Sequential prediction with coded side information under logarithmic loss

Yanina Shkel yshkel@princeton.edu
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

Maxim Raginsky maxim@illinois.edu
Department of Electrical and Computer Engineering
Coordinated Science Laboratory
University of Illinois
Urbana, IL 61801, USA

Sergio Verdú verdu@princeton.edu
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544, USA

Editor:

Abstract

We study the problem of sequential prediction with coded side information under logarithmic loss (log-loss). We show an operational equivalence between this setup and lossy compression with log-loss distortion. Using this insight, together with recent work on lossy compression with log-loss, we connect prediction strategies with distributions in a certain subset of the probability simplex. This allows us to derive a Shtarkov-like bound for regret and to evaluate the regret for several illustrative classes of experts. In the present work, we mainly focus on the "batch" side information setting with sequential prediction.

1. Introduction

We study the problem of sequential prediction, or sequential probability assignment, with logarithmic loss (log-loss). In this setting, the predictor sees realizations $x^{t-1} \in \mathcal{X}^{t-1}$ of some sequence taking values on a finite or countably infinite alphabet $\mathcal{X}$ and aims to predict the next realization of the sequence, x_t. The predictor outputs *soft information*; that is, it provides a distribution $\hat{P} \in \mathcal{P}(\mathcal{X})$ over the possible values of x_t. The loss incurred by the predictor at time t is given by

$$\ell(x_t, \hat{P}) = \log \frac{1}{\hat{P}(x_t)}, \tag{1}$$

thereby incurring a high loss if the actual outcome is unlikely under the predicted distribution $\hat{P}$. This problem has been well studied in information-theoretic and learning-theoretic literature, see, for example, Merhav and Feder (1998), Cesa-Bianchi and Lugosi (2006, Chapter 9) and references therein. In this work, we generalize the problem and allow the

predictor access to coded side information prior to making its prediction. In our setting there are two agents: a compressor and a predictor. The compressor knows the true sequence, but is only allowed a rate-limited connection to the predictor; thus it needs to compress the sequence to facilitate the prediction task.

It is well known (e.g. Merhav and Feder (1998)) that the problem of sequential prediction with log-loss is equivalent to the problem of *lossless* compression. Given a good compressor for an information source, we can use this compressor to design a good predictor. Conversely, given a good predictor we can use it to design a good compressor. This connection between lossless compression and prediction has far-reaching implications: it has given rise to statistical inference frameworks such as the Minimum Description Length principle, see, for example, Grünwald (2007). This work demonstrates an analogous correspondence between universal *lossy* compression with log-loss and sequential prediction with coded side information. Such an equivalence establishes a nice duality between the two problems: in universal lossy compression, the aim is to learn the distribution in order to lower distortion, while in sequential prediction with coded side information the goal is to compress in order to improve learning.

Universal lossy compression with log-loss has been previously studied in Shkel et al. (2017) and tight bounds on redundancy with respect to a class of distributions were derived. The operational equivalence stated here shows that these bounds also hold for sequential prediction with coded side information. In this work, we focus on a learning-theoretic style of analysis for these problems and characterize the regret with respect to a reference class of experts. Because of the established equivalence our analysis also extends to lossy compression with log-loss with respect to a reference class of experts. Moreover, since conventional sequential prediction is equivalent to sequential compression, the present set up also covers the problem of lossless sequential compression with coded side information. For example, suppose a compression system assigns a fixed-length tag to the data file and then adapts its lossless compression strategy according to this tag. In this case, the rate of the lossy compressor is the number of distinct values the tag can take, while the loss of the predictor is the length of the compressed file.

2. Preliminaries

We consider a variation on the conventional sequential prediction where there are two agents: a compressor and a predictor. The compressor is omniscient: it sees the entire sequence to be predicted, but can only communicate a rate-limited version of this sequence to the predictor at time zero. At each time step, the predictor tries to guess the next observation using the previously observed values and the message communicated by the compressor.

2.1 Compressor-predictor system

Definition 1 *A batch compressor-predictor system with M messages is a collection of mappings:*

$$\textit{Compressor:} \quad \mathsf{f_c} : \mathcal{X}^n \to \{1, \ldots, M\} \tag{2}$$

$$\textit{Predictor:} \quad \mathsf{f}_t : \{1, \ldots, M\} \times \mathcal{X}^{t-1} \to \mathcal{P}(\mathcal{X}), \quad 1 \leq t \leq n. \tag{3}$$

Note that even though the side information encoding is batch, the prediction is still sequential. We discuss an extension to the sequential side information in Section 5.

We use the notation $\mathsf{f}_t(\cdot|m, x^{t-1})$ to denote the distribution selected by the predictor at time t on the basis of the side information index m, and the sequence past x^{t-1}. At each time step t the predictor incurs the loss

$$\ell\left(x_t, \mathsf{f}_t(\cdot|m, x^{t-1})\right) = \log \frac{1}{\mathsf{f}_t\left(x_t|m, x^{t-1}\right)}. \tag{4}$$

The cumulative loss incurred by a batch compressor-predictor system $\mathsf{f} = \{\mathsf{f}_\mathsf{c}, \mathsf{f}_1, \ldots, \mathsf{f}_n\}$ with M messages for a sequence x^n is

$$L_\mathsf{f}(x^n) = \sum_{t=1}^{n} \ell\left(x_t, \mathsf{f}_t\left(\cdot|m, x^{t-1}\right)\right) \text{ where } m = \mathsf{f}_\mathsf{c}(x^n). \tag{5}$$

Next, recall the definition of a lossy code given in Shkel et al. (2017).

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

$$\begin{aligned} \textit{Compressor:} \quad & \mathsf{f}_\mathsf{c} : \mathcal{X}^n \to \{1, \ldots, M\} \\ \textit{Decompressor:} \quad & \mathsf{f}_\mathsf{d} : \{1, \ldots, M\} \to \mathcal{P}(\mathcal{X}^n). \end{aligned}$$

The distortion incurred by a lossy code $(\mathsf{f}_\mathsf{c}, \mathsf{f}_\mathsf{d})$ on a sequence x^n is denoted by

$$\mathsf{d}(x^n, \mathsf{f}_\mathsf{d}(\mathsf{f}_\mathsf{c}(x^n))) = \log \frac{1}{\hat{P}(x^n)}, \quad \hat{P} = \mathsf{f}_\mathsf{d}(\mathsf{f}_\mathsf{c}(x^n)). \tag{6}$$

It turns out that the compressor-predictor systems in Definition 1 are operationally equivalent to lossy compressors in Definition 2, as is shown in the following lemma.

