Randomness-efficient Curve Sampling

Thesis by
Zeyu Guo

In Partial Fulfillment of the Requirements
for the
Master Degree in Computer Science

California Institute of Technology
Pasadena, California

2014
(Submitted February 5, 2014)

Abstract

Curve samplers are sampling algorithms that proceed by viewing the domain as a vector space over a finite field, and randomly picking a low-degree curve in it as the sample. Curve samplers exhibit a nice property besides the sampling property: the restriction of low-degree polynomials over the domain to the sampled curve is still low-degree. This property is often used in combination with the sampling property and has found many applications, including PCP constructions, local decoding of codes, and algebraic PRG constructions.

The randomness complexity of curve samplers is a crucial parameter for its applications. It is known that (non-explicit) curve samplers using $O(\log N + \log(1/\delta))$ random bits exist, where N is the domain size and δ is the confidence error. The question of explicitly constructing randomness-efficient curve samplers was first raised in [TSU06] where they obtained curve samplers with near-optimal randomness complexity.

In this thesis, we present an explicit construction of low-degree curve samplers with *optimal* randomness complexity (up to a constant factor) that sample curves of degree $\big(m \log_q(1/\delta)\big)^{O(1)}$ in $\mathbb{F}_q^m$. Our construction is a delicate combination of several components, including extractor machinery, limited independence, iterated sampling, and list-recoverable codes.

Contents

Chapter 1

Introduction

Overview

Randomness has numerous uses in computer science, and sampling is one of its most classical applications: Suppose we are interested in the size of a particular subset A lying in a large domain D. Instead of counting the size of A directly by enumeration, one can randomly draw a small sample from D and calculate the density of A in the sample. The approximated density is guaranteed to be close to the true density (measured by a parameter ϵ, the *accuracy error*) with probability $1 - \delta$ where δ is very small, known as the *confidence error*. This sampling technique is extremely useful both in practice and in theory.

One class of sampling algorithms, known as *curve samplers*, proceed by viewing the domain as a vector space over a finite field, and picking a random low-degree curve in it. Curve samplers exhibit the following nice property besides the sampling property: the restriction of low-degree polynomials over the domain to the sampled curve is still low-degree. This special property, combined with the sampling property, turns out to be useful in many settings, e.g local decoding of Reed-Muller codes and hardness amplification [STV01], PCP constructions [AS98, ALM+98, MR08], algebraic constructions of pseudorandom-generators [SU05, Uma03], extractor constructions [SU05, TSU06], and some pure complexity results (e.g. [SU06]).

The problem of explicitly constructing low-degree curve samplers was raised in [TSU06]. Typically, we are looking for low-degree curve samplers with small sample complexity (poly-logrithmic in the domain size) and confidence error (polynomially small in the domain size), and we focus on minimizing the randomness complexity. The simplest way is picking a completely random low-degree curve whose sampling properties are guaranteed by tail bounds for limited independence. The randomness complexity of this method, however, is far from being optimal. The probabilistic method guarantees the existence of (non-explicit) low-degree curve samplers using $O(\log N + \log(1/\delta))$ random bits where N is the domain size and δ is

the confidence error. The real difficulty, however, is to find an explicit construction matching this bound.

Previous work

Randomness-efficient samplers (without the requirement that the sample points form a curve) are constructed in [CG89, Gil98, BR94, Zuc97]. In particular, [Zuc97] obtains explicit samplers with optimal randomness complexity (up to a $1+\gamma$ factor for arbitrary small $\gamma > 0$) using the connection between samplers and extractors. See [Gol11] for a survey of samplers.

Degree-1 curve samplers are also called *line samplers*. Explicit randomness-efficient line samplers are constructed in the PCP literature [BSSVW03, MR08], motivated by the goal of constructing almost linear sized PCPs. In [BSSVW03] line samplers are derandomized by picking a random point and a direction sampled from an ϵ-biased set, instead of two random points. An alternative way is suggested in [MR08] where directions are picked from a subfield. It is not clear, however, how to apply these techniques to higher degree curves.

In [TSU06] it was shown how to explicitly construct derandomized curve samplers with near-optimal parameters by employing an iterated sampling technique. Formally they obtained

- curve samplers picking curves of degree $(\log\log N + \log(1/\delta))^{O(\log\log N)}$ using randomness $O(\log N + \log(1/\delta)\log\log N)$, and
- curve samplers picking curves of degree $(\log(1/\delta))^{O(1)}$ using randomness $O(\log N + \log(1/\delta)(\log\log N)^{1+\gamma})$ for any constant $\gamma > 0$

for domain size N, field size $q \geq (\log N)^{\Theta(1)}$ and confidence error $\delta = N^{-\Theta(1)}$. Their work left the problem of explicitly constructing low-degree curve samplers (ideally picking curves of degree $O(\log_q(1/\delta))$) with essentially optimal $O(\log N + \log(1/\delta))$ random bits as a prominent open problem.

Main results

It is known that curve samplers in $\mathbb{F}_q^m$ must have sample complexity $\Omega\left(\frac{\log(\epsilon/\delta)}{\epsilon^2}\right)$ and randomness complexity $(m-1)\log q + \log(1/\epsilon) + \log(1/\delta) - O(1)$ [RTS00]. It is also not hard to show that the degree of the sampled curves has to be $\Omega\left(\log_q(1/\delta)\right)$ (c.f. Theorem 3.3.3). We construct explicit curve samplers with parameters that match or are close to these lower bounds. In particular, we show how to sample degree-$\left(m\log_q(1/\delta)\right)^{O(1)}$ curves in $\mathbb{F}_q^m$ using $O(\log N + \log(1/\delta))$ random bits for domain size $N = |\mathbb{F}_q^m|$ and confidence error $\delta = N^{-\Theta(1)}$.

Before stating our main theorem, we first present the formal definition of samplers and curve samplers.

Samplers. Given a finite set $\mathcal{M}$ as the domain, the *density* of a subset $A \subseteq \mathcal{M}$ is $\mu(A) \stackrel{\text{def}}{=} \frac{|A|}{|\mathcal{M}|}$. For a collection of elements $\mathcal{T} = \{t_i : i \in I\} \in \mathcal{M}^I$ indexed by set I, the density of A in $\mathcal{T}$ is $\mu_{\mathcal{T}}(A) \stackrel{\text{def}}{=} \frac{|A \cap \mathcal{T}|}{|\mathcal{T}|} = \Pr_{i \in I}[t_i \in A]$.

Definition 1.0.1 (sampler)**.** A *sampler* is a function $S : \mathcal{N} \times \mathcal{D} \to \mathcal{M}$ where $|\mathcal{D}|$ is its *sample complexity* and $\mathcal{M}$ is its *domain*. We say S samples $A \subseteq \mathcal{M}$ with *accuracy error* ϵ and *confidence error* δ if

$$\Pr_{x \in \mathcal{N}}[|\mu_{S(x)}(A) - \mu(A)| > \epsilon] \leq \delta$$

where $S(x) \stackrel{\text{def}}{=} \{S(x, y) : y \in \mathcal{D}\}$. We say S is an (ϵ, δ) sampler if it samples all subsets $A \subseteq \mathcal{M}$ with accuracy error ϵ and confidence error δ. The *randomness complexity* of S is $\log(|\mathcal{N}|)$.

Lines, curves, and manifolds. To define curve sampler, we first define curves, lines, and more generally manifolds. Let $f : \mathbb{F}_q^d \to \mathbb{F}_q^D$ be a map. We may view f as D individual functions $f_i : \mathbb{F}_q^d \to \mathbb{F}_q$ describing its operation on each output coordinate, i.e., $f(x) = (f_1(x), \ldots, f_D(x))$ for all $x \in \mathbb{F}_q^d$.

Definition 1.0.2 (manifold)**.** A manifold in $\mathbb{F}_q^D$ is a function $C : \mathbb{F}_q^d \to \mathbb{F}_q^D$ where $C_1, \ldots, C_D$ are d-variate polynomials over $\mathbb{F}_q$. We call d the *dimension* of C. We say a manifold C has *degree* t if each polynomial C_i has degree t. An 1-dimensional manifold is also called a *curve*. A curve of degree 1 is also called a *line*.

Now we are ready to define curve samplers, the central objects studied in this thesis.

Definition 1.0.3 (curve/line sampler)**.** Let $\mathcal{M} = \mathbb{F}_q^D$ and $\mathcal{D} = \mathbb{F}_q$. The sampler $S : \mathcal{N} \times \mathcal{D} \to \mathcal{M}$ is a *degree-t curve sampler* if for all $x \in \mathcal{N}$, the function $S(x, \cdot) : \mathcal{D} \to \mathcal{M}$ is a curve of degree at most t over $\mathbb{F}_q$. When $t = 1$, S is also called a *line sampler*.

The main result of this thesis is as follows.

Theorem 1.0.1 (main)**.** *For any $\epsilon, \delta > 0$, integer $m \geq 1$, and sufficiently large prime power $q \geq \left(\frac{m \log(1/\delta)}{\epsilon}\right)^{\Theta(1)}$, there exists an explicit degree-t curve sampler for the domain $\mathbb{F}_q^m$ with $t = \left(m \log_q(1/\delta)\right)^{O(1)}$, accuracy error ϵ, confidence error δ, sample complexity q, and randomness complexity $O\left(m \log q + \log(1/\delta)\right) = O(\log N + \log(1/\delta))$ where $N = q^m$ is the domain size. Moreover, the curve sampler itself has degree $\left(m \log_q(1/\delta)\right)^{O(1)}$ as a polynomial map.*

Theorem 1.0.1 has better degree bound and randomness complexity compared with the constructions in [TSU06]. The degree bound, being $\left(m \log_q(1/\delta)\right)^{O(1)}$, is still sub-optimal compared with the lower bound $\log_q(1/\delta)$. However, we remark that in typical settings it is satisfying to to achieve such a degree bound.

As an example, consider the following setting of parameters: domain size $N = q^m$, field size $q = (\log N)^{\Theta(1)}$, confidence error $\delta = N^{-\Theta(1)}$, and accuracy error $\epsilon = (\log N)^{-\Theta(1)}$. Note that this is the typical setting in PCP and other literature [ALM+98, AS98, STV01, SU05]. In this setting, we have the following corollary in which the randomness complexity is logarithmic and the degree is polylogarithmic.

Corollary 1.0.1. *Given domain size $N = |\mathbb{F}_q^m|$, accuracy error $\epsilon = (\log N)^{-\Theta(1)}$, confidence error $\delta = N^{-\Theta(1)}$, and large enough field size $q = (\log N)^{\Theta(1)}$, there exists an explicit degree-t curve sampler for the domain $\mathbb{F}_q^m$ with accuracy error ϵ, confidence error δ, randomness complexity $O(\log N)$, sample complexity q, and $t \leq (\log N)^c$ for some constant $c > 0$ independent of the field size q.*

It remains an open problem to explicitly construct curve samplers that have optimal randomness complexity $O(\log N + \log(1/\delta))$ (up to a constant factor), and sample curves with optimal degree bound $O(\log_q(1/\delta))$. It is also an interesting problem to achieve the optimal randomness complexity up to a $1+\gamma$ factor for any constant $\gamma > 0$ (rather than just an $O(1)$ factor), as achieved by [Zuc97] for general samplers. The standard techniques in [Zuc97] are not directly applicable as they increase the dimension of samples and only yield $O(1)$-dimensional manifold samplers.

Techniques

Extractor machinery. It was shown in [Zuc97] that samplers are equivalent to *extractors*, objects that convert weakly random distributions into almost uniform distributions. Therefore the techniques of constructing extractors are extremely useful in constructing curve samplers. Our construction employs the technique of block source extraction [NZ96, Zuc97, SZ99]. In addition, we also use the techniques appeared in [GUV09], especially their constructions of *condensers*.

Limited independence. It is well known that points on a random degree-$(t-1)$ curve are t-wise independent. So we may simply pick a random curve and use tail inequalities to bound the confidence error. However, the sample complexity is too high, and hence we need to use the technique of *iterated sampling* to reduce the number of sample points.

Iterated sampling. Iterated sampling is a useful technique for explicitly constructing randomness-efficient samplers [BR94, TSU06]. The idea is first picking a large sample from the domain and then draw a sub-sample from the previous sample. The drawback of iterated sampling, however, is that it invests randomness twice while the confidence error does not shrink correspondingly. To remedy this problem, we add another ingredient into our construction, namely the technique of *error reduction*.

Error reduction via list-recoverable codes. We will use explicit list-recoverable codes (a strengthening of list-decodable codes [GI01]). More specifically, we will employ the list-recoverability from (folded) Reed-Solomon codes [GR08, GUV09]. List-recoverable codes provide a way of obtaining samplers with very small confidence error from those with mildly small confidence error. We refer to this transformation as *error reduction*, which plays a key role in our construction.

Sketch of the construction

Our curve sampler is the composition of two samplers which we call the *outer sampler* and the *inner sampler* respectively. The outer sampler picks manifolds of dimension $O(\log m)$ from the domain $\mathcal{M} = \mathbb{F}_q^m$. The outer sampler has near-optimal randomness complexity but the sample complexity is large. To fix this problem, we employ the idea of iterated sampling. Namely we regard the manifold picked by the outer sampler as the new domain $\mathcal{M}'$, and then construct an inner sampler picking a curve from $\mathcal{M}'$ with small sample complexity.

The outer sampler is obtained by constructing an extractor and then using the extractor-sampler connection [Zuc97]. We follow the approach in [NZ96, Zuc97, SZ99]: Given an arbitrary random source with enough min-entropy, we will first use a *block source converter* to convert it into a *block source*, and then feed it to a *block source extractor*. In addition, we need to construct these components carefully so as to maintain the low-degree-ness. The way we construct the block source converter is different from those in [NZ96, Zuc97, SZ99] (as they are not in the form of low-degree polynomial maps), and is based on the Reed-Solomon condenser proposed in [GUV09]: To obtain one block, we simply feed the random source and a fresh new seed into the condenser, and let the output be the block. We show that this indeed gives a block source.

The inner sampler is constructed using techniques of iterated sampling and error reduction. We start with the basic curve samplers picking totally random curves, and then apply the error reduction as well as iterated sampling techniques repeatedly to obtain the desired inner sampler. Either of the two operations improves one parameter while worsening some other one: Iterated sampling reduces sample complexity but increases the randomness

complexity, whereas error reduction reduces the confidence error but increases the sample complexity. Our construction applies the two techniques alternately such that (1) we keep the invariant that the confidence error is always exponentially small in the randomness complexity, and (2) the sample complexity is finally brought down to q. We remark that the idea of sandwiching several operations to get the desired parameters without spoiling other ones is reminiscent of Reingold's proof that $\mathsf{SL} = \mathsf{L}$ [Rei08] and Dinur's proof of the PCP theorem [Din07].

Outline

The organization of this thesis is as follows: In Chapter 2 we introduce the preliminary definitions and notions as well as some basic facts that will be used later. In Chapter 3 we present some basic results about samplers and curve samplers. Chapter 4 is devoted to the main result of this thesis which describes an explicit construction of curve samplers. We divide this construction into two parts, the outer sampler (Section 4.1) and the inner sampler (Section 4.2). We then put it together and finish the construction of the curve samplers in Section 4.3. We present an alternative and simpler construction of outer samplers in Section 4.4.

Chapter 2

Preliminaries

Notations and basic definitions

We denote the set of numbers $\{1, 2, \ldots, n\}$ by $[n]$. Given a prime power q, write $\mathbb{F}_q$ for the finite field of size q. Write U_S for the uniform distribution over a finite set S, $U_{n,q}$ for the uniform distribution over $\mathbb{F}_q^n$, and U_n for the uniform distribution over $\{0,1\}^n$. Logarithms are taken with base 2 unless the base is explicitly specified.

Random variables and distributions are represented by upper-case letters whereas their specific values are represented by lower-case letters. Write $x \leftarrow X$ if x is sampled according to X. Write $X \subset S$ if X is a distribution over a set S. The support of a distribution $X \subset S$ is $\mathrm{supp}(X) \stackrel{\text{def}}{=} \{x \in S : \Pr[X = x] > 0\}$. We use the *statistical distance* $\Delta(\cdot, \cdot)$ to measure the closeness of two distributions. The statistical distance between $X, Y \subset S$ is defined as

$$\Delta(X, Y) = \max_{T \subseteq S} |\Pr[X \in T] - \Pr[Y \in T]|.$$

Then $\Delta(\cdot, \cdot)$ defines a metric. We say X is ϵ-*close* to Y if $\Delta(X, Y) \leq \epsilon$.

