6.7800 Spring 2025: Bonus Lecture 04

Universal RoShamBo Machine via Universal Probability

Jongha (Jon) Ryu
jongha@mit.edu

May 2, 2025

Contents

1 Introduction

You may remember that, in the very beginning of the class, Prof. Wornell showed Emin Martinian's machine[1]. If you tried the game yourself, you would have found that it is quite hard to beat the machine! As we are now equipped with the notion of "universal prediction", we are ready to peek and understand the behind the scene of Emin's RoShamBo machine.

In this session, we will introduce how a universal prediction scheme can be turned into a sequential decision maker for e.g., RoShamBo machine, along with a guarantee. We will then learn a special universal prediction scheme deployed in Emin's RoShamBo machine, which was

[1] https://web.mit.edu/6.7800/www/rps.html

originally developed for universal compression by Lempel and Ziv in 1978. This will demonstrate how information theory, especially compression side of story, is tightly connected to the theory of online learning.

For what is to be developed, we extend the definition of universal prediction beyond i.i.d. processes. Consider observations $x_1, \dots, x_n$ generated from a data generating distribution $p(x^n)$. Recall the definition of universal prediction from the lecture:

Definition 1. Let $p(x^n) \in \mathscr{P}$ be a distribution for data x^n, where $\mathscr{P}$ is a class of distributions. A distribution $q(y^n)$ is said to be *universal* if, as $n \to \infty$,

$$\max_{p \in \mathscr{P}} D(p(x^n) \parallel q(x^n)) = o(n). \tag{1}$$

2 Bayesian Decision Theory

In the standard decision theoretic setup, we consider an action space $\mathcal{A}$, outcome space $\mathcal{X}$, and loss function $\ell(a, x) \colon \mathcal{A} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. The player chooses an action $a \in \mathcal{A}$, the nature reveals an outcome $x \in \mathcal{X}$, and the player suffers loss of $\ell(a, x)$.

Example 2 (M-ary prediction)**.** *Set $\mathcal{A} = \mathcal{X} = \{1, \dots, M\}$ and $\ell(a, x) = \mathbb{1}\{a \neq x\}$.*

Example 3 (RoShamBo machine)**.** *Set $\mathcal{A} = \mathcal{X} = \{R, P, S\}$ and $\ell(a, x)$ as in Table 1.*

$\mathcal{A} \setminus \mathcal{X}$	R	P	S
R	0	+1	-1
P	-1	0	+1
S	+1	-1	0

Table 1: Loss function for RoShamBo.

Example 4 (Probability assignment)**.** *Set $\mathcal{A} = \mathcal{P}^{\mathcal{X}}$ for some finite alphabet $\mathcal{X}$ and $\ell(a, x) = \log \frac{1}{a(x)}$. Note that all the development with this setting (i.e., under log loss) recovers the universal prediction setting we study in the lectures.*

2.1 Bayes Loss and Bayes Action

If we assume that $\mathsf{x} \sim p(x)$, it is reasonable to consider the *expected* loss $\mathbb{E}_{\mathsf{x} \sim p(x)}[\ell(a, \mathsf{x})]$ of an action a with respect to p, as a measure of goodness of actions. We define the best action under this criterion

$$a^*_{\mathsf{B}}(p) := \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\mathsf{x} \sim p(x)}[\ell(a, \mathsf{x})]$$

and call it the *Bayes action*. The minimum expected loss

$$L^*_{\mathsf{B}}(p) := \mathbb{E}_{\mathsf{x} \sim p(x)}[\ell(a^*_{\mathsf{B}}(p), \mathsf{x})]$$

is called the *Bayes loss*. We remark that as a function of p, $L^*_{\mathsf{B}}(p)$ is concave.

Example 5 (M-ary prediction)**.**

$$a^*_{\mathsf{B}}(p) = \arg\max_{x \in \mathcal{X}} p(x),$$
$$L^*_{\mathsf{B}}(p) = 1 - \max_{x \in \mathcal{X}} p(x).$$

Example 6 (Probability assignment)**.**

$$a^*_{\mathsf{B}}(p) = p,$$
$$L^*_{\mathsf{B}}(p) = H(p).$$

2.2 Regret and Generalized Divergence

If we were given the underlying distribution $p(x)$ of nature's outcome, we could play the Bayes action $a^*_{\mathsf{B}}(p)$. In practice, however, most likely we do not know which distribution generates the nature's outcome. If we play some action b, we define the *regret* of playing b compared to the best action as

$$R(b, p) := \mathbb{E}_{\mathsf{x} \sim p(x)}[\ell(b, \mathsf{x}) - \ell(a^*_{\mathsf{B}}(p), \mathsf{x})].$$

