Randomness Extraction from Somewhat Dependent Sources

Marshall Ball

Computer Science Department, Columbia University, New York, NY, USA

Oded Goldreich

Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel

Tal Malkin

Computer Science Department, Columbia University, New York, NY, USA

Abstract

We initiate a comprehensive study of the question of randomness extractions from two somewhat dependent sources of defective randomness. Specifically, we present three natural models, which are based on different natural perspectives on the notion of bounded dependency between a pair of distributions. Going from the more restricted model to the less restricted one, our models and main results are as follows.

1. *Bounded dependence as bounded coordination*: Here we consider pairs of distributions that arise from independent random processes that are applied to the outcome of a single global random source, which may be viewed as a mechanism of coordination (which is adversarial from our perspective).

   We show that if the min-entropy of each of the two outcomes is larger than the length of the global source, then extraction is possible (and is, in fact, feasible). We stress that the extractor has no access to the global random source nor to the internal randomness that the two processes use, but rather gets only the two dependent outcomes.

   This model is equivalent to a setting in which the two outcomes are generated by two independent sources, but then each outcome is modified based on limited leakage (equiv., communication) between the two sources.

   (Here this leakage is measured in terms of the number of bits that were communicated, but in the next model we consider the actual influence of this leakage.)
2. *Bounded dependence as bounded cross influence*: Here we consider pairs of outcomes that are produced by a pair of sources such that each source has bounded (worst-case) influence on the outcome of the other source. We stress that the extractor has no access to the randomness that the two processes use, but rather gets only the two dependent outcomes.

   We show that, while (proper) randomness extraction is impossible in this case, randomness condensing is possible and feasible; specifically, the randomness deficiency of condensing is linear in our measure of cross influence, and this upper bound is tight. We also discuss various applications of such condensers, including for cryptography, standard randomized algorithms, and sublinear-time algorithms, while pointing out their benefit over using a seeded (single-source) extractor.
3. *Bounded dependence as bounded mutual information*: Due to the average-case nature of mutual information, here there is a trade-off between the error (or deviation) probability of the extracted output and its randomness deficiency. Loosely speaking, for joint distributions of mutual information t, we can condense with randomness deficiency $O(t/\epsilon)$ and error ϵ, and this trade-off is optimal.

All positive results are obtained by using a standard two-source extractor (or condenser) as a black-box.

2012 ACM Subject Classification Theory of computation

Keywords and phrases Randomness Extraction, min-entropy, mutual information, two-source extractors, two-source condenser

Digital Object Identifier 10.4230/LIPIcs.ITCS.2022.12

13th Innovations in Theoretical Computer Science Conference (ITCS 2022).
Editor: Mark Braverman; Article No. 12; pp. 12:1–12:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Related Version Full versions posted on ECCC as TR19-183.
Full Version: https://eccc.weizmann.ac.il/report/2019/183/

Funding This work was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via Contract No. 2019-1902070006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.
Marshall Ball: Partially supported by an IBM Research PhD Fellowship.
Oded Goldreich: Partially supported by an ISF grant number (Nr. 1146/18) and received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702); research was conducted while he enjoyed the hospitality of the computer science department at Columbia University.

1 Introduction

The problem of extracting almost perfect randomness from sources of highly defective randomness is of great theoretical and practical importance, since perfect randomness is essential to cryptography and has numerous applications in algorithmic design, whereas natural sources of randomness are quite defective. In other words, the randomness extraction problem addresses the discrepancy between the perfect randomness that is postulated in various applications and the quite defective randomness that seems available to us in reality.

The foregoing problem has been the focus of much research in the last decades, where research has branched according to how the defective sources of randomness are modelled (see, e.g., [17]). Needless to say, an adequate modelling of such sources is pivotal to such studies. The two main branches, reviewed below, focus on sources of randomness with a worst-case guarantee asserting that no outcome appears with too high a probability. Specifically, the logarithm of the reciprocal of this probability, called *min-entropy*, is a main parameter in these studies.[1]

The two branches differ by the question of whether we have at our disposal a single source of the foregoing type or a few (say, two) *independent* sources of this type. In the first case, a randomness extractor cannot be deterministic; it must use a short random *seed*, where the length of the seed may be logarithmic in the length of the source (and the reciprocal of the desired error probability). Shaltiel's classical survey is focused on this case [16]; see also more recent accounts such as [19, Chap. 6].[2]

In some "off-line" algorithmic applications, seeded extractors provide a good solution to the extraction problem, since one can emulate the use of a short random seed by a deterministic enumeration of all possibilities. This emulation is not possible in "on-line" applications that dominate the areas of cryptography and distributed computing. This reality is the main motivation for the second branch of studies, which considers extraction from several *independent* sources (of defective randomness). Typically, one seeks to minimize

[1] Hence, the postulate that no n-bit long string occurs with probability greater than p takes the form of saying that the distribution has min-entropy at least $\log_2(1/p)$. Note that a perfectly random n-bit long string has min-entropy $\log_2(1/2^{-n}) = n$.

[2] The focus of Shaltiel's survey [16] on seeded extractors reflects the focus of research at that period, and specifically the quest for explicit constructions of extractors with optimal parameters such as seed-length, min-entropy bound, output length, and error probability (see, e.g., [16, Sec. 1.4] and [16, Table 1]).

the number of independent sources, and it is best to use only two. The rather recent breakthrough result of Chattopadhyay and Zuckerman [5], which provides a two-source extractor for polylogarithmic min-entropy, and its follow-ups (e.g., [12]) belong to this branch.

While some amount of independence between the two sources is definitely necessary for (seedless) extraction, it is desirable to allow as much dependence as possible. In other words, one should seek to relax the postulate of perfectly independent (defective) sources, and study the possibility of extracting randomness from somewhat dependent sources.

