Samplers and Extractors for Unbounded Functions

Rohit Agrawal
John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, MA 02138, USA
`https://rohitagr.com`
rohitagr@seas.harvard.edu

Abstract

Błasiok (SODA'18) recently introduced the notion of a subgaussian sampler, defined as an averaging sampler for approximating the mean of functions $f : \{0,1\}^m \to \mathbb{R}$ 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 ε and failure probability δ 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.

2012 ACM Subject Classification Theory of computation → Expander graphs and randomness extractors; Theory of computation → Pseudorandomness and derandomization; Mathematics of computing → Information theory

Keywords and phrases averaging samplers, subgaussian samplers, randomness extractors, Kullback–Leibler divergence

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2019.59

Category RANDOM

Related Version The full version of this paper is available at `https://arxiv.org/abs/1904.08391` [1].

Funding *Rohit Agrawal*: Supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

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

1 Introduction

1.1 Averaging samplers

Averaging (or oblivious) samplers, introduced by Bellare and Rompel [6], 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** ([6]). *A function* Samp : $\{0,1\}^n \to (\{0,1\}^m)^D$ *is a* (δ,ε) averaging sampler *if for all* $f : \{0,1\}^m \to [0,1]$*, it holds that*

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

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* $\mathrm{Samp}(x)_i$ *can be computed in time* $\mathrm{poly}(n, m, \log D)$*.*

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 Table 1. We recommend the survey of Goldreich [17] for more details, especially regarding non-averaging samplers[1].

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

Key Idea	Randomness complexity n	Sample complexity D	Best regime
Pairwise-independent Expander Neighbors [19]	$m + O(\log(1/\delta) + \log(1/\varepsilon))$	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	$\delta = \Omega(1)$
Ramanujan Expander Neighbors[a)] [22, 19]	m	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	$\delta = \Omega(1)$
Extractors [40, 19, 30, 20]	$m + (1+\alpha)\cdot\log(1/\delta)$ any constant $\alpha > 0$	$\mathrm{poly}(\log(1/\delta), 1/\varepsilon)$	$\varepsilon, \delta = o(1)$
Expander Walk Chernoff [16]	$m + O(\log(1/\delta)/\varepsilon^2)$	$O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$	$\varepsilon = \Omega(1)$
Pairwise Independence [12]	$O(m)$	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	None, but simple
Non-Explicit [40]	$m + \log(1/\delta) - \log\log(1/\delta) + O(1)$	$O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$	All
Lower Bound [11, 40, 27]	$m + \log(1/\delta) + \log(1/\varepsilon) - \log(D) - O(1)$	$\Omega\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$	N/A

a) Requires explicit constructions of Ramanujan graphs.

However, averaging samplers can also have uses beyond bounded functions: Błasiok [9], 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 contained in an interval of length $O(\sqrt{m})$, and thus one can construct a subgaussian sampler from a $[0,1]$-sampler by simply scaling the error ε by a factor of $O(\sqrt{m})$. Unfortunately, looking at Table 1 one sees that this

[1] A non-averaging sampler is an algorithm 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.

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 [12] based on pairwise independence (as its analysis requires only bounded variance) and (as we will show) the Ramanujan Expander Neighbor sampler of [22, 19], but Błasiok showed [8] 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 [5] 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[1], 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 ε and δ are both subconstant, and in particular has no dependence on m in the sample complexity. We also show non-constructively that subexponentially samplers exist with essentially the same parameters as $[0,1]$-valued samplers.

▶ **Theorem 1.2** (Informal version of Theorem 6.1)**.** *For every integer $m \in \mathbb{N}$ and $1 > \delta, \varepsilon > 0$, there is an explicit subgaussian (in fact subexponential) sampler with randomness complexity $n = m + O(\log(1/\delta))$ and sample complexity $D = \mathrm{poly}(\log(1/\delta), 1/\varepsilon)$.*

In the full version of this work [1], we show also that for every $m \in \mathbb{N}$, $1 > \delta, \varepsilon > 0$, and $\alpha > 0$, there is a function $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ that is:

- *an explicit subexponential sampler with randomness complexity $n = m + (1+\alpha)\cdot\log(1/\delta)$ and sample complexity $D = \mathrm{poly}(\log(1/\delta), 1/\varepsilon)$.*
- *a non-constructive subexponential sampler with randomness complexity $n = m+\log(1/\delta) - \log\log(1/\delta) + O(1)$ and sample complexity $D = O(\log(1/\delta)/\varepsilon^2)$.*

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 [40] showed that averaging samplers 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 [26])**.** *A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k, ε)* extractor *if for every distribution X over $\{0,1\}^m$ satisfying $2^{-k} \geq \max_{x\in\{0,1\}^n} \Pr[X = x]$, the distributions $\mathrm{Ext}(X, U_d)$ and U_m are ε-close in total variation distance. Equivalently, for all $f : \{0,1\}^m \to [0,1]$ it holds that $\mathbb{E}[f(\mathrm{Ext}(X,U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon$. 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 $f : \{0,1\}^m \to \mathbb{R}$ which need not be bounded.

▶ **Definition 1.4** (Special case of Definition 3.1 using Definition 2.5)**.** *A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k, ε) 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 $\mathbb{E}[f(\mathrm{Ext}(X,U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon$.*

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 [40] to show that extractors for a class of functions of $\mathcal{F}$ are also samplers for that class, 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* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is said to be a* (k,ε) *KL-extractor if for every distribution* X *over* $\{0,1\}^m$ *satisfying* $\max_{x\in\{0,1\}^n} \Pr[X = x] \le 2^{-k}$ *it holds that* $\mathrm{KL}(\mathrm{Ext}(X,U_d) \parallel U_m) \le \varepsilon$, *or equivalently* $\mathrm{H}(\mathrm{Ext}(X,U_d)) \ge m - \varepsilon$.

A strong form of Pinsker's inequality (e.g. [10, Lemma 4.18]) implies that a (k,ε^2) KL-extractor is also a (k,ε) 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 show in the full version of this paper that the zig-zag product for extractors of Reingold, Wigderson, and Vadhan [30] also works for KL-extractors, and therefore 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 δ 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 6.5). *For all integers* m *and* $1 > \delta, \varepsilon > 0$ *there is an explicit* (k,ε) *KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $n = m + O(\log(1/\delta))$, $k = n - \log(1/\delta)$, *and* $d = O(\log\log(1/\delta) + \log(1/\varepsilon))$.

In the full version, we show that n *can be as small as* $m + (1+\alpha)\cdot\log(1/\delta)$ *for any constant* $\alpha > 0$.

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 [20] 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** (Formal statement and proof in full version [1]). *For any integers* $k < n \in \mathbb{N}$ *and* $1 > \varepsilon > 0$ *there is a* (k,ε) *KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *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 [27]. In particular, by Pinsker's inequality, (k, ε^2) KL-extractors with the above parameters are also optimal (k, ε) standard (total variation) extractors [27], 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 argument that a random function should directly be a subgaussian sampler, as it avoids the need to take a union bound over $O(M^M) = O(2^{M \log M})$ test functions (for $M = 2^m$) which results in additional additive $\log \log$ factors in the randomness complexity.

