Samplers and extractors for unbounded functions

Rohit Agrawal*

April 17, 2019

Abstract

Błasiok (SODA’18) recently introduced the notion of a subgaussian sampler, defined as an averaging sampler for approximating the mean of functions \( f : \{0,1\}^m \rightarrow \mathbb{R} \) such that \( f(U_m) \) has subgaussian tails, and asked for explicit constructions. In this work, we give the first explicit constructions of subgaussian samplers (and in fact averaging samplers for the broader class of subexponential functions) that match the best-known constructions of averaging samplers for \([0,1]\)-bounded functions in the regime of parameters where the approximation error \( \varepsilon \) and failure probability \( \delta \) are subconstant. Our constructions are established via an extension of the standard notion of randomness extractor (Nisan and Zuckerman, JCSS’96) where the error is measured by an arbitrary divergence rather than total variation distance, and a generalization of Zuckerman’s equivalence (Random Struct. Alg.’97) between extractors and samplers. We believe that the framework we develop, and specifically the notion of an extractor for the Kullback–Leibler (KL) divergence, are of independent interest. In particular, KL-extractors are stronger than both standard extractors and subgaussian samplers, but we show that they exist with essentially the same parameters (constructively and non-constructively) as standard extractors.

1 Introduction

1.1 Averaging samplers

Averaging (or oblivious) samplers, introduced by Bellare and Rompel [BR94], are one of the main objects of study in pseudorandomness. Used to approximate the mean of a \([0,1]\)-valued function with minimal randomness and queries, an averaging sampler takes a short random seed and produces a small set of correlated points such that any given \([0,1]\)-valued function will (with high probability) take approximately the same mean on these points as on the entire space. Formally,

**Definition 1.1** ([BR94]). A function \( \text{Samp} : \{0,1\}^n \rightarrow \{\{0,1\}^m\}^D \) is a \((\delta, \varepsilon)\) averaging sampler if for all \( f : \{0,1\}^m \rightarrow [0,1] \), it holds that

\[
\Pr_{x \sim U_n} \left[ \left| \frac{1}{D} \sum_{i=1}^{D} f(\text{Samp}(x)_i) - \mathbb{E}[f(U_m)] \right| > \varepsilon \right] \leq \delta,
\]

where \( U_n \) is the uniform distribution on \( \{0,1\}^n \). The number \( n \) is the randomness complexity of the sampler, and \( D \) is the sample complexity. A sampler is explicit if \( \text{Samp}(x, i) \) can be computed in time \( \text{poly}(n, m, \log D) \).

Traditionally, averaging samplers have been used in the context of randomness-efficient error reduction for algorithms and protocols, where the function \( f \) is the indicator of a set \((\{0,1\}\)-valued), or more generally the acceptance probability of an algorithm or protocol \((\{0,1\}\)-valued). There has been significant effort in the literature to establish optimal explicit and non-explicit constructions of samplers, which we summarize in

---

* John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. Email: rohitagr@seas.harvard.edu. Supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.
Table 1. We recommend the survey of Goldreich [Gol11b] for more details, especially regarding non-averaging samplers

<table>
<thead>
<tr>
<th>Key Idea</th>
<th>Randomness complexity</th>
<th>Sample complexity</th>
<th>Best regime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairwise Independence [CG89]</td>
<td>$m + \log(1/\delta) + 2 \log(1/\epsilon)$ + $O(1)$</td>
<td>$O\left(\frac{1}{\epsilon^2}\right)$</td>
<td>$\delta = \Omega(1)$</td>
</tr>
<tr>
<td>Extractors</td>
<td>$m + (1 + \alpha) \cdot \log(1/\delta)$ any constant $\alpha &gt; 0$</td>
<td>$\text{poly}(\log(1/\delta), 1/\epsilon)$</td>
<td>$\epsilon, \delta = o(1)$</td>
</tr>
<tr>
<td>Zuc97, GW97, RVW00, GUV09</td>
<td>$m + O(\log(1/\delta)/\epsilon^2)$</td>
<td>$O\left(\frac{\log(1/\delta)}{\epsilon \delta}\right)$</td>
<td>$\epsilon = \Omega(1)$</td>
</tr>
<tr>
<td>Expander Walks [Gil98]</td>
<td>$m + \log(1/\delta) - \log(1/\delta)$ + $O(1)$</td>
<td>$O\left(\frac{\log(1/\delta)}{\epsilon \delta}\right)$</td>
<td>All</td>
</tr>
<tr>
<td>Non-Explicit [Zuc97]</td>
<td>$m + \log(1/\delta) - \log(1/\delta)$</td>
<td>$\Omega\left(\frac{\log(1/\delta)}{\epsilon \delta}\right)$</td>
<td>N/A</td>
</tr>
<tr>
<td>Lower Bound [CEG95, Zuc97, RT00]</td>
<td>$m + \log(1/\delta) - \log(1/\delta)$ - $\log(1/\epsilon) - O(1)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Best known constructions of averaging samplers for $[0, 1]$-valued functions

However, averaging samplers can also have uses beyond bounded functions: Blasiok [Bla18b], motivated by an application in streaming algorithms, introduced the notion of a subgaussian sampler, which he defined as an averaging sampler for functions $f : \{0, 1\}^m \to \mathbb{R}$ such that $f(U_m)$ is a subgaussian random variable. Since subgaussian random variables have strong tail bounds, subgaussian functions from $\{0, 1\}^m$ have a range of size $O(\sqrt{m})$, and thus one can construct a subgaussian sampler from a $[0, 1]$-sampler by simply scaling the error $\epsilon$ by a factor of $O(\sqrt{m})$. Unfortunately, looking at Table 1 one sees that this induces a multiplicative dependence on $m$ in the sample complexity, and for the expander-walk sampler induces a dependence of $m \log(1/\delta)$ in the randomness complexity. This loss can be avoided for some samplers, such as the sampler of Chor and Goldreich [CG89] based on pairwise independence (as its analysis requires only bounded variance), but Blasiok showed [Bla18b] that the expander-walk sampler does not in general act as a subgaussian sampler without reducing the error to $o(1)$. We remark briefly that the median-of-averages sampler of Bellare, Goldreich, and Goldwasser [BGG93] still works and is optimal up to constant factors in the subgaussian setting (since the underlying pairwise independent sampler works), but it is not an averaging sampler and matching its parameters with an averaging sampler remains open in general even for $[0, 1]$-valued functions.

One of the contributions of this work is to give explicit averaging samplers for subgaussian functions (in fact even for subexponential functions that satisfy weaker tail bounds) matching the extractor-based samplers for $[0, 1]$-valued functions in Table 1 (up to the hidden polynomial in the sample complexity). This achieves the best parameters currently known in the regime of parameters where $\epsilon$ and $\delta$ are both subconstant, and in particular has no dependence on $m$ in the sample complexity. We also show non-constructively that subgaussian samplers exist with essentially the same parameters as $[0, 1]$-valued samplers.

**Theorem 1.2** (Informal version of Theorem 6.1 and Corollary 6.5). For every integer $m \in \mathbb{N}$, $1 > \delta, \epsilon > 0$, and $\alpha > 0$, there is an explicit subgaussian (in fact subexponential) sampler $\text{Samp} : \{0, 1\}^n \to ([0, 1]^m)^D$ with randomness complexity $n = m + (1 + \alpha) \cdot \log(1/\delta)$ and sample complexity $D = \text{poly}(\log(1/\delta), 1/\epsilon)$. Furthermore, there is a non-explicit subgaussian sampler with $n = m + \log(1/\delta) - \log \log(1/\delta) + O(1)$ and $D = O(\log(1/\delta)/\epsilon^2)$.

### 1.2 Randomness extractors

To prove Theorem 1.2, we develop a corresponding theory of generalized randomness extractors which we believe is of independent interest. For bounded functions, Zuckerman [Zuc97] showed that averaging samplers

---

1. A non-averaging sampler is an algorithm $\text{Samp}$ which makes oracle queries to $f$ and outputs an estimate of its average which is good with high probability, but need not simply output the average of $f$’s values on the queried points.
are essentially equivalent to randomness extractors, and in fact several of the best-known constructions of such samplers arose as extractor constructions. Formally, a randomness extractor is defined as follows:

**Definition 1.3** (Nisan and Zuckerman [NZ90]). A function \( \text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m \) is said to be a \((k,\varepsilon)\) extractor if for every distribution \( X \) over \( \{0,1\}^m \) satisfying \( \max_{x \in \{0,1\}^n} \Pr[X = x] \leq 2^{-k} \), the distributions \( \text{Ext}(X, U_d) \) and \( U_m \) are \( \varepsilon \)-close in total variation distance. Equivalently, for all \( f : \{0,1\}^m \rightarrow [0,1] \) it holds that \( \mathbb{E}[f(\text{Ext}(X, U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon \). The number \( d \) is called the seed length, and \( m \) the output length.

The formulation of Definition 1.3 in terms of \([0,1]\)-valued functions implies that extractors produce an output distribution that is indistinguishable from uniform by all bounded functions \( f \). It is therefore natural to consider a variant of this definition for a different set \( \mathcal{F} \) of test functions \( f : \{0,1\}^m \rightarrow \mathbb{R} \) which need not be bounded.

**Definition 1.4** (Special case of Definition 1.3 using Definition 2.5). A function \( \text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m \) is said to be a \((k,\varepsilon)\) extractor for a set of real-valued functions \( \mathcal{F} \) from \([0,1]^{m}\) if for every distribution \( X \) over \([0,1]^{m}\) satisfying \( \max_{x \in \{0,1\}^n} \Pr[X = x] \leq 2^{-k} \) and every \( f \in \mathcal{F} \), it holds that \( \mathbb{E}[f(\text{Ext}(X, U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon \).

We show that much of the theory of extractors and samplers carries over to this more general setting. In particular, we generalize the connection of Zuckerman [Zuc97] to show that extractors are samplers for any class of functions, along with the converse (though as for total variation distance, there is some loss of parameters in this direction). Thus, to construct a subgaussian sampler it suffices (and is preferable) to construct a corresponding extractor for subgaussian test functions, which is how we prove Theorem 1.2.

Unfortunately, the distance induced by subgaussian test functions is not particularly pleasant to work with: for example the point masses on 0 and 1 in \([0,1]\) are \( O(1) \) apart, but embedding them in the larger universe \([0,1]^{m}\) leads to distributions which are \( \Theta(\sqrt{m}) \) apart. We solve this problem by constructing extractors for a stronger notion, the **Kullback–Leibler (KL)**-divergence, equivalently, extractors whose output is required to have very high Shannon entropy.

**Definition 1.5** (Special case of Definition 3.1 using KL divergence). A function \( \text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m \) is said to be a \((k,\varepsilon)\) KL-extractor if for every distribution \( X \) over \([0,1]^{m}\) satisfying \( \max_{x \in \{0,1\}^n} \Pr[X = x] \leq 2^{-k} \) it holds that \( \text{KL}(\text{Ext}(X, U_d) \parallel U_m) \leq \varepsilon \), or equivalently \( \text{H}(\text{Ext}(X, U_d)) \geq m - \varepsilon \).

A strong form of Pinsker’s inequality (e.g. [BLM13, Lemma 4.18]) implies that a \((k,\varepsilon^2)\) KL-extractor is also a \((k,\varepsilon)\) extractor for subgaussian test functions. The KL divergence has the advantage that is nonincreasing under the application of functions (the famous data-processing inequality), and although it does not satisfy a traditional triangle inequality, it does satisfy a similar inequality when one of the segments satisfies stronger \( L_2 \) bounds. These properties allow us to use composition techniques from the literature due to Goldreich and Wigderson [GW97] and Reingold, Wigderson, and Vadhan [RVW00] to construct KL-extractors with seed length depending on \( n \) and \( k \) only through the entropy deficiency \( n - k \) of \( X \) rather than \( n \) itself, which in the sampler perspective corresponds to a sampler with sample complexity depending on the failure probability \( \delta \) rather than the universe size \( 2^m \). Hence, we prove Theorem 1.2 by constructing corresponding KL-extractors.

**Theorem 1.6** (Informal version of Theorem 6.2). For all integers \( m \), \( 1 > \delta, \varepsilon > 0 \), and \( \alpha > 0 \) there is an explicit \((k,\varepsilon)\) KL-extractor \( \text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m \) with \( n = m + (1 + \alpha) \cdot \log(1/\delta) \), \( k = n - \log(1/\delta) \), and \( d = O(\log(1/\delta) + \log(1/\varepsilon)) \).

Though the above theorem is most interesting in the high min-entropy regime where \( n - k = o(n) \), we also show the existence of KL-extractors matching most of the existing constructions of total variation extractors. In particular, we note that extractors for \( L_2 \) are immediately KL-extractors without loss of parameters, and also that any extractor can be made a KL-extractor by taking slightly smaller error, so that the extractors of Guruswami, Umans, and Vadhan [GUU09] can be taken to be KL-extractors with essentially the same parameters.

Furthermore, in addition to our explicit constructions, we also show non-constructively that KL-extractors (and hence subgaussian extractors) exist with very good parameters:
Theorem 1.7 (Informal version of Theorem 5.31). For any integers $k < n \in \mathbb{N}$ and $1 > \varepsilon > 0$ there is a $(k, \varepsilon)$ KL-extractor $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m$ with $d = \log(n - k) + \log(1/\varepsilon) + O(1)$ and $m = k + d - \log(1/\varepsilon) - O(1)$.

One key thing to note about the nonconstructive KL extractors of the above theorem is that they incur an entropy loss of only $1 \cdot \log(1/\varepsilon)$, whereas total variation extractors necessarily incur entropy loss $2 \cdot \log(1/\varepsilon)$ by the lower bound of Radhakrishnan and Ta-Shma [RT00]. In particular, by Pinsker’s inequality, $(k, \varepsilon^2)$ KL-extractors with the above parameters are also optimal $(k, \varepsilon)$ standard (total variation) extractors [RT00], so that one does not lose anything by constructing a KL-extractor rather than a total variation extractor. We also remark that the above theorem gives subgaussian samplers with better parameters than a naive

Theorem 1.7 (Informal version of Theorem 5.31).

The other alternative is to use the generic reduction of Raz, Reingold, and Vadhan [RRV02] which turns variation extractors: for example it seems likely that there are applications (perhaps in coding or cryptography, c.f. [BDK+11]) where it is more important to have high Shannon entropy in the output than small total variation distance to uniform, in which case one may be able to use $(k, \varepsilon)$ KL-extractors with entropy loss only $1 \cdot \log(1/\varepsilon)$ directly, rather than a total variation extractor or $(k, \varepsilon^2)$ KL-extractor with entropy loss $2 \cdot \log(1/\varepsilon)$.

1.3 Future directions

Broadly speaking, we hope that the perspective of KL-extractors will bring new tools (perhaps from information theory) to the construction of extractors and samplers. For example, since KL-extractors can have seed length with dependence on $\varepsilon$ of only $1 \cdot \log(1/\varepsilon)$, trying to explicitly construct a KL-extractor with seed length $1 \cdot \log(1/\varepsilon) + o(\min(n,k))$ may also shed light on how to achieve optimal dependence on $\varepsilon$ in the total variation setting.

In the regime of constant $\varepsilon = \Omega(1)$, we do not have explicit constructions of subgaussian samplers matching the expander-walk sampler of Gillman [Gil98] for $[0,1]$-valued functions, which achieves randomness complexity $m + O(\log(1/\delta))$ and sample complexity $O(\log(1/\delta))$. From the extractor point-of-view, it would suffice (by the reduction of [GW97, RVW00] that we analyze for KL-extractors) to construct explicit linear degree KL-extractors with parameters matching the linear degree extractor of Zuckerman [Zuc07], i.e. with seed length $d = \log(n) + O(1)$ and $m = \Omega(k)$ for $\varepsilon = \Omega(1)$. A potentially easier problem, since the Zuckerman linear degree extractor is itself based on the expander-walk sampler, could be to instead match the parameters of the near-linear degree extractors of Ta-Shma, Zuckerman, and Safra [TZS06] based on Reed–Muller codes, thereby achieving sample complexity $O(\log(1/\delta) \cdot \text{poly log log}(1/\delta))$.

