z}_x$ is supported on $[0, 1]$. To see why, observe that only one term $\mathbf{Z}_{x,v}$ in the sum that defines $\mathbf{Z}_x$ can be nonzero, since f' is injective on S. Then, note that such a nonzero term $\mathbf{Z}_{x,v}$ is always in the range $[0, 1]$: for the first definition of $\mathbf{Z}_{x,v}$, this is clear. For the second definition, simply note that all elements $x \in$ support$(\mathbf{X})$ that are not among the $2K$ most probable must be hit with probability $< 1/(2K)$, because otherwise the sum of $\Pr[\mathbf{X} = x]$ over all elements x will exceed 1 - a contradiction.

Next, observe that

$$\begin{aligned} \mathbb{E}[\mathbf{Z}] &= \sum_{v \in S, x \in (\mathsf{SL}_v \setminus S) \cap X^\star} \mathbb{E}[\mathbf{Z}_{x,v}] + \sum_{v \in S, x \in (\mathsf{SL}_v \setminus S) \setminus X^\star} \mathbb{E}[\mathbf{Z}_{x,v}] \\ &\leq |S||X^\star|/M + 2K|S|/M \\ &= 4\mu K. \end{aligned}$$

Finally, for our last step before applying the Chernoff bound, we must relate $\mathbf{Z}$ to the event in Equation (16). Towards this end, fix some $v \in S$ and suppose the following *inequality* holds (note that the *equality* always holds, by the injectivity of f' on S):

$$\Pr_{\mathbf{X}}[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] = \Pr_{\mathbf{X}}[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v \setminus S] + \Pr_{\mathbf{X}}[\mathbf{X} = v] > 1/K'.$$

Then, since $\mathbf{X}$ has min-entropy $\geq k$, and $1/K' = L/K$, f' must send at least $L+1$ elements from SL_v to $f'(v)$. We consider two cases. First, if at least L of these elements occur in $X^\star$, then there must be at least L elements that f' maps from $(\mathsf{SL}_v \setminus S) \cap X^\star$ to $f'(v)$. As such, we have

$$\sum_{x \in \mathsf{SL}_v \setminus S} \mathbf{Z}_{x,v} \geq \sum_{x \in (\mathsf{SL}_v \setminus S) \cap X^\star} \mathbf{Z}_{x,v} \geq L = K \cdot 1/K'.$$

On the other hand, if less than L of these elements occur in $X^\star$, then there must be at least 2 elements that f' maps from $(\mathsf{SL}_v \setminus S) \setminus X^\star$ to $f'(v)$. In this case, by definition of SL_v, we have that

$$\begin{aligned}\sum_{x \in \mathsf{SL}_v \setminus S} \mathbf{Z}_{x,v} &= \sum_{x \in (\mathsf{SL}_v \setminus S) \cap X^\star} \mathbf{Z}_{x,v} + \sum_{x \in (\mathsf{SL}_v \setminus S) \setminus X^\star} \mathbf{Z}_{x,v} \\ &\geq K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in (\mathsf{SL}_v \setminus S) \cap X^\star] \\ &+ 2K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in (\mathsf{SL}_v \setminus S) \setminus X^\star] \\ &= K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in (\mathsf{SL}_v \setminus \{v\}) \cap X^\star] \\ &+ 2K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in (\mathsf{SL}_v \setminus \{v\}) \setminus X^\star] && \text{(by injectivity of } f' \text{ on } S) \\ &\geq K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in (\mathsf{SL}_v \setminus \{v\}) \cap X^\star] \\ &+ K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in (\mathsf{SL}_v \setminus \{v\}) \setminus X^\star] \\ &+ K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} = v] \\ &= K \cdot \Pr[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] \\ &> K \cdot \frac{1}{K'}.\end{aligned}$$

Combining these two cases, we get that

$$\Pr_{\mathbf{X}}[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > 1/K' \implies \sum_{x \in \mathsf{SL}_v \setminus S} \mathbf{Z}_{x,v} \geq K \cdot 1/K',$$

and moreover,

$$\Pr_{\mathbf{X}}[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > 1/K' \textit{ for all } v \in S \implies \mathbf{Z} = \sum_{v \in S} \sum_{x \in \mathsf{SL}_v \setminus S} \mathbf{Z}_{x,v} \geq K|S|/K'.$$

With all of these observations in hand, we are finally ready to apply a Chernoff bound to upper bound our desired probability. Towards this end, we have

$$\begin{aligned}&\Pr_{f'}\left[\forall v \in S : \Pr_{\mathbf{X}}[f'(\mathbf{X}) = f'(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > 1/K'\right] \\ &\leq \Pr_{f'}\left[\mathbf{Z} \geq K|S|/K'\right] \\ &= \Pr_{f'}[\mathbf{Z} \geq (4\mu K)(G/4)].\end{aligned}$$

Since we showed that $4\mu K \geq \mathbb{E}[\mathbf{Z}]$, the Chernoff bound (Theorem 4) tells us that the above is

$$\leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{4\mu K} \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)4\mu K}$$

for $\delta := G/4 - 1$.[30] Plugging this value of δ into the expression above yields

$$(4e/G)^{\mu G K},$$

as desired. $\square$

Using these claims, it is easy to prove Lemma 15.

Proof of Lemma 15. Fix a function $f : \{0,1\}^n \to \{0,1\}^m$, and suppose there is a set $S \subseteq [M]$ with density $\mu := |S|/M$ such that $\Pr[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] > \mu G + \varepsilon$. We may assume each $v \in S$ has a preimage in X, since otherwise we could remove v from S, while keeping the probability guarantee. For the same reason, we may assume that $\Pr[f(\mathbf{X}) = v \text{ and } \mathbf{X} \in X] > 1/K'$ for each $v \in S$.

Now, let $\tau \in [M]$ be a threshold we will set later. Observe that one of the following must hold:

- $\exists S \subseteq [M]$ with size $< \tau$ and density $\mu := |S|/M$ such that $\Pr[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] > \mu G + \varepsilon$.
- $\exists S \subseteq [M]$ with size τ and density $\mu := \tau/M$ such that $\Pr[f(\mathbf{X}) = v \text{ and } \mathbf{X} \in X] > 1/K', \forall v \in S$.

Indeed, this follows immediately from the discussion above, since if the original set S had size $< \tau$, then the first bullet holds, and if it had size $\geq \tau$, then any subset of S of size τ satisfies the second bullet.

