Discrete Universal Filtering via Hidden Markov Modelling

Taesup Moon
Information Systems Laboratory
Stanford University
Stanford, CA 94305, USA
Email: tsmoon@stanford.edu

Tsachy Weissman
Information Systems Laboratory
Stanford University
Stanford, CA 94305, USA
Email: tsachy@stanford.edu

***Abstract*— We consider the discrete universal filtering problem, where the components of a discrete signal emitted by an unknown source and corrupted by a known DMC are to be causally estimated. We derive a family of filters which we show to be universally asymptotically optimal in the sense of achieving the optimum filtering performance when the clean signal is stationary, ergodic, and satisfies an additional mild positivity condition. Our schemes are based on approximating the noisy signal by a hidden Markov process (HMP) via maximum-likelihood (ML) estimation, followed by use of the well-known forward recursions for HMP state estimation. We show that as the data length increases, and as the number of states in the HMP approximation increases, our family of filters attain the performance of the optimal distribution-dependent filter.**

I. Introduction

The formulation of the filtering problem is the following: A source sequence $x_1, x_2, \cdots$ is corrupted by the discrete memoryless channel (DMC) and only the noisy sequence $z_1, z_2, \cdots$ is observed. The observer wants to generate a reconstruction sequence $\hat{x}_1, \hat{x}_2, \cdots$, where $\hat{x}_t$ is an estimate of x_t based on the observation $z^t = (z_1, \cdots, z_t)$ and the knowledge of the DMC.

The overall performance of filtering is measured by the expectation of the normalized sum of the losses incurred for each estimation. Therefore, the optimal filter which minimizes this expectation bases its estimation at time t on the conditional probability of x_t given z^t. Furthermore, when the DMC is invertible, this conditional probability can be deduced from the conditional probability of z_t given z^{t-1} and the inverse of the channel [3,5]. Thus, the invertibility of the DMC enables implementing the optimal filter from the mere knowledge of probability law of the noisy output process.

However, in the universal setting, where nothing is known about the probability law of the source, the probability law of the noisy source is also not available. Therefore, we need to learn the statistics of the output process and approximate the true probability law as data size increases.

In this paper, we try to use a hidden Markov process (HMP) model to learn the statistics of the output process. When the original clean source is generated from a finite-state Markov chain, the output process is an HMP, and the consistency of maximum likelihood (ML) estimation, [8], guarantees that the Kullback-Leibler (KL) divergence between the true probability law of the output process and the ML estimator converges to zero as data size increases. The question is whether this is also going to be true when the original clean source is a general stochastic process, and whether the induced filtering scheme will be optimal for the approximated source. We show the asymptotic result that this indeed is true under a mild assumption.

The remainder of the paper is organized as follows. Section II introduces some notation and preliminaries. In Section III, the universal filtering problem is defined. In Section IV, our universal filtering scheme is devised and our main theorem is given, and proved. Omitted details in the proofs are given in [9].

II. Notation and preliminaries

A. General notation

Let $\mathcal{X}, \mathcal{Z}, \hat{\mathcal{X}}$ denote, respectively, the finite alphabets of the clean, noisy, and reconstructed source. For simplicity, here we will assume that all alphabets are the same, i.e., $\mathcal{X} = \mathcal{Z} = \hat{\mathcal{X}} = \mathcal{A}$. The channel transition probability is denoted by a $|\mathcal{A}| \times |\mathcal{A}|$ matrix Π, with the x,z-th entry $\Pi(x,z)$ specifying the probability of an output z given that the input is x. A $|\mathcal{A}| \times |\mathcal{A}|$ matrix Λ denotes the loss function, with the x,$\hat{x}$-th entry $\Lambda(x, \hat{x})$ specifying the loss incurred when estimating the clean symbol x by $\hat{x}$. The maximum single-letter loss will be denoted by $\Lambda_{max} = \max_{x,\hat{x} \in \mathcal{M}} \Lambda(x, \hat{x})$.

The invertibility of the DMC is crucial throughout the paper since it enables us to deduce the probability law of X from that of Z. A detailed argument can be found in [3]. The i-th column of Π^{-1} will be denoted by Π_i^{-1}.

$E[\cdot]$ is used as usual expectation. When the subscript of probability law of Z is put and the expectation is over both X and Z, it means the joint distribution is induced from the probability law of Z by inverting the channel, and the expectation is calculated.