Actually, this problem was considered already in the early work of Chor and Goldreich [6], who suggested a simple definition and outlined the possibility and limitation of (proper) extraction under it (see [6, Sec. 3.3]). Their focus was on a small "amount of dependence" (i.e., the sources are almost independent (in a strong sense)), but – in that regime – their definition (i.e., [6, Def. 10]) does not meet our intuition. On the one hand, it is too rigid, as reflected by the fact that it does not cover natural cases that do allow for good extraction and are covered by our first model (described in Section 2.1). On the other hand, their definition led to a negative result regarding extraction (i.e., [6, Thm. 19]), which seems to have discouraged further research in this important direction. (See Section 5 for a more detailed account.)

In this work, we initiate a more comprehensive study of the question of randomness extractions from two somewhat dependent sources of defective randomness. Specifically, we present three natural models, which are based on different natural perspectives on the notion of dependency between a pair of distributions. Indeed, the most general model is based on the notion of mutual information, but we believe that the more restricted models are also natural. While the most restricted model (described in Section 2.1) allows for (proper) randomness extraction, the other two models (described in Sections 2.2 and 2.3) do not allow for (proper) extraction, but do allow for randomness condensing (defined below). In Section 4 we argue that these randomness condensers may be of value for various applications, including in cryptography. But we shall discuss the models themselves first.

The current version. This is an extended abstract of our work, which is posted on ECCC as TR19-183 and is hereafter cited as our report [2]. In the sequel, we make a few references to specific results that appear in our report [2]; however, not all results of [2] are mentioned in this extended abstract.

Organization. The rest of this extended abstract is organized as follow. In Section 2 we present the three models of bounded dependency sources, and in Section 3 we review our main results regarding these models. In Section 4 we discuss the usefulness of condensed sources of randomness: While the applications to cryptography (discussed in Section 4.1) are well known (cf. [14, 7]), this does not seem to be the case with respect to the aplications to standard and sublinear algorithms (discussed in Sections 4.2 and 4.3, resp.). Lastly, related works are discussed in 5.

2 Our models

Our three models draw on three different notions of (the amount of) "dependency" between sources. In each model, the "amount of dependence" is specified by a parameter that upper-bounds this amount (as well as by the standard parameter that lower-bounds the min-entropy of the individual sources).

The models are presented in order of generality, going from the more restricted to the less restricted. When the amount of dependency is zero, all models coincide with the standard model of two independent sources. On the other hand, when the amount of dependence reaches the min-entropy of the individual sources, all models include the case of identical sources (which coincides with the single source case). Hence, the amount of dependence that we consider is between these two extreme bounds.

We refrain form taking a categorical position regarding which of the models is "right"; we believe that each of them may be adequate in some settings and less so in others. Indeed, different models may suit different settings, and the fact that the models support different levels of extraction (or condensing) is a good reason to present them all.

2.1 The coordinated sources model

A *special case* (equiv., restricted version) of the "coordinated sources" model postulates that the two sources have access to many independent "micro-sources" of randomness that are each extremely defective (i.e., having min-entropy that is too low to be of any use). Furthermore, most of these micro-sources may have no entropy at all. Each source corresponds to a subsequence of the micro-sources, and each such subsequence contains a significant amount of min-entropy, but the subsequences corresponding to the two sources have a small (non-empty) intersection that is not known to us. Of course, if we knew the intersection, then we could have ignored the micro-sources in it when extracting from the two (residual) sources (i.e., extract from the non-intersecting parts), but the intersection is not known to us (i.e., to the extractor).[3]

More generally, the two sources may be two random processes that are each fed by the outcome of some global random process. We have no access to "underlying" global random source nor can we access the randomness "added" by each of the two sources (on top of the outcome of the global source). All we get is the outcomes of the two processes (i.e., sources). Specifically, consider the randomized processes A, B and C, each taking a "somewhat random" n-bit long string, denoted r_A, r_B and r_C. Then, we get $A(r_A, C(r_C))$ and $B(r_B, C(r_C))$ only, and we are only guaranteed that their min-entropy is k' bits larger than the length of the outcome of C. Note that we do not require that r_A, r_B and r_C be uniformly distributed, but the min-entropy requirement made regarding the outcomes of A and B does imply that r_A and r_B have min-entropy at least k' each. Here, $C(r_C)$ represents the coordination between the sources.

A different scenario, which is actually equivalent to the above, is that the two sources are initially independent, but become dependent due to (bounded) leakage. Specifically, suppose that each source starts having a somewhat random state (i.e., s_A and s_B), but their state changes in time and may depend also on few bits that are leaked between the states; that is, at each point in time, each source modifies its state based on its current state and bits that are leaked from the current state of the other source. We only see the states at a later time, and after a bounded number of bits were "communicated" (via leakage) between them. Here, this leakage represents the (adversarial) coordination between the sources.

[3] Likewise, if we knew which of the micro-sources included in one subsequence contain a sufficient amount of min-entropy, then we could partition this subsequence into two good parts, viewed as two independent auxiliary sources, and extract from these two auxiliary sources.

2.2 Sources of bounded cross influence

In the bounded "cross influence" model, we envision a setting akin the prior one, except that here we do not upper-bound the *potential* "influence" of each source on the other – as reflected in the amount of leakage (or communication) – but rather the *actual* influence as embodied in the two outcomes. Specifically, consider randomized processes A and B, each taking a sequence of n random bits, denoted r_A and r_B, and outputting $A(r_A, r_B)$ and $B(r_A, r_B)$; that is, each source is given both r_A and r_B. What we shall bound is the influence of r_B on A's output, and likewise the influence of r_A on B's output.