In the total variation setting, there are only a couple of methods known to explicitly achieve optimal entropy loss $2 \cdot \log(1/\varepsilon)$, the easiest of which is to use an extractor which natively has this sort of loss, of which only three are known: An extractor from random walks over Ramanujan Graphs due to Goldreich and Wigderson [19], the Leftover Hash Lemma due to Impagliazzo, Levin, and Luby [21] (see also [23, 7]), and the extractor based on almost-universal hashing of Srinivasan and Zuckerman [33]. Unfortunately, all of these are ℓ_2 extractors and so must have seed length linear in $\min(n-k, m)$ (cf. [35, Problem 6.4]), rather than logarithmic in $n-k$ as known non-constructively. The other alternative is to use the generic reduction of Raz, Reingold, and Vadhan [28] which turns any extractor Ext with entropy loss Δ into one with entropy loss $2 \cdot \log(1/\varepsilon) + O(1)$ by paying an additive $O(\Delta + \log(n/\varepsilon))$ in seed length. We show in the full version of this paper that all of these ℓ_2 extractors and the [28] transformation also work to give KL-extractors with entropy loss $1 \cdot \log(1/\varepsilon) + O(1)$, so that applications which require minimal entropy loss can also use explicit constructions of KL-extractors.

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 ε of only $1 \cdot \log(1/\varepsilon)$, trying to explicitly construct a KL-extractor with seed length $1 \cdot \log(1/\varepsilon) + o(\min(n-k, k))$ may also shed light on how to achieve optimal dependence on ε 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 [16] for $[0,1]$-valued functions, which achieves randomness complexity $m + O(\log(1/\delta))$ and sample complexity $O(\log(1/\delta))$, as asked for by Błasiok [9]. From the extractor point-of-view, it would suffice (by the reduction of [19, 30] that we analyze for KL-extractors) to construct explicit *linear degree* KL-extractors with parameters matching the linear degree extractor of Zuckerman [41], 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 [34] based on Reed–Muller codes, thereby achieving sample complexity $O(\log(1/\delta) \cdot \operatorname{poly} \log \log(1/\delta))$.

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, cf. [4]) 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, ε) KL-extractors with entropy loss only $1 \cdot \log(1/\varepsilon)$ directly, rather than a total variation extractor or (k, ε^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* X *to* $\mathbb{R} \cup \{\pm\infty\}$*. We write* $\mathrm{D}(P \parallel Q)$ *for the value of* D *on distributions* P *and* Q*. Furthermore*

1. *If* $\mathrm{D}(P \parallel Q) \geq 0$ *with equality iff* $P = Q$*, then* D *is* positive-definite*, and we simply call* D *a* divergence*.*
2. *If* $\mathrm{D}(P \parallel Q) = \mathrm{D}(Q \parallel P)$*, then* D *is* symmetric*.*
3. *If* $\mathrm{D}(P \parallel R) \leq \mathrm{D}(P \parallel Q) + \mathrm{D}(Q \parallel R)$*, then* D *satisfies the* triangle inequality*.*
4. *If* $\mathrm{D}(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda\,\mathrm{D}(P_1 \parallel Q_1) + (1-\lambda)\,\mathrm{D}(P_2 \parallel Q_2)$ *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* $\mathrm{D}(f(P) \parallel f(Q)) \leq \mathrm{D}(P \parallel Q)$*, 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 ℓ_p *distance* for $p > 0$ between probability distributions over $\mathcal{X}$ is

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

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 ℓ_p distance is nonincreasing in p.

▶ **Example 2.3.** The *total variation distance* is

$$d_{TV}(P,Q) \stackrel{\text{def}}{=} \frac{1}{2} d_{\ell_1}(P,Q) = \sup_{S\subseteq\mathcal{X}} \Big|\Pr[P \in S] - \Pr[Q \in S]\Big| = \sup_{f\in[0,1]^{\mathcal{X}}} \Big(\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]\Big)$$

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

▶ **Example 2.4** (Rényi Divergences [31])**.** For two probability distributions P and Q over a finite set $\mathcal{X}$, the *Rényi α-divergence* or *Rényi divergence of order α* is defined for real $0 < \alpha \neq 1$ by

$$\mathrm{D}_\alpha(P \parallel Q) \stackrel{\text{def}}{=} \frac{1}{\alpha - 1} \log\left(\sum_{x\in\mathcal{X}} \frac{P_x^\alpha}{Q_x^{\alpha-1}}\right)$$

where the logarithm is in base 2 (as are all logarithms in this paper unless noted otherwise). The Rényi divergence is continuous in α and so is defined by taking limits for $\alpha \in \{0, 1, \infty\}$, giving for $\alpha = 0$ the divergence $\mathrm{D}_0(P \parallel Q) \stackrel{\text{def}}{=} \log(1/\Pr_{x\sim Q}[P_x \neq 0])$, for $\alpha = 1$ the *Kullback–Leibler (or KL) divergence*

$$\mathrm{KL}(P \parallel Q) \stackrel{\text{def}}{=} \mathrm{D}_1(P \parallel Q) = \sum_{x\in X} P_x \log \frac{P_x}{Q_x},$$

and for $\alpha = \infty$ the *max-divergence* $\mathrm{D}_\infty(P \parallel Q) \stackrel{\text{def}}{=} \max_{x\in X} \log \frac{P_x}{Q_x}$. The Rényi divergence is nondecreasing in α. Furthermore, when $\alpha \le 1$ the Rényi divergence is jointly convex, and for all α the Rényi divergence satisfies the data-processing inequality [37].

When $Q = U_{\mathcal{X}}$ is the uniform distribution over the set $\mathcal{X}$, then for all α, $\mathrm{D}_\alpha(P \parallel U_{\mathcal{X}}) = \log|\mathcal{X}| - \mathrm{H}_\alpha(P)$ where $0 \le \mathrm{H}_\alpha(P) \le \log|\mathcal{X}|$ is called the *Rényi α-entropy of P*. For $\alpha = 0$, $\mathrm{H}_0(P) = \log|\mathrm{Supp}(P)|$ is the *max-entropy of P*, for $\alpha = 1$, $\mathrm{H}_1(P) = \sum_{x\in\mathcal{X}} P_x \log(1/P_x)$ is the *Shannon entropy of P*, and for $\alpha = \infty$, $\mathrm{H}_\infty(P) = \min_{x\in\mathcal{X}} \log(1/P_x)$ is the *min-entropy of P*.

For $\alpha = 2$, the Rényi 2-entropy can be expressed in terms of the ℓ_2-distance to uniform:

$$\log|\mathcal{X}| - H_2(P) = \mathrm{D}_2(P \parallel U_{\mathcal{X}}) = \log\left(1 + |\mathcal{X}| \cdot d_{\ell_2}(P, U_{\mathcal{X}})^2\right)$$

2.2 Statistical weak divergences from test functions

Zuckerman's connection [40] 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 $\mathrm{D}^{\mathcal{F}}$ and is defined as*

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

where we use a superscript to avoid confusion with the Csiszár-Morimoto-Ali-Silvey f-divergences [13, 24, 2].