Finally, we hope that KL-extractors can also find uses beyond being subgaussian samplers and total variation extractors: for example it seems likely that there are applications (perhaps in coding or cryptography, c.f. [BDK+11]) where it is more important to have high Shannon entropy in the output than small total variation distance to uniform, in which case one may be able to use $(k, \varepsilon)$ KL-extractors with entropy loss only $1 \cdot \log(1/\varepsilon)$ directly, rather than a total variation extractor or $(k, \varepsilon^2)$ KL-extractor with entropy loss $2 \cdot \log(1/\varepsilon)$.
2 Preliminaries

2.1 (Weak) statistical divergences and metrics

Our results in general will require very few assumptions on notions of “distance” between probability distributions, so we will give a general definition and indicate in our theorems when we need which assumptions.

Definition 2.1. A weak statistical divergence (or simply weak divergence) on a finite set $\mathcal{X}$ is a function $D$ from pairs of probability distributions over $\mathcal{X}$ to $\mathbb{R} \cup \{\pm \infty\}$. We write $D(P \parallel Q)$ for the value of $D$ on distributions $P$ and $Q$. Furthermore

1. If $D(P \parallel Q) \geq 0$ with equality iff $P = Q$, then $D$ is positive-definite, and we simply call $D$ a divergence.
2. If $D(P \parallel Q) = D(Q \parallel P)$, then $D$ is symmetric.
3. If $D(P \parallel R) \leq D(P \parallel Q) + D(Q \parallel R)$, then $D$ satisfies the triangle inequality.
4. If $D(\lambda P_1 + (1 - \lambda)P_2 \parallel \lambda Q_1 + (1 - \lambda)Q_2) \leq \lambda D(P_1 \parallel Q_1) + (1 - \lambda)D(P_2 \parallel Q_2)$ for all $\lambda \in [0, 1]$, then $D$ is jointly convex. If this holds only when $Q_1 = Q_2$ then $D$ is convex in its first argument.
5. If $D$ is defined on all finite sets $\mathcal{Y}$ and for all functions $f : \mathcal{X} \to \mathcal{Y}$ the divergence is nonincreasing under $f$, that is $D(f(P) \parallel f(Q)) \leq D(P \parallel Q)$, then $D$ satisfies the data-processing inequality.

If $D$ is positive-definite, symmetric, and satisfies the triangle inequality, then it is called a metric.

Example 2.2. The $\ell_p$ distance for $p > 0$ between probability distributions over $\mathcal{X}$ is

$$d_{\ell_p}(P, Q) \overset{\text{def}}{=} \left( \sum_{x \in \mathcal{X}} |P_x - Q_x|^p \right)^{1/p}$$

and is positive-definite and symmetric. Furthermore, for $p \geq 1$ it satisfies the triangle inequality (and so is a metric), and is jointly convex. The $\ell_p$ distance is nonincreasing in $p$.

Example 2.3. The total variation distance is

$$d_{TV}(P, Q) \overset{\text{def}}{=} \frac{1}{2} d_1(P, Q) = \sup_{S \subseteq \mathcal{X}} \left| \Pr[P \in S] - \Pr[Q \in S] \right| = \sup_{f \in [0, 1]^\mathcal{X}} \left( \mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \right)$$

and is a jointly convex metric that satisfies the data-processing inequality.

Example 2.4 (Rényi Divergences [Ren61]). For two probability distributions $P$ and $Q$ over a finite set $\mathcal{X}$, the Rényi $\alpha$-divergence or Rényi divergence of order $\alpha$ is defined for real $0 < \alpha \neq 1$ by

$$D_\alpha(P \parallel Q) \overset{\text{def}}{=} \frac{1}{\alpha - 1} \log \left( \sum_{x \in \mathcal{X}} P_x^{\alpha} / Q_x^{\alpha - 1} \right)$$

where the logarithm is in base 2 (as are all logarithms in this paper unless noted otherwise). The Rényi divergence is continuous in $\alpha$ and so is defined by taking limits for $\alpha \in (0, 1, \infty)$, giving for $\alpha = 0$$D_0(P \parallel Q) \overset{\text{def}}{=} \log(1/\Pr_{x \sim Q}[P_x \neq 0])$, for $\alpha = 1$ the Kullback–Leibler (or KL) divergence

$$\text{KL}(P \parallel Q) \overset{\text{def}}{=} D_1(P \parallel Q) = \sum_{x \in \mathcal{X}} P_x \log \frac{P_x}{Q_x},$$

and for $\alpha = \infty$ the max-divergence $D_\infty(P \parallel Q) \overset{\text{def}}{=} \max_{x \in \mathcal{X}} \log \frac{P_x}{Q_x}$. The Rényi divergence is nondecreasing in $\alpha$. Furthermore, when $\alpha \leq 1$ the Rényi divergence is jointly convex, and for all $\alpha$ the Rényi divergence satisfies the data-processing inequality [VEHI].
When \( Q = U_\mathcal{X} \) is the uniform distribution over the set \( \mathcal{X} \), then for all \( \alpha \), \( \text{D}_\alpha(P \parallel U_\mathcal{X}) = \log|\mathcal{X}| - H_\alpha(P) \) where \( 0 \leq H_\alpha(P) \leq \log|\mathcal{X}| \) is called the Rényi \( \alpha \)-entropy of \( P \). For \( \alpha = 0 \), \( H_0(P) = \log|\text{Supp}(P)| \) is the max-entropy of \( P \), for \( \alpha = 1 \), \( H_1(P) = \sum_{x \in \mathcal{X}} P_x \log(1/P_x) \) is the Shannon entropy of \( P \), and for \( \alpha = \infty \), \( H_\infty(P) = \min_{x \in \mathcal{X}} \log(1/P_x) \) is the min-entropy of \( P \).

For \( \alpha = 2 \), the Rényi 2-entropy can be expressed in terms of the \( t_2 \)-distance to uniform:

\[
\log|\mathcal{X}| - H_2(P) = \text{D}_2(P \parallel U_\mathcal{X}) = \log(1 + |\mathcal{X}| \cdot d_{t_2}(P, U_\mathcal{X})^2)
\]

### 2.2 Statistical weak divergences from test functions

Zuckerman’s connection [Zuc97] between samplers for bounded functions and extractors for total variation distance is based on the following standard characterization of total variation distance as the maximum distinguishing advantage achieved by bounded functions,

\[
d_{TV}(P, Q) = \sup_{f \in [0,1]^\mathcal{X}} \mathbb{E}[f(P)] - \mathbb{E}[f(Q)].
\]

By considering an arbitrary class of functions in the supremum, we get the following weak divergence:

**Definition 2.5.** Given a finite \( \mathcal{X} \) and a set of real-valued functions \( \mathcal{F} \subseteq \mathbb{R}^\mathcal{X} \), the \( \mathcal{F} \)-distance on \( \mathcal{X} \) between probability measures on \( \mathcal{X} \) is denoted by \( \text{D}^\mathcal{F} \) and is defined as

\[
\text{D}^\mathcal{F}(P \parallel Q) = \sup_{f \in \mathcal{F}} (\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]) = \sup_{f \in \mathcal{F}} \text{D}^{(f)}(P \parallel Q),
\]

where we use a superscript to avoid confusion with the Csiszár-Morimoto-Ali-Silvey \( f \)-divergences [CS63, Mor62, AS66].

We call the set of functions \( \mathcal{F} \) symmetric if for all \( f \in \mathcal{F} \) there is \( c \in \mathbb{R} \) and \( g \in \mathcal{F} \) such that \( g = c - f \), and distinguishing if for all \( P \neq Q \) there exists \( f \in \mathcal{F} \) with \( \text{D}^{(f)}(P \parallel Q) > 0 \).

**Example 2.6.** If \( \mathcal{F} = \{0,1\}^\mathcal{X} \) or \( \mathcal{F} = [0,1]^\mathcal{X} \), then \( \text{D}^\mathcal{F} \) is exactly the total variation distance.

**Remark 2.7.** An equivalent definition of \( \mathcal{F} \) being symmetric is that for all \( f \in \mathcal{F} \) there exists \( g \in \mathcal{F} \) with \( \text{D}^{(g)}(P \parallel Q) = -\text{D}^{(f)}(P \parallel Q) = \text{D}^{(f)}(Q \parallel P) \) for all distributions \( P \) and \( Q \). Hence, one might also consider a weaker notion of symmetry that reverses quantifiers, where \( \mathcal{F} \) is “weakly-symmetric” if for all \( f \in \mathcal{F} \) and distributions \( P \) and \( Q \) there exists \( g \in \mathcal{F} \) such that \( \text{D}^{(g)}(P \parallel Q) = -\text{D}^{(f)}(P \parallel Q) = \text{D}^{(f)}(Q \parallel P) \). However, such a class \( \mathcal{F} \) gives exactly the same weak divergence \( \text{D}^\mathcal{F} \) as its “symmetrization” \( \overline{\mathcal{F}} = \mathcal{F} \cup \{-f \mid f \in \mathcal{F}\} \), so we do not need to introduce this more complex notion.

**Remark 2.8.** By identifying distributions with their probability mass function, one can realize \( \mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \) as an inner product \( \langle P - Q, f \rangle \). Definition 2.5 can thus be written as \( \text{D}^\mathcal{F}(P \parallel Q) = \sup_{f \in \mathcal{F}} \langle P - Q, f \rangle \), which is essentially the notion of indistinguishability considered in several prior works, (see e.g. the survey of Reingold, Trevisan, Tulsiani, and Vadhan [RTTV08]), but without requiring all \( f \) to be bounded.

**Remark 2.9.** For simplicity, all our probabilistic distributions are given only for random variables and distributions over finite sets as this is all we need for our application. A more general version of Definition 2.5 has been studied by e.g. Zolotarev [Zol84] and Müller [Mül97] and is commonly used in developments of Stein’s method in probability.

We now establish some basic properties of \( \text{D}^\mathcal{F} \).

**Lemma 2.10.** Let \( \mathcal{F} \subseteq \mathbb{R}^\mathcal{X} \) be a set of real-valued functions over a finite set \( \mathcal{X} \). Then \( \text{D}^\mathcal{F} \) satisfies the triangle inequality and is jointly convex, and

1. if \( \mathcal{F} \) is symmetric then \( \text{D}^\mathcal{F} \) is symmetric and

\[
\text{D}^\mathcal{F}(P \parallel Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]| \geq 0,
\]
2. if $\mathcal{F}$ is distinguishing then $D^\mathcal{F}$ is positive-definite,

so that if $\mathcal{F}$ is both symmetric and distinguishing then $D^\mathcal{F}$ is a jointly convex metric on probability distributions over $X$, in which case we also use the notation $d_{\mathcal{F}}(P, Q) \equiv D^\mathcal{F}(P \parallel Q)$.

Proof. The triangle inequality and joint convexity both follow from the linearity of each $D^{(f)}$, as by linearity of expectation, for all $f : X \to \mathbb{R}$ it holds that

$$D^{(f)}(P \parallel R) = D^{(f)}(P \parallel Q) + D^{(f)}(Q \parallel R)$$

$$D^{(f)}(\lambda P_1 + (1 - \lambda) P_2 \parallel \lambda Q_1 + (1 - \lambda) Q_2) = \lambda D^{(f)}(P_1 \parallel Q_1) + (1 - \lambda) D^{(f)}(P_2 \parallel Q_2).$$

Upper bounding the terms on the right-hand side by $D^\mathcal{F}$ and taking the supremum of the left hand side over $f \in \mathcal{F}$ then gives the claims. The symmetry and positive-definite claims are immediate from the definitions.

Furthermore, the notion of dual norm has an appealing interpretation in this framework via Remark 2.8, generalizing the fact that total variation distance corresponds to $[0, 1]$-valued test functions (or equivalently that $\ell_1$ distance corresponds to to $[-1, 1]$-valued functions).

**Proposition 2.11.** Let $1 \leq p, q \leq \infty$ be Hölder conjugates (meaning $1/p + 1/q = 1$), and let

$$\mathcal{M}_q \overset{def}{=} \left\{ f : \{0, 1\}^m \to \mathbb{R} \left| \|f(U_m)\|_q \overset{def}{=} \mathbb{E}[|f(U_m)|^q]^{1/q} \leq 1 \right. \right\}$$

be the set of real-valued functions from $\{0, 1\}^m$ with bounded $q$-th moments. Then $d_{\ell_p} = 2^{-m/q} \cdot d_{\mathcal{M}_q}$, in the sense that for all probability distributions $A$ and $B$ over $\{0, 1\}^m$ it holds that $d_{\ell_p}(A, B) = 2^{-m/q} \cdot d_{\mathcal{M}_q}(A, B)$.

In particular, taking $p = 1$ and $q = \infty$ recovers the result for $\ell_1$ (equivalently total variation) distance.

Proof. As mentioned this is just the standard fact that the $\ell_p$ and $\ell_q$ norms are dual, but for completeness we include a proof in our language using the extremal form of Hölder’s inequality (note that since we are dealing with finite probability spaces the extremal equality holds even for $p = \infty$ and $q = 1$). Given probability distributions $A$ and $B$ over $\{0, 1\}^m$, we have that

$$d_{\ell_p}(A, B) = \left( \sum_x |A_x - B_x|^p \right)^{1/p}$$

$$= 2^{m/p} \mathbb{E}_{x \sim U_m} \left|A_x - B_x\right|^p \left(\mathbb{E}_{x \sim U_m} \left[f(x)(A_x - B_x)\right] \right) \left(\text{Hölder’s extremal equality}\right)$$

$$= 2^{m/p} \max_{f : \{0, 1\}^m \to \mathbb{R} \left| \mathbb{E}_{x \sim U_m} \left[f(x)(A_x - B_x)\right] \right| \leq 1} \mathbb{E}_{x \sim U_m} \left[f(x)(A_x - B_x)\right] \left(\text{by symmetry of } \mathcal{M}_q\right)$$

as desired.

3 Extractors for weak divergences and connections to samplers

3.1 Definitions

We now use this machinery to extend the notion of an extractor due to Nisan and Zuckerman [NZ96] and the average-case variant of Dodis, Ostrovsky, Reyzin, and Smith [DORS08].
Definition 3.1 (Extends Definition 1.4). Let D be a weak divergence on the set \(\{0,1\}^m\), and Ext : \(\{0,1\}^m \times \{0,1\}^d \rightarrow \{0,1\}^m\). Then if for all distributions X over \(\{0,1\}^m\) with \(\tilde{H}_\infty(X) \geq k\) it holds that

1. \(D(\text{Ext}(X, U_d) \mid U_m) \leq \epsilon\), then Ext is said to be a \((k, \epsilon)\) extractor for D, or a \((k, \epsilon)\) D-extractor.

2. \(\mathbb{E}_{x \sim U_d}[D(\text{Ext}(X, s) \mid U_m)] \leq \epsilon\), then Ext is said to be a \((k, \epsilon)\) strong extractor for D, or a \((k, \epsilon)\) strong D-extractor.

Furthermore, if for all joint distributions \((Z, X)\) where X is distributed over \(\{0,1\}^m\) with \(\tilde{H}_\infty(X|Z) \leq k\), it holds that

3. \(\mathbb{E}_{z \sim Z}[D(\text{Ext}(X|Z=z, U_d) \mid U_m) \leq \epsilon]\), then Ext is said to be a \((k, \epsilon)\) average-case extractor for D, or a \((k, \epsilon)\) average-case D-extractor.

4. \(\mathbb{E}_{z \sim Z, s \sim U_d}[D(\text{Ext}(X|Z=z, s) \mid U_m) \leq \epsilon]\), then Ext is said to be a \((k, \epsilon)\) average-case strong extractor for D, or a \((k, \epsilon)\) average-case strong D-extractor.

Remark 3.2. By taking D to be the total variation distance we recover the standard definitions of extractor and strong extractor due to [NZ96] and the definition of average-case extractor due to [DORS08].

However, our definitions are phrased slightly differently for strong and average-case extractors as an expectation rather than a joint distance, that is, for strong average-case extractors we require a bound on the expectation \(\mathbb{E}_{z \sim Z, s \sim U_d}[D(\text{Ext}(X|Z=z, s) \mid U_m)]\) rather than a bound on \(D(Z, U_d, \text{Ext}(X, U_d) \mid Z, U_d, U_m)\). In our setting, the weak divergence D need not be defined over the larger joint universe, but it is defined for all random variables over \(\{0,1\}^m\). In the case of \(d_{TV}\) and KL divergence, both definitions are equivalent (for KL divergence, this is an instance of the chain rule).

Remark 3.3. The strong variants of Definition 3.1 are also non-strong extractors assuming the weak divergence D is convex in its first argument, as it is for most weak divergences of interest, including the \(\ell_p\) norms for \(p \geq 1\), all \(D^F\) defined by test functions, the KL divergence, Rényi divergences for \(\alpha \leq 1\), and all Csiszár-Morimoto-Ali-Silvey \(f\)-divergences. The average-case variants are always non-average-case extractors by taking Z to be independent of X.

