Computable Bounds for Rate Distortion With Feed Forward for Stationary and Ergodic Sources

Iddo Naiss and Haim H. Permuter, *Member, IEEE*

***Abstract*—In this paper, we consider the rate distortion problem of discrete-time, ergodic, and stationary sources with feed forward at the receiver. We derive a sequence of achievable and computable rates that converge to the feed-forward rate distortion. We show that for ergodic and stationary sources, the rate**

$$R_n(D) = \frac{1}{n} \min I(\hat{X}^n \to X^n)$$

is achievable for any n, where the minimization is performed over the transition conditioning probability $p(\hat{x}^n|x^n)$ such that $\mathbb{E}\left[d(X^n, \hat{X}^n)\right] \leq D$. We also show that the limit of $R_n(D)$ exists and is the feed-forward rate distortion. We follow Gallager's proof where there is no feed forward and, with appropriate modification, obtain our result. We provide an algorithm for calculating $R_n(D)$ using the alternating minimization procedure and present several numerical examples. We also present a dual form for the optimization of $R_n(D)$ and transform it into a geometric programming problem.

***Index Terms*—Alternating minimization procedure, Blahut–Arimoto (BA) algorithm, causal conditioning, concatenating code trees, directed information, ergodic and stationary sources, ergodic modes, geometric programming (GP), rate distortion with feed forward.**

I. Introduction

THE rate distortion function for memoryless sources is well known and was given by Shannon in his seminal work [1]. Shannon [1] showed that the rate distortion function is the minimum of the mutual information between the source X and its reconstruction $\hat{X}$, where the minimization is over transition probabilities $p(\hat{x}|x)$ such that the distortion constraint is satisfied, i.e., $\mathbb{E}\left[d(X, \hat{X})\right] \leq D$. In the case where the source is stationary and ergodic, Gallager [2] showed that the rate distortion is the limit of the following sequence of rates. Each member of the sequence is the nth order rate distortion function, which is the solution of the following minimization problem:

$$\frac{1}{n} \min I(X^n; \hat{X}^n).$$

The minimization is over all conditional probabilities $p(\hat{x}^n|x^n)$ such that the distortion constraint is satisfied, i.e., $\mathbb{E}\left[d(X^n, \hat{X}^n)\right] \leq D$. Gallager showed that the limit of the sequence $\frac{1}{n} \min I(X^n; \hat{X}^n)$ exists and is equal to the infimum of the sequence.

The problem of source coding with feed forward was introduced by Weissman and Merhav [3] and by Venataramanan and Pradhan [4] and is depicted in Fig. 1. The encoder operates on X^n at once and sends a message to the decoder. The decoder, when reconstructing $\hat{X}_i$, has access to the message sent by the encoder and to all sources symbols with a delay $s \geq 1$, i.e., X^{i-s}.

Weissman and Merhav [3] named the problem competitive predictions. In their work, they defined a set of functions F_i that predict the following X_i given the previous X^{i-1}. After defining the *loss function*, $\rho : \mathcal{X} \to [0, \infty)$, between X_i and the prediction, the objective was to minimize the expected loss over all sets of predictors of size M. An important result in [3] is that in the case where the innovation process $W_i = X_i - F_i(X^{i-1})$ is i.i.d., the distortion rate with feed-forward function is the same as the distortion-rate function of W_i where there is no feed forward. In particular, if X_i is an i.i.d. process, then $W_i = X_i$, and thus, the distortion rate with feed forward for the source X_i is the same as if there is no feed forward.

Venkataramanan and Pradhan [4] gave an explicit definition of the feed-forward rate distortion for an arbitrary normalized distortion function and a general source. Their goal was to characterize the minimum rate R of a source given a distortion D using causal conditioning and directed information. The source of information is modeled as the process X_n and is encoded in blocks of length n into a message $T \in \{1, 2, \ldots, 2^{nR}\}$. The message T (after n time units) is sent to the decoder that has to reconstruct the process $\{X_n\}$ using the message T and causal information of the source with some delay s, as in Fig. 1.