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 $P \ne Q$ there exists $f \in \mathcal{F}$ with $\mathrm{D}^{\{f\}}(P \parallel Q) > 0$.*

▶ **Example 2.6.** If $\mathcal{F} = \{0,1\}^{\mathcal{X}}$ or $\mathcal{F} = [0,1]^{\mathcal{X}}$, then $\mathrm{D}^{\mathcal{F}}$ 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 $\mathrm{D}^{\{g\}}(P \parallel Q) = -D^{\{f\}}(P \parallel Q) = \mathrm{D}^{\{f\}}(Q \parallel P)$ 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 $\mathrm{D}^{\{g\}}(P \parallel Q) = -\mathrm{D}^{\{f\}}(P \parallel Q) = \mathrm{D}^{\{f\}}(Q \parallel P)$. However, such a class $\mathcal{F}$ gives exactly the same weak divergence $\mathrm{D}^{\mathcal{F}}$ 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 $\mathrm{D}^{\mathcal{F}}(P \parallel Q) = \sup_{f\in\mathcal{F}}\langle P - Q, f\rangle$, which is essentially the notion of indistinguishability considered in several prior works, (see e.g. the survey of Reingold, Trevisan, Tulsiani, and Vadhan [29]), 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 [39] and Müller [25] and is commonly used in developments of Stein's method in probability.

We now note some basic properties of $\mathrm{D}^{\mathcal{F}}$.

▶ **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 $\mathrm{D}^{\mathcal{F}}$ satisfies the triangle inequality and is jointly convex, and*

1. *if $\mathcal{F}$ is symmetric then $\mathrm{D}^{\mathcal{F}}$ is symmetric and*

$$\mathrm{D}^{\mathcal{F}}(P \parallel Q) = \sup_{f \in \mathcal{F}} \Big| \mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \Big| \geq 0,$$

2. *if $\mathcal{F}$ is distinguishing then $\mathrm{D}^{\mathcal{F}}$ is positive-definite,*

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

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 ℓ_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*

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

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 ℓ_1 (equivalently total variation) distance.

Proof Sketch. As mentioned this is just the standard fact that the ℓ_p and ℓ_q norms are dual, but for completeness we include a proof in Appendix A. ◀

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 [26] and the average-case variant of Dodis, Ostrovsky, Reyzin, and Smith [14].

▶ **Definition 3.1** (Extends Definition 1.4). *Let D be a weak divergence on the set $\{0,1\}^m$, and $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$. Then if for all distributions X over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \geq k$ it holds that*

1. $\mathrm{D}(\mathrm{Ext}(X, U_d) \parallel U_m) \leq \varepsilon$, *then* Ext *is said to be a* (k, ε) extractor for D, *or a* (k, ε) D-extractor.
2. $\mathbb{E}_{s \sim U_d}[\mathrm{D}(\mathrm{Ext}(X, s) \parallel U_m)] \leq \varepsilon$, *then* Ext *is said to be a* (k, ε) strong extractor for D, *or a* (k, ε) strong D-extractor.

Furthermore, if for all joint distributions (Z, X) where X is distributed over $\{0,1\}^n$ with $\tilde{\mathrm{H}}_\infty(X|Z) \stackrel{\text{def}}{=} \log\big(1/\mathbb{E}_{z \sim Z}\big[2^{-\mathrm{H}_\infty(X|_{Z=z})}\big]\big) \geq k$, it holds that

3. $\mathbb{E}_{z \sim Z}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, U_d) \parallel U_m) \leq \varepsilon]$, *then* Ext *is said to be a* (k, ε) average-case extractor for D, *or a* (k, ε) average-case D-extractor.
4. $\mathbb{E}_{z \sim Z, s \sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)] \leq \varepsilon$, *then* Ext *is said to be a* (k, ε) average-case strong extractor for D, *or a* (k, ε) 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 [26] and the definition of average-case extractor due to [14].

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 $\mathbb{E}_{z\sim Z, s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)]$ rather than a bound on $\mathrm{D}(Z, U_d, \mathrm{Ext}(X, U_d) \parallel Z, U_d, U_m)$. 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*).

In the full version of this work [1] we include more discussion about this definition, and also generalize a result of Vadhan [35, Problem 6.8] showing that all $\mathrm{D}^{\mathcal{F}}$-extractors are average-case with only a constant factor loss in the error parameter.

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 [40].

▶ **Definition 3.3.** *Given a class of functions* $\mathcal{F} : \{0,1\}^m \to \mathbb{R}$*, a function* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *is said to be a* (δ,ε) strong averaging sampler for $\mathcal{F}$ *or a* (δ,ε) strong averaging $\mathcal{F}$-sampler *if for all* $f \in \mathcal{F}$*, it holds that*

$$\Pr_{x\sim U_n}\left[\mathop{\mathbb{E}}_{i\sim U_{[D]}}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon\right] \le \delta$$

where $[D] = \{1,\ldots,D\}$*. If this holds only when* $f_1 = \cdots = f_D$*, then it is called a* (non-strong) (δ,ε) averaging sampler for $\mathcal{F}$ *or* (δ,ε) averaging $\mathcal{F}$-sampler*. We say that* Samp *is a* (δ,ε) strong absolute averaging sampler for $\mathcal{F}$ *if it also holds that*

$$\Pr_{x\sim U_n}\left[\left|\mathop{\mathbb{E}}_{i\sim U_{[D]}}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]\Big]\right| > \varepsilon\right] \le \delta.$$

with the analogous definition for non-strong samplers.

▶ Remark 3.4. We separated a single-sided version of the error bound in Definition 3.3 as in [35], as it makes the connection between extractors and samplers cleaner and allows us to be specific about what assumptions are needed. Note that if $\mathcal{F}$ is symmetric then every (δ,ε) (strong) sampler for $\mathcal{F}$ is a $(2\delta,\varepsilon)$ (strong) absolute sampler for $\mathcal{F}$, recovering the standard notion up to a factor of 2 in δ.

3.2 Equivalence of extractors and samplers

We now show that Zuckerman's connection [40] does indeed generalize to this broader setting as promised.

▶ **Theorem 3.5.** *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be an* $(n-\log(1/\delta),\varepsilon)$*-extractor (respectively strong extractor) for the weak divergence* $\mathrm{D}^{\mathcal{F}}$ *defined by a class of test functions* $\mathcal{F} : \{0,1\}^m \to \mathbb{R}$ *as in Definition 2.5. Then the function* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *for* $D = 2^d$ *defined by* $\mathrm{Samp}(x)_i = \mathrm{Ext}(x,i)$ *is a* (δ,ε)*-sampler (respectively strong sampler) for* $\mathcal{F}$*.*

Proof sketch. The proof is given in Appendix A and is similar to that of Zuckerman [40]. The key idea is that for any function $f \in \mathcal{F}$, the set of seeds B_f which are bad for Samp with respect to f must be small, as otherwise $\mathbb{E}\big[f(\mathrm{Ext}(U_{B_f}, U_d))\big] - \mathbb{E}[f(U_m)] > \varepsilon$ contradicting the extractor property, where U_{B_f} is uniform over the set B_f. ◀

▶ Remark 3.6. Hölder's inequality implies that an extractor for ℓ_p with error $\varepsilon \cdot 2^{-m(p-1)/p}$ is also an ℓ_1 extractor and thus $[-1,1]$-averaging sampler with error ε. Proposition 2.11 and Theorem 3.5 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 ∞ moments), also with error ε.