We say that the influence of r_B on A's output is at most t (bits) if for every two values r_A and r_B it holds that $\Pr_r[A(r_A,r) \neq A(r_A,r_B)] \leq 1-2^{-t}$. In other words, the value $A(r_A,r_B)$ is maintained, with probability at least 2^{-t}, when r_B is "re-randomized" (i.e., replaced by a random input r). Hence, the influence of r_B on A's output measures the actual effect that r_B has on A's output, regardless of how this effect comes about. The influence of r_A on B's output is defined analogously, and the cross influence of the two sources is defined as the sum of the two (opposite) influences.

To see that the cross influence can be much lower than the amount of coordination, consider the processes $A(r_A,r_B) = r_A$ and $B(r_A,r_B) = (r_B, \mathtt{IP}_2(r_A,r_B))$, where $\mathtt{IP}_2$ denotes the inner product (mod 2) function. These two sources have a single bit of cross influence (i.e., r_B has no influence on A's output, whereas the influence of r_A on B's output is confined to $\mathtt{IP}_2(r_A,r_B)$). It can be shown that the amount of coordination between these sources exceeds $0.499 \cdot |r_A|$ (see Theorems 3.4 and 6.1 in our report [2]).[4]

We also comment that, as proved in Proposition 4.2 in our report [2], *if a joint distribution* (X,Y) *has cross influence at most* t*, then* $\max_{x,y}\{\ell(x,y)\} \leq t$*, where*

$$\ell(x,y) = \log_2\left(\frac{\Pr[(X,Y)=(x,y)]}{\Pr[X=x]\cdot\Pr[Y=y]}\right). \tag{1}$$

This means that the min-entropy of the joint distribution (X,Y) is at most t units smaller than the sum of the min-entropies of the individual distributions X and Y. Note that the mutual information of X and Y equals $\mathrm{E}[\ell(X,Y)]$.

2.3 Sources of bounded mutual information

The mutual information of the joint distribution (X,Y), denoted $I(X;Y)$, is a well-established measure of the dependence between the two variables. It quantifies the amount of information obtained about one random variable through observing the other random variable; indeed, $I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. It seems that considering this measure in the current context requires no justification. Still, we note that this measure coincides with the *information* communicated between the parties in a generation protocol as considered in the case of coordinated sources.[5]

[4] Indeed, it is well-known that the communication complexity of computing $\mathtt{IP}_2$ on n-bit inputs is $\Omega(n)$, see [6, Thm. 21(ii)], but this does not mean that sampling the distribution $(r_A, (r_B, \mathtt{IP}_2(r_A,r_B)))$ requires $\Omega(n)$ bits of communication (since there may be other ways of sampling this distribution). In other words, the strategies in the coordination protocol need not compute $\mathtt{IP}_2$ on their disjoint inputs; they may sample the desired distribution arbitrarily. Nevertheless, Theorem 6.1 in our report [2] implies that a two-party protocol for sampling the distribution $(r_A, (r_B, \mathtt{IP}_2(r_A,r_B)))$ requires $\Omega(n)$ bits of communication, whereas Theorem 3.4 in [2] relates the communication complexity of sampling a joint distribution to the amount of coordination in it.

[5] Recall that the number of bits communicated in such a protocol captures the measure of coordination considered in our first model. But here we consider the *information about one party's random input*

As stated at the beginning of the current section, although the model of bounded mutual information is well-known and easiest to state, we do not confine our study of the extraction problem to it. The main reason for not confining the study to this model is that extraction under this model is extremely limited (see Theorem 3.4).

3 Our main results

For the sake of stating our results, we use the following notation, where in all cases n denotes the length of the source's outcome, k denotes the min-entropy, and t the bound on the relevant dependency measure.

$\mathtt{STD}_n(k) =$ the standard two-source model, where each (independent) source has min-entropy k.

$\mathtt{COOR}_n(k,t) =$ the model of *t-coordinated sources*, where each source has min-entropy k. (See Section 2.1 and Definition 3.1 in our report [2].)

$\mathtt{CRI}_n(k,t) =$ the model of joint distributions of *cross influence* t, where each source has min-entropy k. (See Section 2.2 and Definition 4.1 in our report [2].)

$\mathtt{MI}_n(k,t) =$ the model of joint distributions of *mutual information* t, where each source has min-entropy k. (See Section 2.3 and Definition 5.1 in our report [2].)

We first state the fact that, essentially, our models are contained in one another, and that this containment is strict.

▶ **Theorem 3.1.** (relations between the models):

1. $\mathtt{COOR}_n(k,t)$ is ϵ-close to being in $\mathtt{CRI}_n(k,t+\log_2(1/\epsilon))$: *For every $\epsilon > 0$ and every t, every t-coordinated joint distribution is ϵ-close to some distribution of cross influence $t+\log_2(1/\epsilon)$. On the other hand, there exists a joint distribution* (of min-entropy $n-4$) *that can be generated with at most six bits of cross influence, but is 0.24-far from any $(n-O(\log n))$-coordinated distribution.*
2. $\mathtt{CRI}_n(k,t)$ is strictly contained in $\mathtt{MI}_n(k,t)$: *For every t, every joint distribution of cross influence t has t bits of mutual information. On the other hand, for every $\epsilon > 0$, there exists a joint distribution* (of min-entropy $n-O(1/\epsilon)$) *that has mutual information at most three, but is $(\epsilon/2)$-far from any distribution that can be generated with $1/\epsilon$ bits of cross influence.*

The separation between cross influence and mutual information seems to be due to the *separation between worst-case and average-case notions of cross influence*, as established in Proposition 4.2 in our report [2]. Indeed, in retrospect, *a worst-case notion seems more adequate in the current setting* (cf., min-entropy versus entropy).

