Universal lossy compression under logarithmic loss

Yanina Shkel, Maxim Raginsky, and Sergio Verdú

***Abstract*—Universal lossy source coding with the logarithmic loss distortion criterion is studied. Bounds on the non-asymptotic fundamental limit of fixed-length universal coding with respect to a family of distributions are derived. These bounds generalize the well-known minimax bounds for universal lossless source coding. The asymptotic behavior of the resulting optimization problem is studied for a family of i.i.d. sources with a finite alphabet size, and is characterized up to a constant. The redundancy of memoryless sources behaves like $\frac{k}{2}\log n$, where n is the blocklength and k is the number of degrees of freedom in the parameter space. The impact of the coding rate is on the constant term: higher compression rate effectively reduces the volume of the parameter uncertainty set.**

I. Introduction

This paper studies fundamental limits of fixed-length universal lossy compression. In this setting the goal is to minimize the distortion of the original data given the target rate of compression. The true parameters of the source distribution are not known; however, it is known that they belong to some family, Λ. We wish to design a compressor which simultaneously performs well for all parameters in Λ. The performance of this compressor is measured in terms of redundancy which captures the additional distortion incurred due to uncertainty about the distribution of the source.

The approach we adopt here is to focus on one particular notion of a distortion criterion: the *logarithmic loss* (log-loss). On the one hand, log-loss distortion criterion has nice mathematical properties which allow for strong theoretical bounds in the non-asymptotic and multiterminal settings [1], [2]. On the other hand, it has deep connections with lossless compression for which universal fundamental limits are well developed. When it comes to lossy universal compression, focusing on logarithmic loss lets us obtain simple and elegant bounds that generalize those in universal lossless coding.

A. Logarithmic-loss distortion criterion

In the setting of this paper, the information source to be compressed is modeled by a random variable X with probability mass function P on either a finite or a countably infinite alphabet $\mathcal{X}$. The reconstruction alphabet is $\mathcal{P}(\mathcal{X})$: the set of all probability mass functions on $\mathcal{X}$.

Yanina Shkel is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA and with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. Maxim Raginsky is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA. Sergio Verdú is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. E-mails: yshkel@princeton.edu, maxim@illinois.edu, verdu@princeton.edu.

This work was supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

Definition 1. *The log-loss distortion between $x \in \mathcal{X}$ and its reconstruction $\hat{P} \in \mathcal{P}(\mathcal{X})$ is given by*

$$\mathsf{d}(x, \hat{P}) = \log \frac{1}{\hat{P}(x)}. \tag{1}$$

The log-loss measures distortion in settings when the reconstructed information is soft, that is, a distribution over possible values is provided. It is natural to construct information processing modules which interface by processing such soft information: A principal example of this is the belief propagation algorithm that works by passing beliefs, or probabilities, between nodes in order to perform statistical inference on graphical models [3].

Log-loss is closely connected to notions of *information* and *entropy*. Given $X \sim P$, the information in $x \in \mathcal{X}$ is given by

$$\imath_P(x) = \log \frac{1}{P(x)} \tag{2}$$

and the entropy of P is given by

$$H(P) = \mathbb{E}\left[\imath_P(X)\right]. \tag{3}$$

When X^n is a stationary and memoryless source, the log-loss rate-distortion function is known to be [1], [2]

$$R(d) = H(P) - d. \tag{4}$$

The linear relationship between rate and distortion in (4) can be attributed to the special structure of (1). Indeed, using (2) we can restate (1) as

$$\mathsf{d}(x, \hat{P}) = \imath_{\hat{P}}(x) \tag{5}$$

and interpret the log-loss distortion between x and its reconstruction $\hat{P}$ as the remaining uncertainty about x given that we know $\hat{P}$.

Finally, log-loss is a common loss function in statistical learning literature, see for example [4], [5]. Universal compression is a natural meeting point of information theory and learning theory, so log-loss is of particular interest in this context.

B. Universal lossless compression

Let the *relative information* in $x \in \mathcal{X}$ according to (P, Q) be given by

$$\imath_{P\|Q}(x) = \log \frac{P(x)}{Q(x)} \tag{6}$$

and the *relative entropy* be given by

$$D(P\|Q) = \mathbb{E}\left[\imath_{P\|Q}(X)\right] \tag{7}$$

where $X \sim P$. The non-asymptotic redundancy of *lossless* source coding is within one bit of the following minimax value [4]:

$$\min_{Q\in\mathcal{P}(\mathcal{X})} \max_{\theta\in\Lambda} D(P_\theta\|Q) \tag{8}$$

where Λ denotes the space of parameters of the possible distributions P_θ of X.