As in [3], we define the extended Bayes response associated with the loss matrix Λ to any $\mathbf{V} \in \mathbb{R}^{|\mathcal{A}|}$ as follows.

$$B(\mathbf{V}) = \arg\min_{a \in \mathcal{A}} \Lambda_a^T \mathbf{V},$$

where Λ_a is the a-th column of Λ, and the minimization resolves ties by taking the letter in the alphabet with the lowest

index.

The n-tuple KL divergence between the two distributions P, Q is denoted by

$$D_n(P||Q) = \sum_{z^n} P(z^n) \log \frac{P(z^n)}{Q(z^n)}$$

Also, when a probability law P is written in a bold face, $\mathbf{P}(\cdot)$, it means it is a simplex vector in $\mathbb{R}^{|\mathcal{A}|}$ with first order marginal of the random variable specified inside the parenthesis. It can also be written as $\mathbf{P}(X_t|Z^t)$ meaning the conditional distribution of X_t given Z^t.

B. Hidden Markov processes (HMP)

1) Definition: The HMPs are generally defined to be a family of stochastic processes that are outputs of a memoryless channel whose inputs are finite state Markov chains. In this paper, we will only consider the case where the alphabet of HMP, $\mathcal{Z}$, and underlying Markov chain, $\mathcal{X}$, are finite and equal, i.e., $\mathcal{Z} = \mathcal{X} = \mathcal{A}$,and the channel is DMC and invertible.

There are three parameters that determine the probability laws of HMP, which are, π, the initial distribution of finite state Markov chain, A, the probability transition matrix of finite state Markov chain, and B, the probability transition matrix of DMC. The triplet $\{\pi, A, B\}$ is called the parameter of HMP, and let Θ be a set of all θ's where $\theta := \{\pi_\theta, A_\theta, B_\theta\}$. For each θ, we can calculate the likelihood function

$$Q_\theta(z^n) = \pi_\theta \prod_{t=1}^{n} (\hat{\mathbf{B}}_{\theta,t} A_\theta)\mathbf{1},$$

where $\hat{\mathbf{B}}_{\theta,t}$ is $|\mathcal{A}| \times |\mathcal{A}|$ diagonal matrix whose (j,j)-th entry is (j, z_t)-th entry of B_θ, and $\mathbf{1}$ is a $|\mathcal{A}| \times 1$ vector with all entries 1.

Now let $\Theta_k \subset \Theta$ be a set of θ's such that the order of underlying Markov chain of HMP is k. Furthermore, for some $\delta > 0$, define $\Theta_k^\delta \subset \Theta_k$ as a set of $\theta \in \Theta_k$ that satisfy the following:

- $a_{ij,\theta} \geq \delta$ if a k-tuple state j is a one shift to right of the k-tuple state i
- $a_{ij,\theta} = 0$ if otherwise
- $b_{jz,\theta} = \Pi(j, z)$, for $\forall j, z$

where $a_{ij,\theta}$ is (i,j)-th entry of A_θ, and $b_{ij,\theta}$ is (j,z)-th entry of B_θ. Hence, if $\theta \in \Theta_k^\delta$, 1) the stochastic matrix A_θ is irreducible and aperiodic, and its stationary distribution π_θ is uniquely determined from A_θ, 2) $B_\theta = \Pi \ \forall \theta$, and therefore, θ is completely specified by A_θ.

For the notational brevity, we omit the subscript θ and denote the probability law $Q \in \Theta_k^\delta$, if $Q = Q_\theta$, and $\theta \in \Theta_k^\delta$.

2) Maximum likelihood (ML) estimation: Generally, suppose a probability law Q is in a certain class Ω, and we have n-tuple signal z^n. Then, the n-th order maximum likelihood (ML) estimator in Ω for z^n, is defined to be

$$\hat{Q}[z^n] = \arg\max_{Q \in \Omega} Q(z^n),$$

resolving ties arbitrarily. Now, if $Q \in \Theta_k^\delta$, then there is an iterative algorithm called expectation-maximization(EM) [4] that iteratively updates the parameter estimates via forward-backward recursion to maximize the likelihood. Thus, when Q is in the class of probability laws of a hidden Markov process, the maximum likelihood estimate can be efficiently attained.[1] We denote the ML estimator in Θ_k^δ based on z^n by

$$\hat{Q}_k[z^n] = \arg\max_{Q \in \Theta_k^\delta} Q(z^n).$$

Obviously, when the n-tuple Z^n is random, $\hat{Q}_k[Z^n]$ is also a random probability law that is a function of Z^n.

III. The universal filtering problem

Consider a stochastic setting, that is, the underlying clean random signal is the double-sided stationary ergodic $\mathbf{X}^\infty$ generated from the probability law P_X. Then, $\mathbf{X}^\infty$ is corrupted by invertible DMC, Π, and we get the noisy random signal $\mathbf{Z}^\infty$.