In particular, if we played the Bayes action with respect to some other distribution q against p is

$$\Delta(p \parallel q) := \mathbb{E}_{\mathsf{x} \sim p(x)}[\ell(a^*_{\mathsf{B}}(q), \mathsf{x}) - \ell(a^*_{\mathsf{B}}(p), \mathsf{x})] \; (= R(a^*_{\mathsf{B}}(q), p)).$$

2.2.1 Minimax Regret

In practice, we can assume that the distribution p belongs to a certain class of distributions $\mathcal{P}$, which we call the *model class* in the lectures. The minimax regret in this case is

$$\min_{b \in \mathcal{A}} \max_{p \in \mathcal{P}} R(b, p).$$

Under some regularity conditions, for any $b \in \mathcal{A}$, there exists a distribution q such that $b = a^*_{\mathsf{B}}(q)$, and thus we can write

$$R(b, p) = R(a^*_{\mathsf{B}}(q), p) = \Delta(p \parallel q).$$

One sufficient condition is the following, which holds for all the examples we consider in this note.

Assumption 7. $|\mathcal{X}| < \infty$ and $a \mapsto \ell(a, x)$ is convex for any $x \in \mathcal{X}$.

Under Assumption 7, we can thus rewrite the minimax regret as follows:

$$\begin{aligned} \min_{b \in \mathcal{A}} \max_{p \in \mathcal{P}} R(b, p) &= \min_{q \in \mathcal{P}^{\mathcal{X}}} \max_{p \in \mathcal{P}} \Delta(p \parallel q) \\ &= \min_{q \in \mathcal{P}^{\mathcal{X}}} \max_{w \in \mathcal{P}^{\mathcal{P}}} \mathbb{E}_{w(p)}[\Delta(\mathsf{p} \parallel q)]. \end{aligned} \tag{2}$$

Here, $w(p)$ is a weight distribution over the model class $\mathcal{P}$. Further, if the min and max can be swapped (which needs to be proved), we finally have

$$\min_{b \in \mathcal{A}} \max_{p \in \mathcal{P}} R(b, p) = \max_{w \in \mathcal{P}^{\mathcal{P}}} \min_{q \in \mathcal{P}^{\mathcal{X}}} \mathbb{E}_{\mathsf{p} \sim w(p)}[\Delta(\mathsf{p} \parallel q)].$$

Provided that this analysis holds under certain conditions, the inner term $\min_{q \in \mathcal{P}^{\mathcal{X}}} \mathbb{E}_{\mathsf{p} \sim w(p)}[\Delta(\mathsf{p} \parallel q)]$ can be viewed as a generalized mutual information, and the minimax value can be understood as a generalized capacity. In this case, we could have developed a tool for bounding the generalized capacity as we studied universal prediction (under log loss).

2.2.2 Bounding Minimax Regret via Universal Predictor

For a bounded loss function, however, we may not need to develop a separate theory! That is, we can reuse the toolkit developed for universal prediction (under log loss) to come up with a universal decision maker. The key observation is the following lemma, which is a simple consequence of Pinsker's inequality.

Lemma 8. *Let* $\ell_{\max} := \max_{a,x} \ell(a, x)$. *Then,*

$$\Delta(p \parallel q) \le 2\ell_{\max} \sqrt{2D(p \parallel q)}.$$

The proof can be found in Appendix. As an immediate corollary of this lemma and (2), we have:

Corollary 9. *If Assumption 7 holds and* $\ell_{\max} < \infty$, *then*

$$\min_{b \in \mathcal{A}} \max_{p \in \mathcal{P}} R(b, p) = \min_{q} \max_{p \in \mathcal{P}} \Delta(p \parallel q) \le 2\sqrt{2}\ell_{\max} \sqrt{\min_{q} \max_{p \in \mathcal{P}} D(p \parallel q)}.$$

This implies that if q is universal with respect to $\mathcal{P}$, then we can play the Bayes action according to q to perform universally well in the decision theoretic setup.

2.3 Sequential Decision Making

So far, we assumed that the game between player and nature is single-round. We need to extend this to a multi-round setting to come up with a sequential decision making scheme.