We next turn to our results regarding randomness extraction (and condensing) for each of the three models. Recall that an extractor with error ϵ for a model $\mathcal{M}$ is a function $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ such that for every joint distribution (X,Y) in $\mathcal{M}$, it holds that $F(X,Y)$ is ϵ-close to the uniform distribution over $\{0,1\}^m$.

▶ **Theorem 3.2.** (extraction for sources of bounded coordination): *If $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ is an extractor of error ϵ for the model $\mathtt{STD}_n(k)$, then F is an extractor of error 2ϵ for the model $\mathtt{COOR}_n(k+t+\log_2(1/\epsilon),t)$.*

that is revealed by the communication. Specifically, consider the trivial protocol in which A generates $(x,y) \leftarrow (X,Y)$ and sends y to B. Then, the information on X communicated to B equals $H(X)-H(X|Y)$, which equals $I(X;Y)$.

We comment that a loss of $t + \log_2(1/\epsilon) - O(1)$ units of min-entropy is unavoidable. But the good news is that modulo this loss, extraction for sources of bounded coordination is possible, let alone via a black-box use of standard extractors, provided that the min-entropy of the individual sources compensates for the amount of coordination. Unfortunately, this good news does not carry over to the other models.

Turning to the other models, the bottom-line is that *proper extraction is not possible, but condensing is possible.* Recall that a condenser with error ϵ and deficiency d for a model $\mathcal{M}$ is a function $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ such that for every joint distribution (X, Y) in $\mathcal{M}$, it holds that $F(X, Y)$ is ϵ-close a distribution that has min-entropy at least $m - d$.

▶ **Theorem 3.3.** (extraction for sources of bounded cross influence):

1. (condensing is possible): *If $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ is a condenser of error ϵ and deficiency d for the model $\texttt{STD}_n(k)$, then F is a condenser of error $2^t \cdot \epsilon$ and deficiency $d + t$ for the model $\texttt{CRI}_n(k, t)$.*
2. (the foregoing is essentially optimal): *There is no condenser with deficiency $o(t)$ and error $2^{-\Omega(t)}$ for $\texttt{CRI}_n(n - O(t), t)$, when the output length exceeds t. Furthermore, there is no extractor with error $1/4$ for $\texttt{CRI}_n(n - 4, 6)$.*

Indeed, $d = 0$ (in Part 1) is a special case that refers to the case that F is an extractor for $\texttt{STD}_n(k)$. In general, the fact that the condensing error (in Part 1) grows by a factor of 2^t does not worry us too much, since we can afford to make ϵ small enough to compensate for that (i.e., we can use $\epsilon \leq 2^{-2t}$, provided $k = \Omega(t)$).[6] So the take-home message here is that *condensing is possible at a cost – in terms of deficiency – that equals the amount of cross influence.* As will be articulated in Section 4, randomness of bounded deficiency is useful in many setting both in cryptography and in algorithmic design.

In light of the foregoing, it is quite disapointing that for the model of mutual information only weaker condensing results are possible. Nevertheless, we can obtain such condensers (again, by using any condenser for the stanadrad model as black-box).

▶ **Theorem 3.4.** (extraction for sources of bounded mutual information): *For every $\beta > 0$, the following holds.*

1. (condensing is possible): *If $F : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ is a condenser of error ϵ and deficiency d for the model $\texttt{STD}_n(k - 1)$, then F is a condenser of error $4\beta + 2^{t'} \cdot \epsilon$ and deficiency $d + t'$ for the model $\texttt{MI}_n(k, t)$, where $t' = (t + O(1))/\beta$.*
2. (the foregoing is essentially optimal): *There is no condenser with deficiency $o(t')$ and error $O(\beta)$ for $\texttt{MI}_n(n - t', t)$, when the output length exceeds $t' \overset{\text{def}}{=} ((t - 2)/\beta)$.*

Indeed, this condensing result is inferior to the one for the cross influence model (i.e., Theorem 3.3), since it involves a trade-off between the deficiency (i.e., t') and the error parameter (typically dominated by β). Specifically, in Theorem 3.4 the product of the error and the deficiency is linear in the mutual information bound (i.e., $\beta \cdot t' = \Theta(t)$; indeed, our issue is with the additive error term of β, not with the multiplicative factor of $2^{t'}$ (which can be tolerated as outlined above)). Still, as will be discussed in Section 4, some applications can benefit even from this trade-off.

Yet another definition of bounded dependence. Part 1 of Theorem 3.3 is proved by showing that if (X, Y) can be generated with t bits of cross influence, then, for every x and y, it holds that

$$\Pr[(X,Y)=(x,y)] \leq 2^t \cdot \Pr[X=x] \cdot \Pr[Y=y]. \tag{2}$$

[6] Currently, explicit constructions of standard extractors with error $\epsilon < 1/n$ are not known, but such extractors do exist. Furthermore, explicit constructions of condensers with error ϵ and deficiency $o(\log(1/\epsilon))$ were recently presented in [3].

(see Proposition 4.2 in our report [2]). In other words, the min-entropy of (X,Y) is at most t units smaller than the sum of the min-entropies of X and Y. One may consider this parametrized upper bound as yet another definition of bounded dependence, and note that it implies that (X,Y) has mutual information at most t. Hence, the parameter t in (2) is sandwiched between cross influence and mutual information. Furthermore, we show *approximate converses* of both inequalities:

- If (X,Y) satisfies (2) with parameter t, then it is ϵ-close to a distribution that can be generated with $t+O(\log(1/\epsilon))$ bits of cross influence (see Proposition 4.13 in our report [2]).
- If (X,Y) has mutual information t, then it is $O(\beta)$-close to a distribution (X',Y') that satisfies (2) with respect to the parameter $t'=(t+O(1))/\beta$. (This is the bulk of the proof of Theorem 5.5 in our report [2], which establishes Part 1 of Theorem 3.4.)

