AICR · OPEN PROBLEM

Randomness Extraction from Unpredictable Random-Oracle Sources

A Leftover-Hash-Lemma Conjecture, Public Seed

Status. Statement: human-written, formalized in Lean (see `lean/Statement.lean`).

Category. Information-Theoretic Cryptography.

Setting. Fix nonempty finite sets $\mathcal{K}$ (seeds), $\mathcal{D}$ (inputs), $\mathcal{R}$ (outputs), and write $K := |\mathcal{K}|$, $D := |\mathcal{D}|$, $R := |\mathcal{R}|$. Let $\mathsf{Fun}(\mathcal{K} \times \mathcal{D}, \mathcal{R})$ be the set of all functions $\mathcal{K} \times \mathcal{D} \to \mathcal{R}$, let $\mathrm{SD}(\cdot,\cdot)$ denote statistical distance and $U_{\mathcal{R}}$ the uniform distribution on $\mathcal{R}$. The source S, the predictor P and the distinguisher D of Figure 1 are computationally unbounded and receive the *entire* function table of H as an explicit input. The seed sd is sampled only after S has terminated, and is given to D: that is what makes this the *public-seed* game.

Game Pred_P^S

$H \xleftarrow{\$} \mathsf{Fun}(\mathcal{K} \times \mathcal{D}, \mathcal{R})$; $\quad (x, z) \xleftarrow{\$} S(H)$; $\quad x' \xleftarrow{\$} P(H, z)$
return $(x = x')$

Game $\mathrm{Ext\text{-}pub}_{\mathsf{D}}^S$ (public seed)

$H \xleftarrow{\$} \mathsf{Fun}(\mathcal{K} \times \mathcal{D}, \mathcal{R})$; $\quad (x, z) \xleftarrow{\$} S(H)$; $\quad sd \xleftarrow{\$} \mathcal{K}$
$y_0 \leftarrow H(sd, x)$; $\quad y_1 \xleftarrow{\$} \mathcal{R}$; $\quad b \xleftarrow{\$} \{0,1\}$
$b' \xleftarrow{\$} \mathsf{D}(H, sd, y_b, z)$; $\quad$ **return** $(b = b')$

Figure 1: Prediction game (top) and the public-seed extraction game (bottom). The first argument of S, P and D is the *full function table* of H; the distinguisher additionally receives the seed sd.

Definition 1. Set $\mathbf{Adv}^{\mathrm{pred}}_{\mathcal{K},\mathcal{D},\mathcal{R},S}(P) := \Pr[\mathrm{Pred}_P^S \Rightarrow 1]$ and

$$\mathbf{Adv}^{\mathrm{ext\text{-}pub}}_{\mathcal{K},\mathcal{D},\mathcal{R}}(S, \mathsf{D}) := 2\Pr[\mathrm{Ext\text{-}pub}_{\mathsf{D}}^S \Rightarrow 1] - 1.$$

A source S is *ϵ-unpredictable* if $\mathbf{Adv}^{\mathrm{pred}}_{\mathcal{K},\mathcal{D},\mathcal{R},S}(P) \le \epsilon$ for every unbounded predictor P.

Conjecture 1 (Public seed). *There is a universal constant $c > 0$, independent of $\mathcal{K}$, $\mathcal{D}$, $\mathcal{R}$ and ϵ, such that for all nonempty finite $\mathcal{K},\mathcal{D},\mathcal{R}$, all $\epsilon \in (0,1]$, every ϵ-unpredictable source S and every unbounded distinguisher* D*,*

$$\mathbf{Adv}^{\text{ext-pub}}_{\mathcal{K},\mathcal{D},\mathcal{R}}(S,\mathsf{D}) \;\le\; \delta_{\mathrm{pub}}(\epsilon,K,D,R) \;:=\; c\cdot\left(\sqrt{\epsilon R} + \sqrt{\frac{\log_2 D}{K}}\right).$$

