Hashing a Source With an Unknown Probability Distribution

Christian Cachin

MIT Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139, USA
cachin@acm.org

May 10, 1998

Abstract

Rényi entropy of order 2 characterizes how many almost uniform random bits can be extracted from a distribution by universal hashing by a technique known as "privacy amplification" in cryptography. We generalize this result and show that if P_S is the assumed distribution of a random variable with true distribution P_X, then the amount of extractable almost uniform randomness corresponds to $-\log \mathrm{P}[X = S]$, when X and S are interpreted as independent random variables.

Keywords. Universal Hashing, Privacy Amplification, Guessing, Inaccuracy, Renyi Entropy.

1 Introduction

We investigate the case of a source with unknown true distribution, when a wrong distribution is assumed instead. For example, the wrong distribution may be the best estimate we can make for the unknown true distribution.

The uncertainty about the outcome of a random variable with unknown distribution consists of two parts that can be related to the uncertainty of the random variable itself and to the lack of knowledge about the correct distribution. More specifically, if it is assumed that the distribution of a random variable X is P_S, the ignorance about X is characterized by $H(X) + D(P_X \| P_S)$, where $H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)$ is the *entropy* of X and $D(P_X \| P_S) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{P_S(x)}$ denotes the *relative entropy* or *discrimination* between two probability distributions P_X and P_S with the same alphabet $\mathcal{X}$. The quantity $H(X) + D(P_X \| P_S)$ is also called *inaccuracy* [6].

Inaccuracy measures the uncertainty about a source with unknown distribution similar to entropy (as demonstrated by the following example from [5]). Consider the construction of an optimal binary prefix-free code for a random variable X. If the distribution of X is known, the optimal code has average length between $H(X)$ and $H(X)+1$, which is one justification of entropy as a fundamental measure of uncertainty. If P_X is unknown and a code optimal for a distribution P_S is used, the average length of the code used for X lies between $H(X) + D(P_X \| P_S)$ and $H(X) + D(P_X \| P_S) + 1$. Thus, the expected number (over X) of binary questions needed to describe an outcome of X is at least $H(X) + D(P_X \| P_S)$, the inaccuracy. It is clear that $\lceil \log |\mathcal{X}| \rceil$ questions are always sufficient

by assuming the uniform distribution P_U over $\mathcal{X}$, corresponding to the equivalence

$$H(X) + D(P_X \| P_U) \;=\; \log |\mathcal{X}|. \tag{1}$$

2 Privacy Amplification by Universal Hashing

Entropy smoothing by universal hashing is a widely-used technique in cryptography and theoretical computer science to concentrate the randomness inherent in a probability distribution, known in different contexts as *entropy smoothing* [7] or *privacy amplification* [2, 1]. The amount of extractable almost uniform randomness is closely related to the Rényi entropy [8] of the distribution. The *Rényi entropy of order* 2 of a random variable X is defined as

$$H_2(X) = -\log \sum_{x \in \mathcal{X}} P_X(x)^2.$$

Universal hashing was introduced by Carter and Wegman [4]. It is a randomized hashing technique involving a family of functions from which one is chosen randomly and applied to the source. Formally, a *universal hash function* is a set $\mathcal{G}$ of functions $\mathcal{X} \to \mathcal{Y}$ if, for all distinct $x_1, x_2 \in \mathcal{X}$, there are at most $|\mathcal{G}|/|\mathcal{Y}|$ functions g in $\mathcal{G}$ such that $g(x_1) = g(x_2)$.

As the following result demonstrates, universal hashing can be used to compress the randomness of a source X into a smaller range Y such that the expected uncertainty of Y is exponentially close to the maximum. The size of the largest Y for which hashing of X yields an almost uniform output is characterized by $H_2(X)$. (The theorem can be extended to Rényi entropy of order α for any $\alpha > 1$ [3].)