Remark 3.4. We gave Definition 3.1 for general weak divergences which need not be symmetric, and made the particular choice that the output of the extractor was on the left-hand side of the weak divergence and that the uniform distribution was on the right-hand side. This is motivated by the standard information-theoretic divergences such as KL divergence, which require the left-hand distribution to have support contained in the support of the right-hand distribution, and putting the uniform distribution on the right ensures this is always the case. Furthermore, the KL divergence to uniform has a natural interpretation as an entropy difference, \(\text{KL}(P \mid U_n) = m - H(P)\) for H the Shannon entropy, so that in particular a KL extractor with error \(\epsilon\) requires the output to have Shannon entropy at least \(m - \epsilon\). If for a weak divergence D the other direction is more natural, one can always reverse the sides by considering the weak divergence \(D'(Q \mid P) = D(P \mid Q)\).

Remark 3.5. Definition 3.1 does not technically need even a weak divergence, as it suffices to simply have a measure of distance to uniform. However, since weak divergences have minimal constraints, one can define a weak divergence from any distance to uniform by ignoring the second component (or setting it to be infinite for non-uniform distributions).

We also give the natural definition of averaging samplers for arbitrary classes of functions \(\mathcal{F}\) extending Definition 1.1 along with the strong variant of Zuckerman [Zuc97].

Definition 3.6. Given a class of functions \(\mathcal{F} : \{0,1\}^m \rightarrow \mathbb{R}\), a function \(\text{Samp} : \{0,1\}^n \rightarrow (\{0,1\}^m)^D\) is said to be a \((\delta, \epsilon)\) strong averaging sampler for \(\mathcal{F}\) or a \((\delta, \epsilon)\) strong averaging \(\mathcal{F}\)-sampler if for all \(f \in \mathcal{F}\), it holds that

\[
\Pr_{x \sim U_n} \left[ \frac{\mathbb{E}_{i \sim U[D]}[f_i(\text{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]] > \epsilon}{\delta} \right] \leq \delta
\]
Then the function extractor for the weak divergence $3.7$

We now show that Zuckerman’s connection [Zuc97] does indeed generalize to this broader setting as promised.

Proof. The proof is essentially the same as that of [Zuc97].

Fix a collection of test functions $f_1, \ldots, f_D \in \mathcal{F}$, where if Ext is not strong we restrict to $f_1 = \cdots = f_D$, and let $B_{f_1, \ldots, f_D} \subseteq \{0, 1\}^n$ be defined as

$$B_{f_1, \ldots, f_D} \overset{\text{def}}{=} \left\{ x \in \{0, 1\}^n \middle| \sum_{i \sim U_{\{D\}}} \left[ f_i(\text{Ext}(x, i)) - E[f_i(U_m)] \right] > \varepsilon \right\}$$

where $U_{\{z\}}$ is the point mass on $z$. Then if $X$ is uniform over $B_{f_1, \ldots, f_D}$, we have

$$\varepsilon < \sum_{x \sim X} \left[ \sum_{i \sim U_{\{D\}}} \left[ f_i(\text{Ext}(x, i)) - E[f_i(U_m)] \right] \right] = \sum_{i \sim U_{\{D\}}} D(f_i)(\text{Ext}(X, i) \parallel U_m) \leq D^\mathcal{F}(\text{Ext}(X, U_d) \parallel U_m)$$

where $D(f_i)(\text{Ext}(X, i) \parallel U_m)$ always.

Since Ext is a $(n - \log(1/\delta), \varepsilon)$-extractor (respectively strong extractor) for $D^\mathcal{F}$ we must have $H_\infty(X) < n - \log(1/\delta)$. But $H_\infty(X) = \log |B_{f_1, \ldots, f_D}|$ by definition, so we have $|B_{f_1, \ldots, f_D}| < \delta^2 n$. Hence, the probability that a random $x \in \{0, 1\}^n$ lands in $B_{f_1, \ldots, f_D}$ is less than $\delta$, and since $B_{f_1, \ldots, f_D}$ is exactly the set of seeds which are bad for Samp, this concludes the proof.

Remark 3.9. Hölder’s inequality implies that an extractor for $\ell_p$, with error $\varepsilon \cdot 2^{-m(p-1)/p}$ is also an $\ell_1$ extractor and thus $[-1, 1]$-averaging sampler with error $\varepsilon$. Proposition 2.11 and Theorem 3.8 show that they are in fact samplers for the much larger class of functions $\mathcal{M}_{p/(p-1)}$ with bounded $p/(p-1)$ moments (rather than just $\infty$ moments), also with error $\varepsilon$.  

9
Furthermore, if all the functions in $F$ have bounded deviation from their mean (for example, subgaussian functions from $f: \{0,1\}^m \to \mathbb{R}$ have such a bound of $O(\sqrt{m})$ by the tail bounds from Lemma 4.3), then we also have a partial converse that recovers the standard converse in the case of total variation distance.

**Theorem 3.10.** Let $F$ be a class of functions $F \subset \{0,1\}^m \to \mathbb{R}$ with finite maximum deviation from the mean, meaning $\max_{f \in F} \max_{x \in \{0,1\}^m} |f(x) - \mathbb{E}[f(U_m)]| < \infty$. Then given a $(\delta, \varepsilon)$ $F$-sampler (respectively $(\delta, \varepsilon)$ strong $F$-sampler) $\text{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$, the function $\text{Ext} : \{0,1\}^n \times \{0,1\}^D \to \{0,1\}^m$ for $d = \log D$ defined by $\text{Ext}(x,i) = \text{Samp}(x)_i$ is a $(k, \varepsilon + \delta \cdot 2^n - \max \text{dev}(F))$ $D^F$-extractor (respectively strong $D^F$-extractor) for every $0 \leq k \leq n$.

In particular, $\text{Ext}$ is an $(n - \log(1/\delta) + \log(1/\eta), \varepsilon + \eta \cdot \max \text{dev}(F))$ average-case $D^F$-extractor (respectively strong average-case $D^F$-extractor) for every $\delta \leq \eta \leq 1$.

**Proof.** Again the proof is analogous to the one in [Zuc97].

Fix a distribution $X$ over $\{0,1\}^m$ with $H_\infty(X) \geq k$ and a collection of test functions $f_1, \ldots, f_D \in F$, where if $\text{Samp}$ is not strong we restrict to $f_1 = \cdots = f_D$. Then since $\text{Samp}$ is a $(\delta, \varepsilon)$ $F$-sampler, we know that the set of seeds for which the sampler is bad must be small. Formally, the set

$$B_{f_1,\ldots,f_D} = \left\{ x \in \{0,1\}^n \mid \mathbb{E}_{i \sim U_d} \left[ f_i(\text{Samp}(x)_i) - \mathbb{E}[f_i(U_m)] \right] > \varepsilon \right\}$$

has size $|B_{f_1,\ldots,f_D}| \leq 2^n$. Thus, since $X$ has min-entropy at least $k$ we know $\Pr[X \in B_{f_1,\ldots,f_D}] \leq 2^{-k} \cdot 2^n$, so we have

$$\mathbb{E}_{i \sim U_d} \left[ f_i(\text{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \right]$$

$$= \mathbb{E}_X \left[ \mathbb{E}_{i \sim U_d} \left[ f_i(\text{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \right] \right]$$

$$= \Pr[X \in B_{f_1,\ldots,f_D}] \cdot \mathbb{E}_X \left[ \mathbb{E}_{i \sim U_d} \left[ f_i(\text{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \right] \mid X \in B_{f_1,\ldots,f_D} \right]$$

$$+ \Pr[X \notin B_{f_1,\ldots,f_D}] \cdot \mathbb{E}_X \left[ \mathbb{E}_{i \sim U_d} \left[ f_i(\text{Ext}(X,i)) - \mathbb{E}[f_i(U_m)] \right] \mid X \notin B_{f_1,\ldots,f_D} \right]$$

$$\leq \Pr[X \in B_{f_1,\ldots,f_D}] \cdot \max \text{dev}(F) + \Pr[X \notin B_{f_1,\ldots,f_D}] \cdot \varepsilon$$

$$\leq 2^{-k} \cdot 2^n \cdot \max \text{dev}(F) + \varepsilon$$

completing the proof of the main claim. The “in particular” statement follows since if $(Z,X)$ are jointly distributed with $H_\infty(X|Z) \geq n - \log(1/\delta) + \log(1/\eta)$ we have

$$\mathbb{E}_{z \sim Z} \left[ \varepsilon + \delta \cdot 2^n - H_\infty(X|Z) \cdot \max \text{dev}(F) \right] = \varepsilon + \delta \cdot 2^n - H_\infty(X|Z) \cdot \max \text{dev}(F) \leq \varepsilon + \eta \cdot \max \text{dev}(F)$$

by definition of conditional min-entropy. □

### 3.3 All extractors are average-case

Under a similar boundedness condition for general weak divergences, we can recover the standard fact that all extractors are average-case extractors under a slight loss of parameters (the same loss as achieved by Dodis, Ostrovsky, Reyzin, and Smith [DORS08] for the case of total variation distance). Most interestingly, if the weak divergence is given by $D^F$ for a symmetric class of (possibly unbounded) functions $F$, we can also generalize and recover the result of Vadhan [Vad12] Problem 6.8 that shows that a $(k, \varepsilon)$ extractor (for total variation) is a $(k, 3\varepsilon)$ average-case extractor without any other loss.
Theorem 3.11. Let $D$ be a bounded weak divergence over $\{0,1\}^m$, meaning that
\[ 0 \leq \|D\|_\infty \overset{\text{def}}{=} \sup_{P \text{ on } \{0,1\}^m} D(P \parallel U_m) < \infty. \]

Then a $(k,\varepsilon)$-extractor for $D$ (respectively strong extractor) $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is also a $(k + \log(1/\eta),\varepsilon + \eta \cdot \|D\|_\infty)$ average-case extractor for $D$ (respectively strong average-case extractor) for any $0 < \eta \leq 1$.

**Proof.** The proof is analogous to that of [DORS08]. We prove it only for non-strong extractors, the proof for strong extractors is completely analogous by adding more expectations.

For jointly distributed random variables $(Z,X)$ such that $\hat{H}_\infty(X|Z) \geq k + \log(1/\eta)$, we have by Lemma 2.2 that the probability that $\Pr_{z \sim Z}[\hat{H}_\infty(X|Z = z) < k] \leq \eta$. Thus
\[ E_{z \sim Z} [D(\text{Ext}(X|Z = z,U_d) \parallel U_m)] = \Pr_{z \sim Z}[\hat{H}_\infty(X|Z = z) < k] \cdot E_{z \sim Z} [D(\text{Ext}(X|Z = z,U_d) \parallel U_m)] + \Pr_{z \sim Z}[\hat{H}_\infty(X|Z = z) \geq k] \cdot E_{z \sim Z} [D(\text{Ext}(X|Z = z,U_d) \parallel U_m)] \hat{H}_\infty(X|Z = z) \geq k]
\leq \eta \cdot \|D\|_\infty + 1 \cdot \varepsilon \]

Theorem 3.12. Let $\mathcal{F}$ be a symmetric class of test functions and $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be a $(k,\varepsilon)$ extractor (respectively strong extractor) for $D^F$, where $k$ is at most $n - 1$. Then $\text{Ext}$ is an $(k,3\varepsilon)$ average-case extractor (respectively strong average-case extractor) for $D^F$.

The proof of Theorem 3.12 follows the strategy outlined by Vadhan [Vad12, Problem 6.8]. We first isolate the following key lemma which shows that any extractor with error that gracefully decays with lower min-entropy is average-case with minimal loss of parameters, as opposed to Theorem 3.11 which used a worst-case error bound when the min-entropy is low.

**Lemma 3.13.** Let $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be a $(k,\varepsilon)$ extractor (respectively strong extractor) for $D$ such that for every $0 \leq t \leq k$, $\text{Ext}$ is also a $(k - t,2^{k+1} \cdot \varepsilon)$ extractor (respectively strong extractor) for $D$. Then $\text{Ext}$ is a $(k,3\varepsilon)$ average-case extractor (respectively strong average-case extractor) for $D$.

**Proof.** We prove this for strong extractors, the non-strong case is analogous. For every $(Z,X)$ with $X$ distributed on $\{0,1\}^n$ and $\hat{H}_\infty(X|Z) \geq k$, we have
\[ E_{z \sim Z,s \sim U_d} [D(\text{Ext}(X|Z = z,s) \parallel U_m)] = E_{z \sim Z} \left[ E_{s \sim U_d} [D(\text{Ext}(X|Z = z,s) \parallel U_m)] \right]
\leq E_{z \sim Z} \left[ 2^{-\hat{H}_\infty(X|Z = z)} \cdot \varepsilon \right] \leq \varepsilon \cdot \left[ 1 + 2^{-\hat{H}_\infty(X|Z = z)} \right] \leq 3 \varepsilon \]
where the last inequality follows from the fact that $E_{z \sim Z}[2^{-\hat{H}_\infty(X|Z = z)}] = 2^{-\hat{H}_\infty(X|Z)}$ by definition of conditional min-entropy. \(\square\)

**Proof of Theorem 3.12.** By the previous lemma, it suffices to prove that for every $t \geq 0$, $\text{Ext}$ is a $(k - t,2^{k+1} + 1 \cdot \varepsilon)$ extractor (respectively strong extractor) for $D^F$. Since $D^F$ is convex in its first argument by Lemma 2.10 following Chor and Goldreich [CG88] it is enough to consider only distributions with min-entropy $k - t$ that are supported on a set of at most $2^{k-1}$. Fix such a distribution $X$ and a collection of test functions $f_1,\ldots,f_D \in \mathcal{F}$ with $f_1 = \cdots = f_D$ if $\text{Ext}$ is not strong. Then since $X$ is supported on a set of size at most $2^{n-1}$, the distribution $Y$ that is uniform over the complement of $\text{Supp}(X)$ has min-entropy at least $n - 1 \geq k$.
and furthermore the mixture $2^{-t}X + (1 - 2^{-t})Y$ has min-entropy at least $k$. Hence, as Ext is a $(k, \varepsilon)$ extractor (respectively strong extractor) for $D^2$, 
\[
\varepsilon \geq \frac{\mathbb{E}}{i \sim U_{|\mathcal{F}|}} \left[ D^{(f_i)} (\text{Ext}(2^{-t}X + (1 - 2^{-t})Y, i) \| U_m) \right] \\
= 2^{-t} \frac{\mathbb{E}}{i \sim U_{|\mathcal{F}|}} \left[ D^{(f_i)} (\text{Ext}(X, i) \| U_m) \right] + (1 - 2^{-t}) \frac{\mathbb{E}}{i \sim U_{|\mathcal{F}|}} \left[ D^{(f_i)} (\text{Ext}(Y, i) \| U_m) \right] \\
= 2^{-t} \frac{\mathbb{E}}{i \sim U_{|\mathcal{F}|}} \left[ D^{(f_i)} (\text{Ext}(X, i) \| U_m) \right] - (1 - 2^{-t}) \frac{\mathbb{E}}{i \sim U_{|\mathcal{F}|}} \left[ D^{(c - f_i)} (\text{Ext}(Y, i) \| U_m) \right] \\
\geq 2^{-t} \frac{\mathbb{E}}{i \sim U_{|\mathcal{F}|}} \left[ D^{(f_i)} (\text{Ext}(X, i) \| U_m) \right] - (1 - 2^{-t}) \varepsilon \\
\geq \frac{2^{t+1} - 1}{\varepsilon} \mathbb{E} \left[ D^{(f_i)} (\text{Ext}(X, i) \| U_m) \right] \\
\]
where $c_i \in \mathbb{R}$ is such that $c_i - f_i \in \mathcal{F}$ as guaranteed to exist by the symmetry of $\mathcal{F}$.

**Remark 3.14.** Theorem 3.12 also applies to extractors for the $\ell_p$ norms via Proposition 2.11

### 4 Subgaussian distance and connections to other notions

Now that we’ve introduced the general machinery we need, we can go back to our motivation of subgaussian samplers. We will need some standard facts about subgaussian and subexponential random variables, we recommend the book of Vershynin [Ver18] for an introduction.