Next, regardless of which bullet holds, we let $S \subseteq [M]$ denote the set referred to in that bullet, and define a new set $S^\star \subseteq X$ as follows. First, for each $v \in S$, let $v^\star$ denote the element in $f^{-1}(v) \cap X$ that receives the least probability under $\mathbf{X}$. Then, define the set $S^\star := \{v^\star : v \in S\}$, and observe the following:

- If S originally referred to the first bullet above, then all of the following hold:
  - $S^\star \subseteq X$ and $|S^\star| < \tau$.
  - $|f(S^\star)| = |S^\star|$.
  - $\Pr_{\mathbf{X}}[\exists v^\star \in S^\star : f(\mathbf{X}) = f(v^\star) \text{ and } \mathbf{X} \in \mathsf{SL}_{v^\star}] = \Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] > \mu G + \varepsilon$.
- If S originally referred to the second bullet above, then all of the following must hold:
  - $S^\star \subseteq X$ and $|S^\star| = \tau$.
  - $|f(S^\star)| = |S^\star|$.
  - $\Pr_{\mathbf{X}}[f(\mathbf{X}) = f(v^\star) \text{ and } \mathbf{X} \in \mathsf{SL}_{v^\star}] = \Pr_{\mathbf{X}}[f(\mathbf{X}) = f(v^\star) \text{ and } \mathbf{X} \in X] > 1/K', \forall v^\star \in S^\star$.

By combining these observations with Claim 7 and Claim 8, we get the following.

$$\begin{aligned}
&\Pr_f\left[\exists S \subseteq [M] : \Pr[f(\mathbf{X}) \in S \text{ and } \mathbf{X} \in X] > \mu G + \varepsilon\right] \\
&\leq \Pr_f\left[\exists S \subseteq X, |S| < \tau : |f(S)| = |S| \text{ and } \Pr_{\mathbf{X}}[\exists v \in S : f(\mathbf{X}) = f(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > \mu G + \varepsilon\right] \\
&+ \Pr_f\left[\exists S \subseteq X, |S| = \tau : |f(S)| = |S| \text{ and } \Pr_{\mathbf{X}}[f(\mathbf{X}) = f(v) \text{ and } \mathbf{X} \in \mathsf{SL}_v] > 1/K', \forall v \in S\right] \\
&\leq \binom{|X|}{< \tau}\left(\frac{e\tau}{\varepsilon M}\right)^{\varepsilon K} + \binom{|X|}{\tau}\left(\frac{4e}{G}\right)^{\tau L}.
\end{aligned}$$

[30] Note that we may assume $\delta > 0$, since otherwise $4/G \geq 1$ and the bound in the claim trivially holds.

Finally, we check if $\varepsilon K/L < 1$. If this holds, we set $\tau = \lceil \varepsilon K/L \rceil \leq 2\varepsilon K/L$, and observe that the probability that produced the term $\binom{|X|}{<\tau}\left(\frac{e\tau}{\varepsilon M}\right)^{\varepsilon K}$ would have actually been 0. If $\varepsilon K/L \geq 1$, we set $\tau = \lfloor \varepsilon K/L \rfloor \geq (\varepsilon K/L)/2$, and observe that $\left(\frac{e\tau}{\varepsilon M}\right)^{\varepsilon K} \leq \left(\frac{e}{G}\right)^{\tau L}$. In either case, we can upper bound the above sum by

$$\leq \binom{|X|}{\leq \tau}\left(\frac{4e}{G}\right)^{\tau L} \leq 2^{-\tau L(g-\frac{1}{L}\log(\frac{2eL}{\varepsilon}\cdot\frac{|X|}{K})-\log(4e))} \leq 2^{-\frac{\varepsilon K}{2}(g-\frac{1}{L}\log(\frac{|X|}{\varepsilon K})-5.886)},$$

as desired. □

Case 3: The subdistributions on $\mathbf{X}_1, \mathbf{X}_2$ have bad joint smooth min-entropy.

Finally, we upper bound the probability that the third bullet in Proposition 2 can occur.

Lemma 16 (A random function jointly condenses the subdistributions on X_1, X_2)**.** *Let $\mathbf{X}$ be an (n, k)-source. For any $\ell \in [0, k]$ and $g \geq 0$ such that $m := k - \ell + g = k' + g$ is an integer, and any $\varepsilon \in (0, 1]$ the following holds. Suppose the support of $\mathbf{X}$ is partitioned into sets X_1, X_2, where X_1 contains the $\min\{\lceil 4KL \rceil, N\}$ highest probability elements, and X_2 the rest. If $f : \{0,1\}^n \to \{0,1\}^m$ is a uniformly random function, then*

$$\Pr_f\left[\exists S \subseteq [M] : \Pr_{\mathbf{X}}[f(\mathbf{X}) = v \wedge \mathbf{X} \in X_1] > \frac{L-1}{L}\cdot\frac{1}{K'}\forall v \in S, \textit{ and } \Pr_{\mathbf{X}}[f(\mathbf{X}) \in S \wedge \mathbf{X} \in X_2] > \varepsilon\right]$$
$$\leq 2 \cdot 2^{-\frac{\varepsilon K}{2}(g-\frac{1}{\lfloor L \rfloor}\log(1/\varepsilon)-11)}.$$

Proof. We start by claiming that we can assume $L \geq 2$, since otherwise the result is easy to prove. To see why, suppose that $L < 2$ (and thus $\lfloor L \rfloor = 1$). In order for the bad event (in the probability expression above) to hold, the random function f must map $> \varepsilon$ weight from X_2 into the set $f(X_1)$. But here, the size of X_1 is at most $\lceil 4KL \rceil < \lceil 8K \rceil$, and thus the size of $f(X_1)$ is also $< \lceil 8K \rceil$. Since f acts independently on X_1, X_2 (as they are disjoint), we get that the bad event above holds with probability at most

$$\Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S^\star \wedge \mathbf{X} \in X_2] > \varepsilon\right],$$

where f is a uniformly random function, and $S^\star$ is an (adversarially) fixed set of size $< \lceil 8K \rceil$. Now, by definition of X_2, each $x \in X_2$ is hit by $\mathbf{X}$ with probability at most $1/(4K)$. Thus, applying Claim 5 (setting parameters appropriately), we get

$$\Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S^\star \wedge \mathbf{X} \in X_2] > \varepsilon\right] \leq 2^{-2\varepsilon K(g-\log(36/\varepsilon))} = 2^{-2\varepsilon K(g-\frac{1}{\lfloor L \rfloor}\log(36/\varepsilon))},$$

as desired. Thus, we can henceforth assume that $L \geq 2$.