Why this is worth asking, and what in it is new. The bound's *shape* is classical. Writing $n := \log_2 D$, $k := \log_2(1/\epsilon)$, $m := \log_2 R$ and $d := \log_2 K$ and balancing each summand against a target error δ gives $m = k - 2\log_2(1/\delta)$ and $d = \log_2 n + 2\log_2(1/\delta)$: the parameters of Theorem 6.17 of Vadhan's *Pseudorandomness*, which Radhakrishnan and Ta-Shma showed are optimal. What is new is the *hypothesis*. Here ϵ-unpredictability bounds $\mathbb{E}_{(H,Z)}\left[\max_x \Pr[X = x \mid H,Z]\right]$, an average over H, so the source may have no entropy at all on an ϵ-fraction of oracles; and z may be H itself. Classical statements quantify per source and conclude with high probability over the extractor, which is not the same thing, and the average-min-entropy leftover hash lemma of Dodis, Ostrovsky, Reyzin and Smith does not close the gap, since there the side information must be independent of the hash key. That the source sees the whole table is not itself the novelty: a union bound over flat sources already runs before the source is fixed.

The interest is what the seed costs. Read the game as key derivation — x a secret with entropy but no uniformity, sd a public salt, $H(sd,x)$ the derived key — and three prices are on offer, scaling in three different ways. An unsalted random oracle needs no salt, but only because it assumes the source cannot look at H; drop that and it fails outright, a source returning a uniform element of the largest fibre of a single row being distinguished with advantage $1 - 1/R$ by an adversary making *no query at all*. A 2-universal family survives such a source, but needs $K \ge D/R$, a salt of $n - m$ bits — Stinson's 1994 seed-length limitation, the second of the two that Barak et al.

open with. This conjecture asks for $\log_2 n + 2\log_2(1/\delta) + O(1)$: the second's robustness at essentially the first's price, and against an *unbounded* distinguisher, where bounding the adversary's queries instead would cost $\log_2 q$ bits. For a Diffie–Hellman element in a 2048-bit group, min-entropy 256, a 128-bit derived key and $\delta = 2^{-32}$, that is 77 bits of salt against 1920 for a universal family, 190 for the leftover hash lemma's own collision route, and 113 at a query budget of 2^{80}. The companion note `lhl-bound.tex` in this folder proves the obstruction and works the comparison out.

Remark 1 (What the leftover hash lemma gives). Condition on the table and the auxiliary output, $(H,Z) = (h,z)$, and write $p_x := \Pr[X = x \mid h,z]$. Because sd is drawn only after S has terminated it is uniform on $\mathcal{K}$ and independent of X given (h,z), so the distinguisher's task is to tell $(sd, h(sd,X))$ from uniform on $\mathcal{K}\times\mathcal{R}$: a strong-extractor question about the *fixed* family $\{h(k,\cdot)\}_{k\in\mathcal{K}}$, to which the leftover hash lemma applies in its almost-universal form. Writing $\mathrm{CP} := \sum_x p_x^2$ for the source's collision probability and $\bar{c} := \frac{1}{K}\sum_{k\in\mathcal{K}} \Pr[h(k,X) = h(k,X')]$ for the family's mean collision probability on that source, it bounds the conditional advantage by

$$\tfrac{1}{2}\sqrt{R\bar{c} - 1}, \qquad R\bar{c} - 1 \;=\; R\cdot\mathrm{CP} + \gamma, \qquad \gamma \;:=\; R\big(\bar{c} - \mathrm{CP}\big) - 1,$$

with $\gamma \le 0$ exactly when the family is universal on the source's support. Averaging over (H,Z) preserves the shape, $\sqrt{\cdot}$ being concave, and $\mathrm{CP} \le \max_x p_x$ turns the first summand into $\frac{1}{2}\sqrt{\epsilon R}$, which is δ_{pub}'s first term with $c = \frac{1}{2}$. On that summand the lemma is already at the conjectured shape.

On the second it is not, and the obstruction is structural rather than statistical. The family $\{h(k,\cdot)\}$ is not universal and cannot be made so — a universal family on D inputs needs a seed of $\Omega(\log D)$ bits, that is $K \ge D^{\Omega(1)}$, whereas this conjecture is of interest at $K \approx \log_2 D$ (Vadhan, *Pseudorandomness*, the remark following Theorem 6.18) — and the source, which reads the whole table before sd exists, can choose a support on which γ is large. One row suffices, and no probability is needed.