The asymptotic behavior of (8) has been extensively studied when X^n is a stationary memoryless source with an unknown distribution $P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Lambda$, where Λ is a compact subset of $\mathbb{R}^k$ [6]–[9]. It can be characterized as

$$\lim_{n\to\infty}\left[\min_{Q_n}\max_{\boldsymbol{\theta}\in\Lambda} D(P^n_{\boldsymbol{\theta}}\|Q_n) - \frac{k}{2}\log\frac{n}{2\pi\mathrm{e}}\right] = \log\int_\Lambda \sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}. \tag{9}$$

where $I(\boldsymbol{\theta})$ denotes the *Fisher Information* matrix.

C. Universal lossy compression

There have been a number of works on the fundamental limits of fixed-length universal lossy compression with general distortion criterion. Notably, [10], [11] report redundancy of $O(\log n)$. To the best of our knowledge, no universal single-shot results have been previously derived for a general distortion criterion.

By focusing on the log-loss distortion we are able to obtain sharp asymptotic and non-asymptotic bounds. To state our results we fix $R \in [0,\infty)$ and define

$$\mathcal{Q}_R(\mathcal{X}) = \{Q \in \mathcal{P}(\mathcal{X}) : H_\infty(Q) \geq R\} \tag{10}$$

where min-entropy, or Rényi entropy of order infinity, is $H_\infty(P) = \min_{x\in\mathcal{X}} \imath_P(x)$. Our first main result, Theorem 1, states that in the non-asymptotic setting the redundancy of universal lossy coding with M codewords is well approximated by

$$\mathcal{D}_\Lambda(R) = \min_{Q\in\mathcal{Q}_R(\mathcal{X})}\max_{\theta\in\Lambda}\left(D(P_\theta\|Q) - \min_{\tilde{Q}\in\mathcal{Q}_R(\mathcal{X})} D(P_\theta\|\tilde{Q})\right) \tag{11}$$

where $R = \log M$. Moreover, we use $\mathcal{D}_\Lambda(n,R)$ to denote (11) when X^n is a stationary memoryless source that is compressed by a code of size $M = \exp(nR)$. Our main asymptotic result, Theorem 2, shows that

$$\lim_{n\to\infty}\left(\mathcal{D}_\Lambda(n,R) - \frac{k}{2}\log\frac{n}{2\pi\mathrm{e}}\right) = \log\int_{\Lambda_R}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \tag{12}$$

where $\Lambda_R = \{\boldsymbol{\theta}\in\Lambda : H(P_{\boldsymbol{\theta}}) \geq R\}$. In other words, the fundamental limits are similar to those of lossless source coding, (8) and (9), and, as far as the limit on the left side of (12) can be compared to the lossless setting, the coding rate R has the net effect of shrinking the parameter uncertainty set from Λ to Λ_R.

The rest of this paper is structured as follows. In Section II we set up the problem and present preliminary results for coding with known distributions. Section III contains our main result, Theorem 1. We derive the redundancy of lossy coding with log-loss for classes of i.i.d. sources in Theorem 2, Section IV.

II. Preliminaries

A. Lossy source-coding with logarithmic loss

Definition 2. *A fixed-length lossy code of size M for the log-loss distortion criterion is a pair of mappings:*

$$\textit{Encoder:}\quad \mathsf{f}:\mathcal{X}\to\{1,\dots,M\}$$
$$\textit{Decoder:}\quad \mathsf{c}:\{1,\dots,M\}\to\mathcal{P}(\mathcal{X}).$$

A lossy source code (f,c) with M codewords is an (M,d)-lossy source code for the source X if

$$\mathbb{E}\left[\mathsf{d}(X,\mathsf{c}(\mathsf{f}(X)))\right] \leq d. \tag{13}$$

The non-asymptotic fundamental limit of compression with log-loss is given by

$$\mathsf{d}^*_P(M) = \inf\{d : \exists(M,d) - \text{lossy source code for } X\}. \tag{14}$$

B. Universal coding

We fix a family of distributions $\{P_\theta : \theta\in\Lambda\}$ and study lossy source codes that are universal over this family. To streamline the notation we use $\mathbb{E}_\theta[\cdot]$ to denote the expectation with respect to P_θ and $\mathsf{d}^*_\theta(\cdot)$ to denote $\mathsf{d}^*_{P_\theta}(\cdot)$.