Furthermore, if all the functions in $\mathcal{F}$ 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 3.7.** *Let $\mathcal{F}$ be a class of functions $\mathcal{F} \subset \{0,1\}^m \to \mathbb{R}$ with finite* maximum deviation from the mean, *meaning* $\max\operatorname{dev}(\mathcal{F}) \stackrel{\text{def}}{=} \sup_{f\in\mathcal{F}} \max_{x\in\{0,1\}^n} \big(f(x) - \mathbb{E}[f(U_m)]\big) < \infty$. *Then given a (δ,ε) $\mathcal{F}$-sampler (respectively (δ,ε) strong $\mathcal{F}$-sampler)* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$, *the function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for $d = \log D$ defined by* $\mathrm{Ext}(x,i) = \mathrm{Samp}(x)_i$ *is a* $\big(k, \varepsilon + \delta \cdot 2^{n-k} \cdot \max\operatorname{dev}(\mathcal{F})\big)$ $\mathrm{D}^{\mathcal{F}}$*-extractor (respectively strong* $\mathrm{D}^{\mathcal{F}}$*-extractor) for every $0 \le k \le n$.*

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

Proof sketch. The proof is given in Appendix A and is again similar to that of Zuckerman [40]. The key idea is that for any function $f \in \mathcal{F}$, since most $x \in \{0,1\}^n$ are good for Samp, for any source X of sufficient min-entropy, the probability over x from X that $\mathbb{E}[f(\mathrm{Ext}(x,U_d))] - \mathbb{E}[f(U_m)] > \varepsilon$ must be at most η, and in this failure case we can fall back on the trivial bound of $\max\operatorname{dev}(\mathcal{F})$. ◀

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 [38] for an introduction.

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

▶ 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 σ 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 σ 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 σ, then

1. *For all $t > 0$,* $\max\big(\Pr[Z > t], \Pr[Z < -t]\big) \le e^{-t^2/2\sigma^2}$.
2. $\|Z\|_p \stackrel{\text{def}}{=} \mathbb{E}[|Z|^p]^{1/p} \le 2\sigma\sqrt{p}$ *for all $p \ge 1$.*
3. *Z is subexponential with parameter σ.*

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

▶ **Definition 4.4.** *For every finite set $\mathcal{X}$, we define the set $\mathcal{G}_{\mathcal{X}}$ of* subgaussian test functions *on $\mathcal{X}$ (respectively the set $\mathcal{E}_{\mathcal{X}}$ of* subexponential test functions *on $\mathcal{X}$) to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that the random variable $f(U_{\mathcal{X}})$ is mean-zero and subgaussian (respectively subexponential) with parameter $1/2$. Then $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}}$ are symmetric and distinguishing, so by Lemma 2.10 the respective distances induced by $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}}$ are jointly convex metrics called the* subgaussian distance *and* subexponential distance *respectively and are denoted as $d_{\mathcal{G}}(P, Q)$ and $d_{\mathcal{E}}(P, Q)$.*

▶ Remark 4.5. We choose subgaussian parameter $1/2$ in Definition 4.4 as by Hoeffding's lemma, all functions $f : \{0,1\}^m \to [0, 1]$ 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 $d_{\mathcal{G}}$.

Note that absolute averaging samplers for $\mathcal{G}_{\{0,1\}^m}$ from Definition 3.3 are exactly subgaussian samplers as defined in the introduction. Thus, by Remark 3.4 and Theorem 3.5, to construct subgaussian samplers it is enough to construct extractors for the subgaussian distance $d_{\mathcal{G}}$.

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 $f : A \to B$ such that

$$d_{\mathcal{G}}(f(P), f(Q)) \not\leq d_{\mathcal{G}}(P, Q).$$

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 $f : \{0,1\}^1 \to \{0,1\}^m$ 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 (cf. Nisan and Zuckerman [26]; Goldreich and Wigderson [19]; Reingold, Vadhan, and Wigderson [30]) in the construction of samplers and extractors for bounded functions is to do the following: given extractors

$$\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m \qquad \mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d,$$

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

$$\mathrm{Ext}\big((x, y), s\big) = \mathrm{Ext}_{out}\big(x, \mathrm{Ext}_{in}(y, s)\big).$$

The reason this works for total variation distance is exactly the data-processing inequality: if Y has enough min-entropy given X, then $\mathrm{Ext}_{in}(Y, U_{d'})$ 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 Ext_{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 Ext_{in} to subsample the outer sampler. On the other hand, for subgaussian distance, the distribution $\mathrm{Ext}_{in}(Y, U_{d'})$ can be ε-close to uniform but still have some element with excess probability mass $\Omega(\varepsilon/\sqrt{d})$, and this element (seed) when mapped by Ext_{out} can retain[2] 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 Ext_{out} is a good subgaussian sampler for $\{0,1\}^m$, there is no reason that a good subgaussian sampler Ext_{in} for $\{0,1\}^d$ the seeds of Ext_{out} will preserve the larger sampler property when $m \gg d$.

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.

4.2 Connections to other weak divergences

Therefore, to aid in extractor construction, we show how $d_{\mathcal{G}}$ relates to other statistical weak divergences (though for space reasons, we defer all proofs to Appendix A).

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

▶ **Lemma 4.6.** *Let P and Q be distributions on $\{0,1\}^m$. Then*

$$d_{TV}(P,Q) \le d_{\mathcal{G}}(P,Q) \le \sqrt{2\ln 2 \cdot m} \cdot d_{TV}(P,Q)$$

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

$$2d_{TV}(P,Q) = d_{\ell_1}(P,Q) \le 2^{m\alpha/(1+\alpha)} \cdot d_{\ell_{1+\alpha}}(P,Q)$$
$$d_{\mathcal{G}}(P,Q) \le 2^{m\alpha/(1+\alpha)}\sqrt{1+\frac{1}{\alpha}} \cdot d_{\ell_{1+\alpha}}(P,Q)$$

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 α. We give the proof in Appendix A.

One downside of starting with bounds on $\ell_{1+\alpha}$ is that, extending a well-known linear seed length linear bound for ℓ_2-extractors (e.g. [35, Problem 6.4]), we show in the full version of this work [1] 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 α to 0 would eliminate this linear lower bound but still bound the subgaussian distance, but phrased this way sending α to 0 just results in a total variation extractor.

However, with a shift in perspective essentially the same approach works: by Example 2.4, $d_{\ell_2}(P, U_m) \le \varepsilon \cdot 2^{-m/2}$ implies $\mathrm{D}_2(P \,\|\, U_m) \le \varepsilon^2/\ln 2$, and there is an analogous linear seed length lower bound on constant error $\mathrm{D}_{1+\alpha}$ extractors for every $\alpha > 0$. In this case, however, sending α 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.