Now, let $\mathcal{E}$ denote the (bad) event in the lemma statement, and let $\tau \in [M]$ be a threshold value that we will set later. Since f acts independently on X_1, X_2 (as they are disjoint), observe that

$$\Pr_f[\mathcal{E}] \leq \Pr_f\left[\exists S \subseteq [M], |S| = \tau : \Pr_{\mathbf{X}}[f(\mathbf{X}) = v \wedge \mathbf{X} \in X_1] > \frac{L-1}{L}\cdot\frac{1}{K'}, \forall v \in S\right]$$
$$+ \Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S^\star \wedge \mathbf{X} \in X_2] > \varepsilon\right],$$

where $S^\star \subseteq [M]$ is an arbitrary fixed set of size $\tau - 1$.[31] By Claim 5 (setting parameters appropriately), and the fact that each $x \in X_2$ is hit by $\mathbf{X}$ with probability at most $1/(4KL)$, we have

$$\Pr_f\left[\Pr_{\mathbf{X}}[f(\mathbf{X}) \in S^\star \wedge \mathbf{X} \in X_2] > \varepsilon\right] \le 2^{-2\varepsilon KL\log(\frac{\varepsilon M}{2\tau})}.$$

Finally, consider any fixed set $S \subseteq [M]$ of size τ. Then, by Claim 8 (used in a similar manner as in the proof to Lemma 15), we have

$$\begin{aligned}
&\Pr_f\left[\exists S \subseteq [M], |S| = \tau : \Pr_{\mathbf{X}}[f(\mathbf{X}) = v \wedge \mathbf{X} \in X_1] > \frac{L-1}{L}\cdot\frac{1}{K'}, \forall v \in S\right]\\
&\le \binom{|X_1|}{\tau}\left(\frac{4eL}{G(L-1)}\right)^{\tau(L-1)}\\
&\le \left(\frac{e\lceil 4KL\rceil}{\tau}\right)^{\tau}\cdot\left(\frac{8e}{G}\right)^{\tau(L-1)}\\
&\le 2^{-\tau L(\frac{L-1}{L}(g-\log(8e))-\frac{1}{L}\log(8eKL/\tau))}.
\end{aligned}$$

And thus, we have

$$\Pr_f[\mathcal{E}] \le 2^{-2\varepsilon KL\log(\frac{\varepsilon M}{2\tau})} + 2^{-\tau L(\frac{L-1}{L}(g-\log(8e))-\frac{1}{L}\log(8eKL/\tau))}.$$

Finally, setting $\tau := \lceil \varepsilon^{(L-1)/L}K'\rceil$ yields

$$\Pr_f\left[\mathcal{E}\right] \le 2\cdot 2^{-\frac{\varepsilon K}{2}(g-\frac{1}{L}\log(1/\varepsilon)-11)},$$

as desired. □

Putting everything together

By combining the necessary conditions for condensing failure (Proposition 2) with the fact that each such condition happens with low probability (Lemmas 14 to 16), we are finally able to prove that a random function is a good condenser (Theorem 11).

Proof of Theorem 11. Before we start, we note that we may assume $\ell \le g/4$. This is because if $\ell > g/4$, then combining Proposition 1 and Lemma 14 (observing that $\log(G)/L \le \log(4\ell)/L \le 2$) yields the result.

Now that we may assume $\ell \le g/4$, the result is almost immediate, via the sketch above. In particular, we first set $k' := k - \ell$, and let X_1 denote the heaviest $\min\{\lceil 4KL\rceil, N\}$ elements in $\text{support}(X)$, and X_2 the rest. Then, by Lemma 15, we know that the first condition in Proposition 2 holds with probability at most

$$\begin{aligned}
&2^{-\frac{\varepsilon K}{6}(g-\frac{1}{L}\log(\frac{3\lceil 4KL\rceil}{\varepsilon K})-5.886)}\\
&\le 2^{-\frac{\varepsilon K}{6}(g-\frac{1}{L}\log(\frac{1}{\varepsilon})-16)}.
\end{aligned}$$

[31] Notice that the second term realizes to 0 if $\tau = 1$.

Next, define $\tilde{G} = G/L$, $\tilde{\varepsilon} = \varepsilon/3$, $\tilde{K} = 4KL$, and $\tilde{L} = \tilde{K}\tilde{G}/M = 4L$. Since each $x \in X_2$ is hit with probability at most $1/(4KL) = 1/\tilde{K}$, Lemma 14 tells us that the second condition in Proposition 2 holds with probability at most

$$\begin{aligned}
&2^{-\frac{\tilde{\varepsilon}\tilde{K}}{2}(\tilde{g}-\frac{1}{\tilde{L}}\log(2e\tilde{G}/\tilde{\varepsilon})-\log e)} \\
&\leq 2^{-\frac{4\varepsilon KL}{6}(g-\ell-\frac{1}{4L}\log(6eG/\varepsilon)-\log e)} \\
&\leq 2^{-\frac{4\varepsilon KL}{6}(3g/4-\frac{1}{4L}\log(6eG/\varepsilon)-\log e)} && \text{(since we assumed } \ell \leq g/4\text{)} \\
&\leq 2^{-\frac{\varepsilon KL}{6}(2g-\frac{1}{L}\log(6e/\varepsilon)-4\log e)} \\
&\leq 2^{-\frac{\varepsilon KL}{6}(2g-\frac{1}{L}\log(1/\varepsilon)-10)}.
\end{aligned}$$

Finally, by Lemma 16, the third condition in Proposition 2 holds with probability at most

$$\begin{aligned}
&2 \cdot 2^{-\frac{\varepsilon K}{6}(g-\frac{1}{\lfloor L\rfloor}\log(3/\varepsilon)-11)} \\
&\leq 2 \cdot 2^{-\frac{\varepsilon K}{6}(g-\frac{1}{\lfloor L\rfloor}\log(1/\varepsilon)-13)}.
\end{aligned}$$

Thus, by a simple union bound, one of the conditions in Proposition 2 holds with probability at most

$$4 \cdot 2^{-\frac{\varepsilon K}{6}(g-\frac{1}{\lfloor L\rfloor}\log(1/\varepsilon)-16)}.$$

By the statement of Proposition 2, the result follows. □

6.2 A random function is a seeded condenser

Using our main existential result from the previous section, it is now straightforward to obtain our existential results for seeded condensers.

Theorem 12 (A random function is a seeded condenser)**.** *There exists a universal constant $C \geq 1$ such that for any $\ell \in [0, k+d]$ and $g \geq 0$ such that $m := k + d - \ell + g$ is an integer, and any $\varepsilon \in (0,1]$, the following holds. If $d \geq \log\left(\frac{n-k}{\varepsilon}\right) + C$ and $g \geq \frac{1}{\lfloor L\rfloor}\log\left(\frac{1}{\varepsilon}\right) + C$, then there exists a seeded condenser* $\mathsf{sCond} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for (n,k)-sources with loss ℓ, gap g, error ε, and seed length d.*

Proof. This is an immediate corollary of our main existential result (Theorem 9), by considering the family $\mathcal{X}$ of sources of the form $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is an (n,k)-source and $\mathbf{Y} \sim \{0,1\}^d$ is a uniform independent seed. □

We remark that a more general theorem can be established (that allows for gap $g \in [0, C]$ and recovers known existential results for seeded extractors), but we only record the one above for simplicity.