**Definition 4.1.** A real-valued mean-zero random variable $Z$ is said to be *subgaussian with parameter $\sigma$* if for every $t \in \mathbb{R}$ the moment generating function of $Z$ is bounded as
\[
\ln \mathbb{E}[e^{tZ}] \leq \frac{t^2 \sigma^2}{2}.
\]
If this is only holds for $|t| \leq b$ then $Z$ is said to be $(\sigma, b)$-*subgamma*, and if $Z$ is $(\sigma, 1/\sigma)$-subgamma then $Z$ is said to be *subexponential with parameter $\sigma$*.

**Remark 4.2.** There are many definitions of subgaussian (and especially subexponential) random variables in the literature, but they are all equivalent up to constant factors in $\sigma$ and only affect constants already hidden in big-$O$’s.

**Lemma 4.3.** Let $Z$ be a real-valued random variable. Then

1. (Hoeffding’s lemma) If $Z$ is bounded in the interval $[0, 1]$, then $Z - \mathbb{E}[Z]$ is subgaussian with parameter $1/2$.

2. If $Z$ is mean-zero, then $Z$ is subgaussian (respectively subexponential) with parameter $\sigma$ if and only if $cZ$ is subgaussian (respectively subexponential) with parameter $|c|\sigma$ for every $c \neq 0$.

Furthermore, if $Z$ is mean-zero and subgaussian with parameter $\sigma$, then

1. For all $t > 0$, $\max(\Pr[Z > t], \Pr[Z < -t]) \leq e^{-t^2/2\sigma^2}$.

2. $\|Z\|_p \overset{\text{def}}{=} \mathbb{E}[|Z|^p]^{1/p} \leq 2\sigma \sqrt[p]{p}$ for all $p \geq 1$.

3. $Z$ is subexponential with parameter $\sigma$.

We are now in a position to formally define the *subgaussian distance*. 

---

12
Remark 4.5. We choose subgaussian parameter $1/2$ in Definition 4.4 as by Hoeffding’s lemma, all functions \( f : \{0, 1\}^m \rightarrow [0, 1] \) have that \( f(U_m) - \mathbb{E}[f(U_m)] \) is subgaussian with parameter $1/2$, so this choice preserves the same “scale” as total variation distance. However, the choice of parameter is essentially irrelevant by linearity, as different choices of parameter simply scale the metric $d_G$.

Note that absolute averaging samplers for $\mathcal{G}_{\{0, 1\}}^m$ from Definition 3.6 are exactly subgaussian samplers as defined in the introduction. Thus, by Remark 3.7 and Theorem 3.8 to construct subgaussian samplers it is enough to construct extractors for the subgaussian distance $d_G$.

4.1 Composition

Unfortunately, the subgaussian distance has a major disadvantage compared to total variation distance that complicates extractor construction: it does not satisfy the data-processing inequality, that is, there are probability distributions $P$ and $Q$ over a set $A$ and a function $f : A \rightarrow B$ such that

\[
d_G(f(P), f(Q)) \not\leq d_G(P, Q).
\]

This happens because subgaussian distance is defined by functions which are required to be subgaussian only with respect to the uniform distribution. A simple explicit counterexample comes from taking $f : \{0, 1\}^1 \rightarrow \{0, 1\}^m$ defined by $x \mapsto (x, 0^{m-1})$ and taking $P$ to be the point mass on 0 and $Q$ the point mass on 1. Their subgaussian distance in $\{0, 1\}^1$ is obviously $O(1)$, but the subgaussian distance of $f(P)$ and $f(Q)$ in $\{0, 1\}^m$ is $\Theta(\sqrt{m})$.

The reason this matters because a standard operation (c.f. Nisan and Zuckerman [NZ96]; Goldreich and Wigderson [GW97]; Reingold, Vadhan, and Wigderson [RVW00]) in the construction of samplers and extractors for bounded functions is to do the following: given extractors

\[
\begin{align*}
\text{Ext}_{\text{out}} &: \{0, 1\}^n \times \{0, 1\}^d \rightarrow \{0, 1\}^m \\
\text{Ext}_{\text{in}} &: \{0, 1\}^{n'} \times \{0, 1\}^{d'} \rightarrow \{0, 1\}^d,
\end{align*}
\]

define Ext : $\{0, 1\}^{n+n'} \times \{0, 1\}^{d'} \rightarrow \{0, 1\}^m$ by

\[
\text{Ext}(x, y, s) = \text{Ext}_{\text{out}}(x, \text{Ext}_{\text{in}}(y, s)).
\]

The reason this works for total variation distance is exactly the data-processing inequality: if $Y$ has enough min-entropy given $X$, then $\text{Ext}_{\text{in}}(Y, U_d)$ will be close in total variation distance to $U_d$, and by the data-processing inequality for total variation distance this closeness is not lost under the application of $\text{Ext}_{\text{out}}$. The assumption that $Y$ has min-entropy given $X$ means that $(X, Y)$ is a so-called block-source, and is implied by $(X, Y)$ having enough min-entropy as a joint distribution. From the sampler perspective, this construction uses the inner sampler $\text{Ext}_{\text{in}}$ to subsample the outer sampler. On the other hand, for subgaussian distance, the distribution $\text{Ext}_{\text{in}}(Y, U_d)$ can be $\varepsilon$-close to uniform but still have some element with excess probability mass $\Omega(\varepsilon/\sqrt{d})$, and this element (seed) when mapped by $\text{Ext}_{\text{out}}$ can retain this excess mass in $\{0, 1\}^m$, which results in subgaussian distance $\Theta(\varepsilon \sqrt{m/d}) \gg \varepsilon$. Similarly, from the sampler perspective, even when the outer sampler $\text{Ext}_{\text{out}}$ is a good subgaussian sampler for $\{0, 1\}^m$, there is no reason that a good subgaussian sampler $\text{Ext}_{\text{in}}$ for $\{0, 1\}^d$ the seeds of $\text{Ext}_{\text{out}}$ will preserve the larger sampler property when $m \gg d$.

\[\text{Given a subgaussian extractor } \text{Ext} \text{ with } d \geq \log(m/\varepsilon), \text{ adding a single extra seed } * \text{ to } \text{Ext} \text{ such that } \text{Ext}(x, *) = 0^m \text{ results in a subgaussian extractor with error at most } 2^{-d} \cdot \sqrt{2m} + \varepsilon \leq 3\varepsilon \text{ by convexity of } d_G \text{ and the fact that } d_G(0^m)_{\infty} < \sqrt{2m}.\]
Thus, since this composition operation is needed to construct high-min entropy extractors with the desired seed length even for total variation distance, to construct such extractors for subgaussian distance we need to bypass this barrier. The natural approach is to construct extractors for a better-behaved weak divergence that bounds the subgaussian distance.

**Remark 4.6.** Similar reasoning shows that if Ext is a strong $(k, \varepsilon)$ subgaussian extractor, then it is not necessarily the case that the function $(x, s) \mapsto (s, \text{Ext}(x, s))$ that prepends the seed to the output is a (non-strong) $(k, \varepsilon)$ subgaussian extractor (in contrast to extractors for total variation distance), though the converse does hold.

### 4.2 Connections to other weak divergences

Therefore, to aid in extractor construction, we show how $d_G$ relates to other statistical weak divergences.

Most basically, the subgaussian distance over $\{0, 1\}^m$ differs from total variation distance up to a factor of $O(\sqrt{m})$.

**Lemma 4.7.** Let $P$ and $Q$ be distributions on $\{0, 1\}^m$. Then

$$d_{TV}(P, Q) \leq d_G(P, Q) \leq \sqrt{2\ln 2 \cdot m} \cdot d_{TV}(P, Q)$$

**Proof.** That $d_{TV} \leq d_G$ is immediate from Hoeffding’s lemma and the discussion in Remark 4.5. The reverse bound holds since any subgaussian function takes values at most $\sqrt{\ln 2/2 \cdot m}$ away from the mean by the tail bounds from part 3 of Lemma 4.3 and so any subgaussian test function $f$ has the property that $1/2 + f/\sqrt{2\ln 2 - m}$ is $[0, 1]$-valued and thus lower bounds the total variation distance. \qed

While this allows constructing subgaussian extractors and samplers from total variation extractors, as discussed in the introduction the fact that the upper bound depends on $m$ leads to suboptimal bounds. By starting with a stronger measure of error, we pay a much smaller penalty.

**Lemma 4.8.** Let $P$ and $Q$ be distributions on $\{0, 1\}^m$. Then for every $\alpha > 0$

$$2d_{TV}(P, Q) = d_{\ell_1}(P, Q) \leq 2^{\frac{m\alpha}{(1+\alpha)}} \cdot d_{\ell_{1+\alpha}}(P, Q)$$

$$d_G(P, Q) \leq 2^{\frac{m\alpha}{(1+\alpha)}} \sqrt{1 + \frac{1}{\alpha}} \cdot d_{\ell_{1+\alpha}}(P, Q)$$

In particular, that there is only an additional $\sqrt{1+1/\alpha}$ factor when moving to subgaussian distance compared to total variation, which in particular does not depend on $m$ and is constant for constant $\alpha$.

**Proof.** By Proposition 2.11 for any function $f : \{0, 1\}^m \to \mathbb{R}$ it holds that

$$D(f)(P \parallel Q) \leq \|f(U_m)\|_{1+\frac{1}{\alpha}} \cdot d_{\ell_{1+\alpha}}(P, Q) = \|f(U_m)\|_1 \cdot 2^{\frac{m\alpha}{(1+\alpha)}} \cdot d_{\ell_{1+\alpha}}(P, Q).$$

The result follows since $[-1, 1]$-valued functions $f$ satisfy moment bounds $\|f(U_m)\|_q \leq 1$ for all $q \geq 1$, and functions $f$ which are subgaussian satisfy moment bounds $\|f(U_m)\|_q \leq \sqrt{q}$ by Lemma 4.3. \qed

One downside of starting with bounds on $\ell_{1+\alpha}$ is that, extending a result of Vadhan [Vad12, Problem 6.4], we show in Corollary 5.30 that for every $1 > \alpha > 0$, there is a constant $c_\alpha > 0$ such any $\ell_{1+\alpha}$ extractor with error smaller than $c_\alpha \cdot 2^{-m\alpha/(1+\alpha)}$ requires seed length linear in $\alpha \cdot \min(n-k, m)$, for $n-k$ the entropy deficiency and $m$ the output length. One might hope that sending $\alpha$ to 0 would eliminate this linear lower bound but still bound the subgaussian distance, but phrased this way sending $\alpha$ to 0 just results in a total variation extractor.

However, with a shift in perspective essentially the same approach works: by Example 2.4, $d_2(P, U_m) \leq 2^{-m/2}$ implies $D_2(P \parallel U_m) \leq \varepsilon^2/\ln 2$, and there is an analogous linear seed length lower bound on constant error $D_{1+\alpha}$ extractors for every $\alpha > 0$. In this case, however, sending $\alpha$ to 0 results in the KL divergence, which does upper bound the subgaussian distance, and in fact with the same parameters as for total variation distance.
Lemma 4.9. Let $P$ and $Q$ be distributions on $\{0,1\}^m$. Then
\[
d_g(P,U_m) \leq \sqrt{\frac{\ln 2}{2}} \cdot \text{KL}(P \parallel U_m)
\]
\[
d_E(P,U_m) \leq \begin{cases} 
\sqrt{\frac{\ln 2}{2}} \cdot \text{KL}(P \parallel U_m) & \text{if } \text{KL}(P \parallel U_m) \leq \frac{1}{2m^2} \\
\ln 2 \cdot \text{KL}(P \parallel U_m) + \frac{1}{2} & \text{if } \text{KL}(P \parallel U_m) > \frac{1}{2m^2}
\end{cases}
\]
where these bounds are concave in $\text{KL}(P \parallel U_m)$. In the reverse direction, it holds that
\[
\text{KL}(P \parallel U_m) \leq m \cdot d_{TV}(P,U_m) + h(d_{TV}(P,U_m))
\]
where $h(x) = x \log(1/x) + (1-x) \log(1/(1-x))$ is the (concave) binary entropy function.

Proof. The upper bound on subgaussian distance follows from a general form of Pinsker’s inequality as in [BLM13, Lemma 4.18], but for the extension to subexponential functions we reproduce its proof here, based on the Donsker–Varadhan “variational” formulation of KL divergence [DV76] (c.f. [BLM13 Corollary 4.15])
\[
\text{KL}(P \parallel U_m) = \frac{1}{2m} \cdot \sup_{g : \{0,1\}^m \to \mathbb{R}} \left( \mathbb{E}[g(P)] - \log \mathbb{E}[e^{g(U_m)}] \right).
\]
Now if $f : \{0,1\}^m \to \mathbb{R}$ satisfies $\mathbb{E}[f(U_m)] = 0$, then by letting $g(x) = t \cdot f(x)$, this implies
\[
\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] = \frac{1}{t} \cdot \mathbb{E}[g(P)] \leq \frac{\ln 2 \cdot \text{KL}(P \parallel U_m) + \log \mathbb{E}[e^{t f(U_m)}]}{t}
\]
for all $t > 0$. Thus, when $\mathbb{E}[e^{t f(U_m)}] \leq t^2/8$, we have $|\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)]| \leq \ln 2 \cdot \text{KL}(P \parallel U_m)/t + t/8$.

Then since subgaussian random variables satisfy such a bound for all $t$, we can make the optimal choice $t = \sqrt{8 \ln 2 \cdot KL(P \parallel U_m)}$ to get the claimed bound on $d_g$. For subexponential random variables, which satisfy such a bound only for $|t| \leq 2$, we choose $t = \min(\sqrt{8 \ln 2 \cdot \text{KL}(P \parallel U_m)}, 2)$, which gives
\[
d_E(P,U_m) \leq \begin{cases} 
\sqrt{\frac{\ln 2}{2}} \cdot \text{KL}(P \parallel U_m) & \text{if } \text{KL}(P \parallel U_m) \leq \frac{1}{2m^2} \\
\ln 2 \cdot \text{KL}(P \parallel U_m) + \frac{1}{2} & \text{if } \text{KL}(P \parallel U_m) > \frac{1}{2m^2}
\end{cases}
\]
as desired. The concavity of this bound follows by noting that it has a continuous and nonincreasing derivative.

For the reverse inequality, we use a bound on the difference in entropy between distributions $P$ and $Q$ on a set of size $S$ which states
\[
|H(P) - H(Q)| \leq \log(S-1) \cdot d_{TV}(P,Q) + h(d_{TV}(P,Q)).
\]
This inequality is a simple consequence of Fano’s inequality as noted by Goldreich and Vadhan [GV99, Fact B.1], and implies the desired result by taking $Q = U_m$ as $\text{KL}(P \parallel U_m) = H(U_m) - H(P)$ and $|\{0,1\}^m| = 2^m$.

Remark 4.10. There are sharper upper bounds on the KL divergence than given in Lemma 4.9, such as the bound of Audenaert and Eisert [AE05 Theorem 6], but the bound we use has the advantage of being defined for the entire range of the total variation distance and being everywhere concave.

5 Extractors for KL divergence

By Lemma 4.9 the subgaussian distance can be bounded in terms of the KL divergence to uniform, so by the following easy lemma to construct subgaussian extractors it suffices to construct extractors for KL divergence.

Lemma 5.1. Let $V_1$ and $V_2$ be weak divergences on the set $\{0,1\}^m$ and $f : \mathbb{R} \to \mathbb{R}$ be a function such that $V_1(P \parallel U_M) \leq f(V_2(P \parallel U_m))$ for all distributions $P$ on $\{0,1\}^m$. Then if $f$ is increasing on $(0,\epsilon)$, every $(k,\epsilon)$ extractor $\text{Ext}$ for $V_1$ is also a $(k,f(\epsilon))$-extractor for $V_2$, and if $f$ is also convex, then if $\text{Ext}$ is strong or average-case as a $V_1$-extractor, it has the same properties as a $(k,f(\epsilon))$ extractor for $V_2$. 

15
Importantly, the KL divergence does not have the flaws of subgaussian distance discussed in Section 4.1. The classic data-processing inequality says that KL divergence is non-increasing under postprocessing by (possibly randomized) functions, and the chain rule for KL divergence says that

$$KL(A, B \| X, Y) = KL(A \| X) + \mathbb{E}_{a \sim A} [KL(B_{A=a} \| Y_{X=a})]$$

for all distributions $A$, $B$, $X$, and $Y$, so that in particular

$$\mathbb{E}_{s \sim U_d} [KL(\text{Ext}(X,s) \| U_m)] = KL(U_d, \text{Ext}(X,U_d) \| U_d, U_m)$$

and prepending the seed of a strong KL-extractor does in fact give a non-strong KL-extractor:

**Lemma 5.2.** A function $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a $(k, \varepsilon)$ strong KL-extractor (respectively strong average-case KL-extractor) if and only if the function $\text{Ext}' : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{d+m}$ defined by $\text{Ext}'(x, s) = (s, \text{Ext}(x, s))$ is a (non-strong) $(k, \varepsilon)$ KL-extractor (respectively average-case KL-extractor).