[2] Given a subgaussian extractor Ext with $d \ge \log(m/\varepsilon)$, adding a single extra seed $*$ to Ext such that $\mathrm{Ext}(x,*) = 0^m$ results in a subgaussian extractor with error at most $2^{-d} \cdot \sqrt{2m} + \varepsilon \le 3\varepsilon$ by convexity of $d_{\mathcal{G}}$ and the fact that $\left\| d_{\mathcal{G}_{\{0,1\}^m}} \right\|_\infty < \sqrt{2m}$.

▶ **Lemma 4.8** (cf. [10, Lemma 4.15], [18, Fact B.1]). *Let P be a distribution on $\{0,1\}^m$. Then*

$$d_{\mathcal{G}}(P,U_m) \le \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)}$$
$$d_{\mathcal{E}}(P,U_m) \le \begin{cases} \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)} & \text{if } \mathrm{KL}(P \parallel U_m) \le \frac{1}{2\ln 2} \\ \frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m) + \frac{1}{4} & \text{if } \mathrm{KL}(P \parallel U_m) > \frac{1}{2\ln 2} \end{cases}$$

where these bounds are concave in $\mathrm{KL}(P \parallel U_m)$. In the reverse direction, it holds that

$$\mathrm{KL}(P \parallel U_m) \le m \cdot d_{TV}(P,U_m) + h(d_{TV}(P,U_m))$$

where $h(x) = x\log(1/x) + (1-x)\log(1/(1-x))$ is the (concave) binary entropy function.

Due to space constraints, we defer the proof to Appendix A.

5 Extractors for KL divergence

Since by Lemma 4.8 the subgaussian distance can be bounded in terms of the KL divergence to uniform, the following easy lemma shows that 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 $V_1(P \parallel U_M) \le f(V_2(P \parallel U_m))$ for all distributions P on $\{0,1\}^m$. Then if f is increasing on $(0,\varepsilon)$, every (k,ε) extractor Ext for V_1 is also a $(k, f(\varepsilon))$-extractor for V_2, and if f is also concave, then if Ext is strong or average-case as a V_1-extractor, it has the same properties as a $(k, f(\varepsilon))$ extractor for V_2.*

Importantly, the KL divergence does not have the flaws of subgaussian distance discussed in Section 4.1. For instance, 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

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

for all distributions A, B, X, and Y, which implies for example that

$$\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{KL}(\mathrm{Ext}(X,s) \parallel U_m)] = \mathrm{KL}(U_d, \mathrm{Ext}(X,U_d) \parallel U_d, U_m).$$

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

▶ **Lemma 5.2** (cf. [36, Lemma 6.6]). *Let P, Q, and R be distributions over a finite set $\mathcal{X}$. Then for all $\alpha > 0$, it holds that*

$$\mathrm{KL}(P \parallel R) \le \left(1 + \frac{1}{\alpha}\right) \cdot \mathrm{KL}(P \parallel Q) + \mathrm{D}_{1+\alpha}(Q \parallel R)$$

We give the proof in Appendix A.

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 5.3** (Composition for high min-entropy Rényi entropy extractors, cf. [19]). *Suppose*

1. $\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is an* $(n - \log(1/\delta), \varepsilon_{out})$ *extractor for* $\mathrm{D}_{1+\alpha}$ *with* $\alpha > 0$,
2. $\mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d$ *is an* $(n' - \log(1/\delta), \varepsilon_{in})$ *average-case KL-extractor,*

and define $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ *by* $\mathrm{Ext}\big((x,y),s\big) = \mathrm{Ext}_{out}(x, \mathrm{Ext}_{in}(y,s))$. *Then* Ext *is an* $(n + n' - \log(1/\delta), \varepsilon_{out} + (1 + 1/\alpha) \cdot \varepsilon_{in})$ *extractor for KL.*

We prove this in Appendix A.

5.2 Further theory

The reader is advised to consult the full version of this paper [1] for a more thorough development of the theory of KL-extractors, including an extension of the zig-zag product for extractors (Reingold, Vadhan, and Wigderson [30]), which allows us to avoid the $\log(1/\delta)$ entropy loss inherent in Theorem 5.3. We also give lower bounds, an optimal non-explicit construction, and interpretations of several existing extractor constructions as KL-extractors.

6 Constructions of subgaussian samplers

We can now establish a weak version of our explicit construction of subgaussian samplers with sample complexity having no dependence on m and sample complexity matching the best-known $[0,1]$-valued sampler when ε and δ are subconstant (up to the hidden polynomial in the sample complexity). Obtaining matching randomness complexity as well requires more technology from KL-extractors to develop, and as such we defer the proof to the full version of this paper [1].

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

▶ Remark 6.2. In the full version of this paper, we show for every constant $\alpha > 0$ the existence of an explicit absolute subexponential sampler with the same sample complexity $D = \mathrm{poly}(\log(1/\delta), 1/\varepsilon)$ and randomness complexity $n = m + (1 + \alpha)\log(1/\delta)$, and also an analogous result for strong subexponential samplers.

We will use essentially the same construction used for bounded samplers in this regime, combining the expander extractor of Goldreich and Wigderson [19] and an extractor with logarithmic seed length. However, as described in Section 4.1, this construction does not work for general subgaussian extractors, so we will instead use the analysis of Theorem 5.3. This requires a $\mathrm{D}_{1+\alpha}$-extractor for $\alpha > 0$, for this we note (following [35]) that the extractor of [19] is already an extractor for D_2 (see the full version of this work [1] for more details).

▶ **Theorem 6.3** ([19] [35, Discussion after Theorem 6.22]). *For all* $k \le n \in \mathbb{N}$ *and* $1/2 \ge \varepsilon > 0$ *there is an explicit* (k, ε) D_2*-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with seed length* $d = O(n - k + \log(1/\varepsilon))$ *and output length* $m = n$.

We also need an average-case KL-extractor, which we can construct by reducing the error in the extractors of Guruswami–Umans–Vadhan [20]:

▶ **Theorem 6.4** (Akin to [20, Theorem 1.5]). *For every $\alpha, \varepsilon > 0$ and integers $k \leq n$, there is an explicit average-case (k, ε)-KL-extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d \leq \log n + O_\alpha(\log(k/\varepsilon))$ and $m \geq (1-\alpha)k$.*

Though Theorem 6.4 has seed length depending on n the input length, this is tolerable for us since we will apply it to Ext_{in} in the composition of Theorem 5.3 with $n = O(\log(1/\delta) + \log(1/\varepsilon))$:

Proof. Let $\varepsilon' = \frac{\min(\varepsilon, 1/2)}{48(m+\log(1/\varepsilon))}$ so that $m \cdot 3\varepsilon' + h(3\varepsilon') \leq \varepsilon$, where $h(x) = x \log(1/x) + (1-x)\log(1/(1-x))$ is the binary entropy function. By [20, Theorem 1.5] and [35, Problem 6.8] there is an explicit $(k, 3\varepsilon')$ extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $d \leq \log n + O_\alpha(\log(k/\varepsilon')) = \log n + O_\alpha(\log(k/\varepsilon))$ and $m \geq (1-\alpha)k$. By Lemmas 4.8 and 5.1, we also have that Ext is a $(k, m \cdot 3\varepsilon' + h(3\varepsilon'))$ average-case KL-extractor, and thus a (k, ε) average-case KL-extractor as desired. ◀