For that purpose, Venkataramanan and Pradhan [4] defined the measures

$$\overline{I}(\hat{X} \to X) \triangleq \limsup_{\text{inprob}} \frac{1}{n} \log \frac{p(X^n, \hat{X}^n)}{p(\hat{X}^n \| X^{n-1}) p(X^n)}$$

and

$$\underline{I}(\hat{X} \to X) \triangleq \liminf_{\text{inprob}} \frac{1}{n} \log \frac{p(X^n, \hat{X}^n)}{p(\hat{X}^n \| X^{n-1}) p(X^n)}$$

where $p(\hat{x}^n \| x^{n-1})$, which is used throughout the paper, denotes the *causal conditioning probability*, and is given by

$$p(\hat{x}^n \| x^{n-1}) \triangleq \prod_{i=1}^{n} p(\hat{x}_i | \hat{x}^{i-1}, x^{i-1}). \tag{1}$$

Manuscript received June 05, 2011; revised July 10, 2012; accepted July 31, 2012. Date of publication October 03, 2012; date of current version January 16, 2013. This work was supported in part by the Marie Curie Reintegration Fellowship Program under U.S.–Israel Binational Science Foundation Grant 2008402 and in part by the German–Israel Foundation for Scientific Research and Development Grant 2275/2010.

I. Naiss was with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel. He is now with Samsung Electronics, Tel-Aviv 61131, Israel (e-mail: naiss@bgu.ac.il).

H. H. Permuter is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, 84105 Beer-Sheva, Israel (e-mail: haimp@bgu.ac.il).

Communicated by S. Diggavi, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2012.2222345

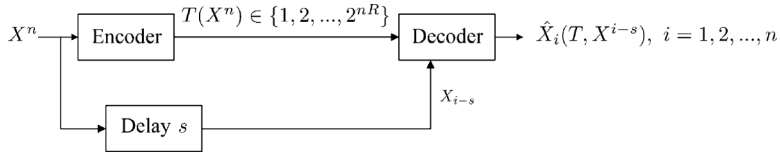

Fig. 1. Source coding with feed forward: the decoder knows the source with delay s and needs to reconstruct the source within the constraint $\mathbb{E}\left[d(X^n, \hat{X}^n)\right] \leq D$.

The limsup in probability of a sequence of random variables $\{X_n\}$ is defined as the smallest extended real number α such that $\forall \epsilon > 0$

$$\lim_{n\to\infty} \Pr[X_n \geq \alpha + \epsilon] = 0$$

and the liminf in probability is the largest extended real number β such that $\forall \epsilon > 0$

$$\lim_{n\to\infty} \Pr[X_n \leq \beta - \epsilon] = 0.$$

The main result in [4] is that for a general source $\{X_n\}$ and distortion D, the rate distortion with feed forward $R(D)$ is given by

$$R(D) = \inf_{\mathcal{P}} \overline{I}(\hat{X} \to X)$$

where the infimum is evaluated over the set $\mathcal{P}$ of probabilities $\{p(\hat{x}^n|x^n)\}_{n\geq 1}$ that satisfy the distortion constraint. Moreover, if

$$\overline{I}(\hat{X} \to X) = \underline{I}(\hat{X} \to X)$$

Venkataramanan and Pradhan showed in [4] that

$$R(D) = \inf_{\mathcal{P}} \lim_{n\to\infty} \frac{1}{n} I(\hat{X}^n \to X^n).$$

The work of Venkataramanan and Pradhan has made a significant contribution since it gives a multiletter characterization for the rate distortion function with feed forward. In [5], they evaluated these formulas for a stock market example and provided an analytical expression for the rate distortion function. However, these types of formulas are still very hard to evaluate for the general case. In this paper, we show that, assuming ergodicity and stationarity of the source, the rate distortion function with feed forward and delay $s = 1$ is upper bounded by $R_n(D)$, where

$$R_n(D) = \frac{1}{n} \min_{p(\hat{x}^n|x^n):\mathbb{E}\left[d(X^n,\hat{X}^n)\right]\leq D} I(\hat{X}^n \to X^n). \quad (2)$$

We further show that the limit of the sequence $\{R_n(D)\}$ exists, is equal to $\inf_n R_n(D)$, and is the feed-forward rate distortion function $R(D)$. These expressions for $R_n(D)$ are computable using a Blahut–Arimoto (BA)-type algorithm or using geometric programming (GP), as demonstrated here.