Furthermore, KL divergence satisfies a type of triangle inequality when combined with higher Rényi divergences:

**Lemma 5.3.** Let $P$, $Q$, and $R$ be distributions over a finite set $X$. Then for all $\alpha > 0$, it holds that

$$KL(P \| R) \leq \left(1 + \frac{1}{\alpha}\right) \cdot KL(P \| Q) + D_{1+\alpha}(Q \| R)$$

**Proof.** This follows from a characterization of Rényi divergence due to van Erven and Harremoës [vEH10, Lemma 6.6], [vEH14, Theorem 30] and Shayevitz [Sha11, Theorem 1], who prove that for every positive real $\beta \neq 1$ and distributions $X$ and $Y$ that

$$(1 - \beta) D_{\beta}(X \| Y) = \inf_{Z} \{\beta KL(Z \| X) + (1 - \beta) KL(Z \| Y)\}.$$ 

In particular, choosing $\beta = 1 + \alpha$, $X = Q$, and $Y = R$ and upper bounding the infimum by the particular choice of $Z = P$ gives the claim. \qed

## 5.1 Composition

These properties imply that composition does work as we want (without any loss depending on the output length $m$) assuming we have extractors for KL and higher divergences.

**Theorem 5.4** (Composition for high min-entropy Rényi entropy extractors, c.f. [GW97]). Suppose

1. $\text{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is an $(n - \log(1/\delta), \varepsilon_{out})$ extractor for $D_{1+\alpha}$ with $\alpha > 0$,

2. $\text{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^d \to \{0,1\}^d$ is an $(n' - \log(1/\delta), \varepsilon_{in})$ average-case extractor for KL,

and define $\text{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^d \to \{0,1\}^m$ by $\text{Ext}((x, y), s) = \text{Ext}_{out}(x, \text{Ext}_{in}(y, s))$. Then $\text{Ext}$ is a $(n + n' - \log(1/\delta), \varepsilon_{out} + (1 + 1/\alpha) \cdot \varepsilon_{in})$ extractor for KL. Furthermore, if $\text{Ext}_{in}$ is a strong average-case KL-extractor, then $\text{Ext}$ is a strong KL-extractor, and if $\text{Ext}_{out}$ is average-case then so is $\text{Ext}$.

**Proof.** Let $(Z, X, Y)$ be jointly distributed random variables with $X$ distributed over $\{0,1\}^n$ and $Y$ over $\{0,1\}^{n'}$ conditionally independent of $X$ given $Z$ such that $H_{\infty}(X, Y | Z) \geq n + n' - \log(1/\delta)$. Let $S'$ be a distribution over $\{0,1\}^d$ which is independent of $X, Y$, and $Z$. Then for every $z \in \text{Supp}(Z)$, we have by
Lemma 5.3 and the data-processing inequality for KL divergence that
\[
\text{KL}( \text{Ext}( (X)_{Z=z} Y_{Z=z}, S') \parallel U_m ) \\
\overset{\text{Def. Ext}}{=} \text{KL}( \text{Ext}_{out}(X|Z=z, \text{Ext}_{in}(Y|Z=z), S') \parallel U_m ) \\
\overset{\text{chain rule}}{\leq} (1 + 1/\alpha) \cdot \text{KL}( \text{Ext}_{out}(X|Z=z, \text{Ext}_{in}(Y|Z=z), S') \parallel U_m ) \\
+ D_{1+\alpha}(\text{Ext}_{out}(X|Z=z, U_d) \parallel U_m ) \\
\leq (1 + 1/\alpha) \cdot \text{KL}(X|Z=z, \text{Ext}_{in}(Y|Z=z, S') \parallel X|Z=z, U_d) + D_{1+\alpha}(\text{Ext}_{out}(X|Z=z, U_d) \parallel U_m ) \\
= (1 + 1/\alpha) \cdot \mathbb{E}_{x \sim X|Z=z} \left[ \text{KL}(\text{Ext}_{in}(Y|X=x, Z=z, S') \parallel U_d) \right] + D_{1+\alpha}(\text{Ext}_{out}(X|Z=z, U_d) \parallel U_m )
\]

where the last equality follows from the chain rule for KL divergence. Now by standard properties of conditional min-entropy (see for example [DORS08 Lemma 2.2]), we know that \( H_\infty(X|Z) \geq H_\infty(X, Y|Z) - \log |\text{Supp}(Y)| \geq n - \log(1/\delta) \) and \( H_\infty(Y|X, Z) \geq H_\infty(X, Y|Z) - \log |\text{Supp}(X)| \geq n' - \log(1/\delta) \).

If \( \text{Ext}_{out} \) is not average-case, take \( Z \) to be a constant independent of \( X \) and \( Y \), and if \( \text{Ext}_{out} \) is average-case then take the average of both sides over \( Z \). The claim for non-strong \( \text{Ext}_{in} \) then follows by taking \( S' = U_d \) which bounds the first term by \( (1 + 1/\alpha) \cdot \varepsilon_{in} \) and the second by \( \varepsilon_{out} \). The claim for strong \( \text{Ext}_{in} \) follows by choosing \( S' = U(s) \) to be the point mass on \( s \in \{0,1\}^d \) and then taking the expectation of both sides over a uniform \( s \in \{0,1\}^d \).

\[ \square \]

Remark 5.5. Theorem 5.1 in fact a construction of a block-source KL-extractor, meaning that the claimed error bounds hold for any joint distributions \( (X, Y, Z) \) such that \( H_\infty(Y) \geq n' - \log(1/\delta) \) and \( H_\infty(X|Y, Z) \geq n - \log(1/\delta) \) rather than just those distributions with \( H_\infty(X, Y) \geq n + n' - \log(1/\delta) \). The extra \( \log(1/\delta) \) entropy loss inherent in the non-block analysis is why Reingold, Wigderson, and Vadhan [RVW00] introduced the zig-zag product for extractors, which we will apply for KL-extractors in Corollary 5.20.

5.2 Existing explicit constructions

The construction of Theorem 5.1 required both a \( D_{1+\alpha} \)-extractor and an average-case KL-extractor, so for the result not to be vacuous we need to show the existence of such extractors. Thankfully, Example 2.4 implies that extractors for \( \ell_2 \) are also extractors for \( D_2 \), so we can use existing \( \ell_2 \) extractors from the literature, such as the Leftover Hash Lemma of Impagliazzo, Levin, and Luby [ILL89] (see also [McI87, BBR88]) and its variant using almost-universal hash functions due to Srinivasan and Zuckerman [SZ99].

Proposition 5.6 ([McI87, BBR88, ILL89, IZ89, SZ99, DORS08]). Let \( \mathcal{H} \) be a collection of \( \varepsilon \)-almost universal hash functions from the set \( \{0,1\}^n \) to the set \( \{0,1\}^m \), meaning that for all \( x \neq y \in \{0,1\}^n \) it holds that \( \Pr_{h \sim \mathcal{H}}[h(x) = h(y)] \leq (1 + \varepsilon)/2^m \). Then the function \( \text{Ext} : \{0,1\}^n \times \mathcal{H} \rightarrow \{0,1\}^m \) defined by \( \text{Ext}(x, h) = (h(x), h(x)) \) is an average-case \( (m + \log(1/\varepsilon), 2/\log 2 - \varepsilon) \) \( D_2 \)-extractor.

In particular, for every \( k, n \in \mathbb{N} \) and \( 1 > \varepsilon > 0 \) there is an explicit strong average-case \((k, \varepsilon)\) extractor for \( D_2 \) (and KL) with seed length \( d = O(k + \log(n/\varepsilon)) \) and \( m = k - \log(1/\varepsilon) - O(1) \), given by \( \text{Ext}'(x, h) = h(x) \) for \( h \) drawn from an appropriate almost-universal hash family.

Proof. The \( D_2 \) claim is implicit in Rackoff’s proof of the Leftover Hash Lemma (see [IZ89]) and Srinivasan and Zuckerman’s proof of the claim for total variation [SZ99], which both analyzed the collision probability of the output, and the average-case claim was proved by Dodis, Ostrovsky, Reyzin, and Smith [DORS08], though we include a proof here for completeness.

Given a joint distribution \((Z, X)\) such that \( X \) is distributed over \( \{0,1\}^n \) with \( H_\infty(X|Z) \geq m + \log(1/\varepsilon) \),
we have

\[
\mathbb{E}_{z \sim Z} \left[ D_2(\text{Ext}(X|Z=z, \mathcal{H}) \parallel \mathcal{H} \times U_m) \right]
= \mathbb{E}_{z \sim Z} \left[ \log \left( 2^m \cdot |\mathcal{H}| \cdot \Pr_{h, h' \sim \mathcal{H}, x, x' \sim X|Z=z} [(h, h(x)) = (h', h'(x))] \right) \right]
\leq \mathbb{E}_{z \sim Z} \left[ \log \left( 2^m \cdot \left( 2^{-H_{\infty}(X|Z=z)} + 1 + \varepsilon \right) \right) \right]
\leq \log \left( \mathbb{E}_{z \sim Z} \left[ 2^{m-H_{\infty}(X|Z)} + 1 + \varepsilon \right] \right) \quad \text{(by Jensen’s inequality)}
= \log \left( 2^{m-H_{\infty}(X|Z)} + 1 + \varepsilon \right)
\leq \log(1 + 2\varepsilon) \leq \frac{2}{\ln 2} \cdot \varepsilon.
\]

The in particular statement follows from Lemma 5.7 below and from the existence of \(\varepsilon\)-almost universal hash families with size \(\text{poly}(2^k, n, 1/\varepsilon)\) as constructed by \[SZ99\].

To establish the claim about strong extractors, we generalize Lemma 5.2 to extractors for \(D_{1+\alpha}\) for \(\alpha > 0\):

**Lemma 5.7.** If \(\text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m\) is a \((k,\varepsilon)\) \(D_{1+\alpha}\)-extractor (respectively average-case \(D_{1+\alpha}\)-extractor) for \(\alpha > 0\) such that \(\text{Ext}(x,s) = (s, \text{Ext}'(x,s))\), then \(\text{Ext}'\) is a strong \((k,\varepsilon)\) \(D_{1+\alpha}\)-extractor (respectively strongly average-case \((k,\varepsilon)\) \(D_{1+\alpha}\)-extractor).

**Proof.**

\[
\mathbb{E}_{s \sim U_d} \left[ D_{1+\alpha}(\text{Ext}'(X,s) \parallel U_m) \right] = \mathbb{E}_{s \sim U_d} \left[ \frac{1}{\alpha} \log \left( 1 + 2^m \alpha \mathbb{E}_{y \sim \{0,1\}^m} \Pr_{X,U} [\text{Ext}'(X,s,y) = 1] \right) \right]
\leq \frac{1}{\alpha} \log \left( 1 + 2^m \mathbb{E}_{s \sim U_d} \left[ \sum_{y \in \{0,1\}^m} \Pr_{X,U} [\text{Ext}'(X,s,y) = 1] \right] \right)
= \frac{1}{\alpha} \log \left( 1 + 2^m \sum_{(s,y) \in \{0,1\}^{d+m}} \Pr_{U_d, \text{Ext}'(X,U_d)} [(s,y) = 1] \right)
= D_{1+\alpha}(\text{Ext}(X,U_d) \parallel U_d, U_m). \quad \Box
\]

Following Vadhan \[Vad12\], we also note that the extractor based on expander walks due to Goldreich and Wigderson \[GW97\], which has the nice property that its seed length depends only on \(n-k\) and the entropy deficiency of the source rather than \(n\) itself, is also an \(\ell_2\) extractor.

**Proposition 5.8** \([GW97, Vad12, Discussion after Theorem 6.22]\). For all \(k \leq n \in \mathbb{N}\) and \(1 \geq \varepsilon > 0\) there is an explicit average-case \((k,\varepsilon/\ln 2)\) \(D_2\)-extractor \(\text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m\) with seed length \(d = O(n-k + \log(1/\varepsilon))\) and output length \(m = n\).

Furthermore, \(\text{Ext}\) has the property that the function \((x,s) \mapsto (s,\text{Ext}(x,s))\) is an injection out of \(\{0,1\}^n \times \{0,1\}^d\).

**Proof.** We sketch the proof for completeness. First, recall from Example 2.4 that for every distribution \(P\) that \(D_2(P \parallel U_m) = \log(1 + 2^m d_2(P,U_m)^2) \leq 2^m d_2(P,U_m)^2/\ln 2\), so it suffices to construct a \((k, \sqrt{\varepsilon}/2^{m/2})\) \(\ell_2\)-extractor.

Let \(G\) be an undirected \(g\)-regular graph on \(\{0,1\}^n\) with transition matrix \(M = \frac{1}{g} A\) where \(A\) is the adjacency matrix of \(G\), and let \(1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1\) be the spectrum of \(M\). Then if \(\lambda = \max\{\lambda_2, -\lambda_n\}\) we know that for every distribution \(D\) on \(\{0,1\}^n\), the \(\ell_2\) distance between \(D\) and uniform decreases by a factor of
After one random step on \( G \), hence, since every distribution \( X \) on \( \{0, 1\}^n \) with \( H_\infty(X) \geq k \) has \( \ell_2 \) distance from uniform at most \( \sqrt{2^{-k} - 2^{-n}} \leq 2^{-k/2} \), a walk of length \( \log_\lambda(\sqrt{\varepsilon} \cdot 2^{-n/2} / 2^{-k/2}) \) suffices to reduce the \( \ell_2 \) distance to uniform to \( \sqrt{\varepsilon} \cdot 2^{-n/2} \), which takes \( d \) random bits for

\[
d = \log g \cdot \log \lambda \frac{\sqrt{\varepsilon} \cdot 2^{-n/2}}{2^{-k/2}} = (n - k + \log(1/\varepsilon)) \cdot \frac{\log g}{\log(1/\lambda^2)}.
\]

Hence, we need \( G \) on \( \{0, 1\}^n \) such that \( \log g / \log(1/\lambda^2) \) is constant; for this, we can take \( G \) to be the explicit constant degree expander of Margulis–Gaber–Galil [Mar75, GGR81] (technically this requires \( n \) even, which following Goldreich [Gol11] we can solve when \( n \) is odd by joining two graphs on \( \{0, 1\}^{n-1} \) by the canonical perfect matching, and we can add self-loops to ensure the degree is a power of 2). This graph has the property that edges are labelled by invertible transformations on the vertex set. Hence, given \( s \) and an output vertex \( v = \text{Ext}(x, s) \), we can recover the input \( x \) by walking back from \( v \) according to the inverses of the transformations associated to the path \( s \), and thus \((x, s) \mapsto (s, \text{Ext}(x, s)) \) is injective as desired.

The average-case claim holds since we have in fact shown that for any distribution \( Y \) on \( \{0, 1\}^n \) that \( D_2(\text{Ext}(Y, U_d) \parallel U_m) \leq \varepsilon \cdot 2^{k - H_\infty(Y)} / \ln 2 \), so for a joint distribution \((Z, X)\) with \( H_\infty(X|Z) \geq k \), letting \( Y = X_{|Z=\bar{Z}} \) and taking the expectation over \( Z \) finishes the proof.

**Remark 5.9.** The fact that \((s, \text{Ext}(x, s)) \) is an injection implies that, unlike for the extractors from hashing of Proposition 5.8, the result of prepending the seed to the output of the expander-walk extractor does not give a \( D_2 \) extractor. However, it will be very useful in concert with Reingold, Vadhan, and Wigderson’s zig-zag product for extractors [RVW00] to avoid the entropy loss in Theorem 5.4.

**Remark 5.10.** Given suitable explicit constructions of Ramanujan graphs (or just with \( \lambda = O(1/\sqrt{y}) \)), one can take the seed length to be \( d = (n - k + \log(1/\varepsilon))(1 + O(\log^{-1} t)) \) which for \( t = 2^{O(n - k + \log(1/\varepsilon))} \) is \( n - k + \log(1/\varepsilon) + O(1) \). Furthermore, such graphs give a \((k, \varepsilon) \) \( D_2 \) extractor (and thus KL-extractor) with seed length \( n - k + \log(1/\varepsilon) + O(1) \) and entropy loss \( \log(1/\varepsilon) + O(1) \), which has optimal dependence on \( \varepsilon \) as we show in Theorem 5.27.