We now assume that the nature generates an outcome sequence $x_1, \ldots, x_n$ according to a certain distribution $p(x^n)$. Our loss function for the multi-round setting is the *cumulative* loss

$$\ell(a^n, x^n) := \sum_{i=1}^{n} \ell(a_i, x_i),$$

Note that in $\ell(a^n, x^n)$, the action a is an abstract notation for a strategy as a sequence of functions. That is, for each round i, a_i is indeed a function of history x^{i-1}, i.e., $a_i = a_i(x^{i-1})$.

Under this stochastic assumption, the best possible strategy is the Bayes action $a^n = a^*_{\mathsf{B}}(p(x^n))$ and the minimum loss achieved is $L^*_{\mathsf{B}}(\mathsf{x}^n) := L^*_{\mathsf{B}}(p(x^n))$.[2] If the distribution $p(x^n)$ is known, the action at round i induced by the Bayes action $a^*_{\mathsf{B}}(p(x^n))$ is

$$(a^*_{\mathsf{B}})_i(x^{i-1}) := a^*_{\mathsf{B}}(p(x_i|x^{i-1})) = \arg\min_{a_i \in \mathcal{A}} \mathbb{E}_{p(x_i|x^{i-1})}[\ell(a_i, \mathsf{x}_i)].$$

[2]This notation is analogous to another conventional notation of entropy $H(\mathsf{x})$ for $\mathsf{x} \sim p(x)$.

One special case is when the underlying distribution is i.i.d., i.e., $p(x^n) = p(x_1) \cdots p(x_n)$, in which case the Bayes loss is given as $L^*_{\mathsf{B}}(\mathsf{x}^n) = nL^*_{\mathsf{B}}(\mathsf{x}_1)$.

As we developed in Section 2.2, under the same condition as in Corollary 9, the minimax cumulative regret can be written as

$$\rho^*_n := \min_{b^n} \max_{p \in \mathscr{P}} R(b^n, p(x^n)) = \min_{q \in \mathcal{P}^{\mathcal{X}}} \max_{p \in \mathscr{P}} \Delta(p(x^n) \parallel q(x^n)). \tag{3}$$

To conclude that we can plug-in a universal predictor (or universal probability assignment) $q(x^n)$ with respect to the model class $\mathscr{P}$ to bound the minimax regret for a sequential decision problem with bounded loss, we need a similar argument as in Lemma 8. We remark that applying Lemma 8 directly by viewing the cumulative loss $\ell(a^n, x^n)$ as a single loss function leads to vacuous bound, since we can only guarantee that $\sup_{a^n, x^n} \ell(a^n, x^n) \le n\ell_{\max}$, which leads to

$$\frac{1}{n}\Delta(p(x^n) \parallel q(x^n)) \le 2\sqrt{2}\ell_{\max}\sqrt{D(p(x^n) \parallel q(x^n))}.$$

Indeed, we can prove a much tighter bound, whose proof is deferred to Appendix.

Lemma 10.

$$\frac{1}{n}\Delta(p(x^n) \parallel q(x^n)) \le 2\sqrt{2}\ell_{\max}\sqrt{\frac{1}{n}D(p(x^n) \parallel q(x^n))}.$$

As a corollary, we have the following theorem:

Theorem 11. *Suppose that Assumption 7 holds and $\ell_{\max} < \infty$. If the model class $\mathscr{P}$ admits a universal predictor under log loss, then it also admits one under $\ell(a, x)$. In particular, if $q(x^n)$ is universal with respect to $\mathscr{P}$, then $(a^*_{\mathsf{B}}(q(x_i|x^{i-1})))_{i=1}^n$ guarantees a diminishing worst case regret against $\mathscr{P}$ as $n \to \infty$.*

3 Examples of Universal Predictors (Under Log Loss)

So far, we learned that a universal predictor (under log loss) can be used as a surrogate for a true data generating distribution, to play the Bayes action with respect to it. This is a simple and elegant plug-and-play strategy for sequential decision making. We now turn our attention to the particular universal strategy used by Emin's RoShamBo machine.

Note that the model class $\mathscr{P}$ is a design choice from the designer (sequential decision maker), capturing the *uncertainty* in the data generating process. On one extreme, if the designer simply assumes that the true data distribution is an i.i.d. categorical process with unknown parameter, then we know that the KT mixture is the minimax optimal universal predictor with convergence rate of the normalized cumulative regret $O(\frac{\log n}{n})$. The downside of this simple model class is that if the true distribution does not belong to the small class, there is a risk that the resulting worst-case regret may not vanish, as we learned in the lectures. On the other extreme, the particular strategy used Emin's RoShamBo machine is a universal predictor with respect to a class of ergodic stationary processes, which is really huge class of distributions. While such a wide coverage is attractive, the cost comes in an extremely slow rate $O(\frac{\log \log n}{\log n})$, as we will see.