Lemma 1 *Given a compressor-predictor system f, we can construct a lossy code $(\mathsf{f}_\mathsf{c}, \mathsf{f}_\mathsf{d})$ such that*

$$L_\mathsf{f}(x^n) = \mathsf{d}(x^n, \mathsf{f}_\mathsf{d}(\mathsf{f}_\mathsf{c}(x^n))) \tag{7}$$

for all $x^n \in \mathcal{X}^n$. The reverse is also true: given a lossy code $(\mathsf{f}_\mathsf{c}, \mathsf{f}_\mathsf{d})$ we can construct a compressor-predictor system f such that (7) holds.

Proof Given a compressor-predictor system $\mathsf{f} = \{\mathsf{f}_\mathsf{c}, \mathsf{f}_1, \ldots, \mathsf{f}_n\}$ we can construct a lossy code $(\mathsf{f}_\mathsf{c}, \mathsf{f}_\mathsf{d})$:

$$\mathsf{f}_\mathsf{d}(m) = \hat{P}_m, \text{ where } \hat{P}_m(x^n) = \prod_{t=1}^{n} \mathsf{f}_t(x_t|m, x^{t-1}). \tag{8}$$

Likewise, given a lossy code $(\mathsf{f}_\mathsf{c}, \mathsf{f}_\mathsf{d})$ we can construct a compressor-predictor system f:

$$\mathsf{f}_t\left(x_t|m, x^{t-1}\right) = \frac{\sum_{\tilde{x}_{t+1}^n \in \mathcal{X}^{n-t}} \hat{P}_m\left(x^t \cdot \tilde{x}_{t+1}^n\right)}{\sum_{\tilde{x}_t^n \in \mathcal{X}^{n-t+1}} \hat{P}_m\left(x^{t-1} \cdot \tilde{x}_t^n\right)}, \tag{9}$$

where $\hat{P}_m = \mathsf{f_d}(m)$ and $a \cdot b$ denotes the concatenation of strings a and b. Then

$$L_{\mathsf{f}}(x^n) = \sum_{t=1}^{n} \log \frac{1}{\mathsf{f}_t\left(x_t | \mathsf{f_c}(x^n), x^{t-1}\right)} = \log \frac{1}{\hat{P}_m(x^n)} = \mathsf{d}\left(x^n, \mathsf{f_d}\left(\mathsf{f_c}\left(x^n\right)\right)\right). \tag{10}$$

■

Note that in Shkel et al. (2017) the definition of lossy codes and subsequent derivation of non-asymptotic bounds uses the so-called "single-shot approach", where everything is derived for an arbitrary alphabet $\mathcal{X}$, and no Cartesian product structure or blocklength n is assumed. In contrast, in the present prediction problem we cannot follow the single-shot approach, as the dynamical aspect of the problem is of the essence. Nevertheless, the two approaches are equivalent from the compression viewpoint. Indeed, we can always take $n = 1$ to recover the single-shot results. Conversely, if a statement is true for an arbitrary alphabet $\mathcal{Y}$, it is also true for $\mathcal{Y} = \mathcal{X}^n$.

2.2 Redundancy and Regret

Lemma 1 shows that, under log-loss, the problem of lossy compression is equivalent to the problem of prediction with coded side information in a very strong sense. That is, we can establish a one-to-one correspondence between lossy compression schemes and compressor-predictor schemes. This correspondence is such that the distortion incurred by the lossy compressor is the same as the loss incurred by the corresponding compressor-predictor system simultaneously for every $x^n \in \mathcal{X}^n$. This, in turn, means that most of the fundamental limits of interest for these two problems are going to be the same.

In particular, suppose that the sequence X^n is randomly generated according to some distribution P_θ. We do not know P_θ exactly, but we do know that it belongs to some family of distributions indexed by $\theta \in \Lambda$. In this case, we can immediately leverage redundancy bounds for lossy compression. Let

$$L^*_{n,\theta}(M) = \inf_{\mathsf{f}\colon |\mathsf{f_c}| \le M} \mathbb{E}\left[L_{\mathsf{f}}(X^n)\right] \tag{11}$$

denote the smallest cumulative loss incurred by any compressor-predictor system where $X^n \sim P_\theta \in \mathcal{P}(\mathcal{X}^n)$ and $|\mathsf{f_c}|$ denotes the cardinality of the image $\mathsf{f_c}(\mathcal{X}^n)$.

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

$$\mathcal{R}_n(M, \Lambda) = \inf_{\mathsf{f}\colon |\mathsf{f_c}| \le M} \sup_{\theta \in \Lambda} \left\{\mathbb{E}\left[L_{\mathsf{f}}(X^n)\right] - L^*_{n,\theta}(M)\right\} \quad \textit{where } X^n \sim P_\theta. \tag{12}$$

To state our results, let $R \in [0, \log|\mathcal{X}|]$ and define

$$\mathcal{Q}_\nu(\mathcal{X}^n) = \{Q \in \mathcal{P}(\mathcal{X}^n)\colon H_\infty(Q) \ge \nu\} \tag{13}$$

$$= \{Q \in \mathcal{P}(\mathcal{X}^n)\colon Q(x^n) \le \exp(-\nu) \quad \forall x^n \in \mathcal{X}^n\} \tag{14}$$

where

$$H_\infty(Q) = \min_{x^n \in \mathcal{X}^n} \log \frac{1}{Q(x^n)} \tag{15}$$

denotes the min-entropy of Q. Recall that the *relative entropy* between P and reference measure Q is given by

$$D(P\|Q) = \mathbb{E}\left[\log \frac{P(X)}{Q(X)}\right], \quad P \ll Q \text{ and } X \sim P. \tag{16}$$

The following is a direct corollary of Shkel et al. (2017, Theorem 1) and Lemma 1.

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

$$\left| \min_{Q \in \mathcal{Q}_{\log M}(\mathcal{X}^n)} \max_{\theta \in \Lambda} \left\{ D(P_\theta\|Q) - \min_{\tilde{Q} \in \mathcal{Q}_{\log M}(\mathcal{X}^n)} D(P_\theta\|\tilde{Q}) \right\} - \mathcal{R}_n(M, \Lambda) \right| \leq \log 2 \tag{17}$$

where the left side is zero whenever $M = 1$ or $M = |\mathcal{X}^n|$.