Definition 3 (Redundancy)**.** *The redundancy for the family of distributions $\{P_\theta : \theta\in\Lambda\}$ is defined to be*

$$\mathcal{R}^*_\Lambda(M) = \min_{(\mathsf{f},\mathsf{c}):|\mathsf{f}|\leq M}\max_{\theta\in\Lambda}\left(\mathbb{E}_\theta\left[\mathsf{d}(X,\mathsf{c}(\mathsf{f}(X)))\right] - \mathsf{d}^*_\theta(M)\right) \tag{15}$$

where $|\mathsf{f}|$ denotes the cardinality of the image $\mathsf{f}(\mathcal{X})$.

C. Bounds on $\mathsf{d}^*_\theta(M)$

We give upper and lower bounds on $\mathsf{d}^*_\theta(M)$ that immediately yield our main result in Section III. The upper bound, Lemma 1, is based on the same greedy construction as [2, Theorem 6]; we omit the proof due to space constraints.

Lemma 1 (Achievability)**.** *Given any $Q \in \mathcal{Q}_{\log M}(\mathcal{X})$ there exists a code (f,c) of size M such that*

$$\mathbb{E}_\theta\left[\mathsf{d}(X,\mathsf{c}(\mathsf{f}(X)))\right] \leq H(P_\theta) - \log M + D(P_\theta\|Q) + \log 2 \tag{16}$$

for all $\theta\in\Lambda$ and

$$\mathbb{E}_\theta\left[\mathsf{d}(X,\mathsf{c}(\mathsf{f}(X)))\right] = H(P_\theta) - \log M + D(P_\theta\|Q) \tag{17}$$

if $M = 1$ or $M = |\mathcal{X}|$.

Note that when Λ is a singleton the following result improves on the converse in [1, Lemma 1] (see also [2, Theorem 3] for the single-shot version).

Lemma 2 (Converse)**.** *Assume $|\mathcal{X}| \geq M$ and let (f,c) be a lossy code of size M. There exists $Q \in \mathcal{Q}_{\log M}(\mathcal{X})$ such that*

$$\mathbb{E}_\theta\left[\mathsf{d}(X,\mathsf{c}(\mathsf{f}(X)))\right] \geq H(P_\theta) - \log M + D(P_\theta\|Q) \tag{18}$$

for all $\theta \in \Lambda$.

Proof. Fix an arbitrary code, (f, c), with M codewords and let $\hat{P}_u = \mathsf{c}(u)$, $U = \mathsf{f}(X)$, and c^* be another decoder given by

$$\mathsf{c}^*(u) = P_u^*, \tag{19}$$

$$P_u^*(x) = \mathbb{P}\left[X = x | U = u\right]. \tag{20}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X)))|\mathsf{f}(X) = u\right] \\
&= \sum_{x \in \mathcal{X}} \mathbb{P}\left[X = x|\mathsf{f}(X) = u\right] \log \frac{1}{\hat{P}_m(x)} && (21)\\
&= \sum_{x \in \mathcal{X}} P_u^*(x) \left(\log \frac{P_u^*(x)}{\hat{P}_u(x)} + \log \frac{1}{P_u^*(x)} \right) && (22)\\
&= D(P_u^* \| \hat{P}_u) + \mathbb{E}\left[\mathsf{d}(X, \mathsf{c}^*(\mathsf{f}(X)))|U = u\right]. && (23)
\end{aligned}$$

Averaging both sides with respect to P_U and noting that relative entropy is non-negative shows

$$\mathbb{E}_\theta\left[\mathsf{d}(X, \mathsf{c}(\mathsf{f}(X)))\right] \geq \mathbb{E}_\theta\left[\mathsf{d}(X, \mathsf{c}^*(\mathsf{f}(X)))\right]. \tag{24}$$

Let the distribution $Q \in \mathcal{Q}_{\log M}$ be given by

$$Q(x) = \frac{1}{M}\mathbb{P}\left[X = x|U = \mathsf{f}(x)\right] = \frac{1}{M} P^*_{\mathsf{f}(x)}(x). \tag{25}$$