Generally, a *filter* is a sequence of probability distributions $\hat{\mathbf{X}} = \{\hat{X}_t\}$, where $\hat{X}_t : \mathcal{A}^t \to \mathcal{M}(\mathcal{A})$. The interpretation is that, upon observing z^t, the reconstruction for the underlying, unobserved, x_t is given by the symbol $\hat{x}$ with probability $\hat{X}_t(z^t)[\hat{x}]$. The *normalized cumulative loss* of the scheme $\hat{\mathbf{X}}$ on the individual pair (x^n, z^n) is defined by

$$L_{\hat{\mathbf{X}}}(x^n, z^n) = \frac{1}{n} \sum_{t=1}^{n} \ell(x_t, \hat{X}_t(z^t)),$$

where $\ell(x_t, \hat{X}_t(z^t)) = \sum_{\hat{x} \in \hat{\mathcal{X}}} \Lambda(x_t, \hat{x}) \hat{X}_t(z^t)[\hat{x}]$. Then, the goal of a filter is to minimize the *expected normalized cumulative loss* $EL_{\hat{\mathbf{X}}}(X^n, Z^n)$. Let $\mathcal{F}$ denote the class of all filters, and define

$$\phi_n(P_X, \Pi) = \min_{\hat{\mathbf{X}} \in \mathcal{F}} E[L_{\hat{\mathbf{X}}}(X^n, Z^n)],$$

where the expectation on the right side assumes the X^n was emitted by the source P_X, and Z^n is its noisy version when corrupted by the channel Π. By stationarity and sub-additivity argument as in [3], we have

$$\lim_{n \to \infty} \phi_n(P_X, \Pi) = \inf_{n \geq 1} \phi_n(P_X, \Pi) \triangleq \mathbf{\Phi}(P_X, \Pi).$$

By definition, $\mathbf{\Phi}(P_X, \Pi)$ is the (distribution-dependent) optimal asymptotic filtering performance attainable when the clean signal is generated by the law P_X and corrupted by Π. This $\mathbf{\Phi}(P_X, \Pi)$ can be achieved by the optimal filter $\hat{\mathbf{X}}_{P_X} = \{\hat{X}_{P_X,t}\}$ where

$$\hat{X}_{P_X,t}(z^t)[\hat{x}] = Pr(B(\mathbf{P}_{\mathbf{X}}(X_t|z^t)) = \hat{x}).$$

For the brevity of notation, we denote $\hat{X}_{P_X}(z^t) = \hat{X}_{P_X,t}(z^t)$. Note that this is a deterministic scheme, i.e., for given z^t, the filter is a unit vector in $\mathbb{R}^{|\mathcal{A}|}$.

As we can see, $\hat{X}_{P_X}(z^t)$ is dependent on the distribution of underlying clean signal. The *universal filtering problem*

[1] We neglect issues of convergence of the EM algorithm and assume the ML estimation is performed perfectly.

is to construct a distribution independent algorithm $\hat{\mathbf{X}}_{univ}$ satisfying

$$\lim_{n\to\infty} E\Big[L_{\hat{\mathbf{X}}_{univ}}(X^n, Z^n)\Big] = \mathbf{\Phi}(P_X, \Pi)$$

for all P_X.

IV. Filtering based on hidden Markov modelling

A. Description of the filter

Before describing our universal filter, we need one more assumption. Suppose for fixed $\delta > 0$, P_X has a property that

$$P_X(X_0|X_{-k}^{-1}) \geq \delta, \quad \forall k \in \mathbb{N}, \forall X_{-k}^0(\omega).$$

This additional assumption is essential in this paper, and the reason will be explained in proving Lemma 3 below.

Now, define the probability law

$$Q_k^t := \hat{Q}_k[Z^{2^{\lfloor \log_2 t \rfloor}}] = \arg\max_{Q\in\Theta_k^\delta} Q(Z^{2^{\lfloor \log_2 t \rfloor}}).$$

Since $Q \in \Theta_k^\delta$, what we only need to do to get Q_k^t is to find the probability transition matrix of underlying Markov chain that maximizes the likelihood of $Z^{2^{\lfloor \log_2 t \rfloor}}$. Once we get Q_k^t, we can efficiently calculate $Q_k^t(x_t|z^t)$ by the forward-recursion formula which can be found in [4]. Also, let $\mathbf{U} \in \mathbb{R}^{|\mathcal{A}|}$ be a random vector uniformly distributed in L_2 ϵ-ball. Then, we define

$$\hat{X}^\epsilon_{Q_k^t,t}(z^t)[\hat{x}] = Pr(B(\mathbf{Q_k^t}(X_t|z^t) + \mathbf{U}) = \hat{x}).$$

For the brevity of notation, we denote $\hat{X}^\epsilon_{Q_k^t}(z^t) = \hat{X}^\epsilon_{Q_k^t,t}(z^t)$.