6.3 Existential condensers for block sources

In this section, we show our existential results for Chor-Goldreich sources, and ultimately prove Corollary 4 from the introduction. As a reminder, we cannot simply invoke our black box result on the existence of seedless condensers for any small family (Corollary 6), because the family of CG sources is not small. Indeed, a rough estimate would indicate that the number of (t,n,k)-CG sources is roughly $\binom{N}{K}^{K^0+K^1+\cdots+K^{t-1}} \approx 2^{gK^t}$. However, since each such source contain kt bits of min-entropy, applying Corollary 6 would only work if we allowed the gap blow-up by a factor of at least $1/\varepsilon$. Here, we aim to do much better, and in fact prove such results for the more general setting of block sources.

6.3.1 Two blocks (via seeded condensers)

We start by showing existential results for condensing block sources that contain only two blocks. As a reminder, we let $g_i := n_i - k_i$ denote the entropy gap in the i^{th} block of the input block source.

Theorem 13 (Existential results for block sources with two blocks)**.** *There is a universal constant $C \geq 1$ such that the following holds. There exists a (non-explicit) condenser $\mathsf{Cond} : \{0,1\}^{n_1} \times \{0,1\}^{n_2} \to \{0,1\}^m$ for $((n_1,k_1),(n_2,k_2))$-block sources with output length $m = k_1 + k_2 - \ell + g$, error ε, loss ℓ, and gap*

$$g \leq g_2 + \frac{1}{\lfloor L \rfloor}(g_2 + \log(1/\varepsilon)) + C,$$

provided that $k_2 \geq \log(g_1/\varepsilon) + C$.

Proof. This is an immediate corollary of our existential result for seeded condensers (Theorem 12), combined with fact that seeded condensers work for CG-correlated seeds (Lemma 8). □

Before we move on to the multi-block setting, a few remarks are in order. First, note that the first bullet in Corollary 4 is an immediate corollary of the above, since CG sources are less general than block sources. Next, we note that when there are not too many blocks (say, $t = O(1)$, and they all have similar lengths), the above result will give the best parameters. This is because one may simply group together the first $t-1$ blocks into a single block, and this will only add about $\log(t)$ onto the min-entropy requirement, which is not bad when t is small. Finally, we mention that using this idea and the above result, one may recover the parameters of the explicit condensers in [DMOZ23] (for constant-sized blocks), by brute-force searching for an excellent block-source condenser (using the above existential result), which condenses to rate 0.99. Then, one can apply the explicit instantiation of the iterative condensing framework, instantiated with the GUV extractor (as in Section 5.3.3).

6.3.2 More than two blocks (via iterative condensing)

We now turn to prove our existential result for the multi-block setting. As above, we do so by combining our existential seeded condensers with the fact that such condensers can handle correlated seeds. This time, however, we'll need to iterate, and apply a sequence of several condensers. We present our main existential result for the multi-block setting below, and remind the reader that we always use $g_i := n_i - k_i$ to denote the entropy gap in the i^{th} block.

Theorem 14 (Existential results for block sources with many blocks)**.** *There is a universal constant $C \geq 1$ such that the following holds. There exists a (non-explicit) condenser $\mathsf{Cond} : \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_t} \to \{0,1\}^m$ for $((n_1,k_1),\ldots,(n_t,k_t =: n_t - g))$-block sources with output length $m = (\sum_{i\in[t]} k_i) - \ell + g'$, error ε, loss ℓ, and gap*

$$g' \leq g + \exp\left(\frac{6\lceil\frac{4t^2}{\ell+1}\rceil}{\lfloor L^{\frac{1}{2t}}\rfloor}\right) \cdot \left(\frac{6\lceil\frac{4t^2}{\ell+1}\rceil}{\lfloor L^{\frac{1}{2t}}\rfloor}\right) \cdot (g_t + \log(1/\varepsilon) + Ct) + Ct$$

provided that $k_{i+1} \geq \log(g_i/\varepsilon) + \ell/t + C$ for all $i \in [t-1]$.

While the above theorem is quite general and can work for nearly any block source, the parameters may be a bit difficult to digest. Soon, we will show exactly what this theorem can yield for the less general (and more standard) setting of CG sources (in Corollary 8, Corollary 9, and Corollary 10). But first, we present its proof, which relies on the following lemma (allowing for a more careful fine-tuning of parameters).

Lemma 17 (Existential results for block sources with many blocks)**.** *There is a universal constant $C \geq 1$ such that for any (not necessarily constant) parameters $\ell \geq 0$ and $\tau \geq 1$, the following holds. There exists a (non-explicit) condenser* $\mathsf{Cond} : \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_t} \to \{0,1\}^m$ *for* $((n_1, k_1), \ldots, (n_t, k_t))$*-block sources with output length* $m = (\sum_{i\in[t]} k_i) - \ell^\star + g^\star$*, error* ε*, loss* $\ell^\star \leq \ell t + \lfloor (t-2)/\tau \rfloor t$*, and gap*

$$g^\star \leq g_t + e^{\frac{6\tau}{\lfloor L \rfloor}} \cdot \frac{6\tau}{\lfloor L \rfloor} \cdot (g_t + \log(1/\varepsilon) + Ct) + Ct$$

provided that $k_{i+1} \geq \log(g_i/\varepsilon) + \ell + \lfloor \frac{t-(i+1)}{\tau} \rfloor + C$ *for all* $i \in [t-1]$.

Given this lemma, it is straightforward to prove our main existential result for block sources with many blocks (Theorem 14). Indeed, it just involves picking the best settings of the parameters ℓ, τ.

Proof of Theorem 14. Let $\ell_0 := \ell/(2t)$, $\tau_0 := \lceil \frac{4t^2}{\ell+1} \rceil$, and set these as the first two parameters in Lemma 17. □

At last, we are ready to prove our core lemma. We do so, below.

Proof of Lemma 17. Let $\mathsf{sCond}_1, \mathsf{sCond}_2, \ldots, \mathsf{sCond}_{t-1}$ be a sequence of functions, where each $\mathsf{sCond}_i : \{0,1\}^{n_i} \times \{0,1\}^{m_{i+1}} \to \{0,1\}^{m_i}$ is a seeded $(n_i, k_i) \to_{\varepsilon_i} (m_i, m_i - g_i')$ condenser. Then, define $m_t := n_t$ and

$$m_i := k_i + m_{i+1} - \ell_i + g_i'$$

for every $i \in [t-1]$, where ℓ_i is some parameter to be set later. Our existential result for seeded condensers (Theorem 12) says that such condensers must exist, provided that each m_i is a positive integer and both of the following hold, for every $i \in [t-1]$ (where $C > 0$ is a universal constant):

- **Seed length requirement:** $m_{i+1} \geq \log(g_i/\varepsilon_i) + C$.
- **Output gap requirement:** $g_i' \geq \frac{1}{\lfloor L_i \rfloor} \log(1/\varepsilon_i) + C$.