To motivate the universal procedure for stationary processes, we will first consider the i.i.d. processes and stationary Markov processes.

3.1 Simpler Cases

We revisit the simplest case of i.i.d. sequences.

Example 12 (i.i.d. processes)**.** *Suppose that* $\mathsf{x}_1, \dots, \mathsf{x}_N \sim$ *i.i.d.* $p(x;\theta) = \theta^x(1-\theta)^{1-x}$ *for* $x \in \{0,1\}$*, where* $\theta \in [0,1]$*. As we do not know the underlying* θ *that generates the data, the idea is to take a* mixture *with respect to :*

$$q_w(x^n) := \int_0^1 p_\theta(x^n) w(\theta)\,\mathrm{d}\theta.$$

For exponential family models as in this case, we know that Jeffreys' prior is the optimal weight distribution, which is in this particular case the Beta distribution with parameters $(\frac{1}{2}, \frac{1}{2})$*. The resulting mixture is specifically called the Krichevsky–Trofimov mixture*

$$q_{\mathsf{KT}}(x^n) := \int_0^1 p_\theta(x^n)\mathsf{Beta}\Big(\theta; \frac{1}{2}, \frac{1}{2}\Big)\,\mathrm{d}\theta = B\Big(\sum_{i=1}^n x_i + \frac{1}{2}, n - \sum_{i=1}^n x_i + \frac{1}{2}\Big),$$

and the sequential distribution is

$$q_{\mathsf{KT}}(x_i = 1 | x^{i-1}) = \frac{q_{\mathsf{KT}}(x^{i-1}1)}{q_{\mathsf{KT}}(x^{i-1})} = \frac{\sum_{i=1}^n x_i + \frac{1}{2}}{n+1}.$$

We know that the KT mixture satisfies

$$\max_{\theta\in[0,1]} D(p(x^n;\theta) \parallel q(x^n)) = \frac{1}{2}\log n + O(1).$$

A slightly more advanced example is a stationary Markov process.

Example 13 (Stationary Markov processes)**.** *Suppose now that* $\mathsf{x}_1, \dots, \mathsf{x}_n | \mathsf{x}_0 \sim p(x^n | x_0; \theta_0, \theta_1)$*, where the distribution factorizes as*

$$p(x^n | x_0; \theta_0, \theta_1) := \prod_{i=1}^n p(x_i | x_{i-1}; \theta_0, \theta_1) = \prod_{i=1}^n p(x_i; \theta_{x_{i-1}}).$$

That is, the symbol x_i *at time* i *is generated from a Bernoulli distribution with parameter* $\theta_{x_{i-1}}$*. The process is thus fully characterized by two unknown parameters* θ_0 *and* θ_1*. We can apply the same idea of mixture for i.i.d. processes by* parsing *the sequence based on the previous symbols into two parts:*

$$[n] := \{1, \dots, n\} = \bigsqcup_{y\in\{0,1\}} I_y(x^n),$$

where, for $y \in \{0,1\}$*,*

$$I_y(x^n) := \{i \in [n] \colon x_{i-1} = y\}.$$

Since $(x_i)_{i\in I_y(x^n)}$ *is an i.i.d. process with parameter* θ_y*, we can apply the KT mixture on each sequence separately, and the resulting KT probability is*

$$q_{\mathsf{KT}}(x^n | x_0) := q_{\mathsf{KT}}((x_i)_{i\in I_0(x^n)}) q_{\mathsf{KT}}((x_i)_{i\in I_1(x^n)}),$$

which guarantees

$$\max_{\theta_0,\theta_1\in[0,1]} D(p(x^n|x_0;\theta_0,\theta_1) \parallel q_{\mathsf{KT}}(x^n|x_0)) = 2\cdot\frac{1}{2}\log n + O(1).$$

This can be extended to stationary k-th order Markov processes in a straightforward manner, and the resulting minimax regret would scale as

$$\max_{\theta_s\in[0,1],s\in\{0,1\}^k} D(p(x^n|x_{-k-1}^0;\{\theta_s : s\in\{0,1\}^k) \parallel q_{\mathsf{KT}}(x^n|x_{-k-1}^0)) = 2^k\cdot\frac{1}{2}\log n + O(1).$$

The scaling behavior 2^k is intuitive as we manage the sequence with 2^k separate states independently. If the alphabet size is m, the regret behaves as $\frac{m^k(m-1)}{2}\log n + O(1)$.