Note that, when $M = 1$, the present problem reduces to the sequential prediction with no side information. In that case, Theorem 1 recovers

$$\mathcal{R}_n(1, \Lambda) = \min_{Q \in \mathcal{P}(\mathcal{X}^n)} \max_{\theta \in \Lambda} D(P_\theta\|Q) \tag{18}$$

which is well known to be the redundancy of sequential prediction with log-loss, see, for example, Merhav and Feder (1998). When $M = |\mathcal{X}^n|$, there is only one reasonable compressor-predictor strategy: to losslessly encode every element of $\mathcal{X}^n$ and to incur no prediction loss. In that case, the distribution of the source is irrelevant, and Theorem 1 recovers

$$\mathcal{R}_n(|\mathcal{X}^n|, \Lambda) = 0. \tag{19}$$

The minimax bound in Theorem 1 is further analyzed in Shkel et al. (2017), while this work focuses on a learning-theoretic analysis of sequential prediction with coded side information, i.e. the regret with respect to a class of experts.

Definition 4 (Regret) *Let $\mathcal{F}$ be an expert class of compressor-predictor systems with M messages. The regret with respect to $\mathcal{F}$ is*

$$\mathcal{V}_n(M, \mathcal{F}) = \inf_{\mathsf{g}\colon |\mathsf{g}_\mathsf{c}| \leq M} \sup_{x^n \in \mathcal{X}^n} \left\{ L_\mathsf{g}(x^n) - \inf_{\mathsf{f} \in \mathcal{F}} L_\mathsf{f}(x^n) \right\}. \tag{20}$$

For $M = 1$, (20) denotes the regret of sequential prediction with no coded side information. In this case, the minimax optimal forecaster can be determined explicitly, and the minimax regret characterized exactly, as

$$\mathcal{V}_n(1, \mathcal{F}) = \log \sum_{x_n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} \exp(-L_\mathsf{f}(x^n)), \tag{21}$$

see (Cesa-Bianchi and Lugosi, 2006, Theorem 9.1) and Shtarkov (1987). In the next section we extend this result for any $1 \leq M \leq |\mathcal{X}^n|$.

Finally, we note that the expert setting for prediction with coded side information is indeed distinct from other side information settings previously studied, for example Cesa-Bianchi and Lugosi (2006, Chapter 9.9). In the present setting, the side information is not assumed to be common to all of the experts; instead, each expert is allowed to design its own side information.

3. Bounds on regret

We begin by deriving sequence-wise versions of Shkel et al. (2017, Lemmas 1 and 2) which connect compressor-predictor schemes to distributions in $\mathcal{Q}_{\log M}(\mathcal{X}^n)$. We then derive a Shtarkov bound for $\mathcal{V}_n(M, \mathcal{F})$ for an arbitrary family of experts $\mathcal{F}$.

3.1 Characterizing compressor-predictor systems

First, we state the following regularity condition on a compressor-predictor system f: the compressor-predictor system satisfies

$$\hat{P}_m(x^n) = 0, \text{ where } \hat{P}_m = \prod_{t=1}^{n} \mathsf{f}_t(\cdot|m, x^{t-1}), \quad \forall m \neq \mathsf{f}_{\mathsf{c}}(x^n) \tag{22}$$

for all $x^n \in \mathcal{X}^n$. Note that (22) imposes an obvious optimality criterion on a compressor-predictor system; for further justification of this regularity condition see Appendix A. We will assume that all compressor-predictor systems in our subsequent discussion satisfy (22).

Lemma 2 *Given a compressor-predictor system f with M messages, there exists $Q \in \mathcal{Q}_{\log M}(\mathcal{X}^n)$ such that*

$$L_{\mathsf{f}}(x^n) = \log \frac{1}{Q(x^n)} - \log M \tag{23}$$

holds for all $x^n \in \mathcal{X}^n$.

Proof Given a compressor-predictor system f, define

$$Q(x^n) = \frac{1}{M} \sum_{m=1}^{M} \prod_{t=1}^{n} \mathsf{f}_t\left(x_t|m, x^{t-1}\right) = \frac{1}{M} \prod_{t=1}^{n} \mathsf{f}_t\left(x_t|\mathsf{f}_{\mathsf{c}}(x^n), x^{t-1}\right) \tag{24}$$

where the second equality in (24) follows from (22) and it is straightforward to verify that Q is indeed a distribution in $\mathcal{Q}_{\log M}(\mathcal{X}^n)$. Then

$$\sum_{t=1}^{n} \ell\left(x^t, \mathsf{f}_t\left(\cdot|\mathsf{f}_{\mathsf{c}}(x^n), x^{t-1}\right)\right) = \log \frac{1}{MQ(x^n)}. \tag{25}$$

■

Lemma 2 shows that every compressor-predictor system can be associated with a distribution in $\mathcal{Q}_{\log M}$. The reverse is not true: that is, given a distribution in $\mathcal{Q}_{\log M}$ it may not be possible to construct a compressor-predictor system that satisfies (23). However, it is possible to construct one that satisfies it approximately, as the following lemma shows.

Lemma 3 *Given any $Q \in \mathcal{Q}_{\log M}(\mathcal{X}^n)$ it is possible to construct a compressor-predictor system f such that*

$$L_{\mathsf{f}}(x^n) \leq \log \frac{1}{Q(x^n)} - \log(M+1) + \log 2 < \log \frac{1}{Q(x^n)} - \log M + \log 2 \tag{26}$$

for all $x^n \in \mathcal{X}^n$. Moreover, (26) can be tighrened to $L_{\mathsf{f}}(x^n) = \log \frac{1}{Q(x^n)} - \log M$ whenever $M = 1$ or $M = |\mathcal{X}^n|$.

Equation (26) follows by using the greedy construction from Shkel and Verdú (2018, Theorem 4) which we include for completeness in Appendix B. The main idea is to construct a compressor that tries to equalize the probability assigned to each message according to distribution Q and outputs a posterior distribution of x^n given the message. If such a compressor exists, then (23) would be satisfied. A simple greedy compressor achieves such equalization up to a factor of $\frac{2M}{M+1}$. This greedy compressor then induces a compressor-predictor scheme per Lemma 1. In general, it is not possible to do better than (26) as the next example shows.