Because Proposition 5.8 has seed length depending only on \( n - k \) the entropy deficiency of the source rather than \( n \) itself, when written as a high min-entropy extractor it has the appealing property that given an entropy deficiency \( \Delta \) and error \( \varepsilon \), there is a single seed length \( d \) that works for all input lengths.

**Corollary 5.11.** There is a universal constant \( C > 0 \) such that for every \( 1 > \varepsilon > 0, \Delta > 0, \) and \( n \in \mathbb{N} \) there is an explicit \((n - \Delta, \varepsilon) \) \( D_2 \)-extractor \( \text{Ext} : \{0, 1\}^n \times \{0, 1\}^d \to \{0, 1\}^n \) with \( d = \lceil C \cdot (\Delta + \log(1/\varepsilon)) \rceil \) such that the function \((x, s) \mapsto (s, \text{Ext}(x, s)) \) is an injection.

**Remark 5.12.** In particular, for \( \Delta, \varepsilon \) fixed, one can take \( n \) in Corollary 5.11 depending on \( d \).

We argued that the above extractors are KL-extractors using the fact they are \( \ell_2 \) (and thus \( D_2 \)) extractors, but one can also show that any total variation extractor with sufficiently small error is a KL-extractor, albeit with some loss of parameters.

**Lemma 5.13.** For every \((k, \varepsilon)\) extractor \( \text{Ext} : \{0, 1\}^n \times \{0, 1\}^d \to \{0, 1\}^m \) for total variation distance such that \( \varepsilon \leq 1/2 \), \( \text{Ext} \) is also a \((k, m \cdot \varepsilon + h(\varepsilon))\) KL-extractor, where \( h(x) = x \log(1/x) + (1 - x) \log(1/(1 - x)) \) is the binary entropy function. Furthermore, if \( \text{Ext} \) is strong, average-case, or both as a total variation extractor, then it has the same properties as a KL-extractor.

In particular, if \( \varepsilon' = \frac{4k^2}{\min(\varepsilon, 1/2)} \), then every \((k, \varepsilon')\) extractor (respectively strong extractor) is an average-case \((k, \varepsilon)\) KL-extractor (respectively strongly average-case \((k, \varepsilon)\) KL-extractor).

**Proof.** The main claim is an immediate corollary of Lemmas 5.9 and 5.11. The in particular statement follows since \( \text{Ext} \) being a \((k, \varepsilon')\) extractor (respectively strongly extractor) implies by Theorem 5.12 that it is a \((k, 3\varepsilon')\) average-case (respectively strongly average-case) extractor, so since we have chosen \( \varepsilon' \) to make \( m \cdot 3\varepsilon' + h(3\varepsilon') \leq \varepsilon \), we know \( \text{Ext} \) is an average-case \((k, \varepsilon)\) KL-extractor (respectively strongly average-case KL-extractor).
Remark 5.14. Reducing \( \varepsilon \) by a factor of \( m + \log(1/\varepsilon) \) increases the seed length and entropy loss of the input extractor. For the former, this is often (but not always) tolerable since the input extractor may already depend suboptimally on \( \log(n/\varepsilon) \). For the latter, we will show in Corollary 5.22 how to use the transform of Raz, Reingold, and Vadhan [RRV02] to recover \( O(\log(m/\varepsilon)) \) bits of lost entropy (at least this much must be lost by Radhakrishnan and Ta-Shma [RT00]) at a cost of \( O(\log(n/\varepsilon)) \) in the seed length.

Instantiating Lemma 5.13 with the Guruswami–Umans–Vadhan [GUV09] extractor for total variation distance, we see that the increased seed length and entropy loss are simply absorbed into the existing hidden constants:

**Theorem 5.15 (KL-analogue of [GUV09 Theorem 1.5]).** For every \( n \in \mathbb{N} \), \( k \leq n \), and \( 1 > \alpha, \varepsilon > 0 \), there is an explicit average-case (respectively strong average-case) \((k, \varepsilon)\) KL-extractor \( \text{Ext} : \{0,1\}^n \times \{0,1\}^d \rightarrow \{0,1\}^m \) with \( d \leq \log n + O_\alpha(\log(k/\varepsilon)) \) and \( m \geq (1-\alpha)k \) (respectively \( m \geq (1-\alpha)k - O_\alpha(\log(n/\varepsilon)) \)).

### 5.3 Reducing the entropy loss of KL-extractors

In this section, we show how to avoid the entropy loss inherent in Theorem 5.4 using the zig-zag product for extractors, introduced by Reingold, Vadhan, and Wigderson [RVW00]. This product combines a technique of Raz and Reingold [RR99] to preserve entropy and the method of Wigderson and Zuckerman [WZ99] to extract entropy left over in a source after an initial extraction, and we show that these techniques extend to the setting of KL-extractors. Furthermore, these techniques (along with the Leftover Hash Lemma) are also the key to the transformation of Raz, Reingold, and Vadhan [RRV02] to convert an arbitrary extractor into one with optimal entropy loss, so we show that this transformation works for KL-extractors as well.

For all of these results, the key is the following lemma:

**Lemma 5.16 (Re-extraction from leftovers).** Let

1. \( \text{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \rightarrow \{0,1\}^{m_1} \) be a \((k_1, \varepsilon_1)\) KL-extractor,
2. \( W_1 : \{0,1\}^n \times \{0,1\}^{d_1} \rightarrow \{0,1\}^w \) be a function such that \((\text{Ext}_1, W_1) : \{0,1\}^n \times \{0,1\}^{d_1} \rightarrow \{0,1\}^{m_1} \times \{0,1\}^w\) is an injective map,
3. \( \text{Ext}_2 : \{0,1\}^w \times \{0,1\}^{d_2} \rightarrow \{0,1\}^{m_2} \) be a \((k_2, \varepsilon_2)\) average-case KL-extractor for \( k_2 \leq k_1 + d_1 - m_1 \).

Then \( \text{Ext} : \{0,1\}^n \times \{0,1\}^{d_1+d_2} \rightarrow \{0,1\}^{m_1+m_2} \) defined by \( \text{Ext}(x, (s, t)) = (\text{Ext}_1(x, s), \text{Ext}_2(W_1(x, s), t)) \) is a \((k_1, \varepsilon_1 + \varepsilon_2)\) KL-extractor. Furthermore, if \( \text{Ext}_1 \) is average-case then so is \( \text{Ext} \).

**Remark 5.17.** The pair \((\text{Ext}_1, W_1)\) is a special case of what Raz and Reingold [RR99] called an extractor-condenser pair. One can think of \( W_1 \) as preserving “leftovers” or “waste,” which is then “re-extracted” or “recycled” by \( \text{Ext}_2 \). The identity function on \( \{0,1\}^n \times \{0,1\}^{d_1} \) is a valid choice of \( W_1 \), but the advantage of the more general formulation is that \( w \) can be much smaller than \( n + d_1 \), and most known explicit constructions of extractors have seed length depending on the input length of the source.

**Proof.** Given any joint distribution \((Z, X)\) such that \( X \) is distributed over \( \{0,1\}^n \) and \( \tilde{H}_\infty(X|Z) \geq k_1 \), we have for every \( z \in \text{Supp}(Z) \) that

\[
\begin{align*}
\text{KL}(\text{Ext}(X|Z=z, (U_{d_1}, U_{d_2})) & \parallel U_{m_1+m_2}) \\
& = \text{KL}(\text{Ext}_1(X|Z=z, U_{d_1}), \text{Ext}_2(W_1(X|Z=z, U_{d_1}), U_{d_2})) \parallel U_{m_1}, U_{m_2}) \\
& = \text{KL}(\text{Ext}_1(X|Z=z, U_{d_1}) \parallel U_{m_1}) \\
& \quad + \mathbb{E}_{o_1 \sim \text{Ext}_1(X|Z=z, s)} \left[ \text{KL}(\text{Ext}_2(W_1(X, U_{d_1})|Z=z, \text{Ext}_1(X, U_{d_1})=o_1, U_{d_2}) \parallel U_{m_2}) \right] \quad (5.17.1)
\end{align*}
\]
where the last line follows from the chain rule for KL divergence. Note that

\[
\tilde{H}_\infty(W_1(X, U_{d_1}) \mid Z, \text{Ext}_1(X, U_{d_1})) \\
\quad = \tilde{H}_\infty(\text{Ext}_1(X, U_{d_1}), W_1(X, U_{d_1}) \mid Z, \text{Ext}_1(X, U_{d_1})) \\
\quad = \tilde{H}_\infty(X, U_{d_1} \mid Z, \text{Ext}_1(X, U_{d_1})) \\
\quad \geq \tilde{H}_\infty(X, U_{d_1} \mid Z) - \log|\text{Supp}(\text{Ext}_1(X, U_{d_1}))| \\
\quad = \tilde{H}_\infty(X \mid Z) + H_\infty(U_{d_1}) - \log|\text{Supp}(\text{Ext}_1(X, U_{d_1}))| \\
\quad \geq k_1 + d_1 - m_1 \geq k_2
\]

where the line \((*)\) follows from standard properties of conditional min-entropy (e.g. [DORS08 Lemma 2.2]). That \text{Ext} is a \((k_1, \varepsilon_1 + \varepsilon_2)\) KL-extractor now follows immediately from Eq. (5.17.1) by taking \(Z\) independent of \(X\), and the average-case claim follows from taking expectations over \(z \sim Z\).

\[
\text{Remark 5.18. } \text{The proof above in fact works any weak divergence } D \text{ such that } D(X, Y \mid U_{m_1}, U_{m_2}) \leq D(X \mid U_{m_1}) + \mathbb{E}_{x \sim X}[D(Y \mid X = x \mid U_{m_2})] \text{ for all joint distributions } (X, Y) \text{ independent of } (U_{m_1}, U_{m_2}). \text{ In particular, the proof also gives Lemma 5.16 for standard (total variation) extractors.}
\]

By Lemma 5.16, we get an analogous result for strong KL-extractors.

\[
\text{Corollary 5.19. } \text{Let}
\]

1. \(\text{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1} \) be a strong \((k_1, \varepsilon_1)\) KL-extractor,

2. \(W_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^w \) be a function such that the map \((x, s) \mapsto (s, \text{Ext}_1(x, s), W_1(x, s))\) is an injection,

3. \(\text{Ext}_2 : \{0,1\}^w \times \{0,1\}^{d_2} \to \{0,1\}^{m_2} \) be a \((k_2, \varepsilon_2)\) strong average-case KL-extractor for \(k_2 \leq k_1 - m_1\).

Then \(\text{Ext} : \{0,1\}^n \times \{0,1\}^{d_1 + d_2} \to \{0,1\}^{m_1 + m_2}\) defined by \(\text{Ext}(x, (s, t)) = (\text{Ext}_1(x, s), \text{Ext}_2(W_1(x, s), t))\) is a strong \((k_1, \varepsilon_1 + \varepsilon_2)\) KL-extractor. Furthermore, if \(\text{Ext}_1\) is average-case then so is \(\text{Ext}\).

\[
\text{The zig-zag product for extractors due to Reingold, Vadhan, and Wigderson [RVW00] (in the special case of injective (Ext, W)-pairs) is an immediate consequence of Lemma 5.16 and Theorem 5.4 our basic composition result. Recall that Theorem 5.4 was able to combine an “outer” extractor, generally taken to have seed length depending only (but linearly) on \(n - k\), with an “inner” extractor to produce seeds for the outer extractor with logarithmic seed length. However, as discussed in Remark 5.5 that basic composition necessarily lost \(\log(1/\delta)\) bits of entropy, so the zig-zag product uses Lemma 5.16 to recover this entropy, using an (Ext, W)-pair to ensure that the re-extraction adds additional seed length depending logarithmically on \(n - k\) rather than \(n\).
\]

\[
\text{Corollary 5.20 (Zig-zag product for KL-extractors, analogous to [RVW00 Theorem 3.6]). Let}
\]

1. \(\text{Ext}_{\text{out}} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m\) be an \((n - \log(1/\delta), \varepsilon_{\text{out}})\) extractor for \(D_{1+\alpha}\) with \(\alpha > 0\),

2. \(W_{\text{out}} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^w\) be a function such that the pair \((\text{Ext}_{\text{out}}, W_{\text{out}})\) is an injection from \(\{0,1\}^n \times \{0,1\}^d\),

3. \(\text{Ext}_{\text{in}} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d\) be an \((n' - \log(1/\delta), \varepsilon_{\text{in}})\) average-case KL-extractor,

4. \(W_{\text{in}} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^{w'}\) be such that the pair \((\text{Ext}_{\text{in}}, W_{\text{in}})\) is an injection from \(\{0,1\}^{n'} \times \{0,1\}^{d'}\),

5. \(\text{Ext}_{\text{waste}} : \{0,1\}^{w + w'} \times \{0,1\}^{d''} \to \{0,1\}^{m''}\) be an average-case \((n + n' - \log(1/\delta) - m, \varepsilon_{\text{waste}})\) KL-extractor,
and define

1. \( \text{Ext}_{\text{comp}} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \rightarrow \{0,1\}^m \) by \( \text{Ext}_{\text{comp}}((x,y),s) = \text{Ext}_{\text{out}}(x, \text{Ext}_{\text{in}}(y,s)) \) as in Theorem 5.4.

2. \( W_{\text{comp}} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \rightarrow \{0,1\}^{w+w'} \) by \( W_{\text{comp}}((x,y),s) = (W_{\text{out}}(x, \text{Ext}_{\text{in}}(y,s)), W_{\text{in}}(y,s)) \).

3. \( \text{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d+d'} \rightarrow \{0,1\}^{m+m''} \) by

\[
\text{Ext}((x,y),(s,t)) = \left( \text{Ext}_{\text{comp}}((x,y),s), \text{Ext}_{\text{waste}}(W_{\text{comp}}((x,y),s),t) \right)
\]

as in Lemma 5.16.

Then \( \text{Ext} \) is an \((n+n'-\log(1/\delta), \varepsilon_{\text{out}} + (1+1/\alpha) \cdot \varepsilon_{\text{in}} + \varepsilon_{\text{waste}})\)-extractor for KL. Furthermore, if \( \text{Ext}_{\text{in}} \) and \( \text{Ext}_{\text{waste}} \) are strong average-case KL-extractors, then \( \text{Ext} \) is a strong KL-extractor, and if \( \text{Ext}_{\text{out}} \) is average-case then so is \( \text{Ext} \).

**Proof.** We claim that \( W_{\text{comp}} \) is such that \( (\text{Ext}_{\text{comp}}, W_{\text{comp}}) \) is an injection: by assumption on \( (\text{Ext}_{\text{out}}, W_{\text{out}}) \) we have that given \( \text{Ext}_{\text{out}}(x, \text{Ext}_{\text{in}}(y,s)) \) and \( W_{\text{out}}(x, \text{Ext}_{\text{in}}(y,s)) \) we can recover \( x \) and \( \text{Ext}_{\text{in}}(y,s) \), and by assumption on \( (\text{Ext}_{\text{in}}, W_{\text{in}}) \) given \( \text{Ext}_{\text{in}}(y,s) \) and \( W_{\text{in}}(y,s) \) we can recover \( (y,s) \), so that \( (\text{Ext}_{\text{comp}}, W_{\text{comp}}) \) has an inverse and is injective as desired. Therefore, since Theorem 5.4 implies \( \text{Ext}_{\text{comp}} \) is an \((n+n'-\log(1/\delta), \varepsilon_{\text{out}} + (1+1/\alpha) \cdot \varepsilon_{\text{in}})\) KL-extractor, the result follows from Lemma 5.16. The furthermore claims follow from the corresponding claims of these lemmas (and Corollary 5.19 for the strong case).

**Remark 5.21.** Corollary 5.20 was presented by Reingold, Vadhan, and Wigderson [RVW00] as a transformation that combined three extractor-condenser pairs into a new extractor-condenser pair. We do not use this generality, so for simplicity we do not present it here, but both Lemma 5.16 and Corollary 5.20 can be easily extended in this manner if required.

The Raz–Reingold–Vadhan [RVV02] transformation to avoid entropy loss follows similarly using the Leftover Hash Lemma (Proposition 5.6).