Basically, this filtering scheme is dividing the output process into exponentially growing sub-blocks, and to filter each sub-block, it is using the ML estimator for the whole observation of output process up to the previous sub-block. Unlike the optimal filter defined in the previous section, this scheme is a randomized scheme and continuous in $\mathbf{Q_k^t}$ due to the random perturbation vector $\mathbf{U}$. This property will be needed in proving Lemma 4 and Lemma 5 below.

The following theorem states the main result of this paper.

Theorem 1: Suppose a stationary, ergodic double-sided sequence $\mathbf{X}^\infty \in \mathcal{A}^\infty$ whose probability law is P_X, and assume P_X has the property $P_X(X_0|X_{-k}^{-1}) \geq \delta, \quad \forall k \in \mathbb{N}, \forall X_{-k}^0(\omega)$, for some $\delta > 0$. Let $\mathbf{Z}^\infty \in \mathcal{A}^\infty$ be the output of the DMC, Π, for $\mathbf{X}^\infty$. Now, for each k, define the filter $\hat{X}^\epsilon_{univ,k} = \{\hat{X}^\epsilon_{Q_k^t,t}\}$. Then:

(a) $\lim_{\epsilon\to 0} \limsup_{k\to\infty} \limsup_{n\to\infty} L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n, Z^n)$ $\leq \mathbf{\Phi}(P_X, \Pi)$ a.s.

(b) $\lim_{\epsilon\to 0} \lim_{k\to\infty} \limsup_{n\to\infty} E\Big[L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n, Z^n)\Big]$ $= \mathbf{\Phi}(P_X, \Pi)$

B. Proof of the theorem

Before proving the theorem, we introduce several lemmas.

Lemma 1: $Q(Z_0|Z_{-t}^{-1})$ converges to a limit $Q(Z_0|Z_{-\infty}^{-1})$ uniformly in $\forall Q \in \Theta_k^\delta$ and $\forall\omega$.

Proof: Let's denote $f_t := Q(Z_0|Z_{-t}^{-1})$, and $f_0 = 0$. Then, the sequence $\{f_t\}$ uniformly converges in $\forall Q \in \Theta_k^\delta$, if following k subsequences,

$$\{f_{jk+l}, j = 0, 1, 2, \cdots, \}, \qquad 0 \leq l \leq k-1,$$

uniformly converge in $\forall Q \in \Theta_k^\delta$, and have the same limit.

First, the uniform convergence of each subsequence $\{f_{jk+l}\}$ can be shown by showing the series $\sum_{j=0}^t (f_{(j+1)k+l} - f_{jk+l})$ converges absolutely. From Lemma 8 in the appendix and setting $m = k$,

$$\begin{aligned}
&\sum_{j=0}^{t} |f_{(j+1)k+l} - f_{jk+l}| \\
&= \sum_{x_0} Q(Z_0|x_0) \sum_{j=1}^{t} |Q(x_0|Z_{-(j+1)k-l}^{-1}) - Q(x_0|Z_{-jk-l}^{-1})| \\
&\leq \sum_{x_0} Q(Z_0|x_0) \sum_{j=1}^{t} \rho^{j+1}.
\end{aligned}$$

Since $\rho < 1$ and ρ does not depend on Q and l, we conclude all k subsequences converges uniformly in $\forall Q \in \Theta_k^\delta$.

Now, to show that k subsequences have the same limit, construct another subsequence, $\{f_{j(k+1)+1}, j = 0, 1, 2, \cdots, \}$. Since this subsequence contains infinitely many terms from all k subsequences, if this subsequence converges uniformly in $\forall Q \in \Theta_k^\delta$, we can conclude that k subsequences have the same limit. The uniform convergence of this subsequence can be shown exactly as above, but setting $m = k+1$ in Lemma 8. Therefore, the original sequence $\{f_t\}$ converges to its limit uniformly in $\forall Q \in \Theta_k^\delta$. ■

Lemma 2: Suppose P is the true probability law of Z, and $Q \in \Theta_k^\delta$. Then,

$$\mathbf{D}(P\|Q) := \lim_{n\to\infty} \frac{1}{n} D_n(P\|Q) = E_P\Big[\log \frac{P(Z_0|Z_{-\infty}^{-1})}{Q(Z_0|Z_{-\infty}^{-1})}\Big]$$

Moreover, we have the following uniform convergence:

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P(Z^n)}{Q(Z^n)} = \mathbf{D}(P\|Q) \text{ a.s.}$$

uniformly in $\forall Q \in \Theta_k^\delta$.