Example 1 *Let $\mathcal{X} = \{0, 1\}$ and $M = 2^k$ for some $1 \le k \le n$. Consider a compressor-predictor system f defined by the following ingredients:*

- *The compressor f_c is a bijection from $\mathcal{X}^k$ to $\{1, \dots, M\}$ that maps the first k bits of x^n to a unique message.*
- *For $1 \le t \le k$ let*

$$\mathsf{f}_t(\cdot|m, x^{t-1}) = \delta_{y_t}, \text{ where } y^k = \mathsf{f}_\mathsf{c}^{-1}(m) \tag{27}$$

  and δ_x denotes a point mass at x.
- *For $k + 1 \le t \le n$ let*

$$\mathsf{f}_t(0|m, x^{t-1}) = \mathsf{f}_t(1|m, x^{t-1}) = \frac{1}{2}. \tag{28}$$

Then the distribution guaranteed to exist by Lemma 2 is given by $Q(x^n) = 2^{-n}$ for all x^n. On the other hand, consider constructing a compressor-predictor system for Q when $M = 2^n - 1$. Using the compressor in Appendix B we can construct a compressor-predictor that will satisfy (26) with equality for two distinct elements in $\mathcal{X}^n$. It is not possible to improve on this. Indeed, by the pigeonhole principle, there exists a message m, such that the pre-image of m contains two distinct elements of $\mathcal{X}^n$. Let x^n and y^n be these elements and let t be the smallest index at which $x_t \ne y_t$. Then, it must be the case that

$$\mathsf{f}_t(x_t|m, x^{t-1}) = \mathsf{f}_t(y_t|m, y^{t-1}) = \frac{1}{2} \tag{29}$$

for (26) to be satisfied. In this case (26) is satisfied with equality for x^n and y^n.

3.2 Shtarkov bound for compressor-predictor strategies

Lemmas 2 and 3 connect compressor-predictor strategies with M messages to distributions in $\mathcal{Q}_{\log M}(\mathcal{X}^n)$. This is a pleasing extension of the usual sequential prediction with log-loss where prediction strategies are associated with elements of the probability simplex. Using Lemma 2 we can rewrite the regret in Definition 4 as

$$\mathcal{V}_n(M, \mathcal{F}) = \inf_{\mathsf{g}:\, |\mathsf{g}_\mathsf{c}| \le M} \sup_{x^n \in \mathcal{X}^n} \log \frac{\sup_{\mathsf{f} \in \mathcal{F}} Q_\mathsf{f}(x^n)}{Q_\mathsf{g}(x^n)} \tag{30}$$

where Q_f denotes the distribution corresponding to the strategy f, see (24). We have now laid the necessary groundwork to characterize the regret for sequential prediction with coded side information.

Theorem 2 (Shtarkov sum) *Let Q_{f} denote the distribution associated with $\mathsf{f} \in \mathcal{F}$ via Lemma 2, see (24). Then,*

$$\log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n) \le \mathcal{V}_n(M, \mathcal{F}) \le \log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n) + \log 2 \tag{31}$$

or, equivalently,

$$\log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} \exp(-L_{\mathsf{f}}(x^n)) - \log M \le \mathcal{V}_n(M, \mathcal{F}) \tag{32}$$

$$\le \log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} \exp(-L_{\mathsf{f}}(x^n)) - \log M + \log 2. \tag{33}$$

Proof Define

$$Q(x^n) = \frac{\sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n)}{\sum_{\tilde{x}^n \in \mathcal{X}^n} \sup_{\tilde{\mathsf{f}} \in \mathcal{F}} Q_{\tilde{\mathsf{f}}}(\tilde{x}^n)} \tag{34}$$

and observe that $Q(x^n) \in \mathcal{Q}_{\log M}$. Let g be the strategy associated with Q guaranteed by Lemma 3 and let Q_{g} be the distribution associated with g via Lemma 2. Then

$$\log \frac{Q(x^n)}{Q_{\mathsf{g}}(x^n)} \le \log 2 \tag{35}$$

and

$$\mathcal{V}_n(M, \mathcal{F}) \le \sup_{x^n \in \mathcal{X}^n} \log \frac{\sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n)}{Q_{\mathsf{g}}(x^n)} \tag{36}$$

$$\le \sup_{x^n \in \mathcal{X}^n} \log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n) + \log 2 \tag{37}$$

$$= \log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n) + \log 2. \tag{38}$$

For the lower bound, let g be the strategy that achieves (20). Then

$$\mathcal{V}_n(M, \mathcal{F}) = \sup_{x^n \in \mathcal{X}^n} \log \frac{\sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n)}{Q_{\mathsf{g}}(x^n)} \tag{39}$$

$$\ge \sup_{x^n \in \mathcal{X}^n} \log \frac{\sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n)}{Q(x^n)} = \log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n) \quad \forall x^n \in \mathcal{X}^n \tag{40}$$

where (40) holds since Q is the minimizer of $\inf_{Q \in \mathcal{P}(\mathcal{X}^n)} \sup_{x^n \in \mathcal{X}^n} \log \frac{\sup_{\mathsf{f} \in \mathcal{F}} Q_{\mathsf{f}}(x^n)}{Q(x^n)}$. ∎

Finally, we note that the gap between the lower and upper bounds in Theorem 2 is an artifact of the compression part of sequential prediction with coded side information. Indeed, the Shtarkov bound suffers from a similar gap in the problem of lossless compression with expert advice.

4. Compressor-predictor expert classes

We have established in Lemmas 2 and 3 that compressor-predictor systems are elements of a subset of the probability simplex. From this perspective, much of the literature on sequential prediction with expert advice carries over to the problem of sequential prediction with coded side-information. This includes, for example, the covering number approaches in Cesa-Bianchi and Lugosi (2006, Chapter 9.10) and Rakhlin and Sridharan (2015).

Rather than treating the expert class as an arbitrary subset of the probability simplex, in this section we define classes of experts which are natural from the operational point of view. The first example considers *naive greedy experts*. We obtain these classes of experts by modifying existing sequential experts to use all of their available compression budget on the first k realizations of the sequence x^n. The second example considers *subset-myopic experts*. Each of these experts uses its available compression budget to minimize prediction loss on a particular subset of $\mathcal{X}^n$. Both of these expert classes allow us to study the effect of the compression budget on the prediction loss and regret.

4.1 Naive greedy experts

Given an existing prediction strategy we can extend it to the coded side-information setting in the following way. We define an expert that spends all of its compression budget early on, and then falls back on its original prediction strategy once the budged is exhausted.

Definition 5 (Greedy experts) *Fix $1 \leq k \leq n$ and let $M = |\mathcal{X}|^k$. Given a sequential prediction strategy $\tilde{\mathsf{f}}$ (with no coded side-information), we define a new compressor-predictor strategy f with M messages in the following way:*

- *Let the compressor f_c be a bijection from $\mathcal{X}^k$ to $\{1, \ldots, M\}$ that maps the first k letters of x^n to a unique message.*
- *For $1 \leq t \leq k$, let*

$$\mathsf{f}_t(\cdot|m, x^{t-1}) = \delta_{y_t}, \text{ where } y^k = \mathsf{f}_\mathsf{c}^{-1}(m) \tag{41}$$

  and δ_x denotes a point mass at x.