Moreover, our iterative condensing framework (Lemma 9) says that given such seeded condensers, there exists a condenser $\mathsf{Cond} : \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_t} \to \{0,1\}^{m_1}$ for $((n_1, k_1), \ldots, (n_t, k_t))$-block sources with output length m_1, output gap $g' = g_t + \sum_{i\in[t-1]} g_i'$, and error $\varepsilon' = \sum_{i\in[t-1]} \varepsilon_i \cdot 2^{g_t + \sum_{j\in(i,t-1]} g_j'}$. Thus, our goal is to set parameters $\varepsilon_i, \ell_i, g_i'$ for every $i \in [t-1]$ such that each seeded condenser sCond_i exists, and so that the final condenser Cond achieves the parameters claimed in the theorem statement.

We start by introducing some intermediate parameters, which will help keep our calculations tidy. In particular, we define $\ell_t := 0$, and for every $i \in [t-1]$, we define

$$k_{\geq i} := \sum_{j\in[i,t]} k_j,$$
$$\ell_{\geq i} := \sum_{j\in[i,t]} \ell_j,$$
$$g_{\geq i}' := g_t + \sum_{j\in[i,t-1]} g_j'.$$

Using these definitions, it is easy to verify that each output length parameter $m_i, i \in [t]$ takes the form

$$m_i = k_{\geq i} - \ell_{\geq i} + g_{\geq i}'.$$

Thus, the final condenser Cond will have output length $m_1 = k_{\geq 1} - \ell_{\geq 1} + g'_{\geq 1}$, output gap $g' = g'_{\geq 1}$, and error $\varepsilon' = \sum_{i \in [t-1]} \varepsilon_i \cdot 2^{g'_{\geq i+1}}$. With these observations in hand, we are ready to start setting parameters.

To start, we focus on setting the error parameters ε_i. We would like to set them so that the overall error ε' is at most some target error ε. Looking at the expression for ε' above, this can be done by setting ε_i to a geometric series. In particular, for every $i \in [t-1]$, we define

$$\varepsilon_i := \varepsilon \cdot 2^{-(t-i)} \cdot 2^{-g'_{\geq i+1}}.$$

In doing so, it is straightforward to verify that the overall error ε' is at most ε, as desired.

Next, before we set each ℓ_i, g'_i, let's see how the setting of ε_i affected the seed length and output length requirements of the seeded condensers. First, plugging in our value of ε_i (and using our observation about the form of each m_i), our seed length requirement becomes the following, for every $i \in [t-1]$:

$$k_{\geq i+1} - \ell_{\geq i+1} \geq \log(g_i/\varepsilon) + (t-i) + C.$$

In fact, by incrementing the universal constant C by 1, it suffices to satisfy the following, for every $i \in [t-1]$:

$$k_{i+1} - \ell_{i+1} \geq \log(g_i/\varepsilon) + C. \tag{17}$$

Let's see how our output gap requirement changed. Plugging in our ε_i, it becomes, for every $i \in [t-1]$:

$$g'_i \geq \frac{1}{\lfloor L_i \rfloor}(\log(1/\varepsilon) + t - i + g'_{\geq i+1}) + C.$$

Moreover, if we add $g'_{\geq i+1}$ to both sides, the output gap requirement becomes:

$$g'_{\geq i} \geq \frac{1}{\lfloor L_i \rfloor}(\log(1/\varepsilon) + t - i) + (1 + \frac{1}{\lfloor L_i \rfloor})g'_{\geq i+1} + C. \tag{18}$$

Finally, recall that each $m_i = k_{\geq i} - \ell_{\geq i} + g'_{\geq i}$ must be a positive integer.

Now, let's turn to setting the loss parameters ℓ_i. We would like to set them so that the overall loss is not too high, but also so that the output gap requirement (which depends on $1/L_i$) stays low. Looking ahead, the final gap $g'_{\geq 1}$ will depend roughly on the sum of the terms $1/L_i$, and thus we set the loss parameters so $\{1/L_i\}$ forms a geometric series. We give ourselves some freedom over the shape of this geometric series, using the parameters $\ell \geq 0$ and $\tau \geq 1$ from the theorem statement. Then, for every $i \in [t-1]$ we define

$$\ell_i := \ell + \left\lfloor \frac{t-(i+1)}{\tau} \right\rfloor.$$

τ should be thought of as a controller for how much additional loss (between $[0,1]$) should be experienced by each successive seeded condenser. Notice that all $\tau > t-2$ yield an additional loss of zero.

Given this setting of loss parameters, observe that the total loss of the final condenser will be

$$\ell^\star = \ell_{\geq 1} = \sum_{i \in [t-1]} \left(\ell + \left\lfloor \frac{t-(i+1)}{\tau} \right\rfloor \right) \leq \ell t + \left\lfloor \frac{t-2}{\tau} \right\rfloor t,$$

as desired. Furthermore, observe that our seed length requirement (Equation (17)) is satisfied if

$$k_{i+1} \geq \log(g_i/\varepsilon) + \ell + \left\lfloor \frac{t-(i+1)}{\tau} \right\rfloor + C$$

for every $i \in [t-1]$, as provided in the theorem statement.

Thus, all that remains is to set the gap parameters g'_i for all $i \in [t-1]$. Towards this end, we pick the smallest values satisfying Equation (18), and so that each $m_i = k_{\geq i} - \ell_{\geq i} + g'_{\geq i}$ is a positive integer. By rounding up, notice that the latter requirement can always be satisfied as long as the former requirement is satisfied with the universal constant C incremented by 1, and so we can safely ignore it. Thus, we henceforth focus on picking the smallest values g'_i satisfying Equation (18). That is, we define each $g'_i, i \in [t-1]$ so that

$$g'_{\geq i} = \frac{1}{\lfloor L_i \rfloor}(\log(1/\varepsilon) + t - i) + (1 + \frac{1}{\lfloor L_i \rfloor})g'_{\geq i+1} + C$$

Then, we observe the following inequality.

$$g'_{\geq i} \leq \frac{1}{\lfloor L_i \rfloor}(\log(1/\varepsilon) + t - 1) + (1 + \frac{1}{\lfloor L_i \rfloor})(g'_{\geq i+1} + C)$$