We also show that *perfect converses* do not hold (see Proposition 4.14 in our report [2] and Part 2 of Theorem 3.1, resp.).

4 On the usefulness of condensers

Randomness condensers (a.k.a condensers), introduced in [15, 14] (in seeded and seedless versions, respectively), transform highly defective sources of randomness, which are only guaranteed to have some minimal amount of min-entropy, into sources of randomness (that are close to) having a relatively small min-entropy *deficiency* (see Definition 2.1 in our report [2]). This falls short of what is done by randomness extractors, which transform highly defective sources of randomness into almost perfect sources of randomness, but it is highly non-trivial and useful in various ways.

Typically, condensers are viewed as a technical tool; specifically, as an intermediate step towards extractors. However, condensers may be useful by themselves, whenever randomness extraction is impossible or too expensive. To be more concrete, we distinguish three possible types of uses of randomness: Its uses in cryptography (and other adversary-ridden settings), in standard algorithmic applications, and in sub-linear time computations.[7]

The distinction boils down to the question of whether or not the paradigm of "derandomizing the seed of a seeded extractor" is applicable. The archetypical case in which this is applicable is when running a randomized decision procedure (or a pseudodeterministic search procedure [8]). In this case, we obtain a single sample from a single source, apply the (seeded) extractor with all possible values of the seed, invoke the original algorithm with its randomness replaced by each of the outcomes, and output the result that is in majority.[8]

(Recall that it is impossible to deterministically perform (seedless) randomness extraction from a single defective source, even if the min-entropy of the source is very high [6, Thm. 1]. Furthermore, the seed length of a seeded extractor (for a single source) must be at least logarithmic in the length of the source [13, Thm. 1.9].)

[7] These different types are discussed in Sections 4.1, 4.3 and 4.2, respectively. Whereas the observations made in Section 4.1 are well known (cf. [14, 7]), this does not seem to be the case with respect to Sections 4.3 and 4.2.

[8] Suppose that A is a decision procedure for the predicate $f:\{0,1\}^*\to\{0,1\}$ such that on input x the algorithm A uses m random bits and it holds that $\Pr_{r\in\{0,1\}^m}[A(x;r)=f(x)]>0.9$. Observe that if $E:\{0,1\}^n\times\{0,1\}^\ell\to\{0,1\}^m$ is a seeded extractor for min-entropy k and error 0.01, then, for every source Z of min-entropy k, it holds that $\Pr[|\{s\in\{0,1\}^\ell : A(x;E(Z,s))=f(x)\}|>2^{\ell-1}]>0.7$, because otherwise $\Pr_s[A(x;E(Z,s))\neq f(x)]\geq 0.3\cdot 0.5$ (which contradicts the error bound of E). Now, on input x and source outcome z, the new algorithm A' outputs the majority vote of $A(x,E(z,s))$, taken over all $s\in\{0,1\}^\ell$. Note that the running-time of A' is 2^ℓ times the running-time of A, where ℓ is the length of the seed used by E.

4.1 Cryptographic applications

Cryptographic settings are the archetypical case in which the paradigm of "de-randomizing the seed of a seeded extractor" is not applicable. Hence, one may either assume that the small amount of (perfect) randomness required for the seed of an extractor can obtained (independently of the main source of randomness) or assume that we have access to two (or more) independent sources of defective randomness. But what if the two sources are dependent in a manner allowing only for condensing (as in Theorems 3.3 and 3.4)?

As noted by Rao [14] (see articulation in [7, Sec. 5]), it turns out that a level of min-entropy deficiency that is sub-logarithmic in the "security error" can be tolerated in "unpredictability applications" (e.g., unforgeable signatures). Specifically, we say that a cryptographic system (of this type) has a security error of μ if adversaries (with specified resources) can break the system with probability at most μ (see [7, Def. 5.1]). The key observation is that *if the system has a security error of μ when using perfect randomness, then implementing the system with a distribution that is ϵ-close to having deficiency at most d yields a system of security error at most $2^d \cdot \mu + \epsilon$.* An analogous bound holds when a system designed for randomness of deficiency d_0 is used with randomness that has deficiency $d_0 + d$ (see [7, Lem. 5.1]).

The foregoing suggestion is appealing when having access to two sources of sufficiently low *cross influence*. Specifically, Theorem 3.3 asserts that, when using a standard two-source condenser (or extractor), a cross influence bound of t translates to an added deficiency of t units (while the error gets multipied by a factor of 2^t). Furthermore, in such a case we can use standard two-source condensers with small deficiency rather than standard two-source extractor.

The latter comment is important because the currently known explicit two-source extractors (for low (i.e., lower than 0.44) min-entropy (e.g., [5, 12])) have noticeable error (i.e., error that is polynomially related to the length of the source), which is unacceptable in most cryptographic applications. In contrast, a recent work of Ben-Aroya et al. [3] provides an explicit two-source condenser with deficiency that is *sub-logarithmic* in the desired error, which in turn may be set to be negligible. Hence, if an "unpredictability applications" has a security error of μ when using perfect randomness, then implementing it with two sources of cross influence t yields a system of security error at most $2^{2t} \cdot \mu^{1-o(1)}$, since for independent sources a condensing error of μ can be achieved with deficiency $o(\log(1/\mu)$.[9]

4.2 Standard algorithmic applications

Somewhat surprisingly, using two-source condensers for dependent sources may also be useful for running standard randomized algorithms. That is, *assuming that we have two somewhat dependent sources of weak randomness at our disposal*, we compare the option of using only one of these sources while employing the paradigm of "de-randomizing the seed of a seeded extractor" to the option of using the two sources with an adequate condenser. We claim that in many cases, the latter option is better.