Then,

$$\begin{aligned}
\mathbb{E}_\theta\left[\mathsf{d}(X, \mathsf{c}^*(\mathsf{f}(X)))\right] &= \sum_{x \in \mathcal{X}} P_\theta(x) \log \frac{1}{P^*_{\mathsf{f}(x)}(x)} && (26)\\
&= H(P_\theta) + \sum_{x \in \mathcal{X}} P_\theta(x) \log \frac{P_\theta(x)}{P^*_{\mathsf{f}(x)}(x)} && (27)\\
&= H(P_\theta) + \sum_{x \in \mathcal{X}} P_\theta(x) \log \frac{P_\theta(x)}{Q(x)} - \log M && (28)\\
&= H(P_\theta) - \log M + D(P_\theta \| Q) && (29)
\end{aligned}$$

and the result is proved. □

III. Main single-shot result

Our main result connects the operational notion of redundancy $\mathcal{R}^*_\Lambda(M)$ given in Definition 3 to the information quantity (11).

Theorem 1. *The redundancy for a family of distributions* Λ *satisfies*

$$\left|\mathcal{R}^*_\Lambda(M) - \mathcal{D}_\Lambda(\log M)\right| \leq \log 2 \tag{30}$$

and

$$\mathcal{R}^*_\Lambda(M) = \mathcal{D}_\Lambda(\log M) \tag{31}$$

whenever $M = 1$ *or* $M = |\mathcal{X}|$.

Proof. For a given P_θ we bound

$$\mathsf{d}^*_\theta(M) \leq H(P_\theta) - R + \min_{\tilde{Q} \in \mathcal{Q}_R(\mathcal{X})} D(P_\theta \| \tilde{Q}) + \log 2 \tag{32}$$

and

$$\mathsf{d}^*_\theta(M) \geq H(P_\theta) - R + \min_{\tilde{Q} \in \mathcal{Q}_R(\mathcal{X})} D(P_\theta \| \tilde{Q}) \tag{33}$$

where $R = \log M$. Equation (32) follows from applying Lemma 1 to $\{P_\theta\}$ and optimizing over $\mathcal{Q}_R$. Likewise, (33) follows from applying Lemma 2 to $\{P_\theta\}$ and optimizing over $\mathcal{Q}_R$.

Taking any $Q \in \mathcal{Q}_R$ and applying Lemma 1 to a family Λ, together with (33), yields

$$\mathcal{R}^*_\Lambda(M) \leq \max_{\theta \in \Lambda} \left(D(P_\theta \| Q) - \min_{\tilde{Q} \in \mathcal{Q}_R(\mathcal{X})} D(P_\theta \| \tilde{Q}) \right) + \log 2. \tag{34}$$

Since this holds for any $Q \in \mathcal{Q}_R$ we can optimize over $\mathcal{Q}_R$ to obtain an upper bound in (30). To see the other direction, let (f, c) be the code at the minimum in (30), and let $Q \in \mathcal{Q}_R$ be the corresponding distribution guaranteed by Lemma 2. Applying Lemma 2 to Λ, together with (32), yields

$$\begin{aligned}
\mathcal{R}^*_\Lambda(M) &\geq \max_{\theta \in \Lambda} \left(D(P_\theta \| Q) - \min_{\tilde{Q} \in \mathcal{Q}_R(\mathcal{X})} D(P_\theta \| \tilde{Q}) \right) - \log 2 && (35)\\
&\geq \mathcal{D}_\Lambda(R) - \log 2 && (36)
\end{aligned}$$

thereby completing the proof of (30). For $M = 1$ and $M = |\mathcal{X}|$ the lower bound in Lemma 2 matches the upper bound in Lemma 1. Repeating the same sequence of steps with a tighter upper bound shows (31). □

Example 1.

$$\mathcal{R}^*_\Lambda(1) = \mathcal{D}_\Lambda(0) = \min_{Q \in \mathcal{P}(\mathcal{X})} \max_{\theta \in \Lambda} D(P_\theta \| Q) \tag{37}$$

which recovers (8).

Example 2.

$$\mathcal{R}^*_\Lambda(|\mathcal{X}|) = \mathcal{D}_\Lambda(\log |\mathcal{X}|) = 0 \tag{38}$$

since $|\mathcal{Q}_{\log |\mathcal{X}|}(\mathcal{X})| = 1$. The redundancy is zero when the message set is large enough to losslessly encode all $x \in \mathcal{X}$.

The rest of this section is dedicated to bounds on relative entropy projections on $\mathcal{Q}_R$ which are given in Lemmas 3 and 4. These will prove useful in applying Theorem 1.