Fact 1. *The statistical distance is half the ℓ_1 distance, i.e., for $X, Y \subset S$, we have*

$$\Delta(X, Y) = \frac{1}{2} \sum_{x \in S} |\Pr[X = x] - \Pr[Y = x]|.$$

For an event A, let $\mathbf{I}[A]$ be the indicator variable that evaluates to 1 if A occurs and 0 otherwise. For a distribution X and an event A that occurs with nonzero probability, define the *conditional distribution* $X|_A$ by $\Pr[X|_A = x] = \frac{\Pr[(X=x) \wedge A]}{\Pr[A]}$.

We use forms like $\{t_x : x \in I\} \in S^I$ to denote a collection of elements indexed by I with each element in the set S. Alternatively, we view $\{t_x : x \in I\}$ as the function from I to S that maps x to t_x. We also slightly abuse the notation and use $\{t_x : x \in I\}$ for an

(unordered) multi-set.

Indeterminates are written as upper-case letters. E.g., we use $\mathbb{F}[X_1, \ldots, X_n]$ to denote the polynomial ring over the field $\mathbb{F}$ with indeterminates $X_1, \ldots, X_n$. We say a polynomial $p(X_1, \ldots, X_n)$ has degree t if the sum of the individual degrees $\sum_{i=1}^{n} a_i$ is at most t for all monomial $\prod_{i=1}^{n} X_i^{a_i}$ of p.

Facts about curves and manifolds

The following facts will be useful:

Fact 2. *For any distinct $x_1, \ldots, x_{t+1} \in \mathbb{F}_q$ and any $y_1, \ldots, y_{t+1} \in \mathbb{F}_q^n$, there exists a unique curve $C : \mathbb{F}_q \to \mathbb{F}_q^n$ of degree t such that $C(x_i) = y_i$ for all $i \in [t+1]$. Indeed, C_i's are given by the Lagrange polynomials:*

$$C_i(X) = \sum_{j=1}^{t+1} y_{j,i} \prod_{k \in [t+1] \setminus \{j\}} \frac{X - x_k}{x_j - x_k} \quad \text{where } y_{j,i} \text{ is the } i\text{-th coordinate of } y_j,\ i \in [n].$$

Fact 3. *Let $f_1 : \mathbb{F}_q^{d_1} \to \mathbb{F}_q^{d_0}$ be a manifold of degree t_1 and $f_2 : \mathbb{F}_q^{d_2} \to \mathbb{F}_q^{d_1}$ a manifold of degree t_2. Then $f_1 \circ f_2 : \mathbb{F}_q^{d_2} \to \mathbb{F}_q^{d_0}$ is a manifold of degree $t_1 t_2$.*

We also need the following lemma, generalizing the one in [TSU06]:

Lemma 2.0.1. *A manifold $f : \left(\mathbb{F}_{q^D}\right)^n \to \left(\mathbb{F}_{q^D}\right)^m$ of degree t, when viewed as a function $f : \left(\mathbb{F}_q^D\right)^n \to \left(\mathbb{F}_q^D\right)^m$, is also of degree t.*

Proof. Write $f = (f_1, \ldots, f_m)$. By symmetry we just show f_1, when viewed as a function $f_1 : \left(\mathbb{F}_q^D\right)^n \to \mathbb{F}_q^D$, has degree t. Suppose

$$f_1(x_1, \ldots, x_n) = \sum_{\substack{d=(d_1,\ldots,d_n) \\ \sum d_i \le t}} c_d \prod_{i=1}^{n} {x_i}^{d_i}.$$

Let $(e_1, \ldots, e_D)$ be the standard basis of $\mathbb{F}_q^D$ over $\mathbb{F}_q$. Writing the i-th variable $x_i \in \mathbb{F}_{q^D}$ as $\sum_{j=1}^{D} x_{i,j} e_j \in \mathbb{F}_q^D$ with $x_{i,j} \in \mathbb{F}_q$, and each coefficient $c_d \in F_{q^D}$ as $\sum_{j=1}^{D} c_{d,j} e_j$ with $c_{d,j} \in \mathbb{F}_q$, we obtain

$$f_1(x_1, \ldots, x_n) = \sum_{\substack{d=(d_1,\ldots,d_n) \\ \sum d_i \le t}} \left(\sum_{j=1}^{D} c_{d,j} e_j \right) \prod_{i=1}^{n} \left(\sum_{j=1}^{D} x_{i,j} e_j \right)^{d_i}.$$

After multiplying out, each monomial has degree at most $\max_d \sum_i d_i \le t$, and their coefficients are polynomials in the e_i elements. Rewriting each of these values in the basis

$(e_1, \ldots, e_D)$ and gathering the coefficients on e_i, we obtain the i-th coordinate function of f_1 that has degree at most d for all $1 \le i \le D$. Therefore $f_1 : \left(\mathbb{F}_q^D\right)^n \to \mathbb{F}_q^D$ is a manifold of degree t, and so is $f : \left(\mathbb{F}_q^D\right)^n \to \left(\mathbb{F}_q^D\right)^m$. □

Tail probability bounds

We say random variables $X_1, \ldots, X_n$ are *independent* if for any specific values $x_1, \ldots, x_n$, it holds that

$$\Pr\left[\bigwedge_{i=1}^{n} X_i = x_i\right] = \prod_{i=1}^{n} \Pr[X_i = x_i].$$

We say $X_1, \ldots, X_n$ are *pairwise independent* if for any specific values x_1, x_2 and any distinct $i_1, i_2 \in [n]$, it holds that

$$\Pr\left[X_{i_1} = x_1 \wedge X_{i_2} = x_2\right] = \Pr[X_{i_1} = x_1] \Pr[X_{i_2} = x_2].$$

In general, for an integer $t > 1$, we say $X_1, \ldots, X_n$ are *t-wise independent* if for any specific values $x_1, \ldots, x_t$ and any distinct $i_1, \ldots, i_t \in [n]$, it holds that

$$\Pr\left[\bigwedge_{j=1}^{t} X_{i_j} = x_j\right] = \prod_{j=1}^{t} \Pr[X_{i_j} = x_j].$$

We consider the behaviour of a fully random curve C of degree t over $\mathbb{F}_q$, i.e., write $C = (C_1, \ldots, C_n)$, then each C_i is a degree-t univariate polynomial whose $t+1$ coefficients are chosen uniformly at random from $\mathbb{F}_q$, and all C_i's are chosen independently. It is known that the points on C are $(t+1)$-wise independent.

Lemma 2.0.2. *Let $C : \mathbb{F}_q \to \mathbb{F}_q^n$ be a random curve of degree t over $\mathbb{F}_q$. Then the random variables $C(x)$'s are $(t+1)$-wise independent where x ranges over $\mathbb{F}_q$.*

Proof. First note that each $C(x)$ is uniformly distributed. By Fact 2, for any distinct $y_1, \ldots, y_{t+1} \in \mathbb{F}_q$ and any $z_1, \ldots, z_{t+1} \in \mathbb{F}_q^n$, there is a unique degree-t curve, out of all $q^{(t+1)n}$ curves, that passes z_i at y_i for all $i \in [t+1]$. So we have

$$\Pr\left[\bigwedge_{i=1}^{t+1} C(y_i) = z_i\right] = \frac{1}{q^{(t+1)n}} = \prod_{i=1}^{t+1} \Pr[C(y_i) = z_i].$$

By definition, the random variables $C(x)$'s with x ranging over $\mathbb{F}_q$ are $(t+1)$-wise independent. □

We need the Chernoff bound of the following form:

Lemma 2.0.3. *Suppose $X_1, \ldots, X_n \subset [0,1]$ are independent random variables. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$, and let $R \geq 6\mu$. Then*

$$\Pr[X \geq R] \leq 2^{-R}.$$

The following bound follows from Chebyshev's inequality:

Lemma 2.0.4. *Suppose $X_1, \ldots, X_n$ are pairwise independent random variables. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$, and let $A > 0$. Then*

$$\Pr[|X - \mu| \geq A] \leq \frac{\sum_{i=1}^n \mathbf{Var}[X_i]}{A^2}.$$

We will also use the following tail bound for t-wise independent random variables:

Lemma 2.0.5 ([BR94])**.** *Let $t \geq 4$ be an even integer. Suppose $X_1, \ldots, X_n \subset [0,1]$ are t-wise independent random variables. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$, and let $A > 0$. Then*

$$\Pr[|X - \mu| \geq A] = O\left(\left(\frac{t\mu + t^2}{A^2}\right)^{t/2}\right).$$

Basic line/curve samplers

The simplest line samplers are those picking completely random lines, as defined below. We call them as *basic line samplers*.

Definition 2.0.4 (basic line sampler)**.** For $m \geq 1$ and prime power q, let $\mathsf{Line}_{m,q} : \mathbb{F}_q^{2m} \times \mathbb{F}_q \to \mathbb{F}_q^m$ be the line sampler that picks a uniformly random line in $\mathbb{F}_q^m$. Formally,

$$\mathsf{Line}_{m,q}((a,b),y) \stackrel{\text{def}}{=} (a_1 y + b_1, \ldots, a_m y + b_m)$$

for $a = (a_1, \ldots, a_m), b = (b_1, \ldots, b_m) \in \mathbb{F}_q^m$ and $y \in \mathbb{F}_q$.

Remark 1. Note that although $\mathsf{Line}_{m,q}(x, \cdot)$ is a line (i.e., a degree-1 curve) for $x \in \mathbb{F}_q^{2m}$, the function $\mathsf{Line}_{m,q}$ itself is a degree-2 manifold.

The basic line samplers are indeed good samplers:

Lemma 2.0.6. *For $\epsilon > 0$, $m \geq 1$ and prime power q, $\mathsf{Line}_{m,q}$ is an $\left(\epsilon, \frac{1}{\epsilon^2 q}\right)$ line sampler.*

Proof. Let A be an arbitrary subset of $\mathbb{F}_q^m$. Note that $\mathsf{Line}_{m,q}$ picks a line uniformly at random. By Lemma 2.0.2, the random variables $\mathsf{Line}_{m,q}(U_{2m,q}, y)$ with y ranging over $\mathbb{F}_q$ are pairwise independent. So the indicator variables $\mathbf{I}[\mathsf{Line}_{m,q}(U_{2m,q}, y) \in A]$ with y ranging over $\mathbb{F}_q$ are also pairwise independent. Applying Lemma 2.0.4, we get

$$\begin{aligned}
&\Pr_{x \leftarrow U_{2m,q}} \left[\left| \mu_{\mathsf{Line}_{m,q}(x,\cdot)}(A) - \mu(A) \right| > \epsilon \right] \\
&= \Pr \left[\left| \sum_{y \in \mathbb{F}_q} \mathbf{I}[\mathsf{Line}_{m,q}(U_{2m,q}, y) \in A] - \mathbb{E} \left[\sum_{y \in \mathbb{F}_q} \mathbf{I}[\mathsf{Line}_{m,q}(U_{2m,q}, y) \in A] \right] \right| > \epsilon q \right] \\
&\leq \frac{\sum_{y \in \mathbb{F}_q} \mathbf{Var}\left[\mathbf{I}[\mathsf{Line}_{m,q}(U_{2m,q}, y) \in A] \right]}{\epsilon^2 q^2} \\
&\leq \frac{1}{\epsilon^2 q}.
\end{aligned}$$

By definition, $\mathsf{Line}_{m,q}$ is an $\left(\epsilon, \frac{1}{\epsilon^2 q}\right)$ line sampler. □

Similarly we consider the simplest low-degree curve samplers that pick completely random curves. We call them *basic curve samplers*.

Definition 2.0.5 (basic curve sampler)**.** For $m \geq 1$, $t \geq 4$ and prime power q, let $\mathsf{Curve}_{m,t,q} : \mathbb{F}_q^{tm} \times \mathbb{F}_q \to \mathbb{F}_q^m$ be the curve sampler that picks a uniformly random curve of degree $t-1$ in $\mathbb{F}_q^m$. Formally,

$$\mathsf{Curve}_{m,t,q}((c_0, \ldots, c_{t-1}), y) \stackrel{\text{def}}{=} \left(\sum_{i=0}^{t-1} c_{i,1} y^i, \ldots, \sum_{i=0}^{t-1} c_{i,m} y^i \right)$$

for each $c_0 = (c_{0,1}, \ldots, c_{0,m}), \ldots, c_{t-1} = (c_{t-1,1}, \ldots, c_{t-1,m})$ and $y \in \mathbb{F}_q$.

Remark 2. Note that $\mathsf{Curve}_{m,t,q}$ is a manifold of degree t.

The basic curve samplers are indeed good samplers:

Lemma 2.0.7. *For $\epsilon > 0$, $m \geq 1$, $t \geq 4$ and sufficiently large prime power $q = (t/\epsilon)^{O(1)}$, $\mathsf{Curve}_{m,t,q}$ is an $\left(\epsilon, q^{-t/4}\right)$ sampler.*

Proof. Let A be an arbitrary subset of $\mathbb{F}_q^m$. Note that $\mathsf{Curve}_{m,t,q}(x, \cdot)$ picks a degree-$(t-1)$ curve uniformly at random. By Lemma 2.0.2, the random variables $\mathsf{Curve}_{m,t,q}(U_{tm,q}, y)$ with y ranging over $\mathbb{F}_q$ are t-wise independent. So the indicator variables $\mathbf{I}[\mathsf{Curve}_{m,t,q}(U_{tm,q}, y) \in A]$

with y ranging over $\mathbb{F}_q$ are also t-wise independent. Applying Lemma 2.0.5, we get

$$\begin{aligned}
&\Pr\left[\left|\mu_{\mathsf{Curve}_{m,t,q}(U_{tm,q},\cdot)}(A) - \mu(A)\right| > \epsilon\right] \\
&= \Pr\left[\left|\sum_{y\in\mathbb{F}_q} \mathbf{I}[\mathsf{Curve}_{m,t,q}(U_{tm,q}, y) \in A] - \mathbb{E}\left[\sum_{y\in\mathbb{F}_q} \mathbf{I}[\mathsf{Curve}_{m,t,q}(U_{tm,q}, y) \in A]\right]\right| > \epsilon q\right] \\
&= O\left(\left(\frac{tq\mu(A) + t^2}{\epsilon^2 q^2}\right)^{t/2}\right) = q^{-t/4}
\end{aligned}$$

provided that $q = (t/\epsilon)^{O(1)}$ is sufficiently large. By definition, $\mathsf{Curve}_{m,t,q}$ is an $\left(\epsilon, q^{-t/4}\right)$ sampler. □

Extractors and condensers

A *(seeded) extractor* is an object that takes an imperfect random variable (i.e., a random variable that contains some randomness but is not uniformly distributed) called the *(weakly) random source*, invests a small amount of randomness called the *seed*, and produces an output whose distribution is very close to the uniform distribution.

We introduce the following notion to measure the amount of randomness contained in a random source.

Definition 2.0.6 (min-entropy)**.** We say a random variable X over a set S has *min-entropy* k and *entropy deficiency* $\log|S| - k$ if for any $x \in S$, it holds that $\Pr[X = x] \le 2^{-k}$. The min-entropy of X is at most $\log|S|$, and it achieves $\log|S|$ iff $X = U_S$.

We say X has *q-ary min-entropy k* if for any $x \in S$, it holds that $\Pr[X = x] \le q^{-k}$ (or equivalently, X has min-entropy $k \log q$).

Lemma 2.0.8 (chain rule for min-entropy)**.** *Let (X, Y) be a joint distribution where $X \subset \mathbb{F}_q^\ell$ and Y has q-ary min-entropy k. We have*

$$\Pr_{x\leftarrow X}\left[Y|_{X=x} \text{ has } q\text{-ary min-entropy } k - \ell - \log_q(1/\epsilon)\right] \ge 1 - \epsilon.$$

Proof. We say $x \in \mathrm{supp}(X)$ is *good* if $\Pr[X = x] \ge \epsilon q^{-\ell}$ and *bad* otherwise. Then $\Pr_{x\leftarrow X}[x \text{ is bad}] \le \mathrm{supp}(X)\epsilon q^{-\ell} \le \epsilon$. Consider arbitrary good x. For any specific value y for Y, we have $\Pr[Y|_{X=x} = y] = \frac{\Pr[(Y=y)\wedge(X=x)]}{\Pr[X=x]} \le \frac{\Pr[Y=y]}{\epsilon q^{-\ell}} \le q^{-\left(k-\ell-\log_q(1/\epsilon)\right)}$. By definition, $Y|_{X=x}$ has q-ary min-entropy $k - \ell - \log_q(1/\epsilon)$ when $X = x$ is good, which occurs with probability at least $1 - \epsilon$. □

Before giving the formal definition of extractors, we first consider a kind of objects called *condensers* that can be seen as a relaxation of extractors. A condenser is weaker than an extractor in the sense that the output is only required to be close to a distribution with a large amount of min-entropy, rather than close to the uniform distribution.