Recall that the standard suggestion is to run such algorithm by extracting almost perfect randomness from a single defective random source, by using a seeded extractor. Following this suggestion, extraction from an n-bit long source requires a seed of length at least $\log_2 n$, which means that under the foregoing paradigm the original algorithm must be invoked $\Omega(n)$ times (while using all possible seeds with the same n-bit long outcome of the source).[10]

[9] Actually, we can get a system of security error at most $2^{t+o(t)} \cdot \mu^{1-o(1)}$ by using a standard condenser with error $2^{-t} \cdot \mu$ and deficiency $o(t + \log(1/\mu))$.

[10] See Footnote 8.

In contrast, using the single outcome extracted from two sources with *cross influence* at most t yields the same performance provided that the error probability of the algorithm is reduced from $\delta < 1/3$ to $2^{-t} \cdot \delta$. The point is that such an error reduction can be obtained by invoking the original algorithm $O(t)$ times, whereas typically $t \ll n$. The bottom-line is that *if the bound on the cross influence of the two sources is good enough* (i.e., $t \ll n$ as well as t being smaller than (say) half the min-entropy of each source), *then we are better off using two somewhat-dependent sources* (via the condenser) rather than one source (via a seeded extractor).[11] (Recall that Theorem 3.3 asserts that a cross influence bound of t translates to an added deficiency of t units, which means that the error probability can grow by a factor of at most 2^t.)

The foregoing holds even when using two sources that have *mutual information* at most t (rather than cross influence at most t), where we assume all along that each source has sufficient amount of min-entropy. In this case we use Theorem 3.4 rather than Theorem 3.3, which means that the deficiency bound we obtain is larger (see next). Specifically, the number of invocations grows by a factor of $O(t/\epsilon)$ (rather than t), where ϵ is the desired error. That is, given a randomized algorithm A of error probability δ, and wishing to utilize A when having access to a pair of sources that have mutual information t, while obtaining error probability ϵ, we need to reduce A's error to $2^{-\Omega(t/\epsilon)} \cdot \epsilon$ (rather than to $2^{-t} \cdot \epsilon$, as when using sources of cross influence t). Hence, we shall invoke A for $O(t/\epsilon)$ times (rather than $O(t)$ times), which is feasible only if we are willing to tolerate a noticeable error probability (e.g., $\epsilon = 0.01$ or so).

Finally, note that for search problems that can be solved in probabilistic polynomial-time but are not BPP-search problems (i.e., valid solutions cannot be efficiently recognized (cf. [10])[12]), the paradigm of "de-randomizing the seed of a seeded extractor" is not applicable at all, since listing the solutions that correspond to all possible seeds does not allow to select a valid one. So in this case, if the search problem is solvable with sufficiently small error probability (when using perfect randomness), then we can use the "condenser path".[13]

4.3 Sublinear-time applications

As noted in [9], the fact that the overhead of the paradigm of "de-randomizing the seed of a seeded extractor" is at least as large as the randomness complexity of the original algorithm limits the applicability of this paradigm in the context of sublinear-time computations. This is particularly the case when seeking sample or query complexity that is independent of the size of the input. Two such cases arise in the context of sampling and property testing.

Consider, for example, the task of estimating the average of a function $f : \{0,1\}^n \to [0,1]$. Any reasonable notion of estimation will require randomness complexity $\Omega(n)$. Hence, employing the "seed de-randomization" paradigm will require making $\Omega(n)$ probes to the function. In contrast, when given access to two sources of cross influence t, it suffices to make $O(t)$ probes to obtain a constant factor approximation with probability 0.999. (This corresponds to using a sampler that, when using perfect randomness, obtains such an approximation with probability $1 - 2^{-t-10}$.)

[11] Specifically, given a randomized algorithm A, we reduce its error by a factor of 2^t by invoking it $O(t)$ times. Next, we run the resulting algorithm, denoted A', when feeding it with the output of the condenser, which was applied to the a pair of outcomes provided by the two sources. Hence, we invoke A' only once, but this invocation generates t invocations of A. Yet, assuming $t \ll n$, this is preferable to the $2^\ell > n$ invocations that take place when using a seeded extractor with seed length ℓ (see Footnote 8).

[12] So, in particular, they do not have pseudodeterministic algorithms (cf. [8]).

[13] Note that error reduction is also not feasible in this case (when valid solutions cannot be efficiently recognized).

The same considerations apply in the context of property testing (see [11]). In particular, many testers have complexity that only depends on the proximity parameter (and is independent of the size of the tested object), hereafter called strong testers. On the other hand, the randomness complexity of any reasonable testing task is at least logarithmic in the size of the object. Hence, strong testability cannot be achieved when using the paradigm of "de-randomizing the seed of a seeded extractor" but it can be achieved when employing a condenser to a pair of sources of cross influence $O(1)$. Furthermore, some testing tasks may have sublinear query and randomness complexities, but the paradigm of "de-randomizing the seed of a seeded extractor" yields query complexity that is the product of the two, which may be more than linear (i.e., worse than the trivial "tester" that just reads the entire object).[14] In contrast, we can obtain sub-linear complexity when employing a condenser to a pair of sources of cross influence $t = o(n/q)$, where n is the size of the tested object and q is the query complexity of the original tester (which uses perfect randomness).