▶ **Theorem 6.5.** *For all integers m and $\delta, \varepsilon > 0$ there is an explicit (k, ε)-KL-extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ with $n = m + O(\log(1/\delta))$, $k = n - \log(1/\delta)$, and $d = O(\log\log(1/\delta) + \log(1/\varepsilon))$.*

Proof. Let $\mathrm{Ext}_{out} : \{0,1\}^m \times \{0,1\}^{d_{out}} \to \{0,1\}^m$ the $(m - \log(1/\delta), \varepsilon/3)$ be the D_2-extractor from Theorem 6.3 with $d_{out} = O(\log(1/\delta) + \log(1/\varepsilon))$, and let$\mathrm{Ext}_{in} : \{0,1\}^{n_{in}} \times \{0,1\}^{d_{in}} \to \{0,1\}^{d_{out}}$ be the $(n_{in} - \log(1/\delta), \varepsilon/3)$ average-case KL-extractor from Theorem 6.4 with output length d_{out}, so that $n_{in} = O(\log(1/\delta) + \log(1/\varepsilon))$ and $d_{in} = O(\log\log(1/\delta) + \log(1/\varepsilon))$.

Then instantiating Theorem 5.3 with Ext_{out} and Ext_{in} gives an $(n' - \log(1/\delta), \varepsilon)$ KL-extractor $\mathrm{Ext}' : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ with $n' = m + n_{in}$ $d' = d_{in}$. The result follows follows from defining $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ by $\mathrm{Ext}(x, (s,t)) = \mathrm{Ext}'((x,s), t)$ for s of length $O(\log(1/\varepsilon))$. ◀

We can now prove Theorem 6.1.

Proof of Theorem 6.1. Let $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be the explicit $(k, \varepsilon^2/2)$ KL-extractor of Theorem 6.5 with $n = O(m + \log(1/\delta') + \log(1/\varepsilon))$, $k = n - \log(1/\delta')$, and $d = O(\log\log(1/\delta') + \log(1/\varepsilon))$ for $\delta' = \delta/2$. Then by Lemmas 4.8 and 5.1, Ext is also a (k, ε) extractor for $d_{\mathcal{E}}$, so by Theorem 3.5 the function $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ for $D = 2^d$ defined by $\mathrm{Samp}(x)_i = \mathrm{Ext}(x, i)$ is a (δ', ε) subexponential sampler. Finally, by Remark 3.4, we have that Samp is a (δ, ε) absolute subexponential sampler as desired. ◀

In the full version [1] of this paper, in addition to proving the stronger version of Theorem 6.1, we also discuss explicit samplers for other ranges of parameters and non-explicit constructions.

References

1 Rohit Agrawal. Samplers and Extractors for Unbounded Functions. *arXiv:1904.08391 [cs]*, July 2019. `arXiv:1904.08391`.

2 Syed Mumtaz Ali and Samuel David Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

3 Koenraad M. R. Audenaert and Jens Eisert. Continuity Bounds on the Quantum Relative Entropy. *Journal of Mathematical Physics*, 46(10):102104, October 2005. doi:10.1063/1.2044667.

4 Boaz Barak, Yevgeniy Dodis, Hugo Krawczyk, Olivier Pereira, Krzysztof Pietrzak, François-Xavier Standaert, and Yu Yu. Leftover Hash Lemma, Revisited. In Phillip Rogaway, editor, *Advances in Cryptology – CRYPTO 2011*, Lecture Notes in Computer Science, pages 1–20. Springer Berlin Heidelberg, 2011.

5 Mihir Bellare, Oded Goldreich, and Shafi Goldwasser. Randomness in Interactive Proofs. *computational complexity*, 3(4):319–354, December 1993. doi:10.1007/BF01275487.

6 Mihir Bellare and John Rompel. Randomness-Efficient Oblivious Sampling. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 276–287, November 1994. doi:10.1109/SFCS.1994.365687.

7 Charles H. Bennett, Gilles Brassard, and Jean-Marc Robert. Privacy Amplification by Public Discussion. *SIAM Journal on Computing*, 17(2):210–229, April 1988. doi:10.1137/0217014.

8 Jarosław Błasiok. Private Communication, 2018.

9 Jarosław Błasiok. Optimal Streaming and Tracking Distinct Elements with High Probability. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, Proceedings, pages 2432–2448. Society for Industrial and Applied Mathematics, January 2018. doi:10.1137/1.9781611975031.156.

10 Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 1 edition edition, February 2013. doi:10.1093/acprof:oso/9780199535255.001.0001.

11 Ran Canetti, Guy Even, and Oded Goldreich. Lower Bounds for Sampling Algorithms for Estimating the Average. *Information Processing Letters*, 53(1):17–25, January 1995. doi:10.1016/0020-0190(94)00171-T.

12 Benny Chor and Oded Goldreich. On the Power of Two-Point Based Sampling. *Journal of Complexity*, 5(1):96–106, March 1989. doi:10.1016/0885-064X(89)90015-0.

13 Imre Csiszár. Eine Informationstheoretische Ungleichung Und Ihre Anwendung Auf Den Beweis Der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 8:85–108, 1963.

14 Yevgeniy Dodis, Rafail Ostrovsky, Leonid Reyzin, and Adam Smith. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data. *SIAM Journal on Computing*, 38(1):97–139, January 2008. doi:10.1137/060651380.

15 Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic Evaluation of Certain Markov Process Expectations for Large Time—III. *Communications on Pure and Applied Mathematics*, 29(4):389–461, 1976. doi:10.1002/cpa.3160290405.

16 David Gillman. A Chernoff Bound for Random Walks on Expander Graphs. *SIAM Journal on Computing*, 27(4):1203–1220, August 1998. doi:10.1137/S0097539794268765.

17 Oded Goldreich. A Sample of Samplers: A Computational Perspective on Sampling. In Oded Goldreich, editor, *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, Lecture Notes in Computer Science, pages 302–332. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-22670-0_24.

18 Oded Goldreich and Salil Vadhan. Comparing Entropies in Statistical Zero Knowledge with Applications to the Structure of SZK. In *Proceedings of the Fourteenth Annual IEEE Conference on Computational Complexity*, pages 54–73, May 1999. doi:10.1109/CCC.1999.766262.

19 Oded Goldreich and Avi Wigderson. Tiny Families of Functions with Random Properties: A Quality-Size Trade-off for Hashing. *Random Structures & Algorithms*, 11(4):315–343, 1997. doi:10.1002/(SICI)1098-2418(199712)11:4<315::AID-RSA3>3.0.CO;2-1.

20 Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced Expanders and Randomness Extractors from Parvaresh–Vardy Codes. *Journal of the ACM*, 56(4):20:1–20:34, July 2009. doi:10.1145/1538902.1538904.

21 Russell Impagliazzo, Leonid A. Levin, and Michael Luby. Pseudo-Random Generation from One-Way Functions. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, STOC '89, pages 12–24, New York, NY, USA, 1989. ACM. doi:10.1145/73007.73009.

22 Richard Karp, Nicholas Pippenger, and Michael Sipser. A Time-Randomness Tradeoff. In *AMS Conference on Probabilistic Computational Complexity*, Durham, New Hampshire, 1985.