Definition 2.0.7 (condenser)**.** Given a function $C : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$, we say C is an $(n, k_1) \to_{\epsilon,q} (m, k_2)$ *condenser* if for every distribution X with q-ary min-entropy k_1, $C(X, U_{d,q})$ is ϵ-close to a distribution with q-ary min-entropy k_2. The second argument of C is its *seed*. The quantities $n \log q$, $d \log q$ and $m \log q$ are called the *input length*, *seed length* and *output length* of C respectively. We call $k_1 \log q$ the *min-entropy threshold* of C and ϵ the *error* of C.

Next we define extractors as the strengthening of condensers.

Definition 2.0.8 (extractor)**.** The function $E : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is a (k, ϵ, q) *extractor* if it is a $(n, k) \to_{\epsilon,q} (m, m)$ condenser. The *seed*, *input length*, *seed length*, *output length*, *min-entropy threshold*, and *error* of the extractor E are the same as the corresponding parameters of E as a condenser.

We say a condenser/extractor $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ has *degree* t if f has degree t as a manifold in $\mathbb{F}_q^{n+d}$.

A random source X is called a *flat source* if it is uniformly distributed over its support, i.e.,

$$\Pr[X = a] = \begin{cases} \frac{1}{|\mathrm{supp}(X)|} & a \in \mathrm{supp}(X), \\ 0 & \text{otherwise.} \end{cases}$$

The following basic fact will be useful:

Fact 4. *A random source X of min-entropy k is a convex combination of flat sources of min-entropy k.*

Write $X = \sum_{i \in I} c_i X_i$ as such a convex combination. Then we have

$$\Delta(X, Y) = \Delta\left(\sum_{i \in I} c_i X_i, \sum_{i \in I} c_i Y\right) \leq \sum_{i \in I} c_i \Delta(X_i, Y) \leq \sup_{i \in I} \left(\Delta(X_i, Y)\right).$$

Thus, we may assume that the input is always a flat source of min-entropy k when proving the extractor or condenser property for min-entropy threshold k.

Block source extraction

One important class of random sources is the class of *block sources*, first introduced in [CG88]. A block source is a random source with the property that conditioning on any prefix of blocks, the remaining blocks still have some min-entropy.

Definition 2.0.9 (block source [CG88])**.** A random source $X = (X_1, \ldots, X_s) \subset \mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}$ with each $X_i \subset \mathbb{F}_q^{n_i}$ is a $(k_1, \ldots, k_s)$ *q-ary block source* if for any $i \in [s]$ and $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1})$, the conditional distribution $X_i|_{X_1=x_1,\ldots,X_{i-1}=x_{i-1}}$ has q-ary min-entropy k_i. Each X_i is called a *block*.

We consider the problem of extracting randomness from block sources:

Definition 2.0.10 (block source extractor)**.** A function $E : (\mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}) \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is called a $((k_1, \ldots, k_s), \epsilon, q)$ *block source extractor* if for any $(k_1, \ldots, k_s)$ q-ary block source $(X_1, \ldots, X_s) \subset \mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}$, $E((X_1, \ldots, X_s), U_{d,q})$ is ϵ-close to $U_{m,q}$.

One nice property of block sources is that their special structures allow us to compose several extractors and get a block source extractor with only a small amount of randomness invested.

Definition 2.0.11 (block source extraction by composition)**.** Let $s \geq 1$ be an integer and for each $i \in [s]$, let $E_i : \mathbb{F}_q^{n_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{m_i}$ be a map. Suppose that $m_i \geq d_{i-1}$ for all $i \in [s]$, where we define $d_0 = 0$. Define $E = \mathsf{BlkExt}(E_1, \ldots, E_s)$ as follows:

$$E : (\mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}) \times \mathbb{F}_q^{d_s} \to (\mathbb{F}_q^{m_1 - d_0} \times \cdots \times \mathbb{F}_q^{m_s - d_{s-1}})$$
$$((x_1, \ldots, x_s), y_s) \mapsto (z_1, \ldots, z_s)$$

where for $i = s, \ldots, 1$, we iteratively define (y_{i-1}, z_i) to be a partition of $E_i(x_i, y_i)$ into the prefix $y_{i-1} \in \mathbb{F}_q^{d_{i-1}}$ and the suffix $z_i \in \mathbb{F}_q^{m_i - d_{i-1}}$.

See Figure 2.1 for an illustration of the above definition.

Lemma 2.0.9. *Let $s \geq 1$ be an integer and for each $i \in [s]$, let $E_i : \mathbb{F}_q^{n_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{m_i}$ be a (k_i, ϵ_i, q) extractor of degree $t_i \geq 1$. Then $E = \mathsf{BlkExt}(E_1, \ldots, E_s)$ is a $((k_1, \ldots, k_s), \epsilon, q)$ block source extractor of degree t where $\epsilon = \sum_{i=1}^s \epsilon_i$ and $t = \prod_{i=1}^s t_i$.*

Proof. Induct on s. When $s = 1$ the claim follows from the extractor property of E_1.

When $s > 1$, assume the claim holds for all $s' < s$. Define $E' = \mathsf{BlkExt}(E_2, \ldots, E_s)$. By the induction hypothesis, E' is a $((k_2, \ldots, k_s), \epsilon', q)$ block source extractor where $\epsilon' = \sum_{i=2}^s \epsilon_i$ and $t' = \prod_{i=2}^s t_i$.

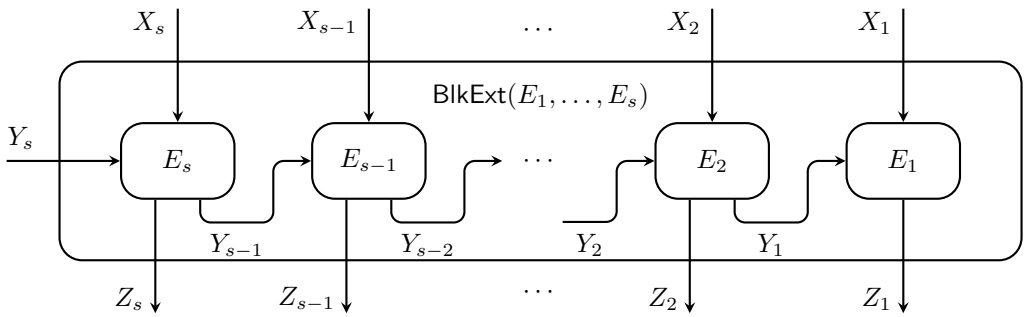

Figure 2.1: The composed block source extractor $\mathsf{BlkExt}(E_1, \ldots, E_s)$

Let $(X_1, \ldots, X_s) \subset \mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}$ be an arbitrary $(k_1, \ldots, k_s)$ q-ary block source. Let $(Y_{i-1}, Z_i) \subset \mathbb{F}_q^{d_{i-1}} \times \mathbb{F}_q^{m_i - d_{i-1}}$ be the output of E_i and $(Z_1, \ldots, Z_s)$ be the the output of E when $(X_1, \ldots, X_s)$ is fed into E as the input and an independent uniform distribution $Y_s = U_{d_s,q}$ is used as the seed (c.f. Figure 2.1). The output of E' is then $(Y_1, Z_2, \ldots, Z_s)$.

Fix $x \in \mathrm{supp}(X_1)$. By the definition of block sources, the distribution $(X_2, \ldots, X_s)|_{X_1=x}$ is a $(k_2, \ldots, k_s)$ q-ary block source. Also note that $Y_s|_{X_1=s} = Y_s$ is an independent uniform distribution. By the induction hypothesis, the distribution

$$(Y_1, Z_2, \ldots, Z_s)|_{X_1=x} = E'((X_2, \ldots, X_s), Y_s)|_{X_1=x}$$

is ϵ'-close to $(U_{d_1,q}, U_{m_2-d_1,q}, \ldots, U_{m_s-d_{s-1},q})$. As this holds for all $x \in \mathrm{supp}(X_1)$, the distribution $(X_1, Y_1, Z_2, \ldots, Z_s)$ is ϵ'-close to $(X_1, U_{d_1,q}, U_{m_2-d_1,q}, \ldots, U_{m_s-d_{s-1},q})$. So the distribution

$$E((X_1, \ldots, X_s), Y_s) = (E_1(X_1, Y_1), Z_2, \ldots, Z_s)$$

is ϵ'-close to $(E_1(X_1, U_{d_1,q}), U_{m_2-d_1,q}, \ldots, U_{m_s-d_{s-1},q})$. Then we know that it is also ϵ-close to $(U_{m_1-d_0,q}, U_{m_2-d_1,q}, \ldots, U_{m_s-d_{s-1},q})$ since E_1 is a (k_1, ϵ_1, q) extractor.

Finally, to see that E has degree t, note that $E'((X_2, \ldots, X_s), Y_s) = (Y_1, Z_2, \ldots, Z_s)$ has degree t' in its variables $X_2, \ldots, X_s$ and Y_s by the induction hypothesis and hence $Y_1, Z_2, \ldots, Z_s$ have degree t' in these variables. Then $Z_1 = E_1(X_1, Y_1)$ has degree $t_1 \cdot \max\{1, t'\} = t$ in $X_1, \ldots, X_s$ and Y_s. So $E((X_1, \ldots, X_s), Y_s) = (Z_1, \ldots, Z_s)$ has degree t in $X_1, \ldots, X_s$ and Y_s. $\square$

Chapter 3

Basic results

3.1 Extractor vs. sampler connection

Our construction of curve samplers relies on the following observation by [Zuc97] which shows the equivalence between extractors and samplers.

Theorem 3.1.1 ([Zuc97], restated)**.** *Given a map $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$, we have the following:*

(1) If f is a (k, ϵ, q) extractor, then it is also an (ϵ, δ) sampler where $\delta = 2q^{k-n}$.

(2) If f is an $(\epsilon/2, \delta)$ sampler where $\delta = \epsilon q^{k-n}$, then it is also a (k, ϵ, q) extractor.

Proof. (Extractor to sampler) Assume to the contrary that f is not an (ϵ, δ) sampler. Then there exists a subset $A \subseteq \mathbb{F}_q^m$ with $\Pr_x[|\mu_{f(x)}(A) - \mu(A)| > \epsilon] > \delta$. Then either $\Pr_x[\mu_{f(x)}(A) - \mu(A) > \epsilon] > \delta/2$ or $\Pr_x[\mu_{f(x)}(A) - \mu(A) < -\epsilon] > \delta/2$. Assume $\Pr_x[\mu_{f(x)}(A) - \mu(A) > \epsilon] > \delta/2$ (the other case is symmetric). Let X be the uniform distribution over the set of x such that $\mu_{f(x)}(A) - \mu(A) > \epsilon$, i.e., $\Pr_y[f(x, y) \in A] - \Pr_x[x \in A] > \epsilon$. Then $|f(X, U_{d,q}) - U_{m,q}| > \epsilon$. But the q-ary min-entropy of X is at least $\log_q((\delta/2)q^n) \geq k$, contradicting the extractor property of f.

(Sampler to extractor) Assume to the contrary that f is not a (k, ϵ, q) extractor. Then there exists a subset $A \subseteq \mathbb{F}_q^m$ and a random source X of q-ary min-entropy k satisfying the property that $|\Pr[f(X, U_{d,q}) \in A] - \mu(A)| > \epsilon$. We may assume X is a flat source (i.e. uniformly distributed over its support) since a general source with q-ary min-entropy k is a convex combination of flat sources with q-ary min-entropy k. Note that $|\mathrm{supp}(X)| \geq q^k$. By the averaging argument, for at least an ϵ-fraction of $x \in \mathrm{supp}(X)$, we have $|\Pr[f(x, U_{d,q}) \in A] - \mu(A)| > \epsilon/2$. But it implies that for x uniformly chosen from $\mathbb{F}_q^n$, with probability at least $\epsilon\mu(\mathrm{supp}(X)) = \delta$ we have $|\mu_{f(x)}(A) - \mu(A)| > \epsilon/2$, contradicting the sampling property of f. □

Table 3.1 shows the rough correspondences between the parameters of extractors and those of samplers.

extractor	sampler
error ϵ	accuracy error ϵ
entropy deficiency $(n-k)\log q$	confidence error δ
seed length $d \log q$	sample complexity q^d
input length $n \log q$	randomness complexity $n \log q$
output length $m \log q$	domain size q^m

Table 3.1: The correspondences between the parameters of extractors and samplers

3.2 Existence of a good curve sampler

In this section, we prove the existence of a (non-explicit) low-degree curve sampler with low randomness complexity and a small number of sample points.

Theorem 3.2.1. *For any $m \geq 1$, $\epsilon, \delta > 0$ and sufficiently large $q \geq \left(\frac{\log(1/\delta)}{\epsilon}\right)^{\Theta(1)}$, there exists a (non-explicit) (ϵ, δ) degree-t curve sampler $S : \mathcal{D} \times \mathbb{F}_q \to \mathbb{F}_q^m$ with randomness complexity $\log |\mathcal{D}| = m \log q + \log(1/\delta) + O(1)$, sample complexity q, and $t = O\left(\log_q (1/\delta)\right)$.*

Proof. We use the probabilistic method. Choose the curve sampler S by choosing the degree-t curve $S(x, \cdot) : \mathbb{F}_q \to \mathbb{F}_q^m$ independently at random for each $x \in \mathcal{D}$.

Let A be an arbitrary subset of $\mathbb{F}_q^m$. Fix $x \in \mathcal{D}$. By Lemma 2.0.2, the random variables $S(x, y)$ with y ranging over $\mathbb{F}_q$ are $(t+1)$-wise independent. So the indicator variables $\mathbf{I}[S(x, y) \in A]$ with y ranging over $\mathbb{F}_q$ are also $(t+1)$-wise independent. Applying Lemma 2.0.5, we get

$$\begin{aligned}\Pr\left[\left|\mu_{S(x,\cdot)}(A) - \mu(A)\right| > \epsilon\right] &= \Pr\left[\left|\sum_{y \in \mathbb{F}_q} \mathbf{I}[S(x,y) \in A] - \mathbb{E}\left[\sum_{y \in \mathbb{F}_q} \mathbf{I}[S(x,y) \in A]\right]\right| > \epsilon q\right] \\ &= O\left(\left(\frac{(t+1)q\mu(A) + (t+1)^2}{\epsilon^2 q^2}\right)^{(t+1)/2}\right) \leq \frac{\delta}{6}\end{aligned}$$

for sufficiently large β. Let $B(x)$ be the event that $\left|\mu_{S(x,\cdot)}(A) - \mu(A)\right| > \epsilon$. Then $0 \leq \Pr[\mathbf{I}[B(x)] = 1] \leq \frac{\delta}{6}$ and hence $\mathbb{E}\left[\sum_x \mathbf{I}[B(x)]\right] \leq \frac{\delta|\mathcal{D}|}{6}$. The indicator variables $\mathbf{I}[B(x)]$ with x ranging over $\mathcal{D}$ are independent. Applying Lemma 2.0.3, we obtain

$$\Pr\left[\sum_x \mathbf{I}[B(x)] \geq \delta|\mathcal{D}|\right] \leq 2^{-\delta|\mathcal{D}|}.$$

There are 2^{q^m} possible $A \subseteq \mathbb{F}_q^m$. So with probability at least $1 - 2^{q^m} 2^{-\delta|\mathcal{D}|} > 0$ (for sufficiently large $\log|\mathcal{D}| = m\log q + \log(1/\delta) + O(1)$), the events $\sum_x \mathbf{I}[B(x)] \leq \delta|\mathcal{D}|$ for all $A \subseteq \mathbb{F}_q^m$ occur by the union bound. Take the curve sampler S that makes all these events occur. Then S is an (ϵ, δ) degree-t curve sampler by definition. □

The most interesting case is when the domain size q^m and the confidence error δ are polynomially related, while the field size q and the degree t are kept small:

Corollary 3.2.1. *Given the domain size $N = |\mathbb{F}_q^m| = q^m$, accuracy error $\epsilon = (\log N)^{-O(1)}$, confidence error $\delta = N^{-O(1)}$, and large enough field size $q = (\log N)^{\Theta(1)}$, there exists a (non-explicit) (ϵ, δ) degree-t curve sampler $S : \mathcal{D} \times \mathbb{F}_q \to \mathbb{F}_q^m$ with randomness complexity $\log|\mathcal{D}| = O(\log N)$, sample complexity q, and $t = \Theta\left(\frac{\log N}{\log\log N}\right)$.*

3.3 Lower bounds

We will use the following optimal lower bound for extractors:

Theorem 3.3.1 ([RTS00], restated)**.** *Let $E : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ be a (k, ϵ, q) extractor. Then*

(a) if $\epsilon < 1/2$ and $q^d \leq q^m/2$, then $q^d = \Omega\left(\frac{(n-k)\log q}{\epsilon^2}\right)$, and

(b) if $q^d \leq q^m/4$, then $q^{d+k-m} = \Omega(1/\epsilon^2)$.