3.2 Lempel–Ziv Incremental Parsing

We now introduce the Lempel–Ziv incremental parsing (LZ parsing) to deal with stationary processes. As we do not know the *order* of the underlying process unlike the k-th order Markov processes, we need an *adaptive* parsing as we observe more symbols. The LZ parsing incrementally parses the sequences so that *the newly parsed phrase has never been observed in the past*. For example, a binary sequence $x^n = 0000000000$ is parsed as $0, 00, 000, 0000$. Another example is $x^n = \texttt{ABRACADABRA}$, which is parsed into `A,B,R,AC,AD,AB,RA`.

How can we assign probability via this procedure? The basic idea of the LZ scheme is to assign equal probability over all possible new phrases given the observations. More precisely, suppose that we have observed c *complete* phrases $y^c (= x^n)$, and let $P(y^c) :=$ (all possible new phrases given y^c). Then the LZ scheme assigns probability

$$q_{\mathsf{LZ}}(y_{c+1}|y^c) = \frac{1}{|P(y^c)|} \quad \text{for } y_{c+1}\in P(y^c).$$

For example, if $x^n = 000000$ as before, then we have $c = 3$ complete phrases, and $P(y^3) = \{1, 01, 001, 0000, 0001\}$,

$$q_{\mathsf{LZ}}(y_4|y^3) = \frac{1}{5} \quad \text{for } y_3\in P(y^c) = \{1, 01, 001, 0000, 0001\}.$$

In general, by an inductive argument, we can easily show that

$$|P(y^c)| = (m-1)c + m,$$

for m-ary processes. Therefore, for $x^n = y^c$ with c complete phrases, the probability assigned is

$$q_{\mathsf{LZ}}(y^c) = \prod_{i=1}^{c}\frac{1}{(m-1)i+m}.$$

Given complete phrases, how should we assign probability over the next *symbol*? You can add the probabilities of all the phrases in $P(y^c)$ that start with the symbol. It can be done similarly for a

sequence with incomplete phrase. Suppose that a given sequence is of the form $y^c z$, where z is a phrase that does not belong to $P(y^c)$. Then, we can define a new set

$$P(y^c z) := \{y \in P(y^c) \colon y \text{ starts with } z\}.$$

Then, given the history $y^c z$, the next symbol probability is

$$q_{\mathsf{LZ}}(x|y^c z) = \frac{1}{|P(y^c z)|} \quad \text{for } x \text{ such that } zx \in P(y^c z).$$

It is easy to show that this procedure is consistent with the probability assignment over complete phrases.

We now conclude the session with the universality guarantee of the LZ process. Let

$$\mathcal{P}_k := (\text{the class of stationary } k\text{-th order binary Markov processes}).$$

Theorem 14. *For any $p \in \mathcal{P}_k$ and $x^n \in \{0,1\}^n$ with c complete phrases,*

$$\log \frac{p(x^n)}{q_{\mathsf{LZ}}(x^n)} \le 2 \log\{e(c+2)\} + c\left(\log\left\{e\left(\frac{n}{c}+1\right)\right\} + k\right).$$

In particular,

$$\frac{1}{n} \max_{x^n} \log \frac{p(x^n)}{q_{\mathsf{LZ}}(x^n)} = O\left(\frac{\log\log n}{\log n}\right).$$

The second part of the theorem is based on the following upper bound on the number of complete phrases for a sequence of length n.

Lemma 15. *Let $m = |\mathcal{X}|$ and let $c(x^n)$ be the number of complete phrases in x^n. Then, we have*

$$(1-\varepsilon_n)\sqrt{2n} \le c(x^n) \le (1+\varepsilon_n) m(m-1) \log\left(m \frac{n}{\log n}\right),$$

where $\varepsilon_n = o(1)$ as $n \to \infty$.