**Corollary 5.22 (KL-extractor analogue of [RVV02, Lemma 28]).** Let \( \text{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \rightarrow \{0,1\}^{m_1} \) be a strong \((k, \varepsilon/2)\) KL-extractor with entropy loss \( \Delta_1 \), meaning \( m_1 = k - \Delta_1 \). Then for every \( d_{\text{extra}} \leq \Delta_1 \) there is a strong \((k, \varepsilon)\) KL-extractor \( \text{Ext}_2 : \{0,1\}^n \times \{0,1\}^{d_2} \rightarrow \{0,1\}^{m_2} \) with seed length \( d_2 = d_1 + O(d_{\text{extra}} + \log(n/\varepsilon)) \) and entropy loss \( \Delta_2 = d_{\text{extra}} + \log(1/\varepsilon) - O(1) \), meaning \( m_2 = k - (\Delta_2 - d_{\text{extra}}) - \log(1/\varepsilon) + O(1) \), which is computable in polynomial time making one oracle call to \( \text{Ext}_1 \). Furthermore, if \( \text{Ext}_1 \) is average-case then so is \( \text{Ext}_2 \).

In particular, by taking \( d_{\text{extra}} = \Delta_1 \) we get an extractor with optimal entropy loss \( \log(1/\varepsilon) + O(1) \) by paying an additional \( O(\Delta + \log(n/\varepsilon)) \) in seed length.

**Proof.** Let \( W_1 : \{0,1\}^n \times \{0,1\}^{d_1} \rightarrow \{0,1\}^n \) be given by \( W_1(x,s) = x \), and let \( \text{Ext}_2 : \{0,1\}^n \times \{0,1\}^{d_2} \rightarrow \{0,1\}^{m_2} \) be the strong average-case \((d_{\text{extra}}/2, \varepsilon/2)\) KL-extractor of Proposition 5.6 using almost-universal hash functions, so that \( d_2 = O(d_{\text{extra}} + \log(n/\varepsilon)) \) and \( m_2 = d_{\text{extra}} - \log(1/\varepsilon) - O(1) \). The result follows from taking Ext to be the extractor of Corollary 5.19.

**Remark 5.23.** An analogous versions of the above claim for non-strong KL-extractors follows by taking \( W_1(x,s) = (x,s) \) and using Lemma 5.16.

We can apply Corollary 5.22 to Theorem 5.15 to get the KL-extractors from the total variation extractors of Gurusswami, Umans, and Vadhan [GUV09], thereby avoiding the extra \( O(\log(n/\varepsilon)) \) entropy loss in the strong extractors.

**Corollary 5.24.** For every \( n \in \mathbb{N} \), \( 1 > \alpha, \varepsilon > 0 \), and \( k, k' \geq 0 \) with \( k + k' \leq n \), there is an explicit strong average-case \((k+k', \varepsilon)\) KL-extractor \( \text{Ext} : \{0,1\}^n \times \{0,1\}^{d} \rightarrow \{0,1\}^m \) with \( d \leq O_{\alpha}(\log(n/\varepsilon)) + O(k') \) and \( m \geq (1-\alpha)k + k' - \log(1/\varepsilon) - O(1) \).
5.4 Lower bounds

In this section, we give lower bounds on extractors for the Rényi divergences $D_\beta$ of all orders, including the special case $\beta = 1$ of KL-extractors. A reader primarily interested in explicit constructions of subgaussian samplers can skip to Section 6.

For Rényi divergences $D_\beta$ with $\beta \leq 1$ we reduce to Radhakrishnan and Ta-Shma’s [RT00] lower bounds for total variation extractors and dispersers, which can be understood as a one-sided relaxation of total variation extractors.

**Definition 5.25** (Sipser [Sip88], Cohen and Wigderson [CW99]). A function $\text{Disp} : \{0, 1\}^n \times \{0, 1\}^d \to \{0, 1\}^m$ is a $(k, \varepsilon)$ disperser if for all random variables $X$ over $\{0, 1\}^n$ with $H_\infty(X) \geq k$, it holds that $|\text{Supp}(\text{Disp}(X, U_d))| \geq (1 - \varepsilon)2^m$.

Dispersers are of interest in the context of Rényi extractors because the Rényi 0-entropy of a random variable is the logarithm of its support size (see Example 2.4), and hence dispersers are equivalent to $D_0$-extractors:

**Lemma 5.26.** Disp is a $(k, \varepsilon)$ disperser if and only if Disp is a $(k, \log(1/(1 - \varepsilon)))$ $D_0$-extractor.

Given Lemma 5.26, we can use the Radhakrishnan and Ta-Shma [RT00] lower bounds to give an optimal lower bound on the seed length of $D_\beta$-extractors for $\beta \leq 1$ in terms of the error $\varepsilon$, input length $n$ and supported entropy $k$ (we will give a matching non-explicit upper bound in the next section), as well as lower bounds on the entropy loss. For the case $\beta = 1$ of KL-extractors, the non-explicit upper bound (Theorem 5.31) also shows that the entropy loss lower bound is optimal.

**Theorem 5.27.** Let $0 \leq \beta \leq 1$ and $\text{Ext} : \{0, 1\}^n \times \{0, 1\}^d \to \{0, 1\}^m$ be a $(k, \varepsilon)$ extractor for $D_\beta$ with $k \leq n - 2$, $d \leq m - 1$, and $2^{2m} \leq \varepsilon < 1/4$. Then $d \geq \log(n - k) + \log(1/\varepsilon - O(1)$ and $m \leq k + d - \log(1/\varepsilon) + O(1)$. Furthermore, if $\varepsilon$ is at most $\beta/(2\ln 2)$ then $m \leq k + d - \log(1/\varepsilon) + \log(1/\beta) + O(1)$.

**Proof.** Since $D_\beta$ is nondecreasing in $\beta$, we have that $\text{Ext}$ is a $(k, \varepsilon)$ extractor for $D_0$, and thus by Lemma 5.26 it is a $(k, 1 - 2^{-\varepsilon})$ disperser. Then the disperser seed length lower bound of Radhakrishnan and Ta-Shma [RT00] tells us that $d \geq \log(n - k) + \log(1/(1 - 2^{-\varepsilon})) - O(1) \geq \log(n - k) + \log(1/\varepsilon) - O(1)$ and $m \leq k + d - \log(1/(1 - 2^{-\varepsilon})) + O(1) \leq k + d - \log(1/\varepsilon) + O(1)$.

For the other entropy loss lower bound, we use Gilardoni’s [Gil10] generalization of Pinsker’s inequality, which shows in particular that $d_{TV}(P, U_m) \leq \sqrt{\ln 2/(2\beta)} \cdot D_\beta(P \| U_m)$. Thus, $\text{Ext}$ is also a $(k, \sqrt{\varepsilon \ln 2/(2\beta)})$ total variation extractor, and if $\sqrt{\varepsilon \ln 2/(2\beta)} \leq 1/2$ (equivalently $\varepsilon \leq \beta/(2\ln 2)$) then the [RT00] total variation extractor entropy loss lower bound implies that $m \leq k + d - 2\log(1/\sqrt{\varepsilon \ln 2/(2\beta)}) + O(1) \leq k + d - \log(1/\varepsilon) + \log(1/\beta) + O(1)$.

**Remark 5.28.** For the case of $0 < \beta < 1$, we do not know whether the entropy loss lower bound of Theorem 5.27 is tight.

It is well-known that $\ell_2$-extractors (which are equivalent to $D_2$-extractors by Example 2.4) require seed length at least linear in $\min(n - k, m)$ (see e.g. [Vad12] Problem 6.4). We generalize this to give a seed length lower bound on $D_\beta$ extractors for all $\beta > 1$, in the regime of constant $\varepsilon$.

**Theorem 5.29.** Let $\text{Ext} : \{0, 1\}^n \times \{0, 1\}^d \to \{0, 1\}^m$ be a $(k, 0.99)$ $D_{1 + \alpha}$-extractor for $\alpha > 0$. Then $d \geq \min\{(n - k - 3) \cdot \alpha, (m - 2) \cdot \alpha/(\alpha + 1)\}$.

**Proof.** We follow the strategy suggested by Vadhan [Vad12] Problem 6.4, and view Ext as a bipartite graph with $N = \{0, 1\}^n$ left-vertices, $M = \{0, 1\}^m$ right-vertices, and $D = 2^d$ edges per left-vertex given by $E = \{(x \in \{0, 1\}^n, y \in \{0, 1\}^m) \mid \exists s \in \{0, 1\}^d : \text{Ext}(x, s) = y\}$.

Assume for the sake of contradiction that $d \leq \alpha/(\alpha + 1) \cdot (m - 2)$ and $d \leq \alpha(n - k - 3)$, so that $M \geq 4D^{1 + 1/\alpha}$ and $N/(8D^{1/\alpha}) \geq K$. Now, we claim there exists a set $T \subseteq \{0, 1\}^m$ of size at most $M/(2D^{1 + 1/\alpha})$ such that $X = \{x \in \{0, 1\}^n \mid \exists s \in \{0, 1\}^d \text{ s.t. } \text{Ext}(x, s) \in T\}$ has size at least $N/(8D^{1/\alpha}) \geq K$. This follows from the
following iterative procedure: until \(|X| \geq N/(8D^{1/\alpha})\), choose the vertex \(y \in \{0,1\}^m\) of highest degree, add it to \(T\), and remove \(y\) and its neighbors from the graph (the neighbors go in \(X\)). Then at each step we will add to \(X\) a number of vertices at least the average degree

\[
\frac{(N - |X|) \cdot D}{M - |T|} \geq \frac{(N - N/(8D^{1/\alpha})) \cdot D}{M} \geq ND/2M,
\]

so that the size of \(T\) will be at most \([N/(8D^{1/\alpha}) \cdot 2M/ND] = [M/(4D^{1+1/\alpha})]\) \(\leq M/(2D^{1+1/\alpha})\) as desired. Now, since \(X\) has size at least \(K\) and \(\text{Ext}\) is a \((k,0.99)\) \(D_{1+\alpha}\)-extractor, we have that

\[
0.99 \geq D_{1+\alpha}(\text{Ext}(U_X,U_d) \parallel U_m)
= \frac{1}{\alpha} \log \left( \sum_{y \in \{0,1\}^m} \Pr[\text{Ext}(U_X,U_d) = y]^{1+\alpha} 2^{-\alpha m} \right)
\geq \frac{1}{\alpha} \log \left( M^\alpha \sum_{y \in T} \Pr[\text{Ext}(U_X,U_d) = y]^{1+\alpha} \right)
\geq \frac{1}{\alpha} \log \left( M^\alpha \cdot |T|^{-\alpha} \cdot \left( \sum_{y \in T} \Pr[\text{Ext}(U_X,U_d) = y] \right)^{1+\alpha} \right) \quad \text{(By Hölder’s inequality)}
\geq \frac{1}{\alpha} \log \left( M^\alpha \cdot (M/(2D^{1+1/\alpha}))^{-\alpha} \cdot (1/D)^{1+\alpha} \right) = 1 \quad \text{(By definition of \(T\)}
\]

which is a contradiction, as desired. \(\square\)

We can also use this lower bound to get a similar lower bound for \(d_{\ell_2}\)-extractors for all \(\alpha > 0\), though in this case the lower bound applies up to an error threshold that depends on \(\alpha\).

**Corollary 5.30.** Let \(\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m\) be a \((k,\varepsilon_\alpha,2^{-\alpha m}/(1+\alpha))\) extractor for \(d_{\ell_2}\) where \(\alpha > 0\) and \(\varepsilon_\alpha = (2/3) \cdot \alpha/(\alpha+1)\). Then \(d \geq \min\{(n-k-3) \cdot \alpha, (m-2) \cdot \alpha/(\alpha+1)\}\).

**Proof.** Note that the proof of Theorem 5.29 gave a lower bound on the sum \(\sum_{y \in \{0,1\}^m} P_{y}^{1+\alpha}\) where \(P = \text{Ext}(U_X,U_d)\), whereas \(d_{\ell_2}(P,U_m)^{1+\alpha} = \sum_{y \in \{0,1\}^m} |P_y - 2^{-m}|^{1+\alpha}\). For \(\ell_2\) these can be related without any loss, but in general we can use the triangle inequality to get

\[
D_{1+\alpha}(P \parallel U_m) \leq \frac{1}{\alpha} \log \left( 2^{\alpha m} \cdot \left( d_{\ell_2}(P,U_m) + 2^{-\alpha m/(\alpha+1)} \right)^{1+\alpha} \right)
\]

so that if \(d_{\ell_2}(P,U_m) \leq \varepsilon_{\alpha} \cdot 2^{-\alpha m/(1+\alpha)}\) where \(\varepsilon_{\alpha} = (2/3) \cdot \alpha/(\alpha+1) \leq 2^{0.99 \cdot \alpha/(\alpha+1)} - 1\), then \(D_{1+\alpha}(P \parallel U_m) \leq 0.99\), and we conclude by Lemma 5.1 and Theorem 5.29. \(\square\)

### 5.5 Non-explicit construction

In this section, we show non-constructively the existence of KL-extractors matching the lower-bound of Theorem 5.27 and in particular implying the optimal parameters of standard extractors for total variation distance. Formally, we will prove:

**Theorem 5.31.** For every \(n \in \mathbb{N}, k \leq n\), and \(1 > \varepsilon > 0\) there is an average-case (respectively strong average-case) \((k,\varepsilon)\) KL-extractor \(\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m\) with seed length \(d = \log(n-k+1) + \log(1/\varepsilon) + O(1)\) and output length \(m = k + d - \log(1/\varepsilon) + O(1)\) (respectively \(m = k - \log(1/\varepsilon) - O(1)\)).

**Remark 5.32.** For \(\varepsilon \gg 1\) the above parameters are not necessarily optimal, and it would be interested to get matching upper and lower bounds in this regime of parameters.
We will prove Theorem 5.31 using the probabilistic method, analogously to Zuckerman [Zuc97] or Radhakrishnan and Ta-Shma [RT00] for total variation extractors. However, rather than using Hoeffding’s inequality, we use the following lemma:

**Lemma 5.33.** Let $X$ be uniform over a subset of $\{0,1\}^n$ of size $K$. Then if $\Ext : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a random function, it holds for every $\varepsilon > 0$ that

$$\Pr_{s \sim U_d} \left[ \mathbb{E}_{\Ext} \left[ \KL(\Ext(X, s) \parallel U_m) \right] > \varepsilon \right] \leq 2^{MD - KD\varepsilon^3/3}$$

where $D = 2^d$ and $M = 2^m$.

**Remark 5.34.** For total variation extractors, the analogous bound is

$$\Pr_{\Ext} \left[ d_{TV} \left( (U_d, \Ext(X, U_d)), (U_d, U_m) \right) > \varepsilon \right] \leq 2^{MD - 2KD\varepsilon^2/\ln 2}.$$  

One sees that the bounds are very similar, except the KL divergence version depends on $\varepsilon$ rather than $\varepsilon^2$. For the regime where $\varepsilon < 1$ the linear dependence is preferable, and is responsible for the $1 \cdot \log(1/\varepsilon)$ seed length for KL-extractors compared to the $2 \cdot \log(1/\varepsilon)$ seed length for total variation extractors.

**Proof of Lemma 5.33.** Note that for each $s \in \{0,1\}^d$ and fixed $\Ext$, the random variable $\Ext(X, s)$ is uniform over the multiset $\{\Ext(x, s) \mid x \in \text{Supp}(X)\}$. Hence, since $\Ext$ is a random function, this multiset is distributed exactly as taking $K$ iid uniform samples from $\{0,1\}^m$, so we wish to bound the KL divergence between this empirical distribution and the true distribution. For this, the author [Agr19] gave the moment generating function bound

$$\mathbb{E}_{\Ext} \left[ 2^{t \cdot \KL(\Ext(X, s) \parallel U_m)} \right] \leq \left( \frac{2^{t/K}}{1 - t/K} \right)^{M-1}$$

for every $0 \leq t < K$, which for $t = K/3$ is at most $2^M$. Then since $\Ext(X, s)$ is independent across $s \in \{0,1\}^d$, we have

$$\Pr_{\Ext} \left[ \mathbb{E}_{s \sim U_d} \left[ \KL(\Ext(X, s) \parallel U_m) \right] > \varepsilon \right] = \Pr_{\Ext} \left[ 2^{K/3 \cdot \sum_{s \in \{0,1\}^d} \KL(\Ext(X, s) \parallel U_m)} > 2^{K/3 \cdot D\varepsilon} \right]$$

$$\leq 2^{-K\varepsilon^3/3} \prod_{i=1}^{D} 2^M \quad \square$$

We can now prove Theorem 5.31.