Theorem 3.3.2. *Let $S : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ be an (ϵ, δ) curve sampler where $\epsilon < 1/2$ and $m \geq 2$. Then*

(a) the sample complexity $q = \Omega\left(\frac{\log(2\epsilon/\delta)}{\epsilon^2}\right)$, and

(b) the randomness complexity $n\log q \geq (m-1)\log q + \log(1/\epsilon) + \log(1/\delta) - O(1)$.

Proof. By Theorem 3.1.1, S is a $(k, 2\epsilon, q)$ extractor where $k = n - \log_q(2\epsilon/\delta)$. The first claim then follows from Theorem 3.3.1 (a). Applying Theorem 3.3.1 (b), we get $(1 + k - m)\log q \geq \Omega(\log(1/\epsilon)) - O(1)$. Therefore

$$n\log q = k\log q + \log(2\epsilon/\delta) \geq (m-1)\log q + \log(1/\delta) + \Omega(\log(1/\epsilon)) - O(1).$$

□

In particular, as $\log(1/\epsilon) = O(\log q)$, the randomness complexity $n\log q$ is at least $\Omega(\log N + \log(1/\delta))$ when the domain size $N = q^m \geq N_0$ for some constant N_0. Therefore the randomness complexity in Theorem 1.0.1 is optimal up to a constant factor.

We also present the following lower bound on the degree of curves sampled by a curve sampler:

Theorem 3.3.3. *Let $S : \mathcal{N} \times \mathbb{F}_q \to \mathbb{F}_q^m$ be an (ϵ, δ) degree-t curve sampler where $m \geq 2$, $\epsilon < 1/2$ and $\delta < 1$. Then $t = \Omega\left(\log_q(1/\delta) + 1\right)$.*

Proof. Clearly $t \geq 1$. Suppose $S = (S_1, \ldots, S_m)$ and define $S' = (S_1, S_2)$. Let $\mathcal{C}$ be the set of curves of degree at most t in $\mathbb{F}_q^2$. Then $|\mathcal{C}| = q^{2(t+1)}$. Consider the map $\tau : \mathcal{N} \to \mathcal{C}$ that sends x to $S'(x, \cdot)$. We can pick $k = \lfloor q/2 \rfloor$ curves $C_1, \ldots, C_k \in \mathcal{C}$ such that the union of their preimages

$$B \stackrel{\text{def}}{=} \bigcup_{i=1}^{k} \tau^{-1}(C_i) = \bigcup_{i=1}^{k} \{x : S'(x, \cdot) = C_i\}$$

has size at least $\frac{k|\mathcal{N}|}{|\mathcal{C}|} = \frac{k|\mathcal{N}|}{q^{2(t+1)}}$.

Define $A \subseteq \mathbb{F}_q^m$ by

$$A \stackrel{\text{def}}{=} \{C_i(y) : i \in [k],\ y \in \mathbb{F}_q\} \times \mathbb{F}_q^{m-2},$$

i.e., let A be the set of points in $\mathbb{F}_q^m$ whose first two coordinates are on at least one curve C_i. We have $|A| \leq kq^{m-1}$ and hence $\mu(A) \leq k/q \leq 1/2 < 1 - \epsilon$. On the other hand, it follows from the definition of A that we have $S(x, y) \in A$ for all $x \in B$ and $y \in \mathbb{F}_q$. So $\mu_{S(x)}(A) = 1$ for all $x \in B$. Then $\delta \geq \Pr\left[|\mu_{S(x)}(A) - \mu(A)| > \epsilon\right] \geq \frac{|B|}{|\mathcal{N}|} \geq \frac{k}{q^{2(t+1)}}$ and hence $t \geq \max\left\{1, \frac{1}{2}\log_q(k/\delta) - 1\right\} = \Omega\left(\log_q(1/\delta) + 1\right)$. □

We remark that the condition $m \geq 2$ is necessary in Theorem 3.3.3 since when $m = 1$, the sampler S with $S(x, y) = y$ for all $x \in \mathcal{N}$ and $y \in \mathbb{F}_q$ is a $(0, 0)$ degree-1 curve sampler.

Chapter 4

Explicit constructions

4.1 Outer sampler

In this section we construct an $O(\log k)$-dimensional manifold sampler, which we called the "outer sampler", with the optimal randomness complexity where $k = n - \log_q(1/\delta)$. In the language of extractors, we construct an extractor with the seed in $\mathbb{F}_q^{O(\log k)}$ for random sources of q-ary min-entropy k.

4.1.1 Block source conversion

Definition 4.1.1 (block source converter [NZ96])**.** A function $C : \mathbb{F}_q^n \times \mathbb{F}_q^d \to (\mathbb{F}_q^{m_1} \times \cdots \times \mathbb{F}_q^{m_s})$ is called a $(k, (k_1, \ldots, k_s), \epsilon, q)$ *block source converter* if for any random source $X \subseteq \mathbb{F}_q^n$ of q-ary min-entropy k, the output $C(X, U_{d,q}) \subset \mathbb{F}_q^{m_1} \times \cdots \times \mathbb{F}_q^{m_s}$ is ϵ-close to a $(k_1, \ldots, k_s)$ q-ary block source. In addition, we say C has *degree* t if C has degree t as a manifold in $\mathbb{F}_q^{n+d}$.

It was shown in [NZ96] that one can obtain a block by choosing a pseudorandom subset of bits of the random source. Yet the proof is pretty delicate and cumbersome. Furthermore the resulting extractor does not have a nice algebraic structure. Here we make the observation that the following condenser from Reed-Solomon codes in [GUV09] can be used to obtain blocks and is a low-degree manifold.

Definition 4.1.2 (condenser from Reed-Solomon codes [GUV09])**.** Let $\zeta \in \mathbb{F}_q$ be a generator of the multiplicative group $\mathbb{F}_q^\times$. Define $\mathsf{RSCon}_{n,m,q} : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ for $n, m \geq 1$ and prime power q:

$$\mathsf{RSCon}_{n,m,q}(x, y) = \big(y, f_x(y), f_x(\zeta y), \ldots, f_x(\zeta^{m-2} y)\big)$$

where $f_x(Y) = \sum_{i=0}^{n-1} x_i Y^i$ for $x = (x_0, x_1, \ldots, x_{n-1}) \in \mathbb{F}_q^n$.

Theorem 4.1.1 ([GUV09])**.** *For any $h \geq 1$, $n \geq m \geq 1$, prime power q and $\epsilon > 0$, $\mathsf{RSCon}_{n,m,q}$ is an $\left(n, \log_q\left(\frac{H}{\epsilon}\right)\right) \to_{2\epsilon,q} \left(m, \log_q\left(\frac{L}{2\epsilon}\right)\right)$ condenser, where $H = (h-1)^{\frac{q^{m-1}-1}{q-1}}$ and $L = (\epsilon q - (n-1)(h-1)(m-1)) \cdot h^{m-1} - 1$. In particular, for large enough $q \geq (n/\epsilon)^{O(1)}$, $\mathsf{RSCon}_{n,m,q}$ is a $m \to_{\epsilon,q} 0.99m$ condenser.*

Remark 3. The condenser $\mathsf{RSCon}_{n,m,q}(x, y)$ is a degree-n manifold, as each monomial in any of its coordinate is of the form y or $x_i(\zeta^j y)^i$ where $i \leq n-1$.

Remark 4. The reason we use the condenser from Reed-Solomon codes rather than the ones from Parvaresh-Vardy codes [GUV09, TSU12] is that we need the condenser to be a low-degree manifold in both the seed and the random source. The known condensers from Parvaresh-Vardy codes are low-degree in the seed, yet we have no good bound on the degree in the random source.

We apply the above condenser on the random source with an independent seed to obtain a new block each time. Formally:

Definition 4.1.3 (block source converter via condensing)**.** For integers $n, m_1, \ldots, m_s \geq 1$ and prime power q, define the function $\mathsf{BlkCnvt}_{n,(m_1,\ldots,m_s),q} : \mathbb{F}_q^n \times \mathbb{F}_q^s \to \mathbb{F}_q^{m_1+\cdots+m_s}$ by

$$\mathsf{BlkCnvt}_{n,(m_1,\ldots,m_s),q}(x, y) = (\mathsf{RSCon}_{n,m_1,q}(x, y_1), \ldots, \mathsf{RSCon}_{n,m_s,q}(x, y_s))$$

for $x \in \mathbb{F}_q^n$ and $y = (y_1, \ldots, y_s) \in \mathbb{F}_q^s$.

The function $\mathsf{BlkCnvt}_{n,(m_1,\ldots,m_s),q}$ is indeed a block source converter, as we show below. The intuition is that conditioning on the values of the previous blocks, the random source X still has enough min-entropy, and hence we may apply the condenser to get the next block.

We need the following technical lemmas:

Lemma 4.1.1. *Let $P, Q \subset I$ be two distributions with $\Delta(P, Q) \leq \epsilon$. Let $\{X_i : i \in \mathrm{supp}(P)\}$ and $\{Y_i : i \in \mathrm{supp}(Q)\}$ be two collections of distributions over the same domain S such that $\Delta(X_i, Y_i) \leq \epsilon'$ for any $i \in \mathrm{supp}(P) \cap \mathrm{supp}(Q)$. Then $X \stackrel{def}{=} \sum_{i \in \mathrm{supp}(P)} \Pr[P = i] \cdot X_i$ is $(2\epsilon + \epsilon')$-close to $Y \stackrel{def}{=} \sum_{i \in \mathrm{supp}(Q)} \Pr[Q = i] \cdot Y_i$.*

Proof. Let T be an arbitrary subset of S and we will prove that $|\Pr[X \in T] - \Pr[Y \in T]| \leq 2\epsilon + \epsilon'$.

Note that we can add dummy distributions X_i for $i \in I \setminus \mathrm{supp}(P)$ and Y_j for $j \in I \setminus \mathrm{supp}(Q)$ such that $\Delta(X_i, Y_i) \leq \epsilon'$ for all $i \in I$, and it still holds that $X = \sum_{i \in I} \Pr[P = i] \cdot X_i$

and $Y = \sum_{i \in I} \Pr[Q = i] \cdot Y_i$. Then we have

$$\begin{aligned}
&|\Pr[X \in T] - \Pr[Y \in T]| \\
&= \left| \sum_{i \in I} \Pr[P = i] \Pr[X_i \in T] - \sum_{i \in I} \Pr[Q = i] \Pr[Y_i \in T] \right| \\
&\leq \sum_{i \in I} |\Pr[P = i] \Pr[X_i \in T] - \Pr[Q = i] \Pr[Y_i \in T]| \\
&\leq \sum_{i \in I} |(\Pr[P = i] - \Pr[Q = i]) \Pr[X_i \in T] + \Pr[Q = i](\Pr[X_i \in T] - \Pr[Y_i \in T])| \\
&\leq \left(\sum_{i \in I} |\Pr[P = i] - \Pr[Q = i]| \right) + \epsilon' \left(\sum_{i \in I} \Pr[Q = i] \right) \\
&\leq 2\epsilon + \epsilon'
\end{aligned}$$

□

Lemma 4.1.2. *Let $X = (X_1, \ldots, X_s) \subset \mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}$ be a distribution such that for any $i \in [s]$ and $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1})$, the conditional distribution $X_i|_{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}}$ is ϵ-close to a distribution $\tilde{X}_i(x_1, \ldots, x_{i-1})$ with q-ary min-entropy k_i. Then X is $2s\epsilon$-close to a $(k_1, \ldots, k_s)$ q-ary block source.*

Proof. Define $X' = (X'_1, \ldots, X'_s)$ as the unique distribution such that for any $i \in [s]$ and any $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X'_1, \ldots, X'_{i-1})$, the conditional distribution $X'_i|_{X'_1 = x_1, \ldots, X'_{i-1} = x_{i-1}}$ equals $\tilde{X}_i(x_1, \ldots, x_{i-1})$ if $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1})$[1] and otherwise equals $U_{n_i, q}$.

For any $i \in [s]$ and $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X'_1, \ldots, X'_{i-1})$, we known $X'_i|_{X'_1 = x_1, \ldots, X'_{i-1} = x_{i-1}}$ is either $\tilde{X}_i(x_1, \ldots, x_{i-1})$ or $U_{n_i, q}$. And in either case it has min-entropy k_i. So X' is a $(k_1, \ldots, k_s)$ q-ary block source.

We will then prove that for any $i \in [s]$ and any $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1}) \cap \mathrm{supp}(X'_1, \ldots, X'_{i-1})$, the conditional distribution $X|_{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}}$ is $2(s - i + 1)\epsilon$-close to $X'|_{X'_1 = x_1, \ldots, X'_{i-1} = x_{i-1}}$. Setting $i = 1$ proves the lemma.

Induct on i. For $i = s$ the claim holds by the definition of X'. For $i < s$, assume the claim holds for $i + 1$ and we will prove that it holds for i as well. Consider any $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1}) \cap \mathrm{supp}(X'_1, \ldots, X'_{i-1})$. Let $A = X_i|_{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}}$ and $B = X'_i|_{X'_1 = x_1, \ldots, X'_{i-1} = x_{i-1}}$. We have

$$X|_{X_1 = x_1, \ldots, X_{i-1} = x_{i-1}} = \sum_{x_i \in \mathrm{supp}(A)} \Pr[A = x_i] \cdot X|_{X_1 = x_1, \ldots, X_i = x_i}$$

[1] $(x_1, \ldots, x_{i-1}) \in \mathrm{supp}(X_1, \ldots, X_{i-1})$ always holds if $i = 1$.

and

$$X'|_{X'_1=x_1,\ldots,X'_{i-1}=x_{i-1}} = \sum_{x_i \in \mathrm{supp}(B)} \Pr[B = x_i] \cdot X'|_{X'_1=x_1,\ldots,X'_i=x_i}.$$

By the induction hypothesis, we have

$$\Delta\left(X|_{X_1=x_1,\ldots,X_i=x_i}, X'|_{X'_1=x_1,\ldots,X'_i=x_i}\right) \leq 2(s-i)\epsilon$$

for $x_i \in \mathrm{supp}(A) \cap \mathrm{supp}(B)$. Also note that B is identical to $\tilde{X}_i(x_1,\ldots,x_{i-1})$ and is ϵ-close to A by definition. The claim then follows from Lemma 4.1.1. □

Now we are ready to prove the following theorem.