Proof: The first part and the pointwise convergence of the second part is a kind of generalization of the Shannon-McMillan-Breiman theorem, and the detailed proof can be found in [1, Theorem 2.3.3].

For the uniform convergence of the second part of the lemma, we need to show $\lim_{n\to\infty} \frac{1}{n} \log Q(Z^n) = E_P[\log Q(Z_0|Z_{-\infty}^0)]$ a.s. uniformly in $\forall Q \in \Theta_k^\delta$, by slightly modifying the argument of [1, Lemma 2.4.1]. Since the pointwise convergence is shown and the parameter set Θ_k^δ is compact, it is enough to show that $\frac{1}{n} \log Q(Z^n)$ is an equicontinuous sequence. That is, we need to show for $\forall \epsilon > 0$, $\exists \delta(\epsilon) > 0$ such that

$$\forall n, |\frac{1}{n} \log Q(Z^n) - \frac{1}{n} \log Q'(Z^n)| \leq \epsilon, \text{ if } \quad \|Q - Q'\|_1 < \delta(\epsilon),$$

where $\|Q-Q^{'}\|_1 := \sum_{i,j}|a_{ij}-a_{ij}^{'}|$ is defined to be the distance between two parameters of Q and $Q^{'}$. To show this, just like in [1, Lemma 2.4.1], we first deal with the Markov process $S_t=(X_t,Z_t)$, and show the equicontinuity of $\frac{1}{n}\log Q(S^n)$. The only difference with [1, Lemma 2.4.1] is that the transition matrix T of S_t has some zero elements, but this can be overcome by just considering sequences s^n that have nonzero probabilities. ■

The following definitions are needed for Lemma 3.

Definition 1: When P,Q are the probability laws of Z, we define

$$S(P,Q) := E_P\Big[\log\frac{P(Z_0|Z_{-\infty}^{-1})}{Q(Z_0|Z_{-\infty}^{-1})}\Big].$$

That is, $S(\cdot,\cdot)$ is a functional of two probability laws of Z. Note that when the probability law of the argument is random, $S(\cdot,\cdot)$ is a random variable.

Definition 2: Define the k-th order Markov approximation of P_X for $n\geq k$ as

$$P_{X_k}(X^n) = P_X(X^k)\prod_{i=k+1}^{n}P_X(X_i|X_{i-k}^{i-1}).$$

Also, denote P_Z and P_{Z_k} as the probability law of the output of DMC, Π, when the probability law of input is P_X and P_{X_k}, respectively.

Lemma 3: We have following inequalities.

(a) $\limsup_{n\to\infty} S(P_Z\|\hat{Q}_k[Z^n])\leq \mathbf{D}(P_Z\|P_{Z_k})$ a.s.

(b) $D_n(P_Z\|P_{Z_k})\leq D_n(P_X\|P_{X_k})$, and thus, $\mathbf{D}(P_Z\|P_{Z_k})\leq \mathbf{D}(P_X\|P_{X_k})$

Proof: From the definition of P_{X_k}, we can see why the assumption of our theorem, $P_X(X_0|X_{-k}^{-1})\geq\delta,\quad \forall k\in\mathbb{N},\forall X_{-k}^{0}(\omega)$ is needed. From the assumption, we have P_{Z_k} guaranteed to be in Θ_k^δ. Thus, since $\hat{Q}_k[Z^n]$ is a ML estimator in Θ_k^δ, we can observe the following:

$$\frac{1}{n}\log\frac{P_Z(Z^n)}{\hat{Q}_k[Z^n](Z^n)}\leq\frac{1}{n}\log\frac{P_Z(Z^n)}{P_{Z_k}(Z^n)}$$

for $\forall\omega$, and thus,

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{P_Z(Z^n)}{\hat{Q}_k[Z^n](Z^n)}\leq\limsup_{n\to\infty}\frac{1}{n}\log\frac{P_Z(Z^n)}{P_{Z_k}(Z^n)}\quad\text{a.s.}$$

By Lemma 2, we get

$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{P_Z(Z^n)}{\hat{Q}_k[Z^n](Z^n)}=\limsup_{n\to\infty}S(P_Z,\hat{Q}_k[Z^n])\text{ a.s. , and}$$
$$\limsup_{n\to\infty}\frac{1}{n}\log\frac{P_Z(Z^n)}{P_{Z_k}(Z^n)}=S(P_Z,P_{Z_k})=\mathbf{D}(P_Z\|P_{Z_k})\text{ a.s.}$$

Thus we have,

$$\limsup_{n\to\infty}S(P_Z\|\hat{Q}_k[Z^n])\leq\mathbf{D}(P_Z\|P_{Z_k})\text{ a.s. ,}$$

which gives the part (a) of Lemma.