Finally, we just need to upper bound $g^\star \leq g'_{\geq 1}$. Recalling that $g'_{\geq t} = g_t$, we solve the recurrence above.

$$\begin{aligned} g'_{\geq 1} &\leq \left(-1 + \prod_{i \in [t-1]} (1 + \frac{1}{\lfloor L_i \rfloor}) \right) (\log(1/\varepsilon) + t - 1) + \left(\prod_{i \in [t-1]} (1 + \frac{1}{\lfloor L_i \rfloor}) \right) (g_t + C(t-1)) \\ &\leq \left(e^{\sum_{i \in [t-1]} \frac{1}{\lfloor L_i \rfloor}} - 1 \right) (\log(1/\varepsilon) + t) + \left(e^{\sum_{i \in [t-1]} \frac{1}{\lfloor L_i \rfloor}} \right) (g_t + Ct) \\ &\leq \left(e^{\sum_{i \in [t-1]} \frac{1}{\lfloor L_i \rfloor}} - 1 \right) \log(1/\varepsilon) + \left(e^{\sum_{i \in [t-1]} \frac{1}{\lfloor L_i \rfloor}} \right) (g_t + C't), \end{aligned}$$

where the last step set $C' := C + 1$. Now, plugging in our parameter setting $\ell_i := \ell + \lfloor \frac{t-(i+1)}{\tau} \rfloor$ (and recalling the convention $L_i = 2^{\ell_i}$), we can bound the term in the exponent as follows.

$$\begin{aligned} \sum_{i \in [t-1]} \frac{1}{\lfloor L_i \rfloor} &\leq \frac{1}{\lfloor L \rfloor} \sum_{i \in [t-1]} \frac{1}{2^{\lfloor \frac{t-(i+1)}{\tau} \rfloor}} \\ &\leq \frac{2}{\lfloor L \rfloor} \sum_{i \in [t-1]} 2^{\frac{i+1-t}{\tau}} \\ &\leq \frac{4}{\lfloor L \rfloor} \sum_{i \in [t-1]} 2^{-\frac{i}{\tau}} \\ &= \frac{4}{\lfloor L \rfloor} \cdot \frac{1 - 2^{-(t-1)/\tau}}{2^{1/\tau} - 1} \\ &\leq \frac{4}{\lfloor L \rfloor} \cdot \frac{\tau}{\ln 2} \\ &\leq \frac{6\tau}{\lfloor L \rfloor}. \end{aligned}$$

Plugging this expression back into our bound for $g'_{\geq 1}$, we get

$$g'_{\geq 1} \leq (e^{6\tau/\lfloor L \rfloor} - 1) \log(1/\varepsilon) + e^{6\tau/\lfloor L \rfloor}(g_t + C't).$$

Now, since $e^x - 1 \leq e^x x$ for all $x \geq 0$, we get

$$g^\star \leq g'_{\geq 1} \leq g_t + e^{6\tau/\lfloor L \rfloor} \cdot \frac{6\tau}{\lfloor L \rfloor}(\log(1/\varepsilon) + g_t + C't) + C't,$$

as desired. This completes the proof. □

Corollaries for Chor-Goldreich sources

Now that we have proven our existential result for multi-block sources, we are ready to see what parameters it yields in the more well-behaved CG-source setting. We present our main existential result for multi-block CG sources, and note that $\log^*()$ denotes the extremely slow-growing iterated logarithm function.

Corollary 8 (Existential results for CG sources with many blocks)**.** *There is a universal constant $C \geq 1$ such that the following holds. There exists a (non-explicit) condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$ for $(t, n, k =: n-g)$-CG sources with output length $m = kt - \ell + g'$, error ε, loss ℓ, and gap*

$$g' \leq g + \exp\left(\frac{6\lceil\frac{4(\log^* t)^2}{\ell+1}\rceil}{\lfloor L^{\frac{1}{2\log^* t}}\rfloor}\right) \cdot \left(\frac{6\lceil\frac{4(\log^* t)^2}{\ell+1}\rceil}{\lfloor L^{\frac{1}{2\log^* t}}\rfloor}\right) \cdot (g + \log(1/\varepsilon) + C\log^* t) + C\log^* t$$

provided that $k \geq \log(g/\varepsilon) + \ell/\log^* t + C$.

Before we present its proof, we take some time to digest its parameters. In particular, we list two immediate corollaries, which are presented as bullet two in Corollary 4. In the first corollary, we show what happens to the gap if one asks for a *lossless* condenser for CG sources. In the second, we show that if one is willing to lose a very small amount of min-entropy, the gap can be very well maintained.

Corollary 9 (Existential results for CG sources with many blocks - lossless regime)**.** *There is a universal constant $C \geq 1$ such that the following holds. There exists a (non-explicit) condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$ for $(t, n, k =: n-g)$-CG sources with output length $m = kt + g'$, error ε, loss $\ell = 0$, and gap*

$$g' \leq g + \exp(C(\log^* t)^2) \cdot (g + \log(1/\varepsilon) + C\log^* t),$$

provided that $k \geq \log(g/\varepsilon) + C$.

Corollary 10 (Existential results for CG sources with many blocks - small gap regime)**.** *There is a universal constant $C \geq 1$ such that the following holds. There exists a (non-explicit) condenser $\mathsf{Cond} : (\{0,1\}^n)^t \to \{0,1\}^m$ for $(t, n, k =: n-g)$-CG sources with output length $m = kt - \ell + g'$, error ε, loss $\ell \leq 2(\log^* t)^2$, and gap*

$$g' \leq g + C \cdot 2^{-\log^* t} \cdot (g + \log(1/\varepsilon)) + C\log^* t,$$

provided that $k \geq \log(g/\varepsilon) + 2\log^* t + C$.

With these results in hand, we turn to prove Corollary 8.

Proof of Corollary 8. Let $t' \in \mathbb{N}$ and $b_1, \ldots, b_{t'} \in \mathbb{N}$ be parameters that we will set later, so that $\sum_i b_i = t$. Then, define $n_1, \ldots, n_{t'}$ and $k_1, \ldots, k_{t'}$ such that $n_i := nb_i$ and $k_i := kb_i$. Notice that any (t, n, k)-CG source is automatically an $((n_1, k_1), \ldots, (n_t, k_t))$-block source, simply by grouping the blocks into buckets.

The goal is to find the smallest number of buckets t' that we can divide the CG source into, while maintaining a relatively modest entropy requirement. In particular, recall that in order to get the strong upper bound on the final gap g' provided in Theorem 14, the min-entropy of the block source must satisfy

$$k_{i+1} \geq \log(g_i/\varepsilon) + \ell/t' + C$$

for all $i \in [t'-1]$, where $g_i := n_i - k_i$. Using our block parameters $b_1, \ldots, b_{t'}$ and the relations described above, this min-entropy requirement becomes

$$kb_{i+1} \geq \log(gb_i/\varepsilon) + \ell/t' + C, \tag{19}$$

for all $i \in [t'-1]$.