Theorem 4.1.2. *For $\epsilon > 0$, integers $s, n, m_1, \ldots, m_s \geq 1$ and sufficiently large prime power $q = (n/\epsilon)^{O(1)}$, the function $\mathsf{BlkCnvt}_{n,(m_1,\ldots,m_s),q}$ is a $(k, (k_1,\ldots,k_s), 3s\epsilon, q)$ block source converter of degree n where $k = \sum_{i=1}^{s} m_i + \log_q(1/\epsilon)$ and each $k_i = 0.99m_i$.*

Proof. The degree of $\mathsf{BlkCnvt}_{n,(m_1,\ldots,m_s),q}$ is n since $\mathsf{RSCon}_{n,m,q}$ has degree n. Let X be a random source that has q-ary min-entropy k. Let $Y_1,\ldots,Y_s$ be independent seeds uniformly distributed over $\mathbb{F}_q$. Let $Z = (Z_1,\ldots,Z_s) = \mathsf{BlkCnvt}_{n,(m_1,\ldots,m_s),q}(X,(Y_1,\ldots,Y_s))$ where each $Z_i = \mathsf{RSCon}_{n,m_i,q}(X,Y_i)$ is distributed over $\mathbb{F}_q^{m_i}$. Define

$$B = \left\{ (z_1,\ldots,z_i) : \begin{array}{l} i \in [s],\ (z_1,\ldots,z_i) \in \mathrm{supp}(Z_1,\ldots,Z_i),\ X|_{Z_1=z_1,\ldots,Z_i=z_i} \text{ does not} \\ \text{have } q\text{-ary min-entropy } k - (m_1+\cdots+m_i) - \log_q(1/\epsilon) \end{array} \right\}.$$

Define a new distribution $Z' = (Z'_1,\ldots,Z'_s)$ as follows: Sample $z = (z_1,\ldots,z_s) \leftarrow Z$ and independently $u = (u_1,\ldots,u_s) \leftarrow U_{m_1+\ldots+m_s,q}$. If there exist $i \in [s]$ such that $(z_1,\ldots,z_{i-1}) \in B$, then pick the smallest such i and let $z' = (z_1,\ldots,z_{i-1},u_i,\ldots,u_s)$. Otherwise let $z' = z$. Let Z' be the distribution of z'.

For any $i \in [s]$ and $(z_1,\ldots,z_{i-1}) \in \mathrm{supp}(Z'_1,\ldots,Z'_i)$, if some prefix of $(z_1,\ldots,z_{i-1})$ is in B then $Z'_i|_{Z'_1=z_1,\ldots,Z'_{i-1}=z_{i-1}}$ is the uniform distribution $U_{m_i,q}$, otherwise $Z'_i|_{Z'_1=z_1,\ldots,Z'_{i-1}=z_{i-1}} = Z_i|_{Z_1=z_1,\ldots,Z_{i-1}=z_{i-1}}$. In the second case, $X|_{Z_1=z_1,\ldots,Z_{i-1}=z_{i-1}}$ has min-entropy $k - (m_1+\cdots+m_{i-1}) - \log_q(1/\epsilon) \geq m_i$ since $(z_1,\ldots,z_{i-1}) \notin B$. In this case, $Z'_i|_{Z'_1=z_1,\ldots,Z'_{i-1}=z_{i-1}}$ is ϵ-close to a distribution of min-entropy k_i by Theorem 4.1.1 and the fact

$$Z'_i|_{Z'_1=z_1,\ldots,Z'_{i-1}=z_{i-1}} = Z_i|_{Z_1=z_1,\ldots,Z_{i-1}=z_{i-1}} = \mathsf{RSCon}_{n,m_i,q}(X|_{Z_1=z_1,\ldots,Z_{i-1}=z_{i-1}}, Y_i).$$

In either cases $Z'_i|_{Z'_1=z_1,\ldots,Z'_{i-1}=z_{i-1}}$ is ϵ-close to a distribution of min-entropy k_i. By Lemma 4.1.2, Z' is $2s\epsilon$-close to a $(k_1,\ldots,k_s)$ q-ary block source.

It remains to prove that Z is $s\epsilon$-close to Z', which implies that it is $3s\epsilon$-close to a $(k_1,\ldots,k_s)$ q-ary block source. By Lemma 2.0.8, for any $i \in [s]$, we have $\Pr[(Z_1,\ldots,Z_{i-1}) \in$

$B] \leq \epsilon$. So the probability that $(Z_1, \ldots, Z_{i-1}) \in B$ for some $i \in [s]$ is bounded by $s\epsilon$. Note that the distribution Z' is obtained from Z by redistributing the weights of $(z_1, \ldots, z_s)$ satisfying $(z_1, \ldots, z_{i-1}) \in B$ for some i. We conclude that $\Delta(Z, Z') \leq s\epsilon$, as desired. □

4.1.2 Block source extraction

We will employ Lemma 2.0.9 and compose some "basic" extractors to get a block source extractor. These basic extractors are given by the basic line samplers $\mathsf{Line}_{m,q}$ (see Definition 2.0.4).

Lemma 4.1.3. *For $\epsilon > 0$, $m \geq 1$ and prime power q, $\mathsf{Line}_{m,q}$ is a (k, ϵ, q) extractor of degree 2 where $k = 2m - 1 + 3\log_q(1/\epsilon)$.*

Proof. Apply Lemma 2.0.6 and Theorem 3.1.1. □

Suppose $\mathbb{F}_Q$ is an extension field of $\mathbb{F}_q$ with $[\mathbb{F}_Q : \mathbb{F}_q] = d$, i.e., $Q = q^d$. By Lemma 2.0.1, $\mathsf{Line}_{m,Q} : \mathbb{F}_Q^{2m} \times \mathbb{F}_Q \to \mathbb{F}_Q^m$, as a degree-2 manifold over $\mathbb{F}_Q$, can also be viewed as a degree-2 manifold over $\mathbb{F}_q$: $\mathsf{Line}_{m,Q} : \mathbb{F}_q^{2md} \times \mathbb{F}_q^d \to \mathbb{F}_q^{md}$.

Now we are ready to state the main result of this section. We first compose the basic line samplers to get a block source extractor. It is then applied to a block source obtained from the block source converter.

Definition 4.1.4 (Outer Sampler)**.** For $\delta > 0$, $m = 2^s$ and prime power q, let $n = 4m + \lceil \log_q(2/\delta) \rceil$, $d = s + 1$, and $d_i = 2^{s-i}$ for $i \in [s]$. For $i \in [s]$, view $\mathsf{Line}_{2,q^{d_i}} : \mathbb{F}_{q^{d_i}}^4 \times \mathbb{F}_{q^{d_i}} \to \mathbb{F}_{q^{d_i}}^2$ as a manifold over $\mathbb{F}_q$: $\mathsf{Line}_{2,q^{d_i}} : \mathbb{F}_q^{4d_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{2d_i}$. Composing these line samplers $\mathsf{Line}_{2,q^{d_i}}$ for $i \in [s]$ gives the function $\mathsf{BlkExt}(\mathsf{Line}_{2,q^{d_1}}, \ldots, \mathsf{Line}_{2,q^{d_s}}) : \mathbb{F}_q^{4d_1 + \cdots + 4d_s} \times \mathbb{F}_q \to \mathbb{F}_q^m$. Finally, define $\mathsf{OuterSamp}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$:

$$\mathsf{OuterSamp}_{m,\delta,q}(x, (y, y')) \stackrel{\text{def}}{=} \mathsf{BlkExt}(\mathsf{Line}_{2,q^{d_1}}, \ldots, \mathsf{Line}_{2,q^{d_s}})\big(\mathsf{BlkCnvt}_{n,(4d_1,\ldots,4d_s),q}(x, y), y'\big)$$

for $x \in \mathbb{F}_q^n$, $y \in \mathbb{F}_q^s$ and $y' \in \mathbb{F}_q$.

See Figure 4.1 for an illustration of the above definition.

Theorem 4.1.3. *For any $\epsilon, \delta > 0$, integer $m \geq 1$, and sufficiently large prime power $q \geq (n/\epsilon)^{O(1)}$, $\mathsf{OuterSamp}_{m,\delta,q}$ is an (ϵ, δ) sampler of degree t where $d = O(\log m)$, $n = O\left(m + \log_q(1/\delta)\right)$ and $t = O\left(m^2 + m\log_q(1/\delta)\right)$.*

Proof. We first show that $\mathsf{OuterSamp}_{m,\delta,q}$ is a $(4m, \epsilon, q)$ extractor. Consider any random source X over $\mathbb{F}_q^n$ with q-ary min-entropy $4m$. Let s, d_i be as in Definition 4.1.4. Let $k_i = 4 \cdot 0.99 \cdot d_i$ for $i \in [s]$. Let $\epsilon_0 = \frac{\epsilon}{4s}$.

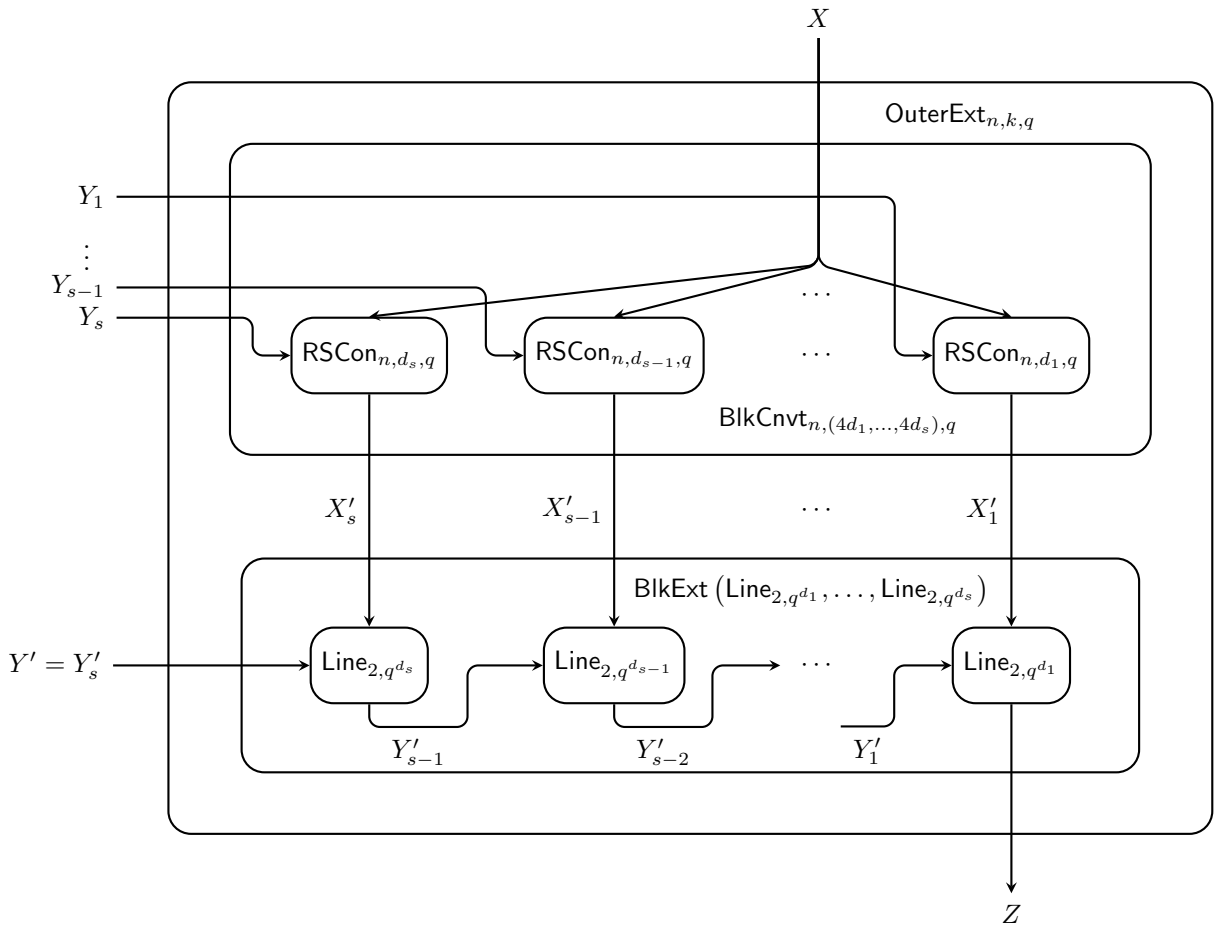

Figure 4.1: The extractor $\mathsf{OuterExt}_{n,k,q}$ that takes the random source X together with the seed $(Y_1, \ldots, Y_s, Y')$ and then outputs Z.

We have $(\sum_{i=1}^{s} 4d_i) + \log_q(1/\epsilon_0) \leq 4m$ for sufficiently large $q \geq (n/\epsilon)^{O(1)}$. So by Theorem 4.1.2, $\mathsf{BlkCnvt}_{n,(4d_1,\ldots,4d_s),q}$ is a $(4m, (k_1, \ldots, k_s), 3s\epsilon_0, q)$ block source converter. Therefore the distribution $\mathsf{BlkCnvt}_{n,(4d_1,\ldots,4d_s),q}(X, U_{s,q})$ is $3s\epsilon_0$-close to a $(k_1, \ldots, k_s)$ q-ary block source X'. Then $\mathsf{OuterSamp}_{m,\delta,q}(X, U_{d,q})$ is $3s\epsilon_0$-close to $\mathsf{BlkExt}(\mathsf{Line}_{2,q^{d_1}}, \ldots, \mathsf{Line}_{2,q^{d_s}})(X', U_{1,q})$.

By Lemma 4.1.3, $\mathsf{Line}_{2,q^{d_i}}$ is a $\left(k_i/d_i, \epsilon_0, q^{d_i}\right)$ extractor for $i \in [s]$ since $3 + 3\log_{q^{d_i}}(1/\epsilon_0) \leq 4 \cdot 0.99 = k_i/d_i$. Equivalently it is a (k_i, ϵ_0, q) extractor.

By Lemma 2.0.9, $\mathsf{BlkExt}(\mathsf{Line}_{2,q^{d_1}}, \ldots, \mathsf{Line}_{2,q^{d_s}})$ is a $((k_1, \ldots, k_s), s\epsilon_0, q)$ block source extractor. Therefore $\mathsf{BlkExt}(\mathsf{Line}_{2,q^{d_1}}, \ldots, \mathsf{Line}_{2,q^{d_s}})(X', U_{1,q})$ is $s\epsilon_0$-close to $U_{m,q}$, which by the previous paragraph, implies that $\mathsf{OuterSamp}_{m,\delta,q}(X, U_{d,q})$ is $4s\epsilon_0$-close to $U_{m,q}$. By definition, $\mathsf{OuterSamp}_{m,\delta,q}$ is a $(4m, \epsilon, q)$ extractor. By Theorem 3.1.1, it is also an (ϵ, δ) sampler.

We have $d = s+1 = O(\log m)$ and $n = O\left(m + \log_q(1/\delta)\right)$. By Lemma 2.0.1, each $\mathsf{Line}_{2,q^{d_i}}$ has degree 2 as a manifold over $\mathbb{F}_q$. Therefore by Lemma 2.0.9, $\mathsf{BlkExt}(\mathsf{Line}_{2,q^{d_1}}, \ldots, \mathsf{Line}_{2,q^{d_s}})$

has degree 2^s. By Theorem 4.1.2, $\mathsf{BlkCnvt}_{n,(4d_1,\ldots,4d_s),q}$ has degree n. Therefore $\mathsf{OuterSamp}_{m,\delta,q}$ has degree $n2^s = O\left(m^2 + m\log_q(1/\delta)\right)$. □

Remark 5. We assume m is a power of 2 above. For general m, simply pick $m' = 2^{\lceil \log m \rceil}$ and let $\mathsf{OuterSamp}_{m,\delta,q}$ be the composition of $\mathsf{OuterSamp}_{m',\delta,q}$ with the projection $\pi : \mathbb{F}_q^{m'} \to \mathbb{F}_q^m$ onto the first m coordinates. It yields an (ϵ, δ) sampler of degree t for $\mathbb{F}_q^m$ since π is linear, and approximating the density of a subset A in $\mathbb{F}_q^m$ is equivalent to approximating the density of $\pi^{-1}(A)$ in $\mathbb{F}_q^{m'}$.

Remark 6. The most important properties of the extractors $\mathsf{Line}_{2,q^{d_i}}$ used here are (1) they work for a certain constant min-entropy rate, and (2) the seed is shorter than the output by a constant factor. As the reader can check, besides the basic line samplers, we may also use the randomness-efficient line samplers given by [MR06], or the (strong) extractors from the universal family of hash functions $\{h_{a,b} : x \mapsto ax + b\}$ [CW79] (operations are performed in a finite field) together with the leftover hash lemma [ILL89], etc.

The sampler $\mathsf{OuterSamp}_{m,\delta,q}$ has optimal randomness complexity $O\left(m \log q + \log(1/\delta)\right)$, yet the sample complexity is sub-optimal, being $q^d = q^{O(\log m)}$ instead of q. We will fix this problem by composing it with an "inner sampler" that has the optimal sample complexity.