Part (b) can be easily proved by using log-sum inequality [6, Thm 2.7.1] and the fact $\sum_{z^n}\prod_{t=1}^{n}\Pi(x_t,z_t)=1$. ■

Lemma 4: Suppose a single letter filtering setting, where P_Z is the true probability law of Z, and Q_Z is some other probability law of Z. Let $\mathbf{U}\in\mathbb{R}^{|\mathcal{A}|}$ be a random vector in L_2 ϵ-ball as before, and $\hat{X}_{P_Z}(z)$ and $\hat{X}^\epsilon_{Q_Z}(z)$ be single letter filters such that

$$\hat{X}_{P_Z}(z)[\hat{x}] = Pr(B(\mathbf{P_Z}(X|z))=\hat{x})$$
$$\hat{X}^\epsilon_{Q_Z}(z)[\hat{x}] = Pr(B(\mathbf{Q_Z}(X|z)+\mathbf{U})=\hat{x}),$$

respectively. Then,

$$E_{P_Z}[\ell(X,\hat{X}^\epsilon_{Q_Z}(Z))]-E_{P_Z}[\ell(X,\hat{X}_{P_Z}(Z))]$$
$$\leq 2\Lambda_{max}K_\Pi\cdot\|\mathbf{P_Z}(Z)-\mathbf{Q_Z}(Z)\|_1+C_\Lambda\cdot\epsilon,$$

where $K_\Pi=\sum_{i=1}^{|\mathcal{A}|}\|\Pi_i^{-1}\|_2$, and $C_\Lambda=\max_{a,b\in\mathcal{A}}\|\Lambda_a-\Lambda_b\|_2$.

Proof: The proof uses an idea from [2,(23)] and uses some simple facts such as $\sum_z\Pi(x,z)=1$, Cauchy-Schwartz inequality and the fact that L_2-norm is less than or equal to L_1-norm. Also, it uses the definition of $\hat{X}^\epsilon_{Q_Z}(z)$, to get the term including ϵ. A more detailed proof can be found in [9]. ■

Lemma 5:

$$\lim_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_Q}(X^n,Z^n)-E[L_{\hat{\mathbf{X}}^\epsilon_Q}(X^n,Z^n)]\Big)=0\quad\text{a.s.}$$

uniformly in $\forall Q\in\Theta_k^\delta$.

Proof: The proof is based on two facts. The first one is

$$|Q(X_0|Z_{-t}^0)-Q(X_0|Z_{-\infty}^0)|<\beta\rho^t,$$

for $\forall\omega$, uniformly in $\forall Q\in\Theta_k^\delta$, where β and ρ are constants that only depend on δ,k, and $|\mathcal{A}|$. We can get this by the same argument as in the proof of Lemma 1. The second one is the fact that $Q(X_0|Z_{-t}^0)$ is continuous in its parameters, i.e., the transition probabilities of underlying Markov chain. A detailed proof is somewhat involved, and can be found in [9]. ■

Proof of Theorem 1: Consider the following inequalities, where $\hat{E}_{P_Z}[\cdot]$ is used as special notation to denote that expectation is over all the random variables, except for the randomness of the probability law inside the bracket:

$$\hat{E}_{P_Z}\Big[L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big]-\phi_n(P_X,\Pi)$$
$$=\frac{1}{n}\sum_{t=1}^{n}\Big(\hat{E}_{P_Z}\Big[\ell(X_t,\hat{X}^\epsilon_{Q_k^t}(Z^t))\Big]-\hat{E}_{P_Z}\Big[\ell(X_t,\hat{X}_{P_Z}(Z^t))\Big]\Big)$$
$$=\frac{1}{n}\sum_{t=1}^{n}\hat{E}_{P_Z}\Big[\hat{E}_{P_Z}\Big[\ell(X_t,\hat{X}^\epsilon_{Q_k^t}(Z_t,Z^{t-1}))|Z^{t-1}\Big]$$
$$-\hat{E}_{P_Z}\Big[\ell(X_t,\hat{X}_{P_Z}(Z_t,Z^{t-1}))|Z^{t-1}\Big]\Big]$$
$$\leq\frac{2K_\Pi\Lambda_{max}}{n}\sum_{t=1}^{n}\hat{E}_{P_Z}\|\mathbf{P_Z}(Z_t|Z^{t-1})-\mathbf{Q_k^t}(Z_t|Z^{t-1})\|_1$$
$$+\tilde{C}_{\Pi,\Lambda}\cdot\epsilon \tag{1}$$
$$\leq\frac{2\sqrt{2\ln 2}K_\Pi\Lambda_{max}}{n}\sum_{t=1}^{n}\hat{E}_{P_Z}\sqrt{\hat{E}_{P_Z}\Big[\log\frac{P_Z(Z_t|Z^{t-1})}{Q_k^t(Z_t|Z^{t-1})}\Big|Z^{t-1}\Big]}$$
$$+\tilde{C}_{\Pi,\Lambda}\cdot\epsilon \tag{2}$$
$$\leq 2\sqrt{2\ln 2}K_\Pi\Lambda_{max}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\hat{E}_{P_Z}\Big[\log\frac{P_Z(Z_t|Z^{t-1})}{Q_k^t(Z_t|Z^{t-1})}\Big]}+\tilde{C}_{\Pi,\Lambda}\cdot\epsilon$$
$$\tag{3}$$

where $\tilde{C}_{\Pi,\Lambda} = 2K_\Pi \Lambda_{max} C_\Lambda$, (1) is from Lemma 4, (2) is from Pinsker's inequality, and (3) is from Jensen's inequality. Therefore, together with Lemma 5, we have almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big)\\
=&\limsup_{n\to\infty}\Big(\hat{E}_{P_Z}\Big[L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big]-\phi_n(P_X,\Pi)\Big)\\
\le&2\sqrt{2\ln 2}K_\Pi\Lambda_{max}\sqrt{\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\hat{E}_{P_Z}\Big[\log\frac{P_Z(Z_t|Z^{t-1})}{Q_k^t(Z_t|Z^{t-1})}\Big]}\\
&+\tilde{C}_{\Pi,\Lambda}\cdot\epsilon
\end{aligned}$$

For the expression inside the square root of the right-hand side of the inequality,

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\hat{E}_{P_Z}\Big[\log\frac{P_Z(Z_t|Z^{t-1})}{Q_k^t(Z_t|Z^{t-1})}\Big]\\
=&\limsup_{t\to\infty}\hat{E}_{P_Z}\Big[\log\frac{P_Z(Z_t|Z^{t-1})}{Q_k^t(Z_t|Z^{t-1})}\Big]\ \text{a.s.} &(4)\\
=&\limsup_{t\to\infty}\hat{E}_{P_Z}\Big[\log\frac{P_Z(Z_0|Z_{-\infty}^{-1})}{Q_k^t(Z_0|Z_{-\infty}^{-1})}\Big]\ \text{a.s.} &(5)\\
=&\limsup_{t\to\infty}S(P_Z,Q_k^t)\ \text{a.s.}\\
=&\limsup_{t\to\infty}S(P_Z,\hat{Q}_k[Z^t])\le\mathbf{D}(P_X\|P_{X_k})\ \text{a.s.} &(6)
\end{aligned}$$

where (4) is from Cesáro's mean convergence theorem, (5) is from the fact that $P(Z_0|Z_{-t}^{-1})\to P(Z_0|Z_{-\infty}^{-1})$ a.s. and Lemma 1, and (6) is from the fact that $2^{\lfloor\log_2 t\rfloor}\to\infty$ as $t\to\infty$ and Lemma 3. Therefore, we have almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big)\\
\le&2\sqrt{2\ln 2}K_\Pi\Lambda_{max}\sqrt{\mathbf{D}(P_X\|P_{X_k})}+\tilde{C}_{\Pi,\Lambda}\cdot\epsilon
\end{aligned}$$

Since this inequality holds for every k, and $\mathbf{D}(P_X\|P_{X_k})\to 0$ as $k\to\infty$, we can now conclude

$$\lim_{k\to\infty}\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big)\le\tilde{C}_{\Pi,\Lambda}\cdot\epsilon,\ \text{a.s.}$$

Finally, sending ϵ to zero, part (a) of the theorem is proved. Part (b) follows directly from (a), and Reverse Fatou's Lemma. That is,

$$\begin{aligned}
&\lim_{k\to\infty}\limsup_{n\to\infty}\Big(E\Big[L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big]-\phi_n(P_X,\Pi)\Big)\\
=&\lim_{k\to\infty}\limsup_{n\to\infty}E\Big[L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big]\\
\le&\lim_{k\to\infty}E\Big[\limsup_{n\to\infty}\Big(L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)-\phi_n(P_X,\Pi)\Big)\Big]\\
\le&\tilde{C}_{\Pi,\Lambda}\cdot\epsilon
\end{aligned}$$