**Proof of Theorem 5.31.** We will show that a random function $\Ext : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a strong average-case $(k,\varepsilon)$ KL-extractor with positive probability, the non-strong version then follows from Lemma 5.2. By Lemma 3.13 it is enough to prove that $\Ext$ is a strong $(k-t, 2^{t+1}/3 \cdot \varepsilon)$ KL-extractor for every $t \geq 0$. To reduce the range of $t$ we need to consider, note that it suffices to be a $(\log(2^{k-t}) \cdot 2^{t+1}/3 \cdot \varepsilon)\text{- extractor for every } t \geq 0$, so that by rounding down it is enough to be a $(k-t, 2^{t}/3 \cdot \varepsilon)$ strong KL-extractor for each $t \geq 0$ such that $2^{k-t}$ is an integer.

Now, consider a fixed $t \geq 0$ such that $2^{k-t}$ is an integer. Since the KL divergence is convex in its first argument and all distributions of min-entropy at least $k-t$ are convex combinations of “flat” distributions which are uniform over a set of size $2^{k-t}$ (Chor and Goldreich [CG88]), it suffices to analyze the behavior of $\Ext$ on such distributions. Then for every subset $X \subseteq \{0,1\}^n$ of size $2^{k-t}$, Lemma 5.33 tells us that

$$\Pr_{\Ext} \left[ \mathbb{E}_{s \sim U_d} \left[ \KL(\Ext(U_X, s) \parallel U_m) \right] > 2^{t}/3 \cdot \varepsilon \right] \leq 2^{MD - 2^{k-t} \cdot D \cdot (2^{t}/3 \cdot \varepsilon)/3} = 2^{MD - KD\varepsilon/9}$$

25
where $M = 2^m$, $D = 2^d$, and $K = 2^k$. There are $\sum_{j=0}^{K} \binom{N}{j}$ such subsets $X$ of $\{0,1\}^n$ for which we simultaneously need to establish that $E_{x \sim U_2} [KL(Ext(U_X, s) \| U_m)] \leq 2^{1/3} \cdot \varepsilon$, so we have by a union bound that the probability that Ext is not a strong average-case $(k, \varepsilon)$ KL-extractor is at most

$$2^{MD - KD\varepsilon/9} \cdot \sum_{j=0}^{K} \binom{N}{j} \leq 2^{MD - KD\varepsilon/9} \cdot \left( \frac{Ne}{K} \right)^{K} = 2^{MD + K \log(Ne/K) - KD\varepsilon/9}. $$

Hence, as long as

$$MD < \frac{KD\varepsilon}{18}$$

$$m \leq k - \log(1/\varepsilon) - O(1)$$

$$K \log \left( \frac{Ne}{K} \right) < \frac{KD\varepsilon}{18}$$

$$d \geq \log(n - k + 1) + \log(1/\varepsilon) + O(1)$$

we know that a random function is a strong average-case $(k, \varepsilon)$ KL-extractor with positive probability as desired.

\[\square\]

6 Constructions of subgaussian samplers

6.1 Subconstant $\varepsilon$ and $\delta$

The goal of this section is to establish the following theorem, which is our explicit construction of subgaussian samplers with sample complexity having no dependence on $m$, and with randomness complexity and sample complexity matching the best-known $[0,1]$-valued sampler when $\varepsilon$ and $\delta$ are subconstant (up to the hidden polynomial in the sample complexity).

**Theorem 6.1.** For all $m \in \mathbb{N}$, $1 > \varepsilon, \delta > 0$, and $\alpha > 0$ there exists an explicit $(\delta, \varepsilon)$ absolute averaging sampler (respectively strong absolute averaging sampler) for subgaussian and subexponential functions $\text{Samp} : \{0,1\}^n \to \{0,1\}^D$ with sample complexity $D = \text{poly}(\log(1/\delta), 1/\varepsilon)$ and randomness complexity $n = m + (1 + \alpha) \cdot \log(1/\delta)$ (respectively $n = m + (1 + \alpha) \cdot \log(1/\delta) + 2 \log(1/\varepsilon) + O(1)$).

We will use essentially the same construction used for bounded samplers in this regime, namely applying the Reingold, Wigderson, and Vadhan [RVW00] zig-zag product for extractors to combine the expander extractor of Goldreich and Wigderson [GW97] and an extractor with logarithmic seed length. However, as described in detail in Section 4.1 even the basic composition used in this construction does not work for general subgaussian extractors, so we will instead use the zig-zag product for KL-extractors (Corollary 5.20) combining extractors for Rényi divergences, specifically the $D_2$-extractor from Proposition 5.8 and the KL-extractor from Corollary 5.24 to get the following high-entropy KL-extractor:

**Theorem 6.2.** For all integers $m$ and $1 > \alpha, \delta, \varepsilon > 0$ there is an explicit average-case (respectively strong average-case) $(k, \varepsilon)$ KL-extractor $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $n = m + (1 + \alpha) \log(1/\delta) - O(1)$ (respectively $n = m + (1 + \alpha) \log(1/\delta) + \log(1/\varepsilon) + O(1)$), $k = n - \log(1/\delta)$, and $d = O_{\alpha}(\log(1/\delta)/\varepsilon)$.

**Proof.** We prove the claim for strong extractors, for the non-strong claim one can simply define $\text{Ext}(x, (s, t)) = \text{Ext}_{\text{strong}}((x, s), t)$ where $t$ has length $\log(1/\varepsilon) + O(1)$.

By Corollary 5.11, there is a universal constant $C > 0$ such that for $d_{\text{out}} = \lceil C \log(1/(\delta \varepsilon)) \rceil \leq C \log(1/\delta) + C \log(1/\varepsilon) + 1$ there is an explicit average-case $n_{\text{out}} = \lceil n_{\text{out}} - \log(1/\delta), \varepsilon/4 \rceil$ $\text{D}_2$-extractor $\text{Ext}_{\text{out}} : \{0,1\}^{n_{\text{out}}} \to \{0,1\}^{d_{\text{out}}}$ with $n_{\text{out}} = m - d_{\text{out}}$. Furthermore, $\text{Ext}_{\text{out}}$ has the property that the function $W_{\text{out}}(x, s) = s$ is such that $(\text{Ext}_{\text{out}}, W_{\text{out}})$ is an injection.

Let $k''_{\text{in}} = C (\log(1/\delta))/(1 - \beta)$, $k''_{\text{in}} = (C + 1) \log(1/\varepsilon) + O(1)$, and $k_{\text{in}} = k''_{\text{in}} + k''_{\text{in}}$ for $0 < \beta < 1$ some parameter to be chosen later. Then by Corollary 5.24 there is an explicit $(k_{\text{in}}, \varepsilon/4)$ strong average-case KL-extractor $\text{Ext}_{\text{in}} : \{0,1\}^{n_{\text{in}}} \times \{0,1\}^{d_{\text{in}}} \to \{0,1\}^{m_{\text{in}}}$ with $n_{\text{in}} = k_{\text{in}} + \log(1/\delta), d_{\text{in}} = O_{\beta}(\log(n_{\text{in}}/\varepsilon)) + O(k''_{\text{in}}) = O_{\beta}(\log(1/\varepsilon))$, and $m_{\text{out}} = (1 - \beta) k''_{\text{in}} + k''_{\text{in}} \log(1/\varepsilon) - O(1) = d_{\text{out}}$. Furthermore, the function $W_{\text{in}}(x, s) = (x, s)$ is an injection.
Furthermore, for \( k_{\text{waste}} = (n_{\text{out}} + n_{\text{in}} - \log(1/\delta)) - n_{\text{out}} = n_{\text{in}} - \log(1/\delta) = k_{\text{in}} = k'_{\text{in}} + k''_{\text{in}} \), by Corollary 5.24 there is also an explicit \((k_{\text{waste}}, \varepsilon/4)\) strong average-case KL-extractor \( \text{Ext}_{\text{waste}} : \{0, 1\}^{n_{\text{out}} + n_{\text{in}} + d_{\text{in}}} \times \{0, 1\}^{d_{\text{waste}}} \rightarrow \{0, 1\}^{m_{\text{waste}}} \) such that \( m_{\text{waste}} = d_{\text{out}} \) and \( d_{\text{waste}} = O(\log((d_{\text{out}} + n_{\text{in}} + d_{\text{in}})/\varepsilon)) + O(k''_{\text{in}}) = O(\log(\log(1/\delta)/\varepsilon)) \).

Then by the zig-zag product for KL-extractors (Corollary 5.20), there is an explicit \((n_{\text{out}} + n_{\text{in}} - \log(1/\delta), \varepsilon)\) strong average-case KL-extractor \( \text{Ext} : \{0, 1\}^{n_{\text{out}} + n_{\text{in}}} \times \{0, 1\}^{d_{\text{in}}} + d_{\text{waste}} \rightarrow \{0, 1\}^{n_{\text{out}} + m_{\text{waste}}} \), where we have

\[
\begin{align*}
  n_{\text{out}} + n_{\text{in}} &= (m - d_{\text{out}}) + \left( (C \log(1/\delta)/(1 - \beta) + (C + 1) \log(1/\varepsilon) + O(1)) + \log(1/\delta) \right) \\
  &\leq m + \log(1/\delta) + \log(1/\varepsilon) + \log(1/\delta) \cdot C \cdot (1/(1 - \beta) - 1) + O(1) \\
  d_{\text{in}} + d_{\text{waste}} &= O_{\beta}(\log(1/\delta)/\varepsilon)) \\
  n_{\text{out}} + n_{\text{waste}} &= (m - d_{\text{out}}) + d_{\text{out}} = m.
\end{align*}
\]

Choosing \( \beta = \alpha/(\alpha + C) \leq \alpha/(1 + C) \) so that \( C \cdot (1/(1 - \beta) - 1) \leq \alpha \) gives the claim.

We can now prove Theorem 6.1

**Proof of Theorem 6.1.** Let \( \text{Ext} : \{0, 1\}^{n} \times \{0, 1\}^{d} \rightarrow \{0, 1\}^{m} \) be the explicit \((n - \log(1/(\delta/2)), \varepsilon^2)\) KL-extractor (respectively strong KL-extractor) of Theorem 6.2 so that \( d = O_{\text{a}}(\log \log(1/\delta)/\varepsilon) \) and \( n = m + (1 + \alpha) \log(1/\delta) \) (respectively \( n = m + 2 \log(1/\delta)/\varepsilon + O(1) \)).

Then by Lemmas 4.9 and 5.1 \( \text{Ext} \) is also an \((n - \log(1/(\delta/2)), \varepsilon)\) \( d_{\text{E}} \)-extractor (respectively strong \( d_{\text{E}} \)-extractor), so by Theorem 3.8 the function \( \text{Samp} : \{0, 1\}^{n} \times \{(0, 1)^{m}\}^{D} \rightarrow \{0, 1\}^{m} \) defined as \( \text{Samp}(h)_{d} = h(d) \) where \( h \) is drawn at random from a size \( 2^n \) pairwise-independent hash family \( \mathcal{H} \) of functions from \( [D] \rightarrow \{0, 1\}^{m} \).

**Proof.** The fact that pairwise independence gives rise to a strong bounded-variance sampler is immediate by Chebyshev’s inequality. The existence of pairwise independent hash functions with the claimed parameters is due to Chor and Goldreich [CG89], with similar constructions in the probability literature dating back to Joffe [Jof71].

**6.2 Constant \( \delta \)**

We also recall from the introduction that the pairwise independent sampler of Chor and Goldreich works for subgaussian functions, and in fact the more general class of functions with bounded variance. The sampler has exponentially worse dependence on \( \delta \) than is necessary for subgaussian samplers but optimal randomness complexity and dependence on \( \varepsilon \), so this sampler is optimal for constant \( \delta \).

**Theorem 6.3 (CG89).** For all \( m \in \mathbb{N} \) and \( 1 > \varepsilon, \delta > 0 \) there is an explicit strong sampler \( \text{Samp} : \{0, 1\}^{n} \rightarrow (\{0, 1\}^{m})^{D} \) for functions with bounded variance \( \mathcal{M}_{2} \), with randomness complexity \( n = m + 2 \log(1/\varepsilon) + \log(1/\delta) + O(1) \) and sample complexity \( D = O\left(\frac{1}{\varepsilon^{2}}\right) \) defined as \( \text{Samp}(h)_{d} = h(d) \) where \( h \) is drawn at random from a size \( 2^{n} \) pairwise-independent hash family \( \mathcal{H} \) of functions from \( [D] \rightarrow \{0, 1\}^{m} \).

**Proof.** The fact that pairwise independence gives rise to a strong bounded-variance sampler is immediate by Chebyshev’s inequality. The existence of pairwise independent hash functions with the claimed parameters is due to Chor and Goldreich [CG89], with similar constructions in the probability literature dating back to Joffe [Jof71].

**6.3 Non-explicit construction**

Applying Lemmas 4.9 and 5.1 to Theorem 6.3, our non-explicit construction of KL-extractors gives:

**Corollary 6.4.** For every \( n \in \mathbb{N}, k \leq n, \) and \( 1 > \varepsilon > 0 \) there is an average-case (respectively strong average-case) \((k, \varepsilon)\) \( d_{\text{E}} \)-extractor \( \text{Ext} : \{0, 1\}^{n} \times \{0, 1\}^{d} \rightarrow \{0, 1\}^{m} \) with \( d = \log(n - k + 1) + 2 \log(1/\varepsilon) + O(1) \)

and \( m \geq k + d - 2 \log(1/\varepsilon) - O(1) \) (respectively \( m \geq k - 2 \log(1/\varepsilon) - O(1) \))

Since \( d_{\text{E}} \)-extractors are also total variation extractors, Corollary 6.4 is optimal up to additive constants by the lower bound of Radhakrishnan and Ta-Shma [RT00].

Using the fact that extractors are samplers (Theorem 5.8), we get
Corollary 6.5. For every integer \( m \) and \( 1 > \delta, \varepsilon > 0 \) there is a \((\delta, \varepsilon)\) sampler (respectively strong sampler) \( \text{Samp} : \{0, 1\}^n \rightarrow (\{0, 1\}^m)^D \) for subgaussian and subexponential functions with sample complexity \( D = O\left(\log \frac{1}{\delta \varepsilon} \right) \) and randomness complexity \( n = m + \log(1/\delta) - \log \log(1/\delta) + O(1) \) (respectively \( n = m + \log(1/\delta) + 2 \log(1/\varepsilon) + O(1) \)).

Note that this matches the best-known (non-explicit) parameters of averaging samplers for \([0, 1]\)-valued functions due to Zuckerman \cite{Zuc97}.

7 Acknowledgements

The author would like to thank Jarosław Błasiok for suggesting the problem of constructing subgaussian samplers and for helpful discussions and feedback, and Salil Vadhan for many helpful discussions and his detailed feedback on this writeup.

References

\[ \text{AE05} \quad \text{K. M. R. Audenaert and J. Eisert, “Continuity bounds on the quantum relative entropy,” } \text{Journal of Mathematical Physics,} \text{ vol. 46, no. 10, p. 102104, Oct. 2005.} \]

\[ \text{Agr19} \quad \text{R. Agrawal, “Concentration of the multinomial in Kullback–Leibler divergence near the ratio of alphabet and sample sizes,” } \text{arXiv:1904.02291 [cs, math, stat], Apr. 2019.} \]


\[ \text{BGG93} \quad \text{M. Bellare, O. Goldreich, and S. Goldwasser, “Randomness in interactive proofs,” } \text{computational complexity,} \text{ vol. 3, no. 4, pp. 319–354, Dec. 1993.} \]

\[ \text{Bla18a} \quad \text{J. Blasiok, Private Communication, Cambridge, MA USA, 2018.} \]


\[ \text{BLM13} \quad \text{S. Boucheron, G. Lugosi, and P. Massart, } \text{Concentration Inequalities: A Nonasymptotic Theory of Independence,} \text{ 1st ed. Oxford University Press, Feb. 2013.} \]

\[ \text{BR94} \quad \text{M. Bellare and J. Rompel, “Randomness-efficient oblivious sampling,” in } \text{Proceedings 35th Annual Symposium on Foundations of Computer Science,} \text{ Nov. 1994, pp. 276–287.} \]

\[ \text{CEG95} \quad \text{R. Canetti, G. Even, and O. Goldreich, “Lower bounds for sampling algorithms for estimating the average,” } \text{Information Processing Letters,} \text{ vol. 53, no. 1, pp. 17–25, Jan. 1995.} \]

\[ \text{CG88} \quad \text{B. Chor and O. Goldreich, “Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity,” } \text{SIAM Journal on Computing,} \text{ vol. 17, no. 2, pp. 230–261, Apr. 1988.} \]