In most models with causal constraints, such as feedback channels or feed-forward rate distortion, the causal conditioning probability, as well as the directed information, characterizes the fundamental limits. In order to address these models, the causal conditioning probability was introduced by Massey [6] and Kramer [7] and is defined in (1). The difference between regular and causal conditioning is that in causal conditioning, the dependence of $\hat{x}_i$ on future x_j is not taken into account. Following the causal conditioning probability, Massey [6] (who was inspired by Marko's work [8] on bidirectional communication) introduced the directed information, defined as

$$\begin{aligned} I(\hat{X}^n \to X^n) &\stackrel{\triangle}{=} H(X^n) - H(X^n\|\hat{X}^n) \\ &= \sum_{i=1}^{n} I(\hat{X}^i; X_i|X^{i-1}) \end{aligned}$$

where $H(X^n\|\hat{X}^n)$ is the *causal conditioning entropy* [7], which is defined as

$$H(X^n\|\hat{X}^n) \stackrel{\triangle}{=} \sum_{i=1}^{n} H(X_i|X^{i-1}, \hat{X}^i).$$

The directed information was used by Tatikonda and Mitter [9], Permuter *et al*. [10], and Kim [11] to characterize the point-to-point channel capacity with feedback. It is shown that the capacity of such channels is characterized by the maximization of the directed information over the causal input probability $p(x^n\|y^{n-1})$. In a previous paper [12], we used these results and obtained bounds to estimate the feedback channel capacity using a BA-type algorithm for finding the global maximum of the directed information.

The main contribution of this study lies in extending the achievability proof given by Gallager in [2] to the case where feed forward with delay $s = 1$ exists. The extension is done by using code trees rather than codewords and the causal conditioning distribution, $p(\hat{x}^n\|x^{n-s})$, defined by

$$p(\hat{x}^n\|x^{n-s}) \stackrel{\triangle}{=} \prod_{i=1}^{n} p(\hat{x}_i|\hat{x}^{i-1}, x^{i-s})$$

rather than the regular reconstruction distribution $p(\hat{x}^n)$ in order to construct the code trees. The proof given is for $s = 1$, but can be extended straightforwardly to any delay $s \geq 1$. The difficulty in this modification is that while in [2] the codebook was an ensemble of sequences (code words) from the reconstruction alphabet using $p(\hat{x}^n)$, our codebook is an ensemble of code trees using $p(\hat{x}^n\|x^{n-s})$. This induces a major problem while showing that the probability of error is small, as discussed in Section III. These difficulties are overcome by appropriate modification of Gallager's proofs. We note that we provide only a sequence of upper bounds on $R(D)$, and hence do not know how the convergence of the sequence behaves. However, in Section VIII, where numerical examples are presented, we can see that for $n = 12$, we have a good estimation, i.e., $R_{12}(D)$ is close up to third digit after the point to $R(D)$ that is known analytically.

Another contribution of this paper is the development of two optimization methods for obtaining $R_n(D)$; a BA-type algorithm which is easy to implement and a GP form. The motivation of the GP form is to provide a maximization problem [13, Ch. 4.5] which can be solved using convex optimization packages such as CVX [14]. Furthermore, this GP formulation via its Lagrangian duality gives us a lower bound to the nth order rate distortion with feed forward, which helps us decide when to terminate our BA-type algorithm, i.e., Algorithm 1. Another important advantage of the GP form over BA-type algorithm is that its input is D and the output is R_n, while in the BA-type algorithm, the input is the slope of the curve $R_n(D)$ as explained in Appendix D and the output is R_n and D.

Csiszar and Korner [15] provided a convergence proof for BA algorithm that is specific for the channel capacity and rate distortion problem, based on the properties of the divergence. In this paper, we suggest a different approach based on the well-known alternating minimization procedure [16] and adopt it in our feed-forward rate distortion problem.