Lemma 3. *Let* $X \sim P \in \mathcal{P}(\mathcal{X})$ *be arbitrary and fix* $R = (0, \log |\mathcal{X}|)$. *Then,*

$$\min_{Q \in \mathcal{Q}_R(\mathcal{X})} D(P \| Q) \geq \left[\mathbb{E}\left[R - \imath_P(X)\right]\right]^+ = \left[R - H(P)\right]^+. \tag{39}$$

Due to space constraints and the simplicity of the result, the proof of Lemma 3 is omitted.

Lemma 4. *Let* $X \sim P$, $P \in \mathcal{P}(\mathcal{X})$ *be arbitrary and fix* $R = (0, \log |\mathcal{X}|)$. *Then*

$$\begin{aligned}
\mathbb{E}\left[\left[R - \imath_P(X)\right]^+\right] &\geq \min_{Q \in \mathcal{Q}_R(\mathcal{X})} D(P \| Q) && (40)\\
&\geq \mathbb{E}\left[\left[R - \imath_P(X)\right]^+\right] - \delta_R && (41)
\end{aligned}$$

where $\delta_R \leq \log \mathrm{e}$.

Proof. Let

$$\mathcal{S} = \{a \in \mathcal{X} : \imath_P(a) \geq R\} \tag{42}$$

and fix any $Q \in \mathcal{Q}_R$.

Claim 1: For any $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{Q}_R$ there exists at least one $\tilde{Q} \in \mathcal{Q}_R$ which satisfies the following requirements:

1) $\tilde{Q}(a) = \exp(-R)$ for all $a \in \mathcal{S}^c$,
2) $P(a) \le \tilde{Q}(a)$ for all $a \in \mathcal{S}$,
3) $\tilde{Q}(a) > Q(a) \implies \tilde{Q}(a) = P(a)$ for all $a \in \mathcal{S}$.

To show that such a distribution must exist we begin with distribution P and transform it into $\tilde{Q}$ via the following procedure. First, we set $\tilde{Q}(a) = \exp(-R)$ on $\mathcal{S}^c$; this ensures requirement 1) is satisfied. We initially set $\tilde{Q}(a) = P(a)$ on $\mathcal{S}$; this ensures that requirement 2) is satisfied. To finish the procedure and make $\tilde{Q}$ a valid distribution we need to assign the extra mass removed from $\mathcal{S}^c$ to $\mathcal{S}$. We do this by moving all the mass from the set $\mathcal{S}^c$ to the set $\tilde{\mathcal{S}} = \{a : Q(a) > P(a)\} \subset \mathcal{S}$ while satisfying the constraint $\tilde{Q}(a) \le Q(a), \forall a \in \tilde{\mathcal{S}}$; this will ensure that requirement 3) is satisfied. It remains to show that we are allowed to add enough mass to $\tilde{\mathcal{S}}$ to accommodate the extra mass removed from $\mathcal{S}^c$ without violating the constraint $\tilde{Q}(a) \le Q(a), \forall a \in \tilde{\mathcal{S}}$. Indeed,

$$\sum_{a\in\tilde{\mathcal{S}}} (Q(a) - P(a)) = \sum_{a\in\mathcal{S}} [Q(a) - P(a)]^+ \tag{43}$$

$$\ge \sum_{a\in\mathcal{S}} (Q(a) - P(a)) = \sum_{a\in\mathcal{S}^c} (P(a) - Q(a)) \tag{44}$$

$$\ge \sum_{a\in\mathcal{S}^c} \left(P(a) - \tilde{Q}(a)\right) = \sum_{a\in\mathcal{S}^c} \left[P(a) - \tilde{Q}(a)\right]^+ \tag{45}$$

where the inequality in (45) holds since $\tilde{Q}(a) \ge Q(a)$ on $\mathcal{S}^c$.

Claim 2: For any $Q \in \mathcal{Q}_R$ and a corresponding $\tilde{Q}$ satisfying requirements 1)-3)

$$D(P\|Q) \ge D(P\|\tilde{Q}). \tag{46}$$

Indeed,

$$D(P\|Q) - D(P\|\tilde{Q}) = \sum_{a\in\mathcal{X}} P(a)\imath_{\tilde{Q}\|Q}(a) \tag{47}$$

$$\ge \sum_{a\in\mathcal{X}} \tilde{Q}(a)\imath_{\tilde{Q}\|Q}(a) = D(\tilde{Q}\|Q) \ge 0 \tag{48}$$

where

- $\imath_{\tilde{Q}\|Q}(a) \ge 0$ and $\tilde{Q}(a) < P(a)$ on $\mathcal{S}^c$, thus (48) holds on $\mathcal{S}^c$ by requirement 1),
- if $\imath_{\tilde{Q}\|Q}(a) \le 0$ then (48) holds since $\tilde{Q}(a) \ge P(a)$ on $\mathcal{S}$ by requirement 2),
- if $\imath_{\tilde{Q}\|Q}(a) > 0$ on $\mathcal{S}$ then $\tilde{Q}(a) = P(a)$ by requirement 3) and (48) holds.