- *For $k+1 \leq t \leq n$, let*

$$\mathsf{f}_t(\cdot|m, x^{t-1}) = \tilde{\mathsf{f}}_t(\cdot|x^{t-1}). \tag{42}$$

We call f a greedy M message extension of $\tilde{\mathsf{f}}$. Given a class $\mathcal{F}$ of sequential experts (no coded side-information), $\mathcal{F}_M$ is a greedy extension of $\mathcal{F}$ with M messages if it is obtained by greedy extensions of experts in $\mathcal{F}$.

Next, fix a class of experts $\mathcal{F}$ (no coded side-information) and $x^k \in \mathcal{X}^k$. We define a new class of experts, denoted $\mathcal{F}(x^k)$, on $\mathcal{X}^{n-k}$. This class consists of all experts of the form

$$\mathsf{f}_t(\cdot|y^t) = \tilde{\mathsf{f}}_t(\cdot|x^k y^{t-1}) \tag{43}$$

for $1 \leq t \leq n-k$, $y^t \in \mathcal{X}^t$, and $\tilde{\mathsf{f}} \in \mathcal{F}$.

Lemma 4 *Let $\mathcal{F}$ be a class of experts (no coded side-information) and let $M = |\mathcal{X}|^k$ for some $1 \leq k \leq n$. Then*

$$\inf_{x^k \in \mathcal{X}^k} \mathcal{V}_{n-k}\left(1, \mathcal{F}\left(x^k\right)\right) \leq \mathcal{V}_n\left(M, \mathcal{F}_M\right) \leq \sup_{x^k \in \mathcal{X}^k} \mathcal{V}_{n-k}\left(1, \mathcal{F}\left(x^k\right)\right) \tag{44}$$

Proof Let g be the greedy extension of the strategy that achieves $\mathcal{V}_n(1, \mathcal{F})$ and denote by g_{x^k} the strategy that achieves $\mathcal{V}_{n-k}(1, \mathcal{F}(x^k))$. Note that $\mathcal{F}_M$ and $\mathcal{F}(x^k)$ can be connected via

$$\inf_{\mathsf{f} \in \mathcal{F}_M} L_{\mathsf{f}}(x^n) = \inf_{\mathsf{f} \in \mathcal{F}(x^k)} L_{\mathsf{f}}(x_{k+1}^n) \text{ and, likewise, } L_{\mathsf{g}}(x^n) = L_{\mathsf{g}_{x^k}}(x_{k+1}^n). \tag{45}$$

Then

$$\begin{aligned}
\mathcal{V}_n(M, \mathcal{F}_M) &\leq \sup_{x^n \in \mathcal{X}^n} \left\{ L_{\mathsf{g}}(x^n) - \inf_{\mathsf{f} \in \mathcal{F}_M} L_{\mathsf{f}}(x^n) \right\} && (46)\\
&= \sup_{x^k \in \mathcal{X}^k} \sup_{y^{n-k} \in \mathcal{X}^{n-k}} \left\{ L_{\mathsf{g}_{x^k}}(y^{n-k}) - \inf_{\mathsf{f} \in \mathcal{F}(x^k)} L_{\mathsf{f}}(y^{n-k}) \right\} && (47)\\
&= \sup_{x^k \in \mathcal{X}^k} \mathcal{V}_{n-k}(1, \mathcal{F}(x^k)). && (48)
\end{aligned}$$

The lower bound follows since

$$\begin{aligned}
\mathcal{V}_n(M, \mathcal{F}_M) &\geq \log \sum_{x^n \in \mathcal{X}^n} \sup_{\mathsf{f} \in \mathcal{F}_M} \exp(-L_{\mathsf{f}}(x^n)) - \log M && (49)\\
&= \log \sum_{x^k \in \mathcal{X}^k} \sum_{y^{n-k} \in \mathcal{X}^{n-k}} \sup_{\mathsf{f} \in \mathcal{F}(x^k)} \exp(-L_{\mathsf{f}}(y^{n-k})) - \log M && (50)\\
&\geq \inf_{x^k \in \mathcal{X}^k} \log |\mathcal{X}^k| \sum_{y^{n-k} \in \mathcal{X}^{n-k}} \sup_{\mathsf{f} \in \mathcal{F}(x^k)} \exp(-L_{\mathsf{f}}(y^{n-k})) - \log M && (51)\\
&= \inf_{x^k \in \mathcal{X}^k} \log \sum_{y^{n-k} \in \mathcal{X}^{n-k}} \sup_{\mathsf{f} \in \mathcal{F}(x^k)} \exp(-L_{\mathsf{f}}(y^{n-k})) && (52)\\
&= \inf_{x^k \in \mathcal{X}^k} \mathcal{V}_{n-k}(1, \mathcal{F}(x^k)) && (53)
\end{aligned}$$

where (49) follows from Theorem 2 and (50) follows from (45). ∎

The upper and lower bounds in Lemma 4 turn out to be tight for many expert classes.

Example 2 (Static experts) *Let $\mathcal{F}$ be the set of static experts, that is, $\mathcal{F}$ is the set of all experts such that $\mathsf{f}_t(\cdot|x^{t-1}) = P$ for all t and some $P \in \mathcal{P}(\mathcal{X})$. Then, $\mathcal{F}(x^k)$ is still the set of static experts on $\mathcal{X}^{n-k}$ for all $x^k \in \mathcal{X}^k$ and*

$$\mathcal{V}_n(M, \mathcal{F}_M) = \mathcal{V}_{n-k}(1, \mathcal{F}). \tag{54}$$

Let $\mathcal{X} = \{0, 1\}$ and consider rate $R \in [0, 1)$ compressor-predictor schemes: that is, $M_n = 2^{nR}$. Using (54) together with Cesa-Bianchi and Lugosi (2006, Theorem 9.2) we obtain

$$\mathcal{V}_n\left(M_n, \mathcal{F}_{M_n}\right) = \frac{1}{2}\log n + \frac{1}{2}\log(1 - R) + \frac{1}{2}\log\frac{\pi}{2} + o(1). \tag{55}$$

Example 2 demonstrates that the greedy extension of the optimal strategy for static experts is actually the optimal strategy for greedy extension of static experts. Surprisingly, this even holds for classes with memory, as the next example shows.