Now, define the parameter t' and block parameters $b_1, \ldots, b_{t'}$ such that the following hold:[32]

[32] Note that we may assume that we started off with $t > 2$ blocks, for otherwise this result holds via Theorem 13.

- $b_{t'} := 2$,
- $b_i \leq 2^{b_{i+1}}$ for every $i \in [t'-1]$,
- $b_{t'} \leq b_{t'-1} \leq \cdots \leq b_1$,
- $b_1 + \cdots + b_{t'} = t$,
- $t' \in \mathbb{N}$ is the smallest integer for which there exist $b_1, \ldots, b_{t'}$ satisfying the above constraints.

Notice that for such parameters, the min-entropy requirement (given in Equation (19)) is satisfied if

$$kb_{i+1} \geq \log(g/\varepsilon) + b_{i+1} + \ell/t' + C,$$

or rather

$$b_{i+1}(k-1) \geq \log(g/\varepsilon) + \ell/t' + C$$

for every $i \in [t'-1]$. But observe that if we simply require

$$k \geq \log(g/\varepsilon) + \ell/t' + C,$$

then all of these conditions must hold, as the above implies that $(k-1) \geq k/2$ (when $k \geq 2$), and we know from our constraints that $b_{i+1} \geq 2$.

Thus for any (t, n, k)-CG source and parameters $b_1, \ldots, b_{t'}$ satisfying the above constraints, we know that we can condense (with an output gap as promised in Theorem 14) as long as $k \geq \log(g/\varepsilon) + \ell/t' + C$. All that remains is to check how big t' can be, and in particular provide an upper bound on it. Towards this end, looking at the constraints on our parameters b_i and the minimality of t', it is straightforward to verify that t' cannot exceed the iterated logarithm of t. In other words, $t' \leq \log^* t$, as desired. □

To conclude this section, we note that one may wish for an existential result for CG sources with many blocks, where the output gap has *no dependence* on the number of blocks t. It is straightforward to combine the above ideas to obtain such a result, albeit with significantly more loss. In particular, one can instantiate the iterative condensing framework with optimal seeded *extractors*, instead of seeded condensers, so that the output gap is *exactly equal to* the input gap g, but the loss becomes roughly $O((\log^* t)(\log^* t + g + \log(1/\varepsilon)))$, and more importantly the required starting min-entropy (per block) becomes roughly $k \geq \log(n/\varepsilon) + 0.99n$. This required starting min-entropy can then be reduced to $k \geq C \log(n/\varepsilon)$ (for some constant C) by adding in (at the beginning) a *single* call to an optimal seeded condenser with seed length that has dependence $1 \log(1/\varepsilon)$ on the error. This will not significantly affect the overall loss, and the final gap will be of the form $g + O(1)$.

7 Impossibility results

We conclude the technical portion of the paper with simple, but useful, impossibility results.

7.1 An impossibility result for condensing general sources

First, we show a condenser version of the classic extractor impossibility result.

Theorem 15 (There do not exist condensers for general sources)**.** *Fix any function* $f : \{0,1\}^n \to \{0,1\}^m$ *and gap* g *such that* $0 \le g \le n$. *Then for any* $0 \le \varepsilon < 1$ *there exists a source* $\mathbf{X} \sim \{0,1\}^n$ *with min-entropy gap* g *such that*

$$H_\infty^\varepsilon(f(\mathbf{X})) \le \min\{n,m\} - \min\{m,g\} + \log\left(\frac{1}{1-\varepsilon}\right).$$

The term $c_\varepsilon := \log(\frac{1}{1-\varepsilon})$ is merely an artifact of the definition of smooth min-entropy (see Section 3.2).

Proof. Let $g' := \min\{m,g\}$. By definition of probability, there must be a prefix $\sigma \in \{0,1\}^{g'}$ such that $\Pr[f(\mathbf{U}_n)_{[g']} = \sigma] \ge 2^{-g'}$. Thus there is a set $X \subseteq \{0,1\}^n$ of density exactly $2^{-g'}$ such that $f(X)_{[g']} = \{\sigma\}$. Let $S = f(X)$ be the image of this set, and note it has size $|S| \le 2^{\min\{n,m\}-g'}$, since S is the image of a set of size $2^{n-g'}$, and since S is a subset of $\{0,1\}^m$ where all prefixes of length g' are the same (leaving at most $m - g'$ coordinates unfixed). Now, by the characterization of smooth min-entropy (Lemma 2),

$$\begin{aligned}1 &= \Pr[f(\mathbf{X}) \in S]\\ &\le |S| \cdot 2^{-H_\infty^\varepsilon(f(\mathbf{X}))} + \varepsilon\\ &= 2^{\min\{n,m\}-g'-H_\infty^\varepsilon(f(\mathbf{X}))} + \varepsilon.\end{aligned}$$

Solving for $H_\infty^\varepsilon(f(\mathbf{X}))$ completes the proof. □

7.2 An impossibility result for condensing block sources

Finally, we extend the above argument to show that it is impossible to condense a CG source without the gap of one of the input blocks showing up in the output.

Theorem 16 (Condensers for CG sources must maintain the gap)**.** *Fix any function* $f : (\{0,1\}^n)^t \to \{0,1\}^m$ *and gap* g *such that* $0 \le g \le n$. *Then for any* $0 \le \varepsilon < 1$ *there exists a* $(t, n, n-g)$*-CG source* $\mathbf{X} \sim (\{0,1\}^n)^t$ *such that*

$$H_\infty^\varepsilon(f(\mathbf{X})) \le m - g + \log\left(\frac{1}{1-\varepsilon}\right).$$