Claims 1 and 2 imply that there exists a

$$Q^* \in \operatorname*{arg\,min}_{Q\in\mathcal{Q}_R(\mathcal{X})} D(P\|Q) \tag{49}$$

satisfying requirements 1) and 2).

The upper bound (40) thus follows from

$$D(P\|Q^*) = \sum_{a\in\mathcal{X}} P(a)\imath_{P\|Q^*}(a) \tag{50}$$

$$= \sum_{a\in\mathcal{S}^c} P(a)\imath_{P\|Q^*}(a) + \sum_{a\in\mathcal{S}} P(a)\imath_{P\|Q^*}(a) \tag{51}$$

$$= \mathbb{E}\left[[R - \imath_P(X)]^+\right] + \sum_{a\in\mathcal{S}} P(a)\imath_{P\|Q^*}(a) \tag{52}$$

$$\le \mathbb{E}\left[[R - \imath_P(X)]^+\right] \tag{53}$$

where (52) holds because of requirement 1) and (53) holds because of requirement 2).

Finally, (41) will follow given an upper bound on

$$\delta_R = -\sum_{a\in\mathcal{S}} P(a)\imath_{P\|Q^*}(a). \tag{54}$$

Let $\epsilon_a = Q^*(a) - P(a)$ and observe that

$$\sum_{a\in\mathcal{S}} \epsilon_a = -\sum_{a\in\mathcal{S}^c} \epsilon_a < \sum_{a\in\mathcal{S}^c} P(a) = \mathbb{P}\left[\imath_P(X) < R\right] \le 1. \tag{55}$$

Then,

$$\delta_R = \sum_{a\in\mathcal{S}} P(a) \log \frac{P(a) + \epsilon_a}{P(a)} \tag{56}$$

$$= \sum_{a\in\mathcal{S}} P(a) \log \left(1 + \frac{\epsilon_a}{P(a)}\right) \tag{57}$$

$$\le \log \mathrm{e} \sum_{x\in\mathcal{S}} P(a) \frac{\epsilon_a}{P(a)} \tag{58}$$

$$= \log \mathrm{e} \sum_{a\in\mathcal{S}} \epsilon_a \tag{59}$$

where (58) follows from $\ln(1 + x) \le x$ for $x \ge 0$. □

Lemmas 3 and 4 give two variants of corresponding upper and lower bounds on

$$\min_{Q\in\mathcal{Q}_R(\mathcal{X})} D(P\|Q). \tag{60}$$

First, (39) and (40) give bounds on (60) which match, up to the slack in Jensen's inequality for the function $f(z) = [z]^+$. Secondly, (40) and (41) give bounds on (60) which differ by a small constant, δ_R. Both variants are needed for Theorem 2 in Section IV as we explain next.

IV. Asymptotic bounds

Let X^n be a stationary and memoryless source defined on some finite alphabet $\mathcal{X}$. In this section we assume that $X_i \sim P_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Lambda$ where Λ is a k-dimensional probability simplex. That is, $\theta_j = P_{\boldsymbol{\theta}}(j)$ for $j \in \{1, \ldots, k\}$ and $k = |\mathcal{X}| - 1$. According to Theorem 1, the redundancy of compressing X^n at rate R is characterized by $\mathcal{D}_\Lambda(n, R)$, see (11), and this is the quantity that we analyze in this section.

First, using (39) and (40) together with standard large deviation techniques we can show the following. Let

$$P_{\boldsymbol{\theta}_\alpha}(x) = \frac{P^\alpha_{\boldsymbol{\theta}}(x)}{\sum_{a\in\mathcal{X}} P^\alpha_{\boldsymbol{\theta}}(a)} \tag{61}$$

where $P_{\boldsymbol{\theta}_\alpha}$ is a scaled version of $P_{\boldsymbol{\theta}}$.