Example 3 (Markov experts) *Let $\mathcal{F}$ be the set of all Markov experts. We can show that $\mathcal{F}(x^k)$ is still the set of all Markov experts (see Appendix C) on $\mathcal{X}^{n-k}$ for all $x^k \in \mathcal{X}^k$ and*

$$\mathcal{V}_n(M, \mathcal{F}_M) = \mathcal{V}_{n-k}(1, \mathcal{F}). \tag{56}$$

In fact, the upper and lower bounds in Lemma 4 are tight even for some large non-parametric expert classes such as the monotonically increasing "static" experts introduced in Cesa-Bianchi and Lugosi (2006, Chapter 9.11). However, Lemma 4 is not tight for greedy extensions of arbitrary expert families. In those cases, the next lemma would be more useful.

Lemma 5 *Let $\mathcal{F}$ be a class of experts (no coded side-information) and let $M = |\mathcal{X}|^k$ for some $1 \le k \le n$. Then*

$$\log \frac{1}{M} \sum_{x^k \in \mathcal{X}^m} \exp\left(\mathcal{V}_{n-k}\left(1, \mathcal{F}\left(x^k\right)\right)\right) \le \mathcal{V}_n(M, \mathcal{F}_M) \tag{57}$$

$$\le \log \frac{1}{M} \sum_{x^k \in \mathcal{X}^k} \exp\left(\mathcal{V}_{n-k}\left(1, \mathcal{F}\left(x^k\right)\right)\right) + \log 2 \tag{58}$$

Proof The lemma follows by particularizing Theorem 2 to greedy experts. ∎

4.2 Subset-myopic experts

Next, we consider experts defined with respect to some subset $\mathcal{S} \subset \mathcal{X}^n$. These experts design their compressor-predictor schemes with the aim of minimizing the maximum loss on $\mathcal{S}$. They disregard all other sequences $x^n \notin \mathcal{S}$.

Definition 6 (Subset-myopic experts) *Given a subset $\mathcal{S} \subset \mathcal{X}^n$ an $\mathcal{S}$-myopic expert f with $M \le |\mathcal{S}|$ messages is given by*

$$\mathsf{f} = \operatorname*{arg\,min}_{\mathsf{f}\colon |\mathsf{f_c}| \le M} \sup_{x^n, y^n \in \mathcal{S}} \left| L_{\mathsf{f}}(x^n) - L_{\mathsf{f}}(y^n) \right|. \tag{59}$$

If $M > |\mathcal{S}|$ an $\mathcal{S}$-myopic expert f is given by

$$\mathsf{f} = \operatorname*{arg\,min}_{\mathsf{f}\colon \substack{|\mathsf{f_c}| \le M, \\ \sum_{x^n \in \mathcal{S}} L_{\mathsf{f}}(x^n) = 0}} \sup_{y^n \notin \mathcal{S}} L_{\mathsf{f}}(y^n). \tag{60}$$

Example 4 (Constant composition subsets) *Let $\mathcal{X} = \{0, 1\}$ and $\mathcal{F}_M$ be a class of $\mathcal{S}$-myopic experts with M messages defined by a collection $\{\mathcal{S}_k\}_{k=0}^n$ where*

$$\mathcal{S}_k = \{x^n \colon x^n \text{ has exactly } k \text{ ones }\}. \tag{61}$$

Consider rate $R \in [0, 1)$ compressor-predictor schemes: that is, $M_n = 2^{nR}$. Then using Theorem 2 together with Stirling's approximation we obtain

$$\mathcal{V}_n(M_n, \mathcal{F}) = \log n + \log(1 - 2h^{-1}(R)) + o(1) \tag{62}$$

where h^{-1} denotes the inverse of the binary entropy function on $(0, \frac{1}{2}]$.

5. Discussion

We have developed a framework to address sequential prediction with coded side information under log-loss. There are two main insights that lay the groundwork for our results. First, we show an operational equivalence between the problem of universal lossy compression and the problem of sequential prediction with coded side information. Secondly, we leverage this equivalence, together with insights from Shkel et al. (2017), to connect prediction strategies to distributions on a subset of the probability simplex. Establishing this connection between compressor-predictor schemes and probability distributions lets us transplant the tools from sequential prediction with no side information to the setting with coded side information.

In the present paper we made the assumption that the compressor is "batch"; that is, the compressor sees the whole sequence in a non-causal manner and sends a compressed version of it to the predictor at time zero. In some settings, it is certainly more natural to assume that the compressor might see ahead for some fixed number of time steps and has a compression budget at each time t; in other words, the coded side information is constrained to be sequential. The problem of prediction with sequential side information is equivalent to a corresponding sequential lossy compression problem. Moreover, it is straightforward to show a counterpart of Lemma 2 for this setting: the caveat is that now the subset of the probability simplex would have additional constraints which would depend on the exact setup of the sequential compression aspect of the problem, see Appendix D. The main difficulty in extending the present work to the sequential side information setting is to show a counterpart of Lemma 3: that is, we need to demonstrate a compression scheme for a given Q that allows us to get close to the bound in Lemma 2. For example, it is possible to chain together such a scheme with a repeated application of Lemma 3, but this incurs a penalty of $\log 2$ at each time step. A more promising direction is to extend the arithmetic coding scheme, Rissanen (1976), to the lossy compression (or equivalently compressor-predictor) setting, and this is the focus of our ongoing work.

Finally, we mention another important extension of the current work in which the compressor does not see the sequence to be predicted, but instead sees a correlated sequence. It is again possible to establish an equivalence between this compression-prediction problem and a lossy compression problem. The resulting noisy lossy compression problem has received much attention in the probabilistic setting, see, for example, Nazer et al. (2017) and references therein. A sequential version of the noisy compression problem could further be related to online learning with log-loss studied in Fogel and Feder (2017). It is possible that the tools developed for lossy compression with log-loss could be leveraged for this learning problem as well. Conversely, it would be interesting to see if the tools from learning theory could be used to get further insight for sequential compression and noisy compression with log-loss.