5 Related work

The problem of *dependent sources* of defective randomness was already considered in the early work of Chor and Goldreich [6]. In particular, they suggested a simple definition (i.e., [6, Def. 10]), which allowed for extraction with error proportional to the governing parameter (cf., [6, Lem. 18]), alas this error bound was shown to be essentially tight (i.e., [6, Thm. 19]). Specifically, for $\delta > 0$, the joint distribution (X, Y) was said to be δ-dependent if for every $x, y \in \{0,1\}^n$ it holds that

$$(1+\delta)^{-1} \;\leq\; \frac{\Pr[(X,Y)=(x,y)]}{\Pr[X=x]\cdot\Pr[Y=y]} \;\leq\; 1+\delta. \tag{3}$$

While the formulation intentionally allows to consider also $\delta > 1$, the focus of [6, Sec. 3.3] was on smaller values of δ. They observed that any standard extractor with error ϵ yields an extractor with error $\delta + (1+\delta)\cdot\epsilon$ for δ-dependent sources of the same min-entropy bound (cf., [6, Lem. 18]). On the other hand, they showed that any extractor for the class of δ-dependent sources has error $\Omega(\delta)$, even when the sources have min-entropy $n-2$ (cf., [6, Thm. 19(i)]).

We believe that, for small values of δ (i.e., $\delta \approx 0$), the model captured by (3) is not a satisfactory model of somewhat dependent sources. On the one hand, it is too rigid, as reflected by the fact that it does not cover natural cases that do allow for good extraction and are covered by our first model (described in Section 2.1). Consider, for example, a joint distribution (X, Y) such that $X = (X', Z)$ and $Y = (Y', Z)$ where X', Y' and Z are mutually independent (and have each some min-entropy). Then, every pair $(x, y) = ((x', z), (y', z))$ in the support of (X, Y) violates (3), since the ratio in this case equals $1/\Pr[Z=z]$. Furthermore, for (x', z) and (y', z') (in the support of X and Y, respectively) such that $z \neq z'$, the ratio is not even bounded. However, intuitively, these sources have bounded dependency (i.e., the "dependency" seems $|Z|$), and we should be able to extract good randomness from them, even when the location of the overlapping parts of X and Y are not be known (to the extractor) but $|Z|$ is very small. Indeed, such sources are $|Z|$-coordinated; in fact, they are t-coordinated if 2^t upper-bounds the support of Z (see Definition 3.1 in our report [2]).

[14] E.g., consider the case that both the query and randomness complexities equal a square root of the size of the object.

On the other hand, even for small values of δ, the model captured by (3) does not allow for (proper) randomness extraction, unless δ is extremely small (i.e., smaller than the desired error of the extractor). However, as mentioned above, extraction is possible in the (incomparable) model of bounded coordination (see Theorem 3.2). In conclusion, the point is that that the model captured by (3) does not offer a wide class of joint distributions for which proper extraction is possible. Furthermore, while extraction is impossible, condensing may be possible and is possible in this case, since distributions that satisfy (3) definitely satisfy (2) with parameter $t = \log_2(1+\delta)$.

Indeed, in retrospect, for large $\delta \gg 1$, (3) essentially coincides with (2). On the one hand, as noted above, any joint distribution that satisfies (3) with parameter δ also satisfies (2) with parameter $t = \log_2(1+\delta)$. On the other hand, any joint distribution (X, Y) that satisfies (2) with parameter t also satisfies the upper bound of (3) with parameter $\delta = 2^t - 1$. Furthermore, while (X, Y) may not satisfy the lower bound of (3) (for any parameter δ), it is $(2^t - 1)^{-1}$-close to satisfy (3) with parameter $\delta = 2^t - 1$. (More generally, for any $\epsilon \le (2^t - 1)^{-1}$, it holds that (X, Y) is ϵ-close to a distributed that satisfies (3) with parameter $\delta = 1/\epsilon$.)[15]

Extraction from recognizable distributions. As part of a wider study, Shaltiel [18] showed that two-source extractors sufficed to extract randomness from two sources that output a joint distribution that is uniform over a set that is recognized by a two-party protocol of bounded communication [18, Def. 3.1 & Thm. 4.6]. We note that this is, in fact, a special case of Theorem 3.2, since any distribution that (has a support that) is recognized by a protocol that communicates t bits is a t-coordinated.[16] The converse does not hold, because a distribution is t-coordinated if and only if it can be sampled by a protocol of communication complexity t (see Theorem 3.4 in our report [2]), whereas sampling a distribution may require far less communication than recognizing it (let alone by a deterministic protocol).[17]

[15] Consider a pair of independent random variables (X', Y') such that $X' \equiv X$ and $Y' \equiv Y$, and let (X'', Y'') equal (X', Y') with probability ϵ and equal (X, Y) otherwise. Then, (X, Y) is ϵ-close to (X'', Y''), which satisfies (3) with parameter $\delta = 1/\epsilon$. To verify the latter claim, note that

$$\begin{aligned}\Pr[(X'',Y'')=(x,y)] &\ge \epsilon\cdot\Pr[(X',Y')=(x,y)]\\ &= \epsilon\cdot\Pr[X'=x]\cdot\Pr[Y'=y]\\ &= \frac{1}{\delta}\cdot\Pr[X''=x]\cdot\Pr[Y''=y],\end{aligned}$$

and, on the other hand,

$$\begin{aligned}\Pr[(X'',Y'')=(x,y)] &= \epsilon\cdot\Pr[(X',Y')=(x,y)] + (1-\epsilon)\cdot\Pr[(X,Y)=(x,y)]\\ &\le \epsilon\cdot\Pr[X'=x]\cdot\Pr[Y'=y] + (1-\epsilon)\cdot 2^t\cdot\Pr[X=x]\cdot\Pr[Y=y]\\ &\le 2^t\cdot\Pr[X''=x]\cdot\Pr[Y''=y],\end{aligned}$$

where the first inequality is due to (2).