1) If $R < H(P_{\boldsymbol{\theta}})$, let $\alpha > 1$ be such that $R = \mathbb{E}_{P_{\boldsymbol{\theta}_\alpha}}[\imath_{P_{\boldsymbol{\theta}}}(X)]$. Then,

$$\min_{Q_n\in\mathcal{Q}_{nR}(\mathcal{X}^n)} D(P^n_{\boldsymbol{\theta}}\|Q_n) = 2^{-nD(P_{\boldsymbol{\theta}_\alpha}\|P_{\boldsymbol{\theta}}) + o(n)}. \tag{62}$$

2) If $R > H(P_{\boldsymbol{\theta}})$, let $\alpha < 1$ be such that $R = \mathbb{E}_{P_{\boldsymbol{\theta}_\alpha}}[\imath_{P_{\boldsymbol{\theta}}}(X)]$. Then,

$$\min_{Q_n\in\mathcal{Q}_{nR}(\mathcal{X}^n)} D(P^n_{\boldsymbol{\theta}}\|Q_n) = nR - nH(P_{\boldsymbol{\theta}})$$

$$+ 2^{-nD(P_{\boldsymbol{\theta}_\alpha}\|X)+o(n)}. \quad (63)$$

It is important to note that, given any $\delta > 0$, the right hand side of (62) goes to zero uniformly for all $\boldsymbol{\theta}$ such that $R < H(P_{\boldsymbol{\theta}}) - \delta$. The proofs of this fact and of (62) and (63) are omitted.

Theorem 2. *For any fixed $R \in (0, \log(k+1))$*

$$\lim_{n\to\infty}\left(\mathcal{D}_\Lambda(n,R) - \frac{k}{2}\log\frac{n}{2\pi\mathrm{e}}\right) = \log\int_{\Lambda_R}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \quad (64)$$

where $\Lambda_R = \{\boldsymbol{\theta}\in\Lambda : H(P_{\boldsymbol{\theta}}) \geq R\}$ and $I(\boldsymbol{\theta})$ is the Fisher Information matrix.

Proof. To show the lower bound fix $\delta > 0$ and for convenience denote

$$f_n(\boldsymbol{\theta}) = \min_{\tilde{Q}_n\in\mathcal{Q}_{nR}(\mathcal{X}^n)} D(P^n_{\boldsymbol{\theta}}\|\tilde{Q}_n). \quad (65)$$

Then

$$\begin{aligned}
&\min_{Q_n\in\mathcal{Q}_{nR}(\mathcal{X}^n)}\max_{\boldsymbol{\theta}\in\Lambda}\left(D(P^n_{\boldsymbol{\theta}}\|Q_n) - f_n(\boldsymbol{\theta})\right)\\
&\geq \min_{Q_n\in\mathcal{P}(\mathcal{X}^n)}\max_{\boldsymbol{\theta}\in\Lambda}\left(D(P^n_{\boldsymbol{\theta}}\|Q_n) - f_n(\boldsymbol{\theta})\right) \quad (66)\\
&\geq \min_{Q_n\in\mathcal{P}(\mathcal{X}^n)}\max_{\boldsymbol{\theta}\in\Lambda_{R+\delta}}\left(D(P^n_{\boldsymbol{\theta}}\|Q_n) - f_n(\boldsymbol{\theta})\right) \quad (67)\\
&\geq \min_{Q_n\in\mathcal{P}(\mathcal{X}^n)}\max_{\boldsymbol{\theta}\in\Lambda_{R+\delta}} D(P^n_{\boldsymbol{\theta}}\|Q_n) - \max_{\boldsymbol{\theta}\in\Lambda_{R+\delta}} f_n(\boldsymbol{\theta}) \quad (68)\\
&= \frac{k}{2}\log\frac{n}{2\pi\mathrm{e}} + \log\int_{\Lambda_{R+\delta}}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} + \epsilon_n - \max_{\boldsymbol{\theta}\in\Lambda_{R+\delta}} f_n(\boldsymbol{\theta}) \quad (69)
\end{aligned}$$

where $\lim_{n\to\infty}\epsilon_n = 0$. Equations (66) and (67) follow since $\mathcal{Q}_{nR} \subset \mathcal{P}(\mathcal{X}^n)$ and $\Lambda_{R+\delta} \subset \Lambda$, respectively. Equation (69) follows by [9, Theorem 2]. Subtracting $\frac{k}{2}\log\frac{n}{2\pi\mathrm{e}}$ from both sides and taking the limit in n we obtain