Acknowledgments

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

Appendix A. Regularity condition (22) for compressor-predictor systems

In this section we justify (22) by demonstrating that if f is not regular, it is always possible to construct an f^* such that

$$L_{\mathsf{f}}(x^n) \geq L_{\mathsf{f}^*}(x^n) \quad \forall x^n \in \mathcal{X}^n. \tag{63}$$

Indeed, let $(\mathsf{f}_{\mathsf{c}}, \mathsf{f}_{\mathsf{d}})$ be the lossy compressor that corresponds to f via Lemma 1. We define $(\mathsf{f}^*_{\mathsf{c}}, \mathsf{f}^*_{\mathsf{d}})$ by

$$\mathsf{f}^*_{\mathsf{c}}(x^n) = \mathsf{f}_{\mathsf{c}}(x^n), \tag{64}$$

$$\mathsf{f}^*_{\mathsf{d}}(u) = P^*_u \text{ where } \hat{P}_u = \mathsf{f}_{\mathsf{d}}(u),$$

$$P^*_u(x^n) = \begin{cases} 0, & \mathsf{f}_{\mathsf{d}}(x^n) \neq u, \\ \frac{\hat{P}_u(x^n)}{\sum_{\tilde{x}^n \colon \mathsf{f}_{\mathsf{d}}(\tilde{x}^n)=u} \hat{P}_u(\tilde{x}^n)}, & \mathsf{f}_{\mathsf{d}}(x^n) = u \text{ and} \\ & \sum_{\tilde{x} \colon \mathsf{f}_{\mathsf{d}}(\tilde{x}^n)=u} \hat{P}_u(\tilde{x}^n) > 0, \\ \tilde{P}_u(x^n), & \text{o.w.} \end{cases} \tag{65}$$

and $\tilde{P}_u$ is an arbitrary distribution supported on $\{x^n \colon \mathsf{f}_{\mathsf{d}}(x^n) = u\}$. Note that by construction

$$P^*_{\mathsf{f}^*_{\mathsf{d}}(x^n)}(x^n) \geq \hat{P}_{\mathsf{f}_{\mathsf{d}}(x^n)}(x^n) \, \forall x^n \in \mathcal{X}^n \tag{66}$$

which shows (63). Finally, we let f^* be the compressor-predictor system that corresponds to $(\mathsf{f}^*_{\mathsf{c}}, \mathsf{f}^*_{\mathsf{d}})$ via Lemma 1.

Appendix B. Proof of Lemma 3

Given a $Q \in \mathcal{Q}_{\log M}$ we construct a lossy compressor $(\mathsf{f}_{\mathsf{c}}, \mathsf{f}_{\mathsf{d}})$ that satisfies

$$\mathsf{d}\left(x^n, \mathsf{f}_{\mathsf{d}}(\mathsf{f}_{\mathsf{c}}(x^n))\right) \leq \log \frac{1}{Q(x^n)} - \log(M+1) + \log 2 \tag{67}$$

for all $x^n \in \mathcal{X}^n$. Moreover, if either $M = 1$ or $M = |\mathcal{X}^n|$, then the lossy compressor satisfies

$$\mathsf{d}\left(x^n, \mathsf{f}_{\mathsf{d}}(\mathsf{f}_{\mathsf{c}}(x^n))\right) = \log \frac{1}{Q(x^n)} - \log M. \tag{68}$$

Assume $M = |\mathcal{X}^n|$ and observe that, in this case, the only distribution belonging to $\mathcal{Q}_{\log M}$ is the uniform distribution over $\mathcal{X}^n$. Indeed, from (14) we have

$$Q(x^n) \leq \frac{1}{|\mathcal{X}^n|}, \quad \forall x^n \in \mathcal{X}^n \tag{69}$$

which implies

$$Q(x^n) = \frac{1}{|\mathcal{X}^n|}, \quad \forall x^n \in \mathcal{X}^n. \tag{70}$$

Let the compressor $\mathsf{f_c}$ be a bijection between $\mathcal{X}^n$ and $\{1,\dots,M\}$. Let the decompressor be $\mathsf{f_d}(m)=\delta_{\mathsf{f}^{-1}(m)}$, where δ_{x^n} denotes the point mass at x^n. Then, for every $x^n\in\mathcal{X}^n$,

$$\mathsf{d}(x^n,\mathsf{f_d}(\mathsf{f_c}(x^n)))=\log\frac{1}{\delta_{x^n}(x^n)}=0 \tag{71}$$

which is also the right hand side of (68).

Assume $M=1$. Let $\mathsf{f_c}(x^n)=1$ for all $x^n\in\mathcal{X}^n$ and $\mathsf{f_d}(1)=Q$. Then, for every $x^n\in\mathcal{X}^n$,

$$\mathsf{d}(x^n,\mathsf{f_d}(\mathsf{f_c}(x^n)))=\log\frac{1}{Q(a)} \tag{72}$$

which is also the right hand side of (68).

For $1<M<|\mathcal{X}^n|$ the result is proved by means of the following greedy construction.

Compressor: Fix a distribution $Q\in\mathcal{Q}_{\log M}$. Without loss of generality assume that the elements of $\mathcal{X}^n$ are sorted according to $Q(x^n)$. The compressor is defined sequentially in $|\mathcal{X}^n|$ steps. At the first M steps, $\mathsf{f}(a)=a$. At steps $a=\{M+1,\dots,|\mathcal{X}^n|\}$ assign

$$\mathsf{f_c}(a)=\operatorname*{arg\,min}_{m\in\{1,\dots,M\}}\sum_{b=1}^{a-1}Q(b)1\{\mathsf{f_c}(b)=m\}. \tag{73}$$

In other words, the sequence a is encoded to a message that has accrued the smallest total probability so far according to Q. It is important to note that by construction, at every intermediate step there is at least one message with accumulated probability strictly less than $\frac{1}{M}$. At the end, this remains true unless all M messages are equiprobable.

Decompressor: The decompressor is defined by $\mathsf{f_d}(m)=\hat{P}_m$, where

$$\hat{P}_m(a)=\begin{cases}\frac{Q(a)}{\mathbb{P}_Q[\mathsf{f_c}(X^n)=m]}, & \mathsf{f_c}(a)=m\\ 0, & \text{otherwise}\end{cases} \tag{74}$$

Note that the decompressor assigns $\hat{P}_m(a)$ to be $\mathbb{P}_Q[X^n|\mathsf{f_c}(X^n)=m]$, the posterior distribution of X^n given the message m.

Distortion analysis: Let $a\in\mathcal{X}^n$ be the last element in $\mathcal{X}^n$ assigned to m. If $a\le M$ (that is, a is one of the M most likely elements in $\mathcal{X}^n$) then

$$\mathbb{P}_Q[\mathsf{f_c}(X^n)=m]=Q(a)\le\frac{1}{M}. \tag{75}$$

Otherwise,

$$\mathbb{P}_Q[\mathsf{f_c}(X^n)=m]\le\frac{\sum_{b=1}^{a-1}Q(b)}{M}+Q(a)\le\frac{1-Q(a)}{M}+Q(a)\le\frac{1+(M-1)Q(a)}{M}. \tag{76}$$

Since $Q(a)\le\frac{1}{M+1}$ we obtain

$$\mathbb{P}_Q[\mathsf{f_c}(X^n)=m]\le\frac{2}{M+1}. \tag{77}$$

Thus

$$\hat{P}_{\mathsf{f_c}(x^n)}(x^n)\le Q(x^n)\frac{M+1}{2} \tag{78}$$

and (67) holds for all $x^n\in\mathcal{X}^n$.