4.2 Inner sampler

We will construct a curve sampler of low degree in this section, or what we called the "inner sampler". It might be viewed as an extractor with optimal seed length, even though it only extracts a tiny fraction of min-entropy from the random source. The construction will be based on two techniques called error reduction and iterated sampling.

4.2.1 Error reduction

Condensers are at the core of many extractor constructions [RSW06, TSUZ07, GUV09, TSU12]. In the language of samplers, the use of condensers can be regarded as an error reduction technique, as we shall see below.

Given a function $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$, define $\mathrm{LIST}_f(T, \epsilon) \stackrel{\text{def}}{=} \left\{x \in \mathbb{F}_q^n : \Pr_y[f(x, y) \in T] > \epsilon\right\}$ for any $T \subseteq \mathbb{F}_q^m$ and $\epsilon > 0$. We are interesting in functions f exhibiting a "list-recoverability" property that the size of $\mathrm{LIST}_f(T, \epsilon)$ is kept small when T is not too large.

Definition 4.2.1. A function $f : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ is (ϵ, L, H) *list-recoverable* if $|\mathrm{LIST}_f(T, \epsilon)| \leq H$ for all $T \subseteq \mathbb{F}_q^m$ of size at most L.

Remark 7. The above definition is justified as an extension of *list-recoverable codes* [GI01]: A code $C \subseteq \mathbb{F}_q^n$ is called (ρ, L, H) *list recoverable* if for any sets $S_1, \ldots, S_n \subseteq \mathbb{F}_q$ of size at most L, there are at most H codewords $x = (x_1, \ldots, x_n) \in C$ such that $\Pr_{i \in [n]}[x_i \in S_i] > 1 - \rho$. Let $f : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ be an (ϵ, H, L) list-recoverable function. Assume $f = (f_1, \ldots, f_m)$ has the extra property that $f_1(x, y) = y$ for all $x \in \mathbb{F}_q^n$ and $y \in \mathbb{F}_q$ (in particular, the Reed-Solomon condenser in Definition 4.1.2 has this property). Define code $C_f \subseteq \left(\mathbb{F}_q^{m-1}\right)^q$ as follows:

$$C_f = \left\{(f_2(x, y_1), \ldots, f_m(x, y_q)) : x \in \mathbb{F}_q^n\right\}$$

where $y_1, \ldots, y_q$ are the q distinct elements of $\mathbb{F}_q$ (in any order). Then C_f is $(1-\epsilon, H, L/q)$ list-recoverable: For any $S_1, \ldots, S_q \subseteq \mathbb{F}_q^{m-1}$ of size at most L/q, let $T = \bigcup_{i=1}^{q} (\{y_i\} \times S_i) \subseteq \mathbb{F}_q^m$ be their union that has size at most L. By definition, every codeword $(f_2(x, y_1), \ldots, f_m(x, y_q))$ satisfying $\Pr_{i \in [n]}[x_i \in S_i] > \epsilon$ corresponds to an element $x \in |\mathrm{LIST}_f(T, \epsilon)|$. Therefore the number of such codewords is upper bounded by $|\mathrm{LIST}_f(T, \epsilon)| \leq H$.

The following lemma shows that the condenser property implies the list-recoverability property.

Lemma 4.2.1 ([GUV09])**.** *Suppose $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is an $(n, \log_q H) \to_{\epsilon, q} \left(m, \log_q \left(\frac{L}{\epsilon}\right)\right)$ condenser. Then it holds that $|\mathrm{LIST}_f(T, 2\epsilon)| \leq H$ for any $T \subseteq \mathbb{F}_q^m$ of size at most L, and hence f is $(2\epsilon, L, H)$ list-recoverable.*

We then define an operation $\star$ as follows.

Definition 4.2.2. For functions $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ and $S : \mathbb{F}_q^m \times \mathbb{F}_q^{d'} \to \mathbb{F}_q^{m'}$, define $S \star f : \mathbb{F}_q^n \times (\mathbb{F}_q^d \times \mathbb{F}_q^{d'}) \to \mathbb{F}_q^{m'}$ as follows:

$$S \star f(x, (y, y')) \stackrel{\text{def}}{=} S(f(x, y), y').$$

See Figure 4.2 for an illustration of the above definition.

The following lemma states that a sampler with mildly small confidence error, when composed with a list-recoverable function via the $\star$ operation, gives a sampler with very small confidence error.

Lemma 4.2.2. *Suppose $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is (ϵ_1, L, H) list-recoverable, and $S : \mathbb{F}_q^m \times \mathbb{F}_q^{d'} \to \mathbb{F}_q^{m'}$ is an (ϵ_2, δ) sampler where $\delta = \frac{L}{q^m}$. Then $S \star f$ is an $\left(\epsilon_1 + \epsilon_2, \frac{H}{q^n}\right)$ sampler.*

Proof. Let A be an arbitrary subset of $\mathbb{F}_q^{m'}$. Let $B = \{y : |\mu_{S(y,\cdot)} - \mu(A)| > \epsilon_2\}$. By the sampler property of S, we have $|B| \leq \delta q^m = L$ and hence $|\mathrm{LIST}_f(B, \epsilon_1)| \leq H$. Therefore it suffices to show that for any $x \in \mathbb{F}_q^n \backslash \mathrm{LIST}_f(B, \epsilon_1)$, it holds that $|\mu_{S \star f(x,\cdot)}(A) - \mu(A)| \leq \epsilon_1 + \epsilon_2$.

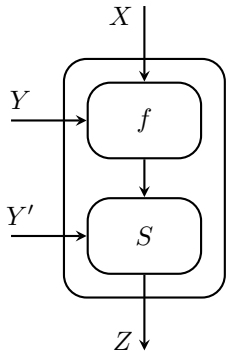

Figure 4.2: The function $S \star f$. The output is denoted by Z.

Fix $x \in \mathbb{F}_q^n \setminus \mathrm{LIST}_f(B, \epsilon_1)$. We have

$$\mu_{S\star f(x,\cdot)}(A) = \Pr_{y,y'}[S \star f(x,(y,y')) \in A] = \Pr_{y,y'}[S(f(x,y),y') \in A] = \mathbb{E}_y\left[\mu_{S(f(x,y),\cdot)}(A)\right].$$

Therefore

$$\begin{aligned}
|\mu_{S\star f(x,\cdot)}(A) - \mu(A)| &= \left|\mathbb{E}_y\left[\mu_{S(f(x,y),\cdot)}(A)\right] - \mu(A)\right| \\
&\leq \mathbb{E}_y|\mu_{S(f(x,y),\cdot)}(A) - \mu(A)| \\
&\leq \Pr_y[f(x,y) \in B] + \epsilon_2 \Pr_y[f(x,y) \notin B] \\
&\leq \epsilon_1 + \epsilon_2.
\end{aligned}$$

To see the last two steps, note that $|\mu_{S(y,\cdot)}(A) - \mu(A)| \leq \epsilon_2$ for $y \notin B$ by definition, and $\Pr_y[f(x,y) \in B] \leq \epsilon_1$ since $x \notin \mathrm{LIST}_f(B, \epsilon_1)$. □

Combining Lemma 4.2.1 and Lemma 4.2.2, we obtain:

Corollary 4.2.1. *Suppose $f : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is an $(n, k_1) \to_{\epsilon,q} (m, k_2)$ condenser, and $S : \mathbb{F}_q^m \times \mathbb{F}_q^{d'} \to \mathbb{F}_q^{m'}$ is an $\left(\epsilon', \epsilon q^{k_2 - m}\right)$ sampler. Then $S \star f$ is an $\left(2\epsilon + \epsilon', q^{k_1 - n}\right)$ sampler.*

Proof. Lemma 4.2.1 implies $|\mathrm{LIST}_f(T, 2\epsilon)| \leq q^{k_1}$ for T of size at most ϵq^{k_2}. Set $H = q^{k_1}$, $L = \epsilon q^{k_2}$, $\epsilon_1 = 2\epsilon$, $\epsilon_2 = \epsilon'$ and apply Lemma 4.2.2. □

Remark 8. It also possible to prove Corollary 4.2.1 using the connection between extractors and samplers (c.f. Theorem 3.1.1), except that some parameters are slightly different, e.g. the resulting accuracy error is $\epsilon + 2\epsilon'$ which has poorer dependence on ϵ', due to the averaging argument used in the proof of Theorem 3.1.1.

We state Corollary 4.2.1 because it offers a way of reducing the confidence error of samplers using condensers in a black-box manner. Nevertheless, for the best known condensers, the condenser properties are actually derived from the list-recoverability properties

[GUV09, TSU12], not the other way around. So we choose to use the list-recoverability properties directly, together with Lemma 4.2.2.

The condenser $\mathsf{RSCon}_{n,m,q}$ from Reed-Solomon codes (see Definition 4.1.2) enjoys the following list-recoverability property:

Theorem 4.2.1 ([GUV09])**.** *For any $h \geq 1$, $n \geq m \geq 1$, prime power q and $\epsilon > 0$, $\mathsf{RSCon}_{n,m,q}$ is (ϵ, L, H) list-recoverable where $H = (h-1)^{\frac{q^{m-1}-1}{q-1}}$ and $L = (\epsilon q - (n-1)(h-1)(m-1)) \cdot h^{m-1} - 1$. In pariticular, for sufficiently large $q \geq (n/\epsilon)^{O(1)}$, $\mathsf{RSCon}_{n,m,q}$ is $(\epsilon, q^{0.99m}, q^m)$ list-recoverable.*

Corollary 4.2.2. *For any $n \geq m \geq 1$, $\epsilon, \epsilon' > 0$ and sufficiently large prime power $q = (n/\epsilon)^{O(1)}$, suppose $S : \mathbb{F}_q^m \times \mathbb{F}_q^d \to \mathbb{F}_q^{m'}$ is an $(\epsilon', q^{-0.01m})$ sampler of degree t, then $S \star \mathsf{RSCon}_{n,m,q}$ is an $(\epsilon + \epsilon', q^{m-n})$ sampler of degree nt.*

Proof. Apply Lemma 4.2.2 and Theorem 4.2.1. Note that $\mathsf{RSCon}_{n,m,q}$ has degree n. Therefore $(S \star \mathsf{RSCon}_{n,m,q})(X, (Y, Y')) = S(\mathsf{RSCon}_{n,m,q}(X, Y), Y')$ has degree nt in its variables X, Y, Y'. □

4.2.2 Iterated sampling

We introduce the operation $\circ$ performed on samplers.

Definition 4.2.3. (composed sampler). Given functions $S_1 : \mathbb{F}_q^{n_1} \times \mathbb{F}_q^{d_1} \to \mathbb{F}_q^{d_0}$ and $S_2 : \mathbb{F}_q^{n_2} \times \mathbb{F}_q^{d_2} \to \mathbb{F}_q^{d_1}$, define $S_1 \circ S_2 : (\mathbb{F}_q^{n_1} \times \mathbb{F}_q^{n_2}) \times \mathbb{F}_q^{d_2} \to \mathbb{F}_q^{d_0}$ such that

$$S_1 \circ S_2((x_1, x_2), y) \stackrel{\text{def}}{=} S_1(x_1, S_2(x_2, y))$$

for all $x_1 \in \mathbb{F}_q^{n_1}$, $x_2 \in \mathbb{F}_q^{n_2}$ and $y \in \mathbb{F}_q^{d_2}$.

See Figure 4.3 for an illustration of the above definition.

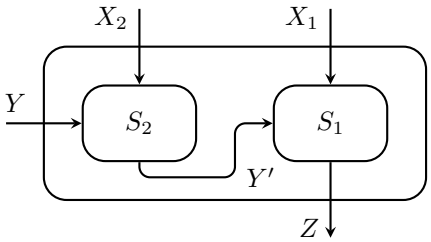

Figure 4.3: The function $S_1 \circ S_2$. The output is denoted by Z.

Think of S_1 and S_2 as samplers. Then $S_1 \circ S_2$ is the composed sampler that first uses its randomness x_1 to get the sample $S_1(x_1, \cdot)$, and then uses its randomness x_2 to get the

subsample $S_2(x_1, S_2(x_2, \cdot)) \subseteq S_1(x_1, \cdot)$. Intuitively, if S_1 and S_2 are good samplers then so is $S_1 \circ S_2$. This is indeed shown by [BR94, TSU06] and we formalize it as follows:

Lemma 4.2.3. *Let $S_1 : \mathbb{F}_q^{n_1} \times \mathbb{F}_q^{d_1} \to \mathbb{F}_q^{d_0}$ be an (ϵ_1, δ_1) manifold sampler of degree t_1. And let $S_2 : \mathbb{F}_q^{n_2} \times \mathbb{F}_q^{d_2} \to \mathbb{F}_q^{d_1}$ be an (ϵ_2, δ_2) manifold sampler of degree t_2. Then $S_1 \circ S_2 : (\mathbb{F}_q^{n_1} \times \mathbb{F}_q^{n_2}) \times \mathbb{F}_q^{d_2} \to \mathbb{F}_q^{d_0}$ is an $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$ manifold sampler of degree $t_1 t_2$.*

Proof. Consider an arbitrary subset $A \subseteq \mathbb{F}_q^{d_0}$. Define $B(x) = \left\{z \in \mathbb{F}_q^{d_2} : S_1(x, z) \in A\right\}$ for each $x \in \mathbb{F}_q^{n_1}$. Pick $x_1 \leftarrow U_{n_1,q}$ and $x_2 \leftarrow U_{n_2,q}$. If $\left|\mu_{S_1 \circ S_2((x_1,x_2),\cdot)}(A) - \mu(A)\right| > \epsilon_1 + \epsilon_2$ occurs, then either $|\mu_{S_1(x_1,\cdot)}(A) - \mu(A)| > \epsilon_1$, or $|\mu_{S_1 \circ S_2((x_1,x_2),\cdot)}(A) - \mu_{S_1(x_1,\cdot)}(A)| > \epsilon_2$ occurs. Call the two events E_1 and E_2 respectively.

Note that E_1 occurs with probability at most δ_1 by the sampler property of S_1. Also note that

$$\mu_{S_1 \circ S_2((x_1,x_2),\cdot)}(A) = \Pr_y[S_1(x_1, S_2(x_2, y)) \in A] = \Pr_y[S_2(x_2, y) \in B(x_1)] = \mu_{S_2(x_2,\cdot)}(B(x_1))$$

whereas

$$\mu_{S_1(x_1,\cdot)}(A) = \Pr_y[S_1(x_1, y) \in A] = \Pr_y[y \in B(x_1)] = \mu(B(x_1)).$$

So the probability that E_2 occurs is $\Pr_{x_1,x_2}[|\mu_{S_2(x_2,\cdot)}(B(x_1)) - \mu(B(x_1))| > \epsilon_2]$ which is bounded by δ_2 by the sampler property of S_2. By the union bound, the event

$$\left|\mu_{S_1 \circ S_2((x_1,x_2),\cdot)}(A) - \mu(A)\right| > \epsilon_1 + \epsilon_2$$

occurs with probability at most $\delta_1 + \delta_2$, as desired.

Finally, we have $S_1 \circ S_2((X_1, X_2), Y)) = S_1(X_1, S_2(X_2, Y))$ which is a manifold of degree $t_1 t_2$ in its variables X_1, X_2, Y since S_1 and S_2 are manifolds of degree t_1 and t_2 respectively. □

A simple induction implies the following generalization of Lemma 4.2.3:

Corollary 4.2.3. *Let $S_i : \mathbb{F}_q^{n_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{d_{i-1}}$ be an (ϵ_i, δ_i) sampler that is a manifold of degree t_i for $i = 1, \ldots, s$. Then $S_1 \circ \cdots \circ S_s : \left(\mathbb{F}_q^{n_1} \times \cdots \times \mathbb{F}_q^{n_s}\right) \times \mathbb{F}_q^{d_s} \to \mathbb{F}_q^{d_0}$[2] is an $(\sum_{i=1}^s \epsilon_i, \sum_{i=1}^s \delta_i)$ sampler that is a manifold of degree $\prod_{i=1}^s t_i$.*

Remark 9. The readers may notice that the composed sampler $S_1 \circ S_2$ has the same form as the composed block source extractor $\mathsf{BlkExt}(S_1, S_2)$ (see Definition 2.0.11), and more generally $S_1 \circ \cdots \circ S_s$ has the same form as $\mathsf{BlkExt}(S_1, \ldots, S_s)$ (with the outputs $Z_s, \ldots, Z_2$ being empty strings, c.f. Figure 2.1). This is not a coincidence. Using the connection between

[2]It is easy to check that $\circ$ is associative, and hence we can write $S_1 \circ \cdots \circ S_s$ with no ambiguity.

extractors and samplers, we see the composed sampler $S_1 \circ \cdots \circ S_s$ is indeed an extractor for random sources with q-ary min-entropy $n_1+\cdots+n_s-\Delta$ where $\Delta \approx \log_q(1/\delta_1 + \cdots + \delta_s)$, and each S_i is an extractor for random sources with q-ary entropy $n_i - \Delta_i$ where $\Delta_i \approx \log_q(1/\delta_i)$. Assume each $\Delta_i \approx \Delta$ for simplicity. It is shown in [GUV09, Lemma 5.8 and Corollary 5.9] that a random source distributed over $\mathbb{F}_q^{n_1+\cdots+n_s}$ with q-ary min-entropy $n_1 + \cdots + n_s - \Delta$ is automatically a $(k_1, \ldots, k_s)$ q-ary block source where $k_i \approx n_i - \Delta$. Then each S_i serves as an extractor for q-ary min-entropy k_i and hence the block source extraction may proceed. Therefore, Lemma 4.2.3 and Corollary 4.2.3 offer an alternative[3], and arguably cleaner view of the extraction of very dense random sources via block source extraction.