[16] Let Π be a protocol of communication complexity t that recognizes the joint distribution (X, Y). Then, each possible accepting t-bit transcript corresponds to the uniform distribution on a combinatorial rectangle. This implies that we can generate (X, Y) via a global process that selects a random transcript (i.e., $z \leftarrow Z$) that is used by each individual process to generate the corresponding conditional distribution (i.e., $X|_{Z=z}$ and $Y|_{Z=z}$).

[17] Consider, for example, a protocol for non-disjointness of $\sqrt{n}$-subsets of $[n]$, in which one party select $i \in [n]$ uniformly at random, sends it to the other party, and each party outputs a uniformly distributed $\sqrt{n}$-subset that contains i. Recall that recognizing the support of this distribution requires $\Omega(\sqrt{n})$ communication.

Concurrent work.[18] In a concurrent work, Chattopadhyay et al. [4] consider the construction of extractors for "adversarial sources" (defined as "somewhat good sources" with "bounded dependency").[19] Specifically, they consider N sources such that at least K of them are good in the sense that they are independent and each contains a considerable amount of min-entropy (i.e., the min-entropy k is $N^{\Omega(1)}$), and each bad source depends on a bounded number of good sources. This seems related to the special case of micro-sounrces discussed in the beginning of Section 2.1, where the differences include

- The micro-sources that we consider may each contain a very small amount of min-entropy (e.g., $k = O(1)$). In contrast, in [4] the min-entropy is related to the number of sources (e.g., $k = N^{\Omega(1)}$).
- We consider a 2-partition of the micro-sources such that the micro-sources on each side have bounded dependency on the micro-sources of the other side, where the bound need only be smaller than the total min-entropy on each side. We can allow arbitrary dependency among the micro-sources that reside on the same side as their total min-entropy is large enough. In contrast, in [4] the bad sources may be coordinated arbitrarily as long as each bad source depends on a bounded number of good sources.

It seems that Chattopadhyay et al. [4] envision sources as being controlled by possibly adversarial parties, whereas we try to model sources that are available in nature.

References

1 Divesh Aggarwal, Maciej Obremski, João L. Ribeiro, Luisa Siniscalchi, and Ivan Visconti. How to extract useful randomness from unreliable sources. In *EUROCRYPT (1)*, volume 12105 of *Lecture Notes in Computer Science*, pages 343–372. Springer, 2020.

2 Marshall Ball, Oded Goldreich, and Tal Malkin. Randomness extraction from somewhat dependent sources. *Electron. Colloquium Comput. Complex.*, page 183, 2019.

3 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 *APPROX-RANDOM*, volume 145 of *LIPIcs*, pages 43:1–43:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

4 Eshan Chattopadhyay, Jesse Goodman, Vipul Goyal, and Xin Li. Extractors for adversarial sources via extremal hypergraphs. In *STOC*, pages 1184–1197. ACM, 2020.

5 Eshan Chattopadhyay and David Zuckerman. Explicit two-source extractors and resilient functions. *Electron. Colloquium Comput. Complex.*, page 119, 2015.

6 Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.*, 17(2):230–261, 1988.

7 Yevgeniy Dodis, Thomas Ristenpart, and Salil P. Vadhan. Randomness condensers for efficiently samplable, seed-dependent sources. In *TCC*, volume 7194 of *Lecture Notes in Computer Science*, pages 618–635. Springer, 2012.

8 Eran Gat and Shafi Goldwasser. Probabilistic search algorithms with unique answers and their cryptographic applications. *Electron. Colloquium Comput. Complex.*, page 136, 2011.

9 Oded Goldreich. Another motivation for reducing the randomness complexity of algorithms. In *Studies in Complexity and Cryptography*, volume 6650 of *Lecture Notes in Computer Science*, pages 555–560. Springer, 2011.

10 Oded Goldreich. In a world of p=bpp. In *Studies in Complexity and Cryptography*, volume 6650 of *Lecture Notes in Computer Science*, pages 191–232. Springer, 2011.

11 Oded Goldreich. *Introduction to Property Testing.* Cambridge University Press, 2017.

[17] A preliminary version of the current work was posted on ECCC in 2019 (see TR19-183).
[19] They follow-up on a very recent work of Aggarwal et al. [1].

12 Xin Li. Improved constructions of two-source extractors. *CoRR*, abs/1508.01115, 2015.

13 Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM J. Discret. Math.*, 13(1):2–24, 2000.

14 Anup Rao. A 2-source almost-extractor for linear entropy. In *APPROX-RANDOM*, volume 5171 of *Lecture Notes in Computer Science*, pages 549–556. Springer, 2008.

15 Ran Raz and Omer Reingold. On recycling the randomness of states in space bounded computation. In *STOC*, pages 159–168. ACM, 1999.

16 Ronen Shaltiel. Recent developments in explicit constructions of extractors. *Bull. EATCS*, 77:67–95, 2002.

17 Ronen Shaltiel. An introduction to randomness extractors. In *ICALP (2)*, volume 6756 of *Lecture Notes in Computer Science*, pages 21–41. Springer, 2011.

18 Ronen Shaltiel. Weak derandomization of weak algorithms: Explicit versions of yao's lemma. *Comput. Complex.*, 20(1):87–143, 2011.

19 Salil P. Vadhan. Pseudorandomness. *Found. Trends Theor. Comput. Sci.*, 7(1-3):1–336, 2012.