$$\begin{aligned}
&\lim_{n\to\infty}\left[\mathcal{D}_\Lambda(n,R) - \frac{k}{2}\log\frac{n}{2\pi\mathrm{e}}\right]\\
&\quad= \log\int_{\Lambda_{R+\delta}}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} - \max_{\boldsymbol{\theta}\in\Lambda_{R+\delta}}\lim_{n\to\infty} f_n(\boldsymbol{\theta}) \quad (70)\\
&\quad= \log\int_{\Lambda_{R+\delta}}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}, \quad (71)
\end{aligned}$$

where (70) follows since $f_n(\boldsymbol{\theta})$ converges to zero uniformly on $\Lambda_{R+\delta}$. Since the choice of δ is arbitrary and $\Lambda_{R+\delta} \subset \Lambda_R$, we obtain

$$\lim_{n\to\infty}\left[\mathcal{D}_\Lambda(n,R) - \frac{k}{2}\log\frac{n}{2\pi\mathrm{e}}\right] \geq \sup_{\delta>0}\log\int_{\Lambda_{R+\delta}}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \quad (72)$$

$$= \log\int_{\Lambda_R}\sqrt{\det I(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta} \quad (73)$$

which completes the proof of the lower bound.

We give a brief sketch for the proof of the upper bound because of space constraints. We pick $Q_n \in \mathcal{Q}_{nR}$ such that

$$Q_n(a^n) \propto \min\left(\exp(-nR), \max_{\boldsymbol{\theta}\in\Lambda} P^n_{\boldsymbol{\theta}}(a^n)\right), \forall a^n \in \mathcal{X}^n. \quad (74)$$

Plugging (74) into $\max_{\boldsymbol{\theta}\in\Lambda}\left(D(P^n_{\boldsymbol{\theta}}\|Q_n) - f_n(\boldsymbol{\theta})\right)$ shows that $\mathcal{D}_\Lambda(n,R) \leq A + B$ where

$$A = \sum_{a^n\in\mathcal{X}^n}\min\left(\exp(-nR), \max_{\boldsymbol{\theta}\in\Lambda} P^n_{\boldsymbol{\theta}}(a^n)\right), \quad (75)$$

$$B = \max_{\boldsymbol{\theta}\in\Lambda}\left(\mathbb{E}_{P^n_{\boldsymbol{\theta}}}\left[\left[nR - \imath_{P^n_{\boldsymbol{\theta}}}(X^n)\right]^+\right] - f_n(\boldsymbol{\theta})\right). \quad (76)$$

Equation (75) can be analyzed using Stirling's approximation and method of types. Bounding (76) with (39) yields the desired answer for almost all $\boldsymbol{\theta} \in \Lambda$. The problematic parameters are those $\boldsymbol{\theta}$ for which $H(P_{\boldsymbol{\theta}})$ is close to R, cf. (62) and (63). If instead we use (41) to bound (76) the desired result follows, up to a $\log\mathrm{e}$ constant gap, for all of Λ. Tweaking (74) appropriately and leveraging the two approaches together yields the desired result over all of Λ and without the constant gap. □

Example 3 (Bernoulli random variable)**.** Let $h^{-1}(R)$ denote the inverse of the binary entropy function on $[0, \frac{1}{2}]$ and let $|\mathcal{X}| = 2$. Then

$$\lim_{n\to\infty}\left[\mathcal{D}_\Lambda(n,R) - \frac{1}{2}\log\frac{n}{2\pi\mathrm{e}}\right] = \log\int_{h^{-1}(R)}^{1-h^{-1}(R)}\frac{\mathrm{d}\theta}{\sqrt{\theta(1-\theta)}} \quad (77)$$

$$= \log\arcsin\left(1 - 2h^{-1}(R)\right) + \log 2. \quad (78)$$

REFERENCES

[1] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, Jan 2014.

[2] Y. Shkel and S. Verdú, "A single-shot approach to lossy source coding under logarithmic loss," submitted.

[3] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 12, pp. 1–305, 2008.

[4] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, Oct 1998.

[5] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[6] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 646–657, Mar 1997.

[7] ——, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, Mar 2000.

[8] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[9] ——, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol. 41, no. 1, pp. 37 – 60, 1994.

[10] T. Linder, G. Lugosi, and K. Zeger, "Fixed-rate universal lossy source coding and rates of convergence for memoryless sources," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 665–676, May 1995.

[11] E.-H. Yang and Z. Zhang, "The redundancy of source coding with a fidelity criterion .II. coding at a fixed rate level with unknown statistics," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 126–145, Jan 2001.