Finally, Lemma 3 holds by letting f be a compressor-predictor system that corresponds to the lossy compressor $(\mathsf{f_c},\mathsf{f_d})$ via Lemma 1.

Appendix C. Markov example

To fix ideas we assume that the Markov predictor initializes by appending zeros to the beginning of the sequence. Let us denote $\mathcal{F}_0 = \mathcal{F}(x^k)$ whenever $x_k = 0$ and $\mathcal{F}_1 = \mathcal{F}(x^k)$ whenever $x_k = 1$. Then $\mathcal{F}_0$ is a class of Markov predictors on $\mathcal{X}^{n-k}$ and $\mathcal{F}_1$ is a class of Markov predictors on $\mathcal{X}^{n-k}$ that initializes by appending ones to the beginning of the sequence. We can show that

$$\mathcal{V}_{n-k}(1, \mathcal{F}_0) = \mathcal{V}_{n-k}(1, \mathcal{F}_1) \tag{79}$$

and this implies (56). Indeed, consider $\mathsf{f} \in \mathcal{F}_0$ and $a^{n-k} \in \mathcal{X}^{n-k}$. Let the predictor $\mathsf{g} \in \mathcal{F}_1$ be given by

$$\mathsf{g}_t(\cdot|0) = 1 - \mathsf{f}_t(\cdot|0), \quad \mathsf{g}_t(\cdot|1) = 1 - \mathsf{f}_t(\cdot|1) \tag{80}$$

and a sequence $b^{n-k} \in \mathcal{X}^{n-k}$ be given by

$$a_t = \mathbf{1}\{b_t = 0\}. \tag{81}$$

Then

$$L_{\mathsf{f}}(a^{n-k}) = L_{\mathsf{g}}(b^{n-k}) \tag{82}$$

which, together with the Shtarkov bound (21), implies (79).

Appendix D. Sequential side information

There are many variations of a sequential compression-prediction problem: these variations include message budget constraints, number of look-ahead time steps and so on. Here, we present an example of how the batch side information results can be generalized to the sequential side information for one such variation.

Definition 7 *A compressor-predictor system with $\{M_t\}_{t=1}^n$-message budget is a collection of mappings:*

$$\textit{Compressor:} \quad \mathsf{f}_t : \mathcal{X}^t \to \{1, \dots, M_t\}, \tag{83}$$

$$\textit{Predictor:} \quad \mathsf{g}_t : \{1, \dots, M_t\} \times \mathcal{X}^{t-1} \to \mathcal{P}(\mathcal{X}), \quad 1 \le t \le n. \tag{84}$$

At each time step t the predictor incurs a loss

$$\ell\left(x_t, \mathsf{g}_t(\cdot|\mathsf{f}_t(x_t), x^{t-1})\right) = \log \frac{1}{\mathsf{g}_t\left(x_t|\mathsf{f}_t(x_t), x^{t-1}\right)}. \tag{85}$$

The cumulative loss incurred by a compressor-predictor system (f, g) for a sequence x^n is

$$L_{\mathsf{f}}(x^n) = \sum_{t=1}^{n} \ell\left(x_t, \mathsf{g}_t\left(\cdot|m_t, x^{t-1}\right)\right) \tag{86}$$

where $m_t = \mathsf{f}_t(x_t)$.

To show a counterpart of Lemma 2 define

$$\mathcal{Q}_{R_1,\ldots,R_n} = \{Q\colon Q_t(\cdot|x^{t-1}) \in \mathcal{Q}_{R_t} \subset \mathcal{P}(\mathcal{X}), \quad \forall x^{t-1} \in \mathcal{X}^{t-1} \text{ and } 1 \le t \le n\}. \tag{87}$$

We can state a regularity condition on a compressor-predictor system f in the spirit of (22). We assume that the compressor-predictor system satisfies

$$\mathsf{g}_t(x_t|m, x^{t-1}) = 0, \quad \forall m \neq \mathsf{g}_t(x_t) \tag{88}$$

for all $x_t \in \mathcal{X}$.

Lemma 6 *Given a compressor-predictor system (f, g) with $\{M_t\}_{t=1}^n$-message budget, there exists $Q \in \mathcal{Q}_{\log M_1,\ldots,\log M_n}$ such that*

$$L_{(\mathsf{f},\mathsf{g})}(x^n) = \log \frac{1}{Q(x^n)} - \sum_{t=1}^{n} \log M_t \tag{89}$$

holds for all $x^n \in \mathcal{X}^n$.

Proof Given a compressor-predictor system (f, g), define

$$Q(x_t|x^{t-1}) = \frac{1}{M_t} \sum_{m=1}^{M_t} \mathsf{g}_t\left(x_t|m, x^{t-1}\right) = \frac{1}{M_t} \mathsf{g}_t\left(x_t|\mathsf{f}_t(x_t), x^{t-1}\right) \tag{90}$$

where the second equality follows from the regularity assumption and it is straightforward to verify that $Q(\cdot|x^{t-1})$ is indeed a distribution in $\mathcal{Q}_{\log M_t}$. Then

$$\ell\left(x_t, \mathsf{g}_t\left(\cdot|\mathsf{f}_t(x_t), x^{t-1}\right)\right) = \log \frac{1}{Q(x_t|x^{t-1})} - \log M_t. \tag{91}$$

The result follows by taking

$$Q(x^n) = \prod_{t=1}^{n} Q(x_t|x^{t-1}) \tag{92}$$

and summing the instantaneous loss. ■

Similarly to the batch side information case, we can show that $\mathcal{Q}_{R_1,\ldots,R_n}$ is convex and derive the lower bound in Theorem 2.

References

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Y. Fogel and M. Feder. On the problem of on-line learning with log-loss. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2995–2999, June 2017.

P. D. Grünwald. *The Minimum Description Length Principle*. Cambridge, Mass. : MIT Press, 2007.

N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, Oct 1998.

B. Nazer, O. Ordentlich, and Y. Polyanskiy. Information-distilling quantizers. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 96–100, June 2017.

A. Rakhlin and K. Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. preprint, 2015.

J. Rissanen. Generalized Kraft's inequality and arithmetic coding. *IBM J. Res. Develop.*, 20(3):198–203, 1976.

Y. Shkel, M. Raginsky, and S. Verdú. Universal lossy compression under logarithmic loss. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1157–1161, June 2017.

Y. Y. Shkel and S. Verdú. A single-shot approach to lossy source coding under logarithmic loss. *IEEE Transactions on Information Theory*, 64(1):129–147, Jan 2018.

Y. M. Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.