23 James Lawrence McInnes. Cryptography Using Weak Sources of Randomness. Technical Report 194/87, University of Toronto, 1987.

24 Tetsuzo Morimoto. Markov Processes and the H-Theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, March 1963. doi:10.1143/JPSJ.18.328.

25 Alfred Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, 1997. doi:10.2307/1428011.

26 Noam Nisan and David Zuckerman. Randomness Is Linear in Space. *Journal of Computer and System Sciences*, 52(1):43–52, February 1996. doi:10.1006/jcss.1996.0004.

27 Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for Dispersers, Extractors, and Depth-Two Superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24, January 2000. doi:10.1137/S0895480197329508.

28 Ran Raz, Omer Reingold, and Salil Vadhan. Extracting All the Randomness and Reducing the Error in Trevisan's Extractors. *Journal of Computer and System Sciences*, 65(1):97–128, August 2002. doi:10.1006/jcss.2002.1824.

29 Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. New Proofs of the Green-Tao-Ziegler Dense Model Theorem: An Exposition. *arXiv:0806.0381 [math]*, June 2008. arXiv:0806.0381.

30 Omer Reingold, Salil Vadhan, and Avi Wigderson. Entropy Waves, the Zig-Zag Graph Product, and New Constant-Degree Expanders and Extractors. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 3–13, November 2000. doi:10.1109/SFCS.2000.892006.

31 Alfréd Rényi. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

32 Ofer Shayevitz. On Rényi Measures and Hypothesis Testing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 894–898, July 2011. doi:10.1109/ISIT.2011.6034266.

33 Aravind Srinivasan and David Zuckerman. Computing with Very Weak Random Sources. *SIAM Journal on Computing*, 28(4):1433–1459, January 1999. doi:10.1137/S009753979630091X.

34 Amnon Ta-Shma, David Zuckerman, and Shmuel Safra. Extractors from Reed–Muller Codes. *Journal of Computer and System Sciences*, 72(5):786–812, August 2006. doi:10.1016/j.jcss.2005.05.010.

35 Salil P. Vadhan. *Pseudorandomness*. Now Publishers Inc, Boston, Mass., October 2012.

36 Tim van Erven. *When Data Compression and Statistics Disagree: Two Frequentist Challenges for the Minimum Description Length Principle*. PhD thesis, Leiden University, 2010. OCLC: 673140651.

37 Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. doi:10.1109/TIT.2014.2320500.

38 Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.

39 Vladimir Mikhailovich Zolotarev. Probability Metrics. *Theory of Probability & Its Applications*, 28(2):278–302, January 1984. doi:10.1137/1128025.

40 David Zuckerman. Randomness-Optimal Oblivious Sampling. *Random Structures & Algorithms*, 11(4):345–367, 1997. doi:10.1002/(SICI)1098-2418(199712)11:4<345::AID-RSA4>3.0.CO;2-Z.

41 David Zuckerman. Linear Degree Extractors and the Inapproximability of Max Clique and Chromatic Number. *Theory of Computing*, 3(1):103–128, August 2007. doi:10.4086/toc.2007.v003a006.

A Missing proofs

In this section, we include some proofs that were omitted from the main text due to space constraints.

Proof of Proposition 2.11. As mentioned this is just the standard fact that the ℓ_p and ℓ_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

$$\begin{aligned}
d_{\ell_p}(A,B) &= \left(\sum_x |A_x - B_x|^p\right)^{1/p} \\
&= 2^{m/p} \mathop{\mathbb{E}}_{x\sim U_m}[|A_x - B_x|^p]^{1/p} \\
&= 2^{m/p} \max_{\substack{f:\{0,1\}^m\to\mathbb{R} \\ \|f(U_m)\|_q\le 1}} \left|\mathop{\mathbb{E}}_{x\sim U_m}[f(x)(A_x - B_x)]\right| && \text{(Hölder's extremal equality)} \\
&= 2^{-m+m/p} \max_{\substack{f:\{0,1\}^m\to\mathbb{R} \\ \|f(U_m)\|_q\le 1}} \left|\mathbb{E}[f(A)] - \mathbb{E}[f(B)]\right| \\
&= 2^{-m/q}\cdot d_{\mathcal{M}_q}(A,B) && \text{(by symmetry of } \mathcal{M}_q)
\end{aligned}$$

as desired. ◀

Proof of Theorem 3.5. The proof is essentially the same as that of [40].

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

$$\begin{aligned}
B_{f_1,\ldots,f_D} &\stackrel{\text{def}}{=} \left\{x\in\{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i\sim U_{[D]}}\Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon\right\} \\
&= \left\{x\in\{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i\sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}\big(U_{\{\mathrm{Ext}(x,i)\}} \,\|\, U_m\big)\Big] > \varepsilon\right\},
\end{aligned}$$

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

$$\begin{aligned}
\varepsilon &< \mathop{\mathbb{E}}_{x\sim X}\left[\mathop{\mathbb{E}}_{i\sim U_{[D]}}\Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)]\Big]\right] \\
&= \mathop{\mathbb{E}}_{i\sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i) \,\|\, U_m)\Big]
\end{aligned}$$

$$\dots = \begin{cases} \mathrm{D}^{\{f_1\}}(\mathrm{Ext}(X, U_d) \parallel U_m) & \text{if } f_1 = \dots = f_D \\ \mathbb{E}_{i\sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X, i) \parallel U_m)\Big] & \text{always} \end{cases}$$
$$\le \begin{cases} \mathrm{D}^{\mathcal{F}}(\mathrm{Ext}(X, U_d) \parallel U_m) & \text{if } f_1 = \dots = f_D \\ \mathbb{E}_{i\sim U_{[D]}}\Big[\mathrm{D}^{\mathcal{F}}(\mathrm{Ext}(X, i) \parallel U_m)\Big] & \text{always} \end{cases}$$

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

Proof of Theorem 3.7. Again the proof is analogous to the one in [40].

Fix a distribution X over $\{0,1\}^m$ with $\mathrm{H}_\infty(X) \ge k$ and a collection of test functions $f_1, \dots, f_D \in \mathcal{F}$, where if Samp is not strong we restrict to $f_1 = \dots = f_D$. Then since Samp is a (δ, ε) $\mathcal{F}$-sampler, we know that the set of seeds for which the sampler is bad must be small. Formally, the set

$$B_{f_1,\dots,f_D} \stackrel{\text{def}}{=} \left\{ x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon \right\}$$
$$= \left\{ x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(x, i)) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon \right\}$$