Proof. By induction. By the proof above, we know that for any function $f : \{0,1\}^n \to \{0,1\}^m$ there is a set $X \subseteq \{0,1\}^n$ of size $2^{n-g'}$ such that the g'-prefix of the set $f(X)$ is a constant σ. Consider now a function $f : (\{0,1\}^n)^t \to \{0,1\}^m$ and all of its restrictions $f_\alpha := f(\alpha, \cdot)$. By induction, for each α there is a $(t-1, n, n-g')$-CG source $\mathbf{X}_\alpha$ such that the g'-prefix of $f(\alpha, \mathbf{X}_\alpha)$ is a constant σ. By averaging, this constant σ must be the same for some $2^{-g'}$ fraction of α's. Let $\mathbf{A}$ be uniform over these, and consider the $(t, n, n-g')$ source $(\mathbf{A}, \mathbf{X}_\mathbf{A})$. By construction, the prefix of f is constantly σ on $(\mathbf{A}, \mathbf{X}_\mathbf{A})$. Moreover, if we define S as the image of this source, we know it has size at most $2^{m-g'}$, since its g'-prefix is fixed. We also know that it has size at most $2^{t(n-g')}$, given the entropy of $(\mathbf{A}, \mathbf{X}_\mathbf{A})$. Thus

$$\begin{aligned}1 &= \Pr[f(\mathbf{A}, \mathbf{X}_\mathbf{A}) \in S]\\ &\le |S| \cdot 2^{-H_\infty^\varepsilon(f(\mathbf{A},\mathbf{X}_\mathbf{A}))} + \varepsilon\\ &\le 2^{\min\{m-g', t(n-g')\}-H_\infty^\varepsilon(f(\mathbf{A},\mathbf{X}_\mathbf{A}))} + \varepsilon.\end{aligned}$$

Solving for $H_\infty^\varepsilon(f(\mathbf{A}, \mathbf{X}_\mathbf{A}))$ completes the proof. □

8 Open problems

The most attractive open problem is to get better explicit seeded condensers. If one could explicitly construct such condensers with seed length that has dependence $1\log(1/\varepsilon)$ on the error (and a reasonably small output gap), then it would become trivial to condense CG sources with even better parameters than in this paper. Indeed, all of the work behind our CG source condensers goes into creating a single block of entropy rate 0.99, and any good enough seeded condenser (i.e., with the above parameters) can do this in a single step.[33]

Even if such seeded condensers remain out of reach, other natural questions remain about condensing CG sources. For example, while we were able to construct explicit condensers for CG sources with very low entropy, we could only do so while blowing up the gap by a polynomial factor.[34] It would be great to see if one could explicitly condense CG sources whose blocks have min-entropy (say) $n^{0.99}$, while keeping the gap blow-up to just a constant factor. This would seem to require completely new techniques.

Finally, it would be interesting to study other natural classes of sources for which we cannot deterministically extract, but *can* deterministically condense, and try to construct the corresponding explicit condensers. Chor-Goldreich sources are just one family in this new category of sources, and we hope that the study of other such families will lead to a long line of fruitful research.

References

[AT19] Nir Aviv and Amnon Ta-Shma. On the entropy loss and gap of condensers. *ACM Transactions on Computation Theory (TOCT)*, 11(3):1–14, 2019.

[BCDT19] Avraham Ben-Aroya, Gil Cohen, Dean Doron, and Amnon Ta-Shma. Two-source condensers with low error and small entropy gap via entropy-resilient functions. In *23rd International Conference on Randomization and Computation (RANDOM 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[BGM22] Marshall Ball, Oded Goldreich, and Tal Malkin. Randomness extraction from somewhat dependent sources. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[BKS$^+$10] Boaz Barak, Guy Kindler, Ronen Shaltiel, Benny Sudakov, and Avi Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. *Journal of the ACM (JACM)*, 57(4):1–52, 2010. Preliminary version in STOC 2005.

[BRSW12] Boaz Barak, Anup Rao, Ronen Shaltiel, and Avi Wigderson. 2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction. *Annals of Mathematics*, pages 1483–1543, 2012. Preliminary version in STOC 2006.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988. Preliminary version in FOCS 1985.

[33] As a reminder, see Lemma 8 for how the parameters of a seeded condenser translate to its performance on CG sources. It is worth noting that for this application, we would also be more than happy with a seeded condenser that is quite lossy.

[34] This blow-up is due to the number of rows produced by the somewhere-condensers used in our constructions.

[CGR24] Eshan Chattopadhyay, Mohit Gurumukhani, and Noam Ringach. On the existence of seedless condensers: Exploring the terrain. In *64th Annual Symposium on Foundations of Computer Science (FOCS 2024, to appear)*. IEEE, 2024.

[CGZ22] Eshan Chattopadhyay, Jesse Goodman, and David Zuckerman. The space complexity of sampling. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *LIPIcs*, pages 40:1–40:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[Coh16] Gil Cohen. Local correlation breakers and applications to three-source extractors and mergers. *SIAM Journal on Computing*, 45(4):1297–1338, 2016. Preliminary version in FOCS 2015.

[DMOZ23] Dean Doron, Dana Moshkovitz, Justin Oh, and David Zuckerman. Almost Chor-Goldreich sources and adversarial random walks. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1–9, 2023.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *Journal of the ACM (JACM)*, 56(4):1–34, 2009. Preliminary version in CCC 2007.

[Li12] Xin Li. Design extractors, non-malleable condensers and privacy amplification. In Howard J. Karloff and Toniann Pitassi, editors, *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 837–854. ACM, 2012.

[Li13] Xin Li. Extractors for a constant number of independent sources with polylogarithmic min-entropy. In *54th Annual Symposium on Foundations of Computer Science (FOCS 2013)*, pages 100–109. IEEE, 2013.

[Li15] Xin Li. Non-malleable condensers for arbitrary min-entropy, and almost optimal protocols for privacy amplification. In Yevgeniy Dodis and Jesper Buus Nielsen, editors, *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part I*, volume 9014 of *Lecture Notes in Computer Science*, pages 502–531. Springer, 2015.

[MW97] Ueli Maurer and Stefan Wolf. Privacy amplification secure against active adversaries. In *17th Annual International Cryptology Conference (CRYPTO 1997)*, pages 307–321. Springer, 1997.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *Journal of Computer and System Sciences*, 52(1):43–52, 1996. Preliminary version in STOC 1993.

[Rao09] Anup Rao. Extractors for a constant number of polynomially small min-entropy independent sources. *SIAM Journal on Computing*, 39(1):168–194, 2009. Preliminary version in STOC 2006.

[Raz05] Ran Raz. Extractors with weak random seeds. In *37th Annual ACM Symposium on Theory of Computing (STOC 2005)*, pages 11–20, 2005.

[RW04] Renato Renner and Stefan Wolf. Smooth rényi entropy and applications. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 233. IEEE, 2004.

[SV86] Miklos Santha and Umesh V Vazirani. Generating quasi-random sequences from semi-random sources. *Journal of computer and system sciences*, 33(1):75–87, 1986. Preliminary version in FOCS 1984.

[Ta-96] Amnon Ta-Shma. On extracting randomness from weak random sources. In *28th Annual ACM Symposium on Theory of Computing (STOC 1996)*, pages 276–285, 1996.

[Vad12] Salil Vadhan. Pseudorandomness. *Foundations and Trends® in Theoretical Computer Science*, 7(1–3):1–336, 2012.

[Zuc07] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3:103–128, 2007. Preliminary version in STOC 2006.