Finally, we can show that the LZ process is universal with respect to any stationary random process:

Theorem 16. *For every stationary ergodic random process $(\mathsf{x}_n)_{n=1}^{\infty} \sim \mathsf{P}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p(\mathsf{x}^n)}{q_{\mathsf{LZ}}(\mathsf{x}^n)} \le 0 \quad \mathsf{P}\text{-almost surely.}$$

In particular,

$$\lim_{n\to\infty} \frac{1}{n} D(p(x^n) \,\|\, q_{\mathsf{LZ}}(x^n)) = 0.$$

A Proof of Lemma 8

Proof. Consider

$$\begin{aligned}
\Delta(p \parallel q) &= \sum_{x\in\mathcal{X}} p(x)(\ell(a^*_{\mathsf{B}}(q),x) - \ell(a^*_{\mathsf{B}}(p),x)) \\
&= \sum_{x\in\mathcal{X}} (p(x)-q(x))(\ell(a^*_{\mathsf{B}}(q),x) - \ell(a^*_{\mathsf{B}}(p),x)) + \sum_{x\in\mathcal{X}} q(x)(\ell(a^*_{\mathsf{B}}(q),x) - \ell(a^*_{\mathsf{B}}(p),x)) \\
&= \sum_{x\in\mathcal{X}} (p(x)-q(x))(\ell(a^*_{\mathsf{B}}(q),x) - \ell(a^*_{\mathsf{B}}(p),x)) - \Delta(q \parallel p) \\
&\le \sum_{x\in\mathcal{X}} (p(x)-q(x))(\ell(a^*_{\mathsf{B}}(q),x) - \ell(a^*_{\mathsf{B}}(p),x)) \\
&\le \sum_{x\in\mathcal{X}} |p(x)-q(x)|(|\ell(a^*_{\mathsf{B}}(q),x)| + |\ell(a^*_{\mathsf{B}}(p),x)|) \\
&\le 2\ell_{\max} \sum_{x\in\mathcal{X}} |p(x)-q(x)| \\
&\le 2\ell_{\max}\sqrt{2D(p \parallel q)}.
\end{aligned}$$

Here, the last inequality follows from Pinsker's inequality $\|p-q\|_1 := \sum_x |p(x)-q(x)| \le \sqrt{2D(p \parallel q)}$ for two distributions p and q. □

B Proof of Lemma 10

Proof. Consider

$$\begin{aligned}
\frac{1}{n}\Delta(p(x^n) \parallel q(x^n)) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p(x^i)}[\ell(a^*_{\mathsf{B}}(q(x_i|\mathsf{x}^{i-1})),\mathsf{x}_i) - \ell(a^*_{\mathsf{B}}(p(x_i|\mathsf{x}^{i-1})),\mathsf{x}_i)] \\
&\overset{(a)}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p(x^{i-1})}\left[\mathbb{E}_{p(x_i|\mathsf{x}^{i-1})}[\ell(a^*_{\mathsf{B}}(q(x_i|\mathsf{x}^{i-1})),\mathsf{x}_i) - \ell(a^*_{\mathsf{B}}(p(x_i|\mathsf{x}^{i-1})),\mathsf{x}_i)]\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p(x^{i-1})}\left[\Delta(p(x_i|\mathsf{x}^{i-1}) \parallel q(x_i|\mathsf{x}^{i-1}))\right] \\
&\overset{(b)}{\le} 2\sqrt{2}\ell_{\max}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p(x^{i-1})}\left[\sqrt{D(p(x_i|\mathsf{x}^{i-1}) \parallel q(x_i|\mathsf{x}^{i-1}))}\right] \\
&\overset{(c)}{\le} 2\sqrt{2}\ell_{\max}\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p(x^{i-1})}\left[D(p(x_i|\mathsf{x}^{i-1}) \parallel q(x_i|\mathsf{x}^{i-1}))\right]} \\
&\overset{(d)}{=} 2\sqrt{2}\ell_{\max}\sqrt{\frac{1}{n}D(p(x^n) \parallel q(x^n))}.
\end{aligned}$$

Here, (a) follows from the tower property of conditional expectation, (b) from Lemma 8 applied to $\Delta(p(x_i|\mathsf{x}^{i-1}) \parallel q(x_i|\mathsf{x}^{i-1}))$ for each $i = 1,\dots,n$, (c) from Jensen's inequality using the concavity of

$\rho \mapsto \sqrt{\rho}$, and (d) from the chain rule of KL divergence

$$D(p(x^n) \parallel q(x^n)) = \sum_{i=1}^{n} \mathbb{E}_{p(x^{i-1})}[D(p(x_i|\mathsf{x}^{i-1}) \parallel q(x_i|\mathsf{x}^{i-1}))].$$

This concludes the proof. □

References

[1] Meir Feder. Gambling using a finite state machine. *IEEE Trans. Inf. Theory*, 37(5):1459–1465, 1991.

[2] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory*, 38(4):1258–1270, 1992.

[3] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, 24(5):530–536, 1978.