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{i\sim U_d}\Big[\mathbb{E}\Big[f_i(\mathrm{Ext}(X, i)) - \mathbb{E}[f_i(U_m)]\Big]\Big] \\
&= \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(X, i)) - \mathbb{E}[f_i(U_m)]\Big]\Big] \\
&= \Pr[X \in B_{f_1,\dots,f_D}] \cdot \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(X, i)) - \mathbb{E}[f_i(U_m)]\Big] \;\Big|\; X \in B_{f_1,\dots,f_D}\Big] \\
&+ \Pr[X \notin B_{f_1,\dots,f_D}] \cdot \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(X, i)) - \mathbb{E}[f_i(U_m)]\Big] \;\Big|\; X \notin B_{f_1,\dots,f_D}\Big] \\
&\le \Pr[X \in B_{f_1,\dots,f_D}] \cdot \max \mathrm{dev}(\mathcal{F}) + \Pr[X \notin B_{f_1,\dots,f_D}] \cdot \varepsilon \\
&\le 2^{-k} \cdot \delta 2^n \cdot \max \mathrm{dev}(\mathcal{F}) + \varepsilon
\end{aligned}$$

completing the proof of the main claim. The "in particular" statement follows since if (Z, X) are jointly distributed with $\tilde{\mathrm{H}}_\infty(X|Z) \ge n - \log(1/\delta) + \log(1/\eta)$ we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim Z}\Big[\varepsilon + \delta \cdot 2^{n-\mathrm{H}_\infty(X|_{Z=z})} \cdot \max \mathrm{dev}(\mathcal{F})\Big] &= \varepsilon + \delta \cdot 2^{n-\tilde{\mathrm{H}}_\infty(X|Z)} \cdot \max \mathrm{dev}(\mathcal{F}) \\
&\le \varepsilon + \eta \cdot \max \mathrm{dev}(\mathcal{F})
\end{aligned}$$

by definition of conditional min-entropy. ◀

Proof of Lemma 4.6. That $d_{TV} \le d_{\mathcal{G}}$ 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 $1/2 + f/\sqrt{2 \ln 2 \cdot m}$ is $[0,1]$-valued and thus lower bounds the total variation distance. ◀

Proof of Lemma 4.7. By Proposition 2.11, for any function $f : \{0,1\}^m \to \mathbb{R}$ it holds that

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

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. ◀

Proof of Lemma 4.8. The upper bound on subgaussian distance follows from a general form of Pinsker's inequality as in [10, 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 [15] (cf. [10, Corollary 4.15])

$$\mathrm{KL}(P \parallel U_m) = \frac{1}{\ln 2} \cdot \sup_{g:\{0,1\}^m \to \mathbb{R}} \Big(\mathbb{E}[g(P)] - \ln \mathbb{E}\Big[e^{g(U_m)}\Big]\Big).$$

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

$$\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] = \frac{1}{t} \cdot \mathbb{E}[g(P)] \leq \frac{\ln 2 \cdot \mathrm{KL}(P \parallel U_m) + \ln \mathbb{E}\big[e^{t \cdot f(U_m)}\big]}{t}$$

for all $t > 0$. Thus, when $\ln \mathbb{E}\big[e^{t \cdot f(U_m)}\big] \leq t^2/8$, we have $\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] \leq \ln 2 \cdot \mathrm{KL}(P \parallel U_m)/t + t/8$.

Then since subgaussian random variables satisfy such a bound for all t, we can make the optimal choice $t = \sqrt{8 \ln 2 \cdot \mathrm{KL}(P \parallel U_m)}$ to get the claimed bound on $d_{\mathcal{G}}$. For subexponential random variables, which satisfy such a bound only for $|t| \leq 2$, we choose $t = \min(\sqrt{8 \ln 2 \cdot \mathrm{KL}(P \parallel U_m)}, 2)$, which gives

$$d_{\mathcal{E}}(P, U_m) \leq \begin{cases} \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)} & \text{if } \mathrm{KL}(P \parallel U_m) \leq \frac{1}{2 \ln 2} \\ \frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m) + \frac{1}{4} & \text{if } \mathrm{KL}(P \parallel U_m) > \frac{1}{2 \ln 2} \end{cases}$$

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

$$|H(P) - H(Q)| \leq \lg(S - 1) \cdot d_{TV}(P, Q) + h(d_{TV}(P, Q)).$$

This inequality is a simple consequence of Fano's inequality as noted by Goldreich and Vadhan [18, Fact B.1], and implies the desired result by taking $Q = U_m$ as $\mathrm{KL}(P \parallel U_m) = H(U_m) - H(P)$ and $|\{0,1\}^m| = 2^m$. ◀

▶ Remark A.1. There are sharper upper bounds on the KL divergence than given in Lemma 4.8, such as the bound of Audenaert and Eisert [3, 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.

Proof of Lemma 5.2. This follows from a characterization of Rényi divergence due to van Erven and Harremoës [36, Lemma 6.6] [37, Theorem 30] and Shayevitz [32, Theorem 1], who prove that for for every positive real $\beta \neq 1$ and distributions X and Y that

$$(1 - \beta)\, \mathrm{D}_\beta(X \parallel Y) = \inf_Z \big\{\beta\, \mathrm{KL}(Z \parallel X) + (1 - \beta)\, \mathrm{KL}(Z \parallel Y)\big\}.$$

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. ◀

Proof of Theorem 5.3. Let (X, Y) be jointly distributed random variables with X distributed over $\{0,1\}^n$ and Y over $\{0,1\}^{n'}$ such that $\tilde{\mathrm{H}}_\infty(X, Y|Z) \geq n + n' - \log(1/\delta)$. Then by Lemma 5.2 and the data-processing inequality for KL divergence we have that

$$\begin{aligned}
&\mathrm{KL}(\mathrm{Ext}((X,Y),U_d) \parallel U_m) \\
&\quad = \mathrm{KL}(\mathrm{Ext}_{out}(X, \mathrm{Ext}_{in}(Y, U_d)) \parallel U_m) \\
&\quad \leq (1 + 1/\alpha) \cdot \mathrm{KL}(\mathrm{Ext}_{out}(X, \mathrm{Ext}_{in}(Y, U_d)) \parallel \mathrm{Ext}_{out}(X, U_d)) \\
&\qquad\quad + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X, U_d) \parallel U_m) \\
&\quad \leq (1 + 1/\alpha) \cdot \mathrm{KL}(X, \mathrm{Ext}_{in}(Y, U_d) \parallel X, U_d) + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X, U_d) \parallel U_m) \\
&\quad = (1 + 1/\alpha) \cdot \mathop{\mathbb{E}}_{x \sim X}[\mathrm{KL}(\mathrm{Ext}_{in}(Y|_{X=x}, U_d) \parallel U_d)] + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X, U_d) \parallel U_m)
\end{aligned}$$

where the last equality follows from the chain rule for KL divergence. Now by standard properties of conditional min-entropy (see for example [14, Lemma 2.2]), we know that $\mathrm{H}_\infty(X) \geq \mathrm{H}_\infty(X, Y) - \log|\mathrm{Supp}(Y)| \geq n - \log(1/\delta)$ and $\tilde{\mathrm{H}}_\infty(Y|X) \geq \mathrm{H}_\infty(X, Y) - \log|\mathrm{Supp}(X)| \geq n' - \log(1/\delta)$. Thus, since by assumption Ext_{in} is an average-case $(n' - \log(1/\delta), \varepsilon_{in})$ KL-extractor the first term is bounded by $(1 + 1/\alpha) \cdot \varepsilon_{in}$, and similarly since Ext_{out} is an $(n - \log(1/\delta), \varepsilon_{out})$ $\mathrm{D}_{1+\alpha}$-extractor we have that the second term is bounded by ε_{out} as desired. ◀