

Schemes for Bi-Directional Modeling of Discrete Stationary Sources

Jiming Yu and Sergio Verdú

Department of Electrical Engineering

Princeton University

Princeton, NJ 08544

e-mail: {jimingyu, verdu}@princeton.edu

Abstract — Adaptive models are developed to deal with bi-directional modeling of unknown discrete stationary sources, which can be generally applied to statistical inference problems such as noncausal universal discrete denoising that exploits bi-directional dependencies. Efficient algorithms for constructing those models are developed and implemented. Denoising is a primary focus of the application of those models, and we compare their performance to that of the DUDE algorithm [1] for universal discrete denoising.

I. INTRODUCTION

Bi-directional models are concerned with estimating the marginal conditional distribution of the current symbol given the rest of the data. To be concrete, we explain why bi-directional models are needed from the denoising example. Denoising is the procedure of recovering the uncoded input to a channel with as high a fidelity as possible, by only observing the output of the channel. Universal discrete denoising [1], deals with discrete input/output channels without knowledge of the input distribution. In contrast to universal prediction in a noisy environment [2–4] and universal (causal) discrete filtering [5], the case when all observations are available to recover the input symbol at any time instant is called *noncausal* universal discrete denoising, which is the main application of this paper. Noncausal denoising provides the main motivation for bi-directional modeling of discrete sources.

Given any sequence $(y_t)_{L_1 \leq t \leq L_2}$, define $y_a^b \triangleq (y_a, y_{a+1}, \dots, y_b)$, $\forall L_1 \leq a \leq L_2, L_1 \leq b \leq L_2$ with the convention that y_a^b is the empty string \mathbf{e} with length 0 if $a > b$ and y_1^n is abbreviated as y^n .

A stationary process $X = (X_t)_{t \in \mathbb{Z}}$ with finite alphabet $\mathcal{A} = \{1, 2, \dots, M\}$ goes through a discrete memoryless channel (DMC) with a channel transition probability matrix $\mathbf{\Pi}$ and a noisy stationary output process $Z = (Z_t)_{t \in \mathbb{Z}}$ with alphabet $\mathcal{B} = \{1, 2, \dots, M'\}$ is observed, where

$$\mathbf{\Pi}(a, b) \triangleq \mathbb{P}\{Z_t = b | X_t = a\}, \forall t \in \mathbb{Z}, a \in \mathcal{A}, b \in \mathcal{B}.$$

A fixed loss function $\Lambda : \mathcal{A} \times \hat{\mathcal{A}} \mapsto [0, \infty)$ such that $\Lambda(a, \hat{a})$ is the loss of estimating $a \in \mathcal{A}$ by \hat{a} in reconstruction alphabet $\hat{\mathcal{A}} \triangleq \{1, 2, \dots, \hat{M}\}$, is represented by an $M \times \hat{M}$ matrix $\mathbf{\Lambda} = \{\Lambda(a, b)\}_{M \times \hat{M}}$. $\mathbf{\Lambda}$ induces an average cumulative loss $\frac{1}{n} \sum_{j=1}^n \Lambda(X_j, \hat{X}_j)$ by giving \hat{X}^n as output of the denoiser, for finite time observations Z^n and finite time input X^n . The denoiser outputs an estimate \hat{X}^n of X^n with as high a fidelity as possible, by only observing Z^n and knowing the channel matrix $\mathbf{\Pi}$ and the loss matrix $\mathbf{\Lambda}$.

Under the assumption that the channel matrix has full row rank, the input distribution is uniquely determined by the output distribution and the channel matrix. This leads to the

following optimal Bayesian denoiser [1]:

$$\hat{X}^{n, \text{opt}}(z^n)[j] = \arg \min_{\hat{x} \in \hat{\mathcal{A}}} \mathbf{P}_{Z_j | z_1^{j-1}, z_{j+1}^n}^T \mathbf{\Pi}^T (\mathbf{\Pi} \mathbf{\Pi}^T)^{-1} [\lambda_{\hat{x}} \odot \pi_{z_j}] \quad (1)$$

for a finite time sample path z^n of the output process $Z, j = 1, 2, \dots, n$, where $\hat{X}^{n, \text{opt}}(z^n) : \mathcal{B}^n \mapsto \hat{\mathcal{A}}^n$ is the denoiser in a vector form, $\hat{X}^{n, \text{opt}}(z^n)[j] = \hat{X}_j$ is its j -th component representing the estimate of X_j , $\mathbf{P}_{Z_j | z_1^{j-1}, z_{j+1}^n} \in [0, 1]^{M'}$ with b -th component specified as

$$\mathbb{P}\{Z_j = b | Z_1^{j-1} = z_1^{j-1}, Z_{j+1}^n = z_{j+1}^n\}, \forall b \in \mathcal{B},$$

and $\lambda_{\hat{x}}$ is the \hat{x} -th column of $\mathbf{\Lambda}$, π_{z_j} is the z_j -th column of $\mathbf{\Pi}$, and

$$u \odot v \triangleq (u_1 v_1, u_2 v_2, \dots, u_d v_d)^T, \forall u, v \in \mathbb{R}^d.$$

In the universal case where input distribution is unknown, the only unknown quantities in (1) are the bi-directional conditional distributions $\mathbf{P}_{Z_j | z_1^{j-1}, z_{j+1}^n}, j = 1, 2, \dots, n$. Thus, through (1), universal discrete denoising (at least under the full row rank condition) reduces to bi-directional modeling of the output process based on the observation of a realization. DUDE [1] uses $\mathbf{P}_{Z_j | z_{j-m}^{j-1}, z_{j+1}^{j+m}}$ to approximate $\mathbf{P}_{Z_j | z_1^{j-1}, z_{j+1}^n}$ for some fixed $m \in \mathbb{N}^* \triangleq \{1, 2, 3, \dots\}$, where $\mathbf{P}_{Z_j | z_{j-m}^{j-1}, z_{j+1}^{j+m}} \in [0, 1]^{M'}$ with b -th component specified as

$$\mathbb{P}\{Z_j = b | Z_{j-m}^{j-1} = z_{j-m}^{j-1}, Z_{j+1}^{j+m} = z_{j+1}^{j+m}\}, \forall b \in \mathcal{B}.$$

Furthermore DUDE [1] replaces all conditional distributions $\mathbf{P}_{Z_j | z_{j-m}^{j-1}, z_{j+1}^{j+m}}$ by their corresponding empirical distributions. Approximating $\mathbf{P}_{Z_j | z_1^{j-1}, z_{j+1}^n}$ by $\mathbf{P}_{Z_j | z_{j-m}^{j-1}, z_{j+1}^{j+m}}$ for some m is not necessarily the best we can do. After all, Z_j may depend on different number of symbols for different j in an asymmetric way, unlike the static symmetric context length used in DUDE. Adaptive bi-directional models that incorporate this feature are considered in this paper.

Sequential models estimate the sequential conditional probabilities $\mathbf{P}_{Z_j | b_1^{j-1}}$ from a realization z^n for a stationary process $Z = (Z_t)_{t \in \mathbb{Z}}$, where $\mathbf{P}_{Z_j | b_1^{j-1}} \in [0, 1]^{M'}$ with a -th component specified as

$$\mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}\}, \forall a \in \mathcal{B}, b_1^{j-1} \in \mathcal{B}^{j-1}, j = 1, 2, \dots, n.$$

They have been subject to extensive research, cf. [2, 6–9] and references therein. Various context tree methods are widely used, also cf. [10]. In this paper, we extend the notion of variable length Markov chain [7–9] to the bi-directional case in a nontrivial way.

A more fundamental problem than denoising (with other applications such as [11]) is the estimation of the marginal

conditional distribution of the input given the output [1]:

$$\mathbf{P}_{X_j|z_1^n} = \frac{1}{\mathbf{P}_{Z_j|z_1^{j-1}, z_{j+1}^n}[z_j]} \pi_{z_j} \odot [(\mathbf{\Pi}\mathbf{\Pi}^T)^{-1}\mathbf{\Pi}\mathbf{P}_{Z_j|z_1^{j-1}, z_{j+1}^n}] \quad (2)$$

where the right hand side of above equation can be estimated via estimates of $\mathbf{P}_{Z_j|z_1^{j-1}, z_{j+1}^n}$ from a bi-directional model for process Z , and $\mathbf{P}_{Z_j|z_1^{j-1}, z_{j+1}^n}[z_j]$ is the z_j -th component of $\mathbf{P}_{Z_j|z_1^{j-1}, z_{j+1}^n}$. Note that (1) is derived from (2) by Bayesian estimation [1]. While the known channel case is solved in [1], the case of a binary symmetric channel with unknown crossover probability is studied in [12, 13] under a minimax criterion. When the input process is a Markov chain, an alternative to compute the marginal conditional input distribution given the output via (2) is backward-forward dynamic programming. Hidden Markov modeling tools [14, 15] are popular in order to estimate both the channel and the input Markov chain transition probability matrix when they are unknown. As in [1], this paper assumes a *known* DMC and an *unknown* input process.

Another reference related to this paper is [16], where an optimal tree under a given cost criterion is found. By estimating the unknown true denoising performance, context trees are constructed to minimize the estimated cost criterion.

II. ADAPTIVE BI-DIRECTIONAL MODELS FOR DISCRETE STATIONARY SOURCES

An adaptive bi-directional model for the \mathcal{B} -valued stationary source $Z = (Z_t)_{t \in \mathbb{Z}}$ is defined as a collection of estimators for conditional distributions¹

$$\mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\}$$

from any given realization z^n of process Z , $\forall n \in \mathbb{N}^*$, $j = 1, 2, \dots, n$, $a \in \mathcal{B}$, $b_1^{j-1} \in \mathcal{B}^{j-1}$, $b_{j+1}^n \in \mathcal{B}^{n-j}$. An adaptive sequential model can be defined analogously by considering sequential conditional distribution estimators.

II.A VARIABLE LENGTH MARKOV CHAINS

A powerful sequential model for discrete stationary sources, the so-called variable length Markov chain, has been developed in [6–9]. The main idea is to model the discrete stationary source $Z = (Z_t)_{t \in \mathbb{Z}}$ as a Markov process with order that depends on the history of the realization:

$$\mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}\} = \mathbb{P}\{Z_j = a | Z_{j-k_n}^{j-1} = b_{j-k_n}^{j-1}\} \quad (3)$$

for any $a \in \mathcal{B}$, $b_1^{j-1} \in \mathcal{B}^{j-1}$ and some $k_n = k_{z^n}(b_1^{j-1}) \leq j-1$ estimated from a realization z^n , $\forall n \in \mathbb{N}^*$, $j = 1, 2, \dots, n$. The estimation of $k_{z^n}(b_1^{j-1})$ is accomplished by a context tree model [7, 9], whose construction (taking a cue from [6]) requires the divergence of conditional distribution estimates from all leaf nodes with respect to that of their father nodes to be larger than some pre-selected parameter. This algorithm has been implemented in [8] with time complexity $O(n \log n)$.

II.B TWO ADAPTIVE BI-DIRECTIONAL MODELS BASED ON ADAPTIVE SEQUENTIAL MODELS

¹Note the difference between arbitrary a , b_1^{j-1} , b_{j+1}^n and the available realization z^n for estimation.

In this section we describe two methods for constructing bi-directional models by amalgamating sequential models running forwards and backwards.

1) *Backward-Forward Product (BFP)*: An extremely simple way to estimate $\mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\}$ from the sequential conditional probabilities is²:

$$\begin{aligned} & \tilde{\mathbb{P}}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\} \propto \\ & \tilde{\mathbb{P}}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}\} \times \tilde{\mathbb{P}}\{Z_j = a | Z_{j+1}^n = b_{j+1}^n\} \end{aligned} \quad (4)$$

for any $a \in \mathcal{B}$, $b_1^{j-1} \in \mathcal{B}^{j-1}$, $b_{j+1}^n \in \mathcal{B}^{n-j}$, $j = 1, 2, \dots, n$, where \propto means "proportional to", and $\tilde{\mathbb{P}}$ stands for the estimated probability law. Once equipped with any sequential model (e.g. the one mentioned in Section II.A), the first term in the right hand side of (4) can be estimated from data z^n , and the second term can be estimated from reversed-order data $z_*^n \triangleq (z_{n-i+1})_{i=1}^n$. (4) can then be used in specifying a corresponding bi-directional model. This ad-hoc approximation obviously has the merit that it is straightforward to implement once we have a computationally efficient sequential model.

2) *Generalized Markov (GM) Scheme*: We make the assumption that $\exists m \in \mathbb{N}^*$, $2m < n$ such that given any Z_{j-m}^{j+m} , the past and the future are conditionally independent, that is, Z_1^{j-m-1} and Z_{j+m+1}^n are conditionally independent given Z_{j-m}^{j+m} , $\forall m+1 \leq j \leq n-m$. Under that assumption, for any $a \in \mathcal{B}$, $b_1^{j-1} \in \mathcal{B}^{j-1}$, $b_{j+1}^n \in \mathcal{B}^{n-j}$, it is easy to show that:

$$\begin{aligned} & \mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\} \propto \\ & \frac{\mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m} | Z_1^{j-m-1} = b_1^{j-m-1}\}}{\mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m}\}} \\ & \times \mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m} | Z_{j+m+1}^n = b_{j+m+1}^n\} \end{aligned} \quad (5)$$

and summing (5) over all $a \in \mathcal{B}$ gives us the reciprocal $f(b_1^{j-1}, b_{j+1}^n)$ of the normalizing constant that does not depend on a .

From a realization z^n , DUDE [1] gives estimates for

$$\mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m}\}, \quad j = m+1, m+2, \dots, n-m \quad (6)$$

and furthermore we only use (5) when $b_{j-m}^{j-1} a b_{j+1}^{j+m}$ actually has appeared in the sequence z^n , otherwise we let the estimate for $\mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\}$ be 0, since the true value is indeed 0 if $\mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m}\} = 0$. By the definition of conditional probabilities, we have:

$$\begin{aligned} & \mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m} | Z_1^{j-m-1} = b_1^{j-m-1}\} = \\ & \mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}\} \mathbb{P}\{Z_{j+1} = b_{j+1} | Z_1^j = b_1^{j-1} a\} \\ & \quad \times \prod_{t=j+2}^{j+m} \mathbb{P}\{Z_t = b_t | Z_1^{t-1} = b_1^{t-1} a b_{j+1}^{t-1}\} \\ & \quad \times \prod_{t=j-m}^{j-1} \mathbb{P}\{Z_t = b_t | Z_1^{t-1} = b_1^{t-1}\} \end{aligned} \quad (7)$$

and:

$$\begin{aligned} & \mathbb{P}\{Z_{j-m}^{j+m} = b_{j-m}^{j-1} a b_{j+1}^{j+m} | Z_{j+m+1}^n = b_{j+m+1}^n\} = \\ & \mathbb{P}\{Z_{j-1} = b_{j-1} | Z_j^n = a b_{j+1}^n\} \mathbb{P}\{Z_j = a | Z_{j+1}^n = b_{j+1}^n\} \\ & \quad \times \prod_{t=j-m}^{j-2} \mathbb{P}\{Z_t = b_t | Z_{t+1}^n = b_{t+1}^n a b_{j+1}^n\} \\ & \quad \times \prod_{t=j+1}^{j+m} \mathbb{P}\{Z_t = b_t | Z_{t+1}^n = b_{t+1}^n\} \end{aligned} \quad (8)$$

²In many applications (e.g. universal discrete denoising) unnormalized conditional marginals are sufficient.

Note terms in (7) can be estimated by a sequential model (e.g. the one mentioned in Section II.A) for data z^n , and terms in (8) can be estimated by a sequential model for reversed-order data $z_*^n \triangleq (z_{n-j+1})_{j=1}^n$. Plugging estimates for (6), (7) and (8) into (5) for $m+1 \leq j \leq n-m$, and assigning arbitrary values to the conditional probabilities for other values of j , we can get a bi-directional model under the above additional assumption.

II.C THE ADAPTIVE BI-DIRECTIONAL CONTEXT TREE (BCT)

Equation (3) inspires us to construct bi-directional models directly instead of only constructing bi-directional models from sequential models. In addition to increased computational complexity, when we consider bi-directional contexts we have to cope with their non-uniqueness.

1) *Contexts*: Let $Z = (Z_t)_{t \in \mathbb{Z}}$ be a stationary process with finite alphabet \mathcal{B} . For a segment Z^n , a pair of strings $(b_{j-s}^{j-1}, b_{j+1}^{j+t})$ for some integers $s, t: 0 \leq s \leq j-1, 0 \leq t \leq n-j$ is called a context of $(b_1^{j-1}, b_{j+1}^n) \in \mathcal{B}^{j-1} \times \mathcal{B}^{n-j}$ if

$$\begin{aligned} \forall a \in \mathcal{B}, \quad & \mathbb{P}\{Z_j = a | Z_{j-s}^{j-1} = b_{j-s}^{j-1}, Z_{j+1}^{j+t} = b_{j+1}^{j+t}\} \\ & = \mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\} \end{aligned} \quad (9)$$

with the *minimality* property that for any $s' \leq s, t' \leq t, s' + t' < s + t$,

$$\begin{aligned} & \mathbb{P}\{Z_j = a | Z_{j-s'}^{j-1} = b_{j-s'}^{j-1}, Z_{j+1}^{j+t'} = b_{j+1}^{j+t'}\} \\ & \neq \mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\} \end{aligned} \quad (10)$$

If $s = j-1, t = n-j$, or equivalently $s+t = n-1$, the context is trivial. Let $S_n(b_1^{j-1}, b_{j+1}^n)$ be the context set for (b_1^{j-1}, b_{j+1}^n) , namely the set of all contexts for (b_1^{j-1}, b_{j+1}^n) . Note that $|S_n(b_1^{j-1}, b_{j+1}^n)|$ is usually larger than 1, in contrast with the sequential model [7,9] where a unique (minimal) context is identified for the same historical data b_1^{j-1} .

2) *Context Trees*: For a segment Z^n of a stationary process Z , let $\mathcal{B}^0 \triangleq \{\mathbf{e}\}$ be the set consisting of only the empty string, we define

$$\begin{aligned} T_n(b_1^{j-1}, b_{j+1}^n) & \triangleq \left\{ (u, v) : u \in \bigcup_{t=0}^{j-1} \mathcal{B}^t, v \in \bigcup_{t=0}^{n-j} \mathcal{B}^t \right. \\ & \text{such that } \exists (x, y) \in S_n(b_1^{j-1}, b_{j+1}^n), \\ & \left. x_{|x|-|u|+1}^{|x|} = u, y_1^{|v|} = v \right\} \end{aligned} \quad (11)$$

That is, $T_n(b_1^{j-1}, b_{j+1}^n)$ is the set of bi-directional substrings of contexts in $S_n(b_1^{j-1}, b_{j+1}^n)$ which include only most recent past and future data.

The context tree is defined as the set:

$$\tau_n = \bigcup_{(b_1^{j-1}, b_{j+1}^n) \in \mathcal{B}^{j-1} \times \mathcal{B}^{n-j}} T_n(b_1^{j-1}, b_{j+1}^n) \quad (12)$$

with a tree structure, in which a node corresponds to a pair of strings (u, v) and vice versa. This correspondence is as follows: the root corresponds to (\mathbf{e}, \mathbf{e}) , and all branches originating from a node are labelled by a pair of symbols in $\mathcal{B} \cup \mathcal{B}^0$, but not both in \mathcal{B}^0 . If a node is obtained by travelling a path from root (\mathbf{e}, \mathbf{e}) consisting of branches labelled by $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ (in that order) for some $x_i, y_i \in \mathcal{B} \cup \mathcal{B}^0, i = 1, 2, \dots, m$, then this node corresponds

to the pair of strings $(x_m x_{m-1} \dots x_2 x_1, y_1 y_2 \dots y_m)$ by juxtapositions. Conversely, for any pair of strings (u, v) , if $|u| = |v|$, then it corresponds to the node obtained by travelling the path $(u_{|u|}, v_1), (u_{|u|-1}, v_2), \dots, (u_1, v_{|u|})$ from root; if $|u| < |v|$, then the path is $(u_{|u|}, v_1), (u_{|u|-1}, v_2), \dots, (u_1, v_{|u|}), (\mathbf{e}, v_{|u|+1}), (\mathbf{e}, v_{|u|+2}), \dots, (\mathbf{e}, v_{|v|})$; similarly for the case $|u| > |v|$.

3) *Estimation of the Context Sets from z^n* : We want (9) to hold approximately for general discrete stationary process Z with an estimate of the context set $S_n(b_1^{j-1}, b_{j+1}^n)$ for any (b_1^{j-1}, b_{j+1}^n) . We use $\hat{S}_{z^n}(b_1^{j-1}, b_{j+1}^n)$ to denote the estimated context set from z^n , which can be obtained by an efficient algorithm (see Section III.B for details). The estimated context tree corresponding to those estimated context sets is denoted by $\hat{\tau}_n(z^n)$.

4) *Two Constraints for an Estimated Context Tree*: The first constraint is the consecutive constraint, which is implicit in the definition of contexts, and it is better made explicit. In addition, this consecutive constraint applies to estimated context trees as well as context trees. For a context (w, v) we have $w = b_{j-s}^{j-1}, v = b_{j+1}^{j+t}$ for some $b_1^{j-1}, b_{j+1}^n, s, t$, i.e. only symbols in consecutive time instants are allowed to exist in either side of a context. In the construction of such a tree, this implicit constraint induces the following for node (w, v) (recall that \mathbf{e} is the empty string with length $|\mathbf{e}| = 0$):

- $|w| = |v|$: Three possible kinds of child nodes with forms $(aw, vb), (aw, v\mathbf{e})$ or $(\mathbf{e}w, vb)$ for some $a, b \in \mathcal{B}$.
- $|w| < |v|$: One possible kind of child node of the form $(\mathbf{e}w, vb)$ for some $b \in \mathcal{B}$.
- $|w| > |v|$: One possible kind of child node of the form $(aw, v\mathbf{e})$ for some $a \in \mathcal{B}$.

This tree structure makes our bi-directional context tree a nontrivial extension of [7,9] in terms of data structure.

To achieve a good performance/complexity tradeoff we place a constraint N on the maximal tree depth (or the maximal single-side context length). In the worst case we have $O((M')^{2N})$ nodes in an estimated context tree, which is a constant not depending on data length n , and we thus have linear complexity in both time and space.

5) *Empirical Distributions from a Realization z^n* : We define $\mathcal{B}^* \triangleq \bigcup_{n \in \mathbb{N}^*} \mathcal{B}^n$ and

$$N_u(w) \triangleq \sum_{t=1}^{|u|-|w|+1} 1_{u_t + |w|-1 = w}, \forall u \in \mathcal{B}^*, w \in \bigcup_{t=1}^{|u|} \mathcal{B}^t \quad (13)$$

$$\hat{P}_u(a|w, v) \triangleq \frac{N_u(wav)}{\sum_{b \in \mathcal{B}} N_u(wbv)} \quad (14)$$

for all $u \in \mathcal{B}^*, a \in \mathcal{B}, w, v \in \bigcup_{t=1}^{|u|-1} \mathcal{B}^t \cup \{\emptyset\}, |w| + |v| + 1 \leq |u|$, and (14) is an estimate for

$$\mathbb{P}\{Z_j = a | Z_{j-|w|}^{j-1} = w, Z_{j+|v|}^{j+|v|} = v\}$$

for any $|w| + 1 \leq j \leq n - |v|$ if $u = z^n$, since Z is stationary.

6) *Estimators for Conditional Distributions from a Realization z^n* : For any (b_1^{j-1}, b_{j+1}^n) , equipped with the context sets $\hat{S}_{z^n}(b_1^{j-1}, b_{j+1}^n)$ estimated from a realization z^n (the estimation algorithm is discussed in Section III.B), let (w, v) be any element in $\hat{S}_{z^n}(b_1^{j-1}, b_{j+1}^n)$. Then by definition of $S_n(b_1^{j-1}, b_{j+1}^n)$ in (9), we know (14) becomes the estimate

$$\hat{P}_{z^n}(a|w, v) = \tilde{\mathbb{P}}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\} \quad (15)$$

2) *Step 2*: Examine every leaf node in $\mathcal{T}(0)(z^n)$ as follows. Select any leaf node (w, v) in $\mathcal{T}(0)(z^n)$, where the order of selection here is irrelevant, prune this node if:

$$\Delta_{(w,v)}(z^n) \leq K_n \quad (20)$$

for some pruning parameter $K_n \geq 0$, where

$$\begin{aligned} \Delta_{(w,v)}(z^n) \triangleq & D(\hat{P}_{z^n}(\cdot|w, v) || \hat{P}_{z^n}(\cdot|w', v')) \\ & \times \sum_{a \in \mathcal{B}} N_{z^n}(wav) \end{aligned} \quad (21)$$

where (w', v') is the father node of (w, v) and it has at most three child nodes by construction. After examining each leaf node in $\mathcal{T}(0)(z^n)$ by the above method, we should get another context tree $\mathcal{T}(1)(z^n)$, which is a subtree of $\mathcal{T}(0)(z^n)$.

3) *Step 3*: Repeat *Step 2*, starting from $\mathcal{T}(i)(z^n)$ to $\mathcal{T}(i+1)(z^n)$, $i \geq 1$ in the same way as *Step 2* does, until no more pruning is possible. This tree is our $\hat{\tau}_n(z^n)$, the estimated context tree from data z^n .

4) *Step 4*: Equipped with the estimated context tree $\hat{\tau}_n(z^n)$, we are ready to specify the estimated context sets $\hat{S}_{z^n}(b_1^{j-1}, b_{j+1}^n)$. Let $(x_i, y_i)_{i=1}^k$ be the $k = |\hat{\tau}_n(z^n)|$ nodes in $\hat{\tau}_n(z^n)$. For any $j = 1, 2, \dots, n$ and any $b_1^{j-1} \in \mathcal{B}^{j-1}, b_{j+1}^n \in \mathcal{B}^{n-j}$, let $\hat{S}_{z^n}(b_1^{j-1}, b_{j+1}^n)$ be those $(x_i, y_i), i \in \{1, 2, \dots, k\}$ such that

$$j \geq |x_i| + 1, \quad j \leq n - |y_i| \quad (22)$$

$$b_{j-|x_i|}^{j-1} = x_i, \quad b_{j+1}^{j+|y_i|} = y_i \quad (23)$$

and no other node in the subtree rooted at (x_i, y_i) satisfies those conditions (22) and (23). If no such (x_i, y_i) exists for (b_1^{j-1}, b_{j+1}^n) , we let $\hat{S}_{z^n}(b_1^{j-1}, b_{j+1}^n) = \{(b_1^{j-1}, b_{j+1}^n)\}$, which only consists of a trivial context.

5) *Step 5*: For any $a \in \mathcal{B}, j = 1, 2, \dots, n, b_1^{j-1} \in \mathcal{B}^{j-1}, b_{j+1}^n \in \mathcal{B}^{n-j}$, estimate $\mathbb{P}\{Z_j = a | Z_1^{j-1} = b_1^{j-1}, Z_{j+1}^n = b_{j+1}^n\}$ by (18) with context function specified in *Step 4* and weights specified in (19).

IV. AN EXPERIMENTAL STUDY OF DENOISING VIA ADAPTIVE BI-DIRECTIONAL MODELS

We now turn our attention to a concrete application of adaptive bi-directional modeling, namely noncausal universal discrete denoising. We compare the four bi-directional models: a) static symmetric contexts (as in DUDE [1]); b) BFP (cf. Section II.B-1); c) GM (cf. Section II.B-2) and d) BCT (cf. Section II.C) by using them in the Bayesian optimal noncausal denoiser (1). We let $\Lambda(a, b) = 1_{a \neq b}, \forall a, b \in \mathcal{A} = \mathcal{B} = \hat{\mathcal{A}} = \{1, 2, \dots, M\}, M = M' = \hat{M}$. We compare the performances of universal discrete denoisers induced from those bi-directional models, with *no* knowledge of the channel input distribution. When the distribution of underlying input to the channel is well-defined and known, the Bayesian optimal denoising performance can be achieved by plugging true conditional distributions into (1). This optimal Bayesian method is denoted by BFDP since it is implemented via the backward-forward dynamic programming algorithm.

Various parameters need to be adjusted accordingly to adapt to real data. DUDE needs to choose the context length m [1]; BFP and GM have a pruning parameter to construct the (uni-directional) context tree [9]; In addition, GM has the parameter m that determines the size of the "present", cf. (5); BCT has a pruning parameter K_n (cf. (20)), an exponent constant β (cf. (19)) and a maximal tree depth N (cf.

$\delta = 0.01$					
p	DUDE	BFP	GM	BCT	BFDP
0.01	0.0007	0.0007	0.0007	0.0007	0.0007
0.05	0.0046	0.0045	0.0043	0.0043	0.0043
0.10	0.0100	0.0116	0.0100	0.0100	0.0100
0.15	0.0100	0.0100	0.0100	0.0100	0.0100
0.20	0.0100	0.0100	0.0100	0.0100	0.0100

Tab. 1: BSC(δ) and First-Order Markov Source(p)

$\delta = 0.10$					
p	DUDE	BFP	GM	BCT	BFDP
0.01	0.0070	0.0063	0.0066	0.0065	0.0058
0.05	0.0306	0.0303	0.0301	0.0302	0.0298
0.10	0.0566	0.0562	0.0567	0.0561	0.0559
0.15	0.0799	0.0758	0.0752	0.0750	0.0750
0.20	0.0924	0.0934	0.0924	0.0924	0.0924

Tab. 2: BSC(δ) and First-Order Markov Source(p)

Section II.C-4). The maximal tree depth N can be chosen rather arbitrarily, and longer depths do not necessarily lead to better performances. Since underlying true input sequence to the channel is *unknown*, in practice it is impossible to choose parameters according to the true loss function induced by a denoiser under given parameters. Instead, the so-called Lempel-Ziv heuristic [1] can be used to tune the parameters. The Lempel-Ziv heuristic says that the optimal parameters are achieved when the denoiser outputs a sequence that has the least length after Lempel-Ziv type universal lossless compression. In the experiments, the performances of Lempel-Ziv heuristic based denoisers and genie-aided denoisers that utilize the knowledge of true loss are very close, strongly supporting the use of the Lempel-Ziv heuristic.

IV.A BINARY MARKOV SOURCE AND BINARY SYMMETRIC CHANNEL (BSC)

This experiment belongs to the stochastic setting, and Bayesian optimal denoiser can be exactly implemented via BFDP. The source is binary Markov with transition probability p , and the channel is a BSC with crossover probability δ . We also let $M = 2, n = 10^6$, and obviously the bit error rate before denoising is δ . Table 1, Table 2, and Table 3 display the bit error rates after denoising for different (p, δ) and different models.

IV.B ENGLISH TEXT CORRECTION

In this individual sequence setting, an English text goes through a DMC called the keyboard channel, which mimics a

$\delta = 0.20$					
p	DUDE	BFP	GM	BCT	BFDP
0.01	0.0333	0.0219	0.0233	0.0245	0.0160
0.05	0.0776	0.0739	0.0739	0.0741	0.0713
0.10	0.1217	0.1203	0.1203	0.1203	0.1192
0.15	0.1542	0.1538	0.1541	0.1539	0.1535
0.20	0.1796	0.1766	0.1765	0.1763	0.1762

Tab. 3: BSC(δ) and First-Order Markov Source(p)

No.	DUDE	BFP	GM	BCT
1	0.022479	0.0137131	0.0112764	0.0170538
2	0.022492	0.0128292	0.0111544	0.0171619
3	0.022555	0.0142284	0.0119915	0.0194086
4	0.024042	0.0159728	0.0129864	0.0208239
5	0.021544	0.0138857	0.0111848	0.0176427
6	0.022778	0.0134215	0.0107607	0.0177708
7	0.024225	0.0161501	0.0133945	0.0214963
8	0.021034	0.0127723	0.0100843	0.0169560
9	0.023044	0.0136409	0.0113392	0.0179127
10	0.021422	0.0128954	0.0105986	0.0169460
11	0.024341	0.0169152	0.0137873	0.0212502
12	0.022794	0.0132631	0.0100639	0.0180168
13	0.024074	0.0141297	0.0112091	0.0193032
14	0.022882	0.0134861	0.0109549	0.0191298
15	0.023871	0.0168967	0.0146290	0.0213615

Tab. 4: Keyboard Channel and English Texts

typist’s most possible errors. Here $M = 128$, and the channel transition probability matrix $\mathbf{\Pi}$ is identical to the one in [1], which corrupts letters with probability 0.05. Fifteen novels have been tried (using data from Project Gutenberg: <http://www.gutenberg.org/>):

- 1. Don Quixote
- 2. David Copperfield
- 3. Great Expectations
- 4. Jane Eyre
- 5. Les Miserables
- 6. Little Dorrit
- 7. Notre-Dame de Paris
- 8. The Count of Monte Cristo
- 9. The Life and Adventures of Nicholas Nickleby
- 10. Household Tales by Brothers Grimm
- 11. Micah Clarke
- 12. The Arabian Nights Entertainments - Volume 01
- 13. The Great Boer War
- 14. The Moonstone
- 15. The White Company

The fifteen novels are of varying sizes, ranging from about 8×10^5 bytes to about 3.3×10^6 bytes. Table 4 shows the error rates after denoising for Lempel-Ziv heuristic based denoisers under different models in the English text case. For the novel No.1 (*Don Quixote*), the work [16] has an error rate about 0.01655 after denoising, which is comparable to our BCT, but worse than BFP and GM.

IV.C DISCUSSION ON EXPERIMENTS FOR DENOISING

The three bi-directional models BFP, GM and BCT are tested via their application in noncausal universal discrete denoising, previous two of which are induced from sequential models, and the remaining one of which is a directly constructed bi-directional model. They perform generally better than DUDE, and they have comparable performances in the

binary Markov source and BSC case. Occasionally BFP performs worse than DUDE for the binary Markov source and BSC, but in general all the proposed models show a slight improvement in that experiment. Much bigger improvements over DUDE are observed in the case of corrupted English text. From the experiments we can see that none of the three proposed methods is uniformly the best. Using a Lempel-Ziv heuristic it is usually possible to choose the best modeler for each particular application.

REFERENCES

- [1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, “Universal discrete denoising: Known channel,” *IEEE Trans. Inform. Theory*, vol. 51, pp. 5–28, Jan. 2005.
- [2] N. Merhav and M. Feder, “Universal prediction,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2124–2147, Oct. 1998.
- [3] T. Weissman and N. Merhav, “Universal prediction of individual binary sequences in the presence of noise,” *IEEE Trans. Inform. Theory*, vol. 47, pp. 2151–2173, Sept. 2001.
- [4] T. Weissman and N. Merhav, “Universal prediction of random binary sequences in a noisy environment,” *Ann. Appl. Probab.*, vol. 14, no. No. 1, pp. 54–89, 2004.
- [5] E. Ordentlich, T. Weissman, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav, “Discrete universal filtering through incremental parsing,” in *Proc. IEEE Data Compression Conf.*, pp. 352–361, Mar. 2004.
- [6] J. Rissanen, “A universal data compression system,” *IEEE Trans. Inform. Theory*, vol. 29, pp. 656–664, Sept. 1983.
- [7] P. Bühlmann and A. J. Wyner, “Variable length markov chains,” *Ann. Statist.*, vol. 27, pp. 480–513, 1999.
- [8] M. Mächler and P. Bühlmann, “Variable length markov chains: Methodology, computing and software,” *Journal of Computational and Graphical Statistics*, pp. 435–455, June 2004.
- [9] F. Ferrari and A. J. Wyner, “Estimation of general stationary processes by variable length markov chains,” *Scandinavian Journal of Statistics*, pp. 459–480, Sept. 2003.
- [10] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, “The context tree weighting method: Basic properties,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.
- [11] E. Ordentlich, G. Seroussi, S. Verdú, K. Viswanathan, M. Weinberger, and T. Weissman, “Channel decoding of systematically encoded unknown redundant sources,” in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, p. 165, Jun./Jul. 2004.
- [12] G. Gemelos, S. Sigurjonsson, and T. Weissman, “Universal discrete denoising under channel uncertainty,” *Submitted to IEEE Trans. Inform. Theory*, 2004.
- [13] G. Gemelos, S. Sigurjonsson, and T. Weissman, “Algorithms for discrete denoising under channel uncertainty,” *Submitted to IEEE Trans. Signal Processing*, 2004.
- [14] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, Feb. 1989.
- [15] Y. Ephraim and N. Merhav, “Hidden markov processes,” *IEEE Trans. Inform. Theory*, vol. 48, no. Vol. IT-48 No. 6, pp. 1518–1569, 2002.
- [16] E. Ordentlich, M. J. Weiberger, and T. Weissman, “Efficient pruning of bi-directional context trees with applications to universal denoising and compression,” in *Proc. IEEE Information Theory Workshop*, Oct. 2004.
- [17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004.