Theorem 1 (Privacy Amplification [1]). *Let X be a random variable over the alphabet $\mathcal{X}$ with Rényi entropy $H_2(X)$, let G be the random variable corresponding to the random choice (with uniform distribution) of a member of a universal hash function $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$, and let $Y = G(X)$. Then*

$$H(Y|G) \;\geq\; \log |\mathcal{Y}| - \frac{2^{\log |\mathcal{Y}| - H_2(X)}}{\ln 2}. \tag{2}$$

In cryptography, for example, this technique can be used to extract a short secret key from information W shared by two parties Alice and Bob, when an eavesdropper Eve has partial information V about W. If Alice and Bob only know a lower bound on $H_2(W|V = v)$ for any particular v held by Eve, no matter what the exact distribution of Eve's knowledge is, they can exchange the description of a randomly chosen function $G \in \mathcal{G} : \mathcal{W} \to \mathcal{Y}$ over a public channel and are guaranteed that Eve has only negligible information about $G(W)$.

3 The Generalization

The following result extends Theorem 1 to universal hashing of a random variable X with unknown distribution. The theorem shows that the ignorance about the hashed value, measured by the inaccuracy, is related to the probability that a value of X can be guessed correctly. More precisely, when S denotes a random variable with the distribution we assume for X and when a shorter value Y is extracted from X by universal hashing, the size of the largest Y with inaccuracy close to the upper bound $\log |\mathcal{Y}|$ is characterized by $-\log \mathrm{P}[X = S]$.

Remember that the conditional relative entropy between P_X and P_Y conditioned on a random variable Z is defined as

$$D(P_{X|Z}\|P_{Y|Z}) = \sum_{z\in\mathcal{Z}} P_Z(z) \sum_{x\in\mathcal{X}} P_{X|Z=z}(x) \log \frac{P_{X|Z=z}(x)}{P_{Y|Z=z}(x)}. \tag{3}$$

Theorem 2. *Let X and S be independent random variables over the same alphabet $\mathcal{X}$ with probability distributions P_X and P_S, respectively, and let G be an independent random variable corresponding to the uniform selection of a member of a universal hash function $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$, and let $Y = G(X)$ and $Z = G(S)$. Then*

$$H(Y|G) + D(P_{Y|G}\|P_{Z|G}) \;\geq\; \log|\mathcal{Y}| - \frac{|\mathcal{Y}| \cdot \mathrm{P}[X = S]}{\ln 2}.$$

Proof. Expanding (3), we obtain

$$\begin{aligned}
&D(P_{G(X)|G}\|P_{G(S)|G}) \\
&= \sum_{g\in\mathcal{G}} P_G(g) \sum_{y\in\mathcal{Y}} P_{G(X)|G=g}(y) \log \frac{P_{G(X)|G=g}(y)}{P_{G(S)|G=g}(y)} \\
&= \sum_{g\in\mathcal{G}} P_G(g) \sum_{y\in\mathcal{Y}} P_{G(X)|G=g}(y) \log P_{G(X)|G=g}(y) \\
&\quad - \sum_{g\in\mathcal{G},y\in\mathcal{Y}} P_G(g) P_{G(X)|G=g}(y) \log P_{G(S)|G=g}(y) \\
&\overset{(a)}{\geq} -H(G(X)|G) - \log \sum_{g\in\mathcal{G},y\in\mathcal{Y}} P_G(g) P_{G(X)|G=g}(y) P_{G(S)|G=g}(y) \\
&\overset{(b)}{=} -H(G(X)|G) - \log \mathrm{P}[G(X) = G(S)] \\
&= -H(G(X)|G) - \log\Big(\mathrm{P}[X = S] + \\
&\qquad\qquad \mathrm{P}[X \neq S] \cdot \mathrm{P}[G(X) = G(S)|X \neq S]\Big) \\
&\overset{(c)}{\geq} -H(G(X)|G) - \log\Big(\mathrm{P}[X = S] + \mathrm{P}[X \neq S] \cdot |\mathcal{Y}|^{-1}\Big) \\
&> -H(G(X)|G) - \log\Big(\mathrm{P}[X = S] + |\mathcal{Y}|^{-1}\Big) \\
&= -H(G(X)|G) - \log\Big(|\mathcal{Y}|^{-1}(1 + |\mathcal{Y}| \cdot \mathrm{P}[X = S])\Big).
\end{aligned}$$