4.2.3 Recursive inner sampler

By Lemma 2.0.7, for $\epsilon > 0$, $m \geq 1$, $t \geq 4$ and sufficiently large prime power $q = (t/\epsilon)^{O(1)}$, the basic curve sampler $\mathsf{Curve}_{m,t,q}$ is an $\left(\epsilon, q^{-t/4}\right)$ sampler. Let $\delta = q^{-t/4}$ is the confidence error of $\mathsf{Curve}_{m,t,q}$. And suppose $m = O(1)$. Then the randomness complexity of $\mathsf{Curve}_{m,t,q}$ is $tm \log q = O(\log \delta)$ which is optimal up to an $O(1)$ factor. So the basic curve samplers sampling $O(1)$-dimensional vector space are randomness-optimal, and we will use them as the building blocks of the inner sampler.

We will recursively construct an inner curve sampler with the optimal sample complexity. The natural idea is applying the technique of iterated sampling to reduce the sample complexity. More specifically, we use the basic curve samplers to reduce the sample complexity polynomially each time. However, sub-sampling increases the randomness complexity while the confidence error does not shrink accordingly. To fix this problem, we also apply the technique of error reduction. Note that error reduction is applicable only when the original confidence error is already exponentially small in the number of random bits invested (cf. Corollary 4.2.2). So we would like to maintain this invariant in the recursive construction. In order to do so, we apply error reduction at each level so that the confidence error shrinks polynomially (except the last step where the confidence error is brought down directly to δ). In summary, we use the basic curve samplers as the building blocks and apply the error reduction as well as iterated sampling techniques repeatedly to obtain the desired inner sampler. The formal construction is as follows:

Definition 4.2.4 (inner sampler)**.** For $m \geq 1$, $\delta > 0$ and prime power q, pick $s = \lceil \log m \rceil$

[3]It is certainly not an exact equivalence since the parameters of extractors and samplers are slightly worsen when they are translated to each other, c.f. Theorem 3.1.1.

and let $d_i = 2^{s-i}$ for $0 \le i \le s$. Also let

$$n_i = \begin{cases} 16^i & 0 \le i \le s-1, \\ \max\left\{16^s, 20\left\lceil \log_q(1/\delta)\right\rceil\right\} & i = s. \end{cases}$$

Define $S_i : \mathbb{F}_q^{n_i d_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^m$ for $i \in [s]$ as follows:

- $S_0 : \mathbb{F}_q \times \mathbb{F}_q^{d_0} \to \mathbb{F}_q^m$ projects (x, y) onto the first m coordinates of y.
- $S_i \stackrel{\text{def}}{=} \left(S_{i-1} \star \mathsf{RSCon}_{n_i/4, 2n_{i-1}, q^{d_i}}\right) \circ \mathsf{Curve}_{3, n_i/4, q^{d_i}}$ for $i = 1, \ldots, s$.

Finally, let $\mathsf{InnerSamp}_{m,\delta,q} \stackrel{\text{def}}{=} S_s$.

Figure 4.4 shows how S_i is obtained from S_{i-1}.

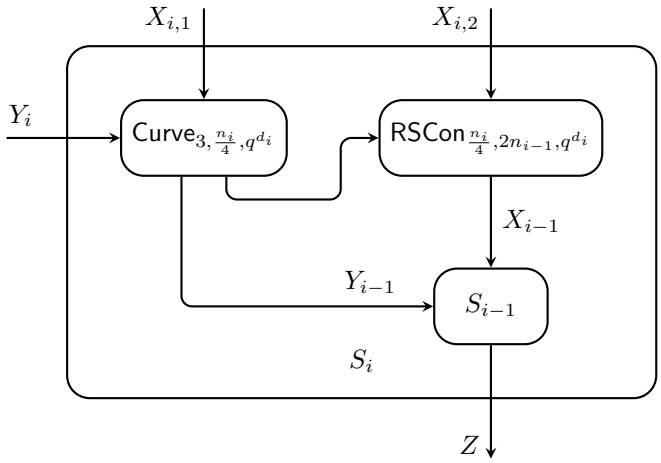

Figure 4.4: The recursive construction of S_i. Here $X_i = (X_{i,1}, X_{i,2})$ (resp. X_{i-1}) and Y_i (resp. Y_{i-1}) are the two arguments of S_i (resp. S_{i-1}). And Z is the common output of S_i and S_{i-1}.

Remark 10. We can check that all S_i's are well-defined. For S_0 this is obvious. For $i > 1$, note that $n_i/4$ is an integer. The function $\mathsf{RSCon}_{n_i/4, 2n_{i-1}, q^{d_i}} : \mathbb{F}_{q^{d_i}}^{n_i/4} \times \mathbb{F}_{q^{d_i}} \to \mathbb{F}_{q^{d_i}}^{2n_{i-1}}$ may be viewed over $\mathbb{F}_q$:

$$\mathsf{RSCon}_{n_i/4, 2n_{i-1}, q^{d_i}} : \mathbb{F}_q^{n_i d_i/4} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{2n_{i-1}d_i}.$$

Given $S_{i-1} : \mathbb{F}_q^{2n_{i-1}d_i} \times \mathbb{F}_q^{2d_i} \to \mathbb{F}_q^m$ (note $d_{i-1} = 2d_i$), we have the function

$$S_{i-1} \star \mathsf{RSCon}_{n_i/4, 2n_{i-1}, q^{d_i}} : \mathbb{F}_q^{n_i d_i/4} \times \mathbb{F}_q^{3d_i} \to \mathbb{F}_q^m.$$

The function $\mathsf{Curve}_{3,n_i/4,q^{d_i}} : \mathbb{F}_{q^{d_i}}^{3n_i/4} \times \mathbb{F}_{q^{d_i}} \to \mathbb{F}_{q^{d_i}}^3$ may also be viewed over $\mathbb{F}_q$:

$$\mathsf{Curve}_{3,n_i/4,q^{d_i}} : \mathbb{F}_q^{3n_i d_i/4} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{3d_i}.$$

Note that $S_i = \left(S_{i-1} \star \mathsf{RSCon}_{n_i/4,2n_{i-1},q^{d_i}}\right) \circ \mathsf{Curve}_{3,n_i/4,q^{d_i}}$. So we have

$$S_i : \mathbb{F}_q^{n_i d_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^m$$

as claimed.

We then have the following theorem:

Theorem 4.2.2. *For any $\epsilon, \delta > 0$, integer $m \geq 1$ and sufficiently large prime power $q \geq \left(\frac{m \log(1/\delta)}{\epsilon}\right)^{O(1)}$, let $\epsilon' = \frac{\epsilon}{2s}$ and d_i, n_i, S_i be as in Definition 4.2.4. Then for each $0 \leq i \leq s$, the function $S_i : \mathbb{F}_q^{n_i d_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^m$ is an (ϵ_i, δ_i) manifold sampler of degree t_i where $\epsilon_i = 2i\epsilon'$, $\delta_i = q^{-n_i d_i/20}$, and $t_i = \prod_{j=1}^{i} (n_j/4)^2$.*

In particular, the function $\mathsf{InnerSamp}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ is an (ϵ, δ) curve sampler of degree t where $n \leq m^{O(1)} + 20 \log_q (1/\delta)$ and $t = O\left(m^{O(\log m)} \log_q^2 (1/\delta)\right)$.

Proof. Induct on i. The claim is trivially true when $i = 0$. Now consider the case $i > 0$ and assume the claim holds for all $i' < i$.

By the induction hypothesis, S_{i-1} is an $(\epsilon_{i-1}, \delta_{i-1})$ manifold sampler of degree t_{i-1}. Then by Corollary 4.2.2, $S_{i-1} \star \mathsf{RSCon}_{n_i/4,2n_{i-1},q^{d_i}}$ is an $\left(\epsilon_{i-1} + \epsilon', q^{(2n_{i-1} - n_i/4)d_i}\right)$ manifold sampler of degree $(n_i/4) \cdot t_{i-1}$.

By Lemma 2.0.7, $\mathsf{Curve}_{3,n_i/4,q^{d_i}}$ is an $\left(\epsilon', q^{-n_i d_i/16}\right)$ curve sampler of degree $n_i/4$. Then by Lemma 4.2.3, the function

$$S_i = \left(S_{i-1} \star \mathsf{RSCon}_{n_i/4,2n_{i-1},q^{d_i}}\right) \circ \mathsf{Curve}_{3,n_i/4,q^{d_i}}$$

is an $\left(\epsilon_{i-1} + 2\epsilon', q^{(2n_{i-1} - n_i/4)d_i} + q^{-n_i d_i/16}\right)$ manifold sampler of degree $(n_i/4)^2 \cdot t_{i-1}$. It is then just a routine to check the following facts:

$$\begin{aligned} \epsilon_{i-1} + 2\epsilon' = 2i\epsilon' &= \epsilon_i, \\ q^{(2n_{i-1} - n_i/4)d_i} + q^{-n_i d_i/16} &\leq q^{-n_i d_i/20} = \delta_i, \\ (n_i/4)^2 \cdot t_{i-1} &= \prod_{j=1}^{i} (n_j/4)^2 = t_i. \end{aligned}$$

Finally, note that $\mathsf{InnerSamp}_{m,\delta,q} = S_s$, $\epsilon = \epsilon_s$ and $\delta \geq \delta_s$. So $\mathsf{InnerSamp}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ is an (ϵ, δ) curve sampler of degree t where $n = n_s d_s \leq m^{O(1)} + 20 \log_q (1/\delta)$ and $t = t_s = O\left(m^{O(\log m)} \log_q^2 (1/\delta)\right)$. □

4.3 Putting it together

The final curve sampler is simply the composition of the outer sampler and the inner sampler.

Definition 4.3.1. For $m \geq 1$, $\delta > 0$ and prime power q, we have the outer sampler

$$\mathsf{OuterSamp}_{m,\delta/2,q} : \mathbb{F}_q^{n_1} \times \mathbb{F}_q^d \to \mathbb{F}_q^m$$

and the inner sampler

$$\mathsf{InnerSamp}_{d,\delta/2,q} : \mathbb{F}_q^{n_2} \times \mathbb{F}_q \to \mathbb{F}_q^d.$$

See Definition 4.1.4 and Definition 4.2.4 for their constructions. Then define

$$\mathsf{Samp}_{m,\delta,q} \stackrel{\text{def}}{=} \mathsf{OuterSamp}_{m,\delta/2,q} \circ \mathsf{InnerSamp}_{d,\delta/2,q}.$$

Theorem 4.3.1 (Theorem 1.0.1 restated)**.** *For any $\epsilon, \delta > 0$, integer $m \geq 1$ and sufficiently large prime power $q \geq \left(\frac{m \log(1/\delta)}{\epsilon}\right)^{O(1)}$, the function $\mathsf{Samp}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ is an (ϵ, δ) curve sampler of degree t where $n \leq O(m) + 21 \log_q (1/\delta)$ and $t = \left(m \log_q (1/\delta)\right)^{O(1)}$.*

Proof. Let n_1, n_2 and d be as in Definition 4.3.1. By Theorem 4.1.3, $\mathsf{OuterSamp}_{m,\delta/2,q} : \mathbb{F}_q^{n_1} \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is an $(\epsilon/2, \delta/2)$ manifold sampler of degree t_1 where $d = O(\log m)$, $n_1 = 8m + \left\lceil \log_q \left(\frac{4}{\delta}\right) \right\rceil$ and $t_1 = O\left(m^2 + m \log_q (1/\delta)\right)$.

By Theorem 4.2.2, $\mathsf{InnerSamp}_{d,\delta/2,q} : \mathbb{F}_q^{n_2} \times \mathbb{F}_q \to \mathbb{F}_q^d$ is an $(\epsilon/2, \delta/2)$ curve sampler of degree t_2 where $n_2 \leq (\log m)^{O(1)} + 20 \log_q (1/\delta)$ and $t_2 = O\left((\log m)^{O(\log \log m)} \log_q^2 (1/\delta)\right)$.

Finally, by Lemma 4.2.3, $\mathsf{Samp}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$ is an (ϵ, δ) curve sampler of degree t where $n = n_1 + n_2 \leq O(m) + 21 \log_q (1/\delta)$ and $t = t_1 t_2 = \left(m \log_q (1/\delta)\right)^{O(1)}$. □

We also obtain an explicit construction of extractors that have low degree and optimal parameters (up to constant factors).

Definition 4.3.2. For $k \leq n$ and prime power q, pick $\delta = q^{-k/50}$ and sufficiently small $m = \Theta(k)$[4] such that we have the function

$$\mathsf{Samp}_{m,\delta,q} : \mathbb{F}_q^{n'} \times \mathbb{F}_q \to \mathbb{F}_q^m$$

with $n' \leq O(m) + 21 \log_q (1/\delta) \leq k - 1$ (c.f. Theorem 4.3.1). Then define $\mathsf{Ext}_{n,k,q} : \mathbb{F}_q^n \times \mathbb{F}_q^2 \to \mathbb{F}_q^m$ as follows:

$$\mathsf{Ext}_{n,k,q} \stackrel{\text{def}}{=} \mathsf{RSCon}_{n,n',q} \star \mathsf{Samp}_{m,\delta,q}.$$

[4] Actually we need $k \geq c$ for some positive constant c here, but otherwise just let the extractor output the seed.

Theorem 4.3.2. *For $k \le n$, $\epsilon > 0$ and sufficiently large prime power $q \ge (n/\epsilon)^{O(1)}$, $\mathsf{Ext}_{n,k,q} : \mathbb{F}_q^n \times \mathbb{F}_q^2 \to \mathbb{F}_q^m$ is a (k, ϵ, q) extractor of degree t where $m = \Theta(k)$ and $t = nk^{O(1)}$.*

Proof. Note that $\delta = q^{-k/50} \le q^{-\alpha(k-1)} \le q^{-\alpha n'}$. By Theorem 4.3.1, $\mathsf{Samp}_{m,\delta,q}$ is an $\left(\epsilon/4, q^{-\alpha n'}\right)$ curve sampler of degree $\left(n' \log_q (1/\delta)\right)^{O(1)} = k^{O(1)}$.

Note that $q^{n'-n} \le q^{k-n-1} \le \epsilon q^{k-n}$. By Corollary 4.2.2, $\mathsf{Ext}_{n,k,q} = \mathsf{RSCon}_{n,n',q} \star \mathsf{Samp}_{m,\delta,q}$ is an $\left(\epsilon/2, q^{n'-n}\right)$ manifold sampler, and hence an $\left(\epsilon/2, \epsilon q^{k-n}\right)$ manifold sampler. And it has degree $nk^{O(1)}$.

Finally, by Theorem 3.1.1, $\mathsf{Ext}_{n,k,q}$ is also a (k, ϵ, q) extractor of degree $nk^{O(1)}$. □

4.4 An alternative outer sampler

We present an alternative outer sampler in this section based on the techniques of error reduction and iterated sampling. It matches the outer sampler that uses the block source extraction in all the parameters except the degree. We do not have a good bound on its degree, though.

We make the observation that the list-recoverability property of the condenser RSCon in Theorem 4.2.1 holds for large subsets if the underlying field is large.