Put $W := \min\{R, \lfloor D\epsilon \rfloor\}$ and let $t := \lfloor 1/\epsilon \rfloor$. Fix any seed k_1 and let $\mathcal{S}$ collect the R/W symbols occurring most often in the row $h(k_1,\cdot)$. Those symbols carry at least $D/W \ge t$ of the D inputs, so a set $T \subseteq \{x : h(k_1,x) \in \mathcal{S}\}$ with $|T| = t$ exists; let S output a

uniform element of it, which is ϵ-unpredictable. On T the law of $h(k_1, X)$ is carried by the R/W symbols of $\mathcal{S}$, so Cauchy–Schwarz gives $c_{k_1} \geq W/R$, and the same inequality gives $c_k \geq 1/R$ for every k. Hence

$$R\bar{c} - 1 \;\geq\; R\left[\tfrac{1}{K} \cdot \tfrac{W}{R} + \tfrac{K-1}{K} \cdot \tfrac{1}{R}\right] - 1 \;=\; \frac{W-1}{K},$$

so the lemma returns at least $\frac{1}{2}\sqrt{(W-1)/K}$. This holds for *every* table, not merely a typical one. Pinning j rows rather than one replaces $W-1$ by about $j(W-1)$ and so buys a further $\log_2 j$ bits, usually one or two; the one-row form is essentially the whole effect.

The rate in K is therefore the same square root δ_{pub} asks for, and what separates them is what sits under it: the lemma pays $\sqrt{W}$, a *set size*, where δ_{pub} pays $\sqrt{\log_2 D}$, a *number of bits*. In seed length that is $\log_2 W$ against $\log_2 \log_2 D$. For $D = 2^{256}$, $R = 2^{64}$ and a source of min-entropy 128, reaching $\delta = 2^{-32}$ needs $\log_2 K \approx 72$ under δ_{pub} and 126 by this route. The gap closes when W does: at $\epsilon = 1/D$ one has $W = 1$ and the bound is 0, correctly, since a source of maximal entropy is uniform on all of $\mathcal{D}$ and has nothing to select; and for a source of min-entropy 254 over the same D and R the construction shows no loss at all.

The cause is the Cauchy–Schwarz step $\|v\|_1 \leq \sqrt{KR}\,\|v\|_2$ inside the lemma. Support selection concentrates the deviation v on one row out of K rather than spreading it over all KR cells, which is far from the even spread that step is tight on. How far is not uniform in the parameters and should not be overstated: against the source above the pinned row contributes about $1/K$ to the true advantage, but the other $K-1$ rows are not uniform either, each contributing $\Theta(\sqrt{R/t})$ on a typical table, so the truth is $\Theta(\sqrt{R/t}) + \Theta(1/K)$ and is dominated by the unpinned rows. The clean $O(1/K)$ reading holds once the distinguisher is query-bounded, where the unpinned rows cost only q_{D}/t; the companion note treats that model. So the conjecture is not a corollary of the leftover hash lemma, and its content is the second summand: the price of letting the source choose its support after seeing the entire table. The solution note in this folder pays that price at the conjectured rate by never passing through a collision probability, bounding the L_1 distance directly instead. (What is stated here is a limitation — the lemma cannot

do better than the displayed bound. Whether it does that well is not claimed.)

The secret-seed companion. This note states one conjecture, the one the accompanying page states. There is a companion statement in the same setting, differing only in that D is called as $\mathsf{D}(H, y_b, z)$ and never receives sd; it conjectures the bound $c\sqrt{(\epsilon R + \log_2 D)/K}$, in which the seed's factor $1/K$ attaches to both summands. It is a genuinely different claim, not a restatement: the solution note in this folder proves the conjecture above and separately shows that transposing the secret-seed expression verbatim into the public-seed game gives a *false* statement. The two were originally posed together in a single write-up and were split so that each note states exactly the conjecture its page states.