The inequality (a) follows from the application of the Jensen inequality to the negative sum. In step (b), the argument of the logarithm corresponds to the probability that $g(X) = g(S)$ if X and S are independent of each other and of G, which is selected randomly according to P_G. The second inequality (c) is a consequence of the universality of the hash function $\mathcal{G}$ from which G is chosen with uniform probability. Because the logarithms are to base 2, the theorem follows from the last expression and the inequality $\log(1 + x) \leq x/\ln 2$. □

The theorem generalizes over Theorem 1 because $D(P_X\|P_S) = 0$ if and only if $P_X(x) = P_S(x)$ for all $x \in \mathcal{X}$ and

$$2^{-H_2(X)} = \mathrm{P}[X_1 = X_2] = \sum_{x\in\mathcal{X}} P_X(x)^2$$

when X_1 and X_2 are independent random variables with distributions equal to P_X.

The theorem applies to situations in which P_X is unknown and the assumed distribution of X is P_S. It shows that if a code optimal for $Z = G(S)$ is used to find a value of $Y = G(X)$ by a series of binary questions, then the average number of questions needed is close to the upper bound $\log|\mathcal{Y}|$ whenever $\mathrm{P}[X = S] < |\mathcal{Y}|^{-1}$. Moreover, the average number of questions can be made arbitrarily close to $\log|\mathcal{Y}|$ by decreasing the size of $\mathcal{Y}$.

Example. Consider a random variable X with alphabet $\mathcal{X} = \{1, \ldots, n\}$ and distribution

$$P_X(i) \;=\; \frac{n+1}{n} \cdot \frac{1}{i(i+1)}.$$

Let P_S be the assumed distribution of X such that $P_S(i) = P_X(n+1-i)$ for $i = 1, \ldots, n$. Thus, the probabilities are assumed to be exactly in reverse order. With $n = 100$, we note that $H(X) = 2.81$ and $H(X) + D(P_X \| P_Y) = 10.36$. The Rényi entropy of order 2 is $H_2(X) = 1.76$, but the guessing probability using P_S satisfies $-\log \mathrm{P}[X = S] = 12.20$.

Let $\mathcal{G}$ be a universal hash function from $\mathcal{X}$ to $\{0, 1\}$. If the distribution of X was known, the average number of binary questions to determine $G(X)$ is $H(G(X)|G) \geq 0.147$ according to Theorem 1. But since the distribution of X is assumed to be P_S, at least $H(G(X)|G) + D(P_{G(X)|G} \| P_{G(S)|G}) \geq 0.999$ are needed on the average as demonstrated by Theorem 2.

4 Conclusions

Our result shows how entropy smoothing by universal hashing extends to random variables with unknown distributions. However, the theorem holds only if G is independent of P_X and of P_S and care has therefore to be taken for applying it in some cases.

Acknowledgment

I am grateful to Ueli Maurer for helpful discussions on this subject.

References

[1] C. H. Bennett, G. Brassard, C. Crépeau, and U. M. Maurer, "Generalized privacy amplification," *IEEE Transactions on Information Theory*, vol. 41, pp. 1915–1923, Nov. 1995.

[2] C. H. Bennett, G. Brassard, and J.-M. Robert, "How to reduce your enemy's information," in *Advances in Cryptology: CRYPTO '85* (H. C. Williams, ed.), vol. 218 of *Lecture Notes in Computer Science*, pp. 468–476, Springer, 1986.

[3] C. Cachin, *Entropy Measures and Unconditional Security in Cryptography*, vol. 1 of *ETH Series in Information Security and Cryptography*. Konstanz, Germany: Hartung-Gorre Verlag, 1997. ISBN 3-89649-185-7 (Reprint of Ph.D. dissertation No. 12187, ETH Zürich).

[4] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *Journal of Computer and System Sciences*, vol. 18, pp. 143–154, 1979.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[7] R. Impagliazzo, L. A. Levin, and M. Luby, "Pseudo-random generation from one-way functions," in *Proc. 21st Annual ACM Symposium on Theory of Computing (STOC)*, pp. 12–24, 1989.

[8] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 547–561, Univ. of Calif. Press, 1961.