Lemma 4.4.1. *For any $n \ge m \ge 1$, $\epsilon, \alpha > 0$, integer $r \ge 1$ and sufficiently large prime power $q \ge (n/\epsilon)^{O(1/\alpha)}$, let $Q = q^r$ and then the function $\mathsf{RSCon}_{n,m,Q} : \mathbb{F}_Q^n \times \mathbb{F}_Q \to \mathbb{F}_Q^m$ is $\left(\epsilon, Q^{(1-\alpha/r)m}, Q^m\right)$ list-recoverable.*

Proof. Choose $h = q^{r-\alpha}$ and let $H = (h-1)\frac{Q^{m-1}-1}{Q-1}$, $L = (\epsilon Q - (n-1)(h-1)(m-1)) \cdot h^{m-1} - 1$. Note that $\epsilon Q - (n-1)(h-1)(m-1) \ge q^{r-\alpha} + 1 = h + 1$ for sufficiently large $q = (n/\epsilon)^{O(1/\alpha)}$, and hence

$$L \ge (h+1)h^{m-1} - 1 \ge h^m = q^{(r-\alpha)m} = Q^{(1-\alpha/r)m}.$$

Also note that $H \le Q^m$. The lemma then follows from Theorem 4.2.1. □

From now on we fix $\alpha > 0$ as a sufficiently small constant and suppress $1/\alpha$ in the $O(\cdot)$ notation. Note that now we can handle sets of size $Q^{(1-\alpha/r)m}$ (which is $Q^{(1-o(1))m}$ for $r = \omega(1)$) in the domain of size Q^m. It should be compared with Theorem 4.2.1 where we can only handle sets of size $q^{0.99m}$ in the domain of size q^m. This stronger list-recoverability property can then be used for error-reduction which is applicable to samplers with relatively large confidence error.

Lemma 4.4.2. *For integers $r \ge 1$, $n \ge m \ge 1$ multiples of r, $\epsilon, \epsilon' > 0$ and sufficiently large prime power $q \ge (n/\epsilon)^{O(1)}$, let $Q = q^r$ and suppose $S : \mathbb{F}_q^m \times \mathbb{F}_q^d \to \mathbb{F}_q^{m'}$ is an $(\epsilon', q^{-\alpha m/r})$ sampler. Then $S \star \mathsf{RSCon}_{n/r,m/r,Q} : \mathbb{F}_q^n \times \mathbb{F}_q^{r+d} \to \mathbb{F}_q^{m'}$ is an $(\epsilon + \epsilon', q^{m-n})$ sampler.*

Remark 11. Note that $\mathsf{RSCon}_{n/r,m/r,Q} : \mathbb{F}_Q^{n/r} \times \mathbb{F}_Q \to \mathbb{F}_Q^{m/r}$ may be viewed over $\mathbb{F}_q$:

$$\mathsf{RSCon}_{n/r,m/r,Q} : \mathbb{F}_q^n \times \mathbb{F}_q \to \mathbb{F}_q^m$$

and hence $S \star \mathsf{RSCon}_{n/r,m/r,Q}$ is well-defined.

Proof. Apply Lemma 4.2.2 and Lemma 4.4.1. □

Now we are ready to present the construction of the alternative outer sampler. The idea is using iterated sampling for $\log m$ levels with the basic curve samplers as the building blocks. At level i we deal with a domain of size $q^{m/2^{i-1}}$ and use a basic curve sampler to halve its dimension. Its randomness complexity would be $O\left(\log\left(q^{m/2^{i-1}}\right) + \log(1/\delta)\right) = O\left(\frac{m}{2^{i-1}} \log q + \log(1/\delta)\right)$. And the total randomness complexity would be

$$O\left(\sum_{i=1}^{\log m} \left(\frac{m}{2^{i-1}} \log q + \log(1/\delta)\right)\right) = O\left(m \log q + \log m \log(1/\delta)\right)$$

which is sub-optimal since we expect $O(m \log q + \log(1/\delta))$ here. To fix this problem, we just set the confidence error to be a relatively large value δ' such that $\log(1/\delta') \approx \log(1/\delta) / \log m$, and then apply error reduction to bring it down to δ. The formal construction is as follows.

Definition 4.4.1. For $m = 2^s \geq 1$[5] and prime power q, pick the following parameters:

- $m' = 2s \cdot \left\lceil \sum_{i=1}^{s} \max\left\{ \left\lceil \frac{2^i}{s} \right\rceil, 4 \right\} 2^{s-i}/s \right\rceil$,
- $n = s \cdot \left\lceil \left(m' + \log_q(1/\delta)\right)/s \right\rceil$,
- $d_i = 2^{s-i}$ for $i \in [s]$,
- $t_i = \max\left\{ \left\lceil \frac{2^i}{s} \right\rceil, 4 \right\}$ for $i \in [s-1]$,
- $t_s = m'/2 - \sum_{i=1}^{s-1} t_i d_i \geq \max\left\{ \left\lceil \frac{2^s}{s} \right\rceil, 4 \right\}$.

Then define $\mathsf{OuterSamp2}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$:

$$\mathsf{OuterSamp2}_{m,\delta,q} \stackrel{\text{def}}{=} \left(\mathsf{Curve}_{2,t_1,q^{d_1}} \circ \cdots \circ \mathsf{Curve}_{2,t_s,q^{d_s}}\right) \star \mathsf{RSCon}_{n/s,m'/s,q^s}.$$

Figure 4.5 illustrates the structure of $\mathsf{OuterSamp2}_{m,\delta,q}$.

[5]We assume m is a power of 2 for simplicity. Otherwise pick a projection $\pi : \mathbb{F}_q^{m'} \to \mathbb{F}_q^m$ where m' is a power of 2. Construct a sampler with domain $\mathbb{F}_q^{m'}$ and compose it with π.

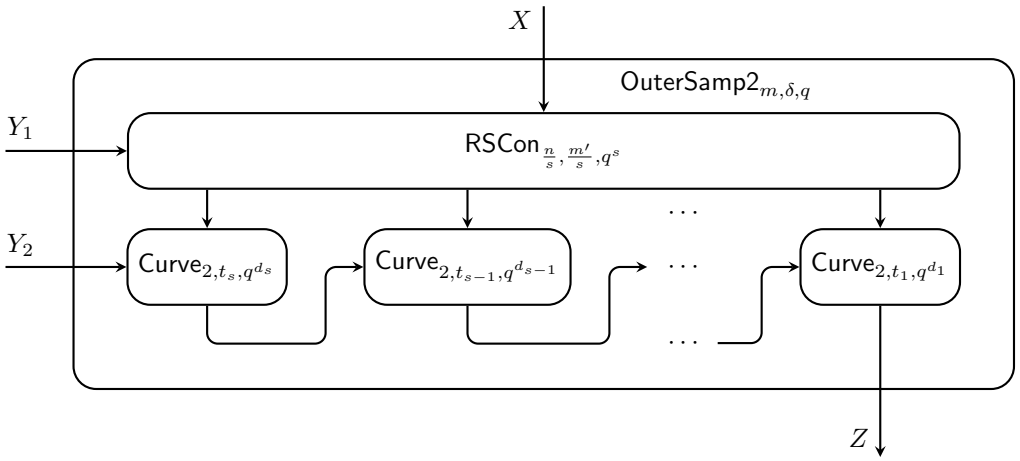

Figure 4.5: The alternative outer sampler $\mathsf{OuterSamp2}_{m,\delta,q}$ with X, $Y = (Y_1, Y_2)$ and Z being its two arguments and output respectively.

Remark 12. To see that $\mathsf{OuterSamp2}_{m,\delta,q}$ is well-defined, note that we may view each curve sampler $\mathsf{Curve}_{2,t_i,q^{d_i}}$ over $\mathbb{F}_q$:

$$\mathsf{Curve}_{2,t_i,q^{d_i}} : \mathbb{F}_q^{2t_i d_i} \times \mathbb{F}_q^{d_i} \to \mathbb{F}_q^{2d_i}.$$

Note that $2d_i = d_{i-1}$. So we have the composition of these curve samplers

$$\mathsf{Curve}_{2,t_1,q^{d_1}} \circ \cdots \circ \mathsf{Curve}_{2,t_s,q^{d_s}} : \mathbb{F}_q^{m'} \times \mathbb{F}_q \to \mathbb{F}_q^m$$

where we use the facts that $m' = \sum_{i=1}^{s} 2t_i d_i$, $d_s = 1$ and $2d_1 = m$. Note that both n and m' are multiples of s. And $\mathsf{RSCon}_{n/s,m'/s,q^s}$ may be viewed over $\mathbb{F}_q$:

$$\mathsf{RSCon}_{n/s,m'/s,q^s} : \mathbb{F}_q^n \times \mathbb{F}_q^s \to \mathbb{F}_q^{m'}$$

So we have the function $\mathsf{OuterSamp2}_{m,\delta,q} = \left(\mathsf{Curve}_{2,t_1,q^{d_1}} \circ \cdots \circ \mathsf{Curve}_{2,t_s,q^{d_s}}\right) \star \mathsf{RSCon}_{n/s,m'/s,q^s}$

$$\mathsf{OuterSamp2}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$$

where $d = s + 1$, as claimed.

Theorem 4.4.1. *For any $\epsilon, \delta > 0$, integer $m \geq 1$ and sufficiently large prime power $q \geq \left(\frac{m \log(1/\delta)}{\epsilon}\right)^{O(1)}$, the function $\mathsf{OuterSamp2}_{m,\delta,q} : \mathbb{F}_q^n \times \mathbb{F}_q^d \to \mathbb{F}_q^m$ is an (ϵ, δ) sampler where $d = O(\log m)$ and $n = O\left(m + \log_q(1/\delta)\right)$.*

Proof. Let s, m', d_i, t_i be as in Definition 4.4.1. Let $\epsilon_0 = \frac{\epsilon}{s+1}$.

By Lemma 2.0.7, for each $i \in [s]$, $\mathsf{Curve}_{2,t_i,q^{d_i}}$ is an $\left(\epsilon_0, q^{-t_i d_i/4}\right)$ sampler and hence an $\left(\epsilon_0, q^{-2^{s-2}/s}\right)$ sampler.

By Lemma 4.2.3, $\mathsf{Curve}_{2,t_1,q^{d_1}} \circ \cdots \circ \mathsf{Curve}_{2,t_s,q^{d_s}} : \mathbb{F}_q^{m'} \times \mathbb{F}_q \to \mathbb{F}_q^m$ is a $\left(s\epsilon_0, sq^{-2^{s-2}/s}\right)$ sampler.

Note that

$$\begin{aligned}
m' &= 2s \cdot \left\lceil \sum_{i=1}^{s} \max\left\{ \left\lceil \frac{2^i}{s} \right\rceil, 4 \right\} 2^{s-i}/s \right\rceil \\
&\leq 2s + 2\sum_{i=1}^{s} \max\left\{ \left\lceil \frac{2^i}{s} \right\rceil, 4 \right\} 2^{s-i} \\
&\leq 2s + 2\sum_{i=1}^{s} \left\lceil \frac{2^i}{s} \right\rceil 2^{s-i} + 8\sum_{i=1}^{s} \cdot 2^{s-i} \\
&\leq 2s + 2\sum_{i=1}^{s} \frac{2^s}{s} + 10 \cdot \sum_{i=1}^{s} 2^{s-i} \\
&\leq 2s + 12 \cdot 2^s \\
&\leq 2^{s+4}.
\end{aligned}$$

Therefore $sq^{-2^{s-2}/s} \leq q^{-\alpha m'/s}$. So

$$\mathsf{OuterSamp2}_{m,\delta,q} = \left(\mathsf{Curve}_{2,t_1,q^{d_1}} \circ \cdots \circ \mathsf{Curve}_{2,t_s,q^{d_s}}\right) \star \mathsf{RSCon}_{n/s,m'/s,q^s}$$

is an $\left((s+1)\epsilon_0, q^{m'-n}\right)$ sampler by Lemma 4.4.2, and hence an (ϵ, δ) sampler.

Finally, note that $n = s \cdot \left\lceil \left(m' + \log_q(1/\delta)\right)/s \right\rceil = O\left(m + \log_q(1/\delta)\right)$ and $d = s + 1 = O(\log m)$, as desired. □

Bibliography

[ALM+98] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, May 1998.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: a new characterization of NP. *J. ACM*, 45(1):70–122, January 1998.

[BR94] M. Bellare and J. Rompel. Randomness-efficient oblivious sampling. In *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, SFCS '94, pages 276–287, Washington, DC, USA, 1994. IEEE Computer Society.

[BSSVW03] Eli Ben-Sasson, Madhu Sudan, Salil Vadhan, and Avi Wigderson. Randomness-efficient low degree tests and short PCPs via epsilon-biased sets. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, STOC '03, pages 612–621, New York, NY, USA, 2003. ACM.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM J. Comput.*, 17:230–261, April 1988.

[CG89] B. Chor and O. Goldreich. On the power of two-point based sampling. *J. Complex.*, 5(1):96–106, April 1989.

[CW79] J. Lawrence Carter and Mark N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18(2):143 – 154, 1979.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *J. ACM*, 54(3), June 2007.

[GI01] V. Guruswami and P. Indyk. Expander-based constructions of efficiently decodable codes. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 658–, Washington, DC, USA, 2001. IEEE Computer Society.

[Gil98] David Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, August 1998.

[Gol11] Oded Goldreich. A sample of samplers: A computational perspective on sampling. In *Studies in Complexity and Cryptography*, volume 6650 of *Lecture Notes in Computer Science*, pages 302–332. Springer Berlin Heidelberg, 2011.

[GR08] V. Guruswami and A. Rudra. Explicit codes achieving list decoding capacity: Error-correction with optimal redundancy. *IEEE Trans. Inf. Theor.*, 54(1):135–150, January 2008.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *J. ACM*, 56:20:1–20:34, July 2009.

[ILL89] R. Impagliazzo, L. A. Levin, and M. 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.

[MR06] Dana Moshkovitz and Ran Raz. Sub-constant error low degree test of almost-linear size. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, STOC '06, pages 21–30, New York, NY, USA, 2006. ACM.

[MR08] Dana Moshkovitz and Ran Raz. Sub-constant error low degree test of almost-linear size. *SIAM J. Comput.*, 38(1):140–180, March 2008.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *J. Comput. Syst. Sci.*, 52:43–52, February 1996.

[Rei08] Omer Reingold. Undirected connectivity in log-space. *J. ACM*, 55(4):17:1–17:24, September 2008.

[RSW06] Omer Reingold, Ronen Shaltiel, and Avi Wigderson. Extracting randomness via repeated condensing. *SIAM J. Comput.*, 35(5):1185–1209, May 2006.

[RTS00] Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM J. Discret. Math.*, 13(1):2–24, January 2000.

[STV01] Madhu Sudan, Luca Trevisan, and Salil Vadhan. Pseudorandom generators without the XOR lemma. *J. Comput. Syst. Sci.*, 62(2):236–266, March 2001.

[SU05] Ronen Shaltiel and Christopher Umans. Simple extractors for all min-entropies and a new pseudorandom generator. *J. ACM*, 52(2):172–216, March 2005.

[SU06] Ronen Shaltiel and Christopher Umans. Pseudorandomness for approximate counting and sampling. *Comput. Complex.*, 15(4):298–341, December 2006.

[SZ99] Aravind Srinivasan and David Zuckerman. Computing with very weak random sources. *SIAM J. Comput.*, 28:1433–1459, March 1999.

[TSU06] Amnon Ta-Shma and Christopher Umans. Better lossless condensers through derandomized curve samplers. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 177–186, Washington, DC, USA, 2006. IEEE Computer Society.

[TSU12] Amnon Ta-Shma and Christopher Umans. Better condensers and new extractors from parvaresh-vardy codes. In *Proceedings of the 27th IEEE Conference on Computational Complexity*, 2012.

[TSUZ07] Amnon Ta-Shma, Christopher Umans, and David Zuckerman. Lossless condensers, unbalanced expanders, and extractors. *Combinatorica*, 27(2):213–240, March 2007.

[Uma03] Christopher Umans. Pseudo-random generators for all hardnesses. *J. Comput. Syst. Sci.*, 67(2):419–440, September 2003.

[Zuc97] David Zuckerman. Randomness-optimal oblivious sampling. In *Proceedings of the Workshop on Randomized Algorithms and Computation*, pages 345–367, New York, NY, USA, 1997. John Wiley & Sons, Inc.