Note that the expectation here is with respect to the randomness of probability law inside the bracket, too. Since it is obvious that

$$\liminf_{n\to\infty}\Big(E\Big[L_{\hat{\mathbf{X}}^\epsilon_{univ,k}}(X^n,Z^n)\Big]-\phi_n(P_X,\Pi)\Big)\ge 0\quad,\ \text{for}\ \forall k,\epsilon$$

and by sending ϵ to zero, part (b) is proved. ■

V. Conclusion and future work

In this paper we proved that a family of filters based on HMPs is universally asymptotically optimal. That is, we showed that a sequence of schemes indexed by k are asymptotically optimal. The future direction of the work would be to find out the relationship between k and n such that we can devise a single scheme that grows k with some rate related to n. Trying to loosen the positivity assumption that we made in our main theorem and extending our discrete universal filtering schemes to discrete universal denoising schemes, are additional future directions.

Appendix

Here we state some revised lemmas from [7]. For the following lemmas, fix k and δ, and suppose $Q\in\Theta_k^\delta$. Also, let fix some $m\in\mathbb{N}$ such that $m\ge k$. Proofs are very similar with [7, Appendix], and details will be given in [9].

Lemma 6: Let Z_t be a HMP with probability law Q, X_t be the underlying Markov process. Then,

$$Q(X_{t+m}=j|X_t=i,Z_{t_\ell},t_\ell\in\mathcal{T})\ge\mu_{\delta,k,m}$$

where $\mu_{\delta,m,k}=(1+\frac{|\mathcal{A}|-1}{\delta^{2(k+m)}})^{-1}$ is independent of $Q,\mathcal{T},Z_{t_\ell},i,j$.

Lemma 7: Let

$$\begin{aligned}
C_t:=&\{X_{t_1}=i_1,\cdots,X_{t_p}=i_p,\ \text{where}\ t_p\ge t,\}\in\mathcal{X}_t^\infty\\
D:=&\{Z_{t_1}=z_1,\cdots,Z_{t_h}=z_h,\ \text{where}\ t_h\text{are arbitrary}\}\in\mathcal{Z}_1^\infty
\end{aligned}$$

and define

$$\begin{aligned}
M^+(d,m,C_t,D):=&\,max_i Q(C_t|X_{t-dm}=i,D)\\
M^-(d,m,C_t,D):=&\,min_i Q(C_t|X_{t-dm}=i,D),
\end{aligned}$$

Then,

$$M^+(d,m,C_t,D)-M^-(d,m,C_t,D)\le\rho^{d-1}$$

where $\rho=1-2\mu_{\delta,k,m}$.

Lemma 8:

$$|Q(C_t|Z_{t-dm-l}^p)-Q(C_t|Z_{t-(d+1)m-l}^p)|\le\rho^{d+1}$$

for $\forall p$, $\forall d\ge 1$, and $0\le l\le m-1$.

References

[1] L. Finesso, "Consistent Estimation of the Order for Markov and Hidden Markov Chains," *Ph.D Dissertation, Univ. Maryland, College Park*, 1990

[2] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Inform. Theory*, 44(6):2124-2147, Oct 1998

[3] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu and M. Weinberger, "Universal Discrete Denoising: Known Channel",*IEEE Trans. Inform. Theory*, 51(1):5-28, Jan 2005

[4] Y. Ephraim and N. Merhav, "Hidden Markov Processes", *IEEE Trans. Inform. Theory*, 48(6):1518-1569, June 2002

[5] E. Ordentlich, T. Weissman, M. Weinberger, Anelia Somekh-Baruch and Neri Merhav, "Discrete Universal Filtering Through Incremental Parsing", *DCC2004*

[6] T.M. Cover and J.A. Thomas, "Elements of Information Theory", New York: Wiley, 1991

[7] L.E. Baum and T. Petrie, "Statistical Inference for probabilistic functions of finite state Markov chains", *Ann. Math. Statist.*, vol.37, 1554-1563, 1966

[8] B.G. Leroux, "Maximum-likelihood estimation for hidden Markov models", *Stochastic Processes Their Appic.*, vol. 40, pp.127-143, 1992

[9] T. Moon and T. Weissman, "Discrete Universal Filtering via Hidden Markov Modelling", *in preparation*