The remainder of this paper is organized as follows. In Section II, we describe the problem model, provide the operational definition of the rate distortion function with feed forward, and state our main theorems. In Section III, we show that $R_n(D)$ is an achievable rate for all n and any distortion D and in Section IV, we show that the limit of $R_n(D)$ exists and is equal to the operational rate distortion function. In Section V, we give a description of the BA-type algorithm for calculating $R_n(D)$ and present the algorithm's complexity and the memory required and in Section VI, we derive the BA-type algorithm and prove its convergence to an optimum value. In Section VII, we present an alternative optimization problem for $R_n(D)$ in a standard GP form that can be solved numerically using convex optimization tools. Numerical examples are given in Section VIII to illustrate the performance of the suggested algorithms.

II. Problem Statement and Main Results

In this section, we present notation, describe the problem model, and summarize the main results of this paper. We first state the definitions of a few quantities that we use in our coding theorems. We denote by X_1^n the vector $(X_1, X_2, \dots X_n)$. Usually, we use the notation $X^n = X_1^n$ for short. Further, when writing a probability mass function (PMF) we simply write $P_X(X = x) = p(x)$. An alphabet of any type is denoted by a calligraphic letter $\mathcal{X}$, and its size is denoted by $|\mathcal{X}|$.

In the rate distortion problem with feed forward of delay $s = 1$, as shown in Fig. 1, we consider a general discrete, stationary, and ergodic source $\{X_n\}$, with the nth order probability distribution $p(x^n)$, alphabet $\mathcal{X}$, and reconstruction alphabet $\hat{\mathcal{X}}$. The normalized bounded distortion measure is defined as $d : \mathcal{X}^n \times \hat{\mathcal{X}}^n \to \mathbb{R}^+$ on pairs of sequences. Normalization is done by dividing the nth distortion by n, i.e.,

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d_i(x_{i-m}^i, \hat{x}_i) \tag{3}$$

where m is fixed and determined by the problem settings. By bounded, we mean that there exists a constant number M such that for every n and every $(x_{i-m}^i, \hat{x}_i)$, $d(x_{i-m}^i, \hat{x}_i) \le M$.

Definition 1 (Code Definition): A $(n, 2^{nR}, D)$ source code with feed forward of block length n and rate R consists of an encoder mapping f, given by

$$f : \mathcal{X}^n \mapsto \{1, 2, \dots, 2^{nR}\}$$

and a sequence of decoder mappings $g_i, i = 1, 2, \dots, n$, where

$$g_i : \{1, 2, \dots, 2^{nR}\} \times \mathcal{X}^{i-1} \mapsto \hat{\mathcal{X}},\ i = 1, 2, \dots, n. \tag{4}$$

The encoder maps a sequence x^n to an index in $\{1, 2, \dots, 2^{nR}\}$. At time i, the decoder has the message that was sent and causal information of the source x^{i-1}, and reconstructs the ith symbol sent, $\hat{x}_i$.

Definition 2 (Achievable Rate): A rate distortion with feed-forward pair (R, D) is achievable if there exists a sequence of $(n, 2^{nR}, D)$ rate distortion codes with

$$\limsup_{n\to\infty} \mathbb{E}\left[d(X^n, \hat{X}^n)\right] \le D.$$

Definition 3 (Operational Definition of Rate Distortion.): The operational rate distortion with feed-forward function $R(D)$ is the infimum of rates R such that (R, D) is achievable.

In this paper, we define the mathematical expression for the rate distortion function as the following limit:

$$R^{(I)}(D) = \lim_{n\to\infty} R_n(D) \tag{5}$$

where $R_n(D)$ is the nth order rate distortion function, given by

$$R_n(D) = \frac{1}{n} \min_{p(\hat{x}^n|x^n):\mathbb{E}\left[d(X^n,\hat{X}^n)\right]\le D} I(\hat{X}^n \to X^n). \tag{6}$$

We show that the limit in (5) exists, $R_n(D)$ is achievable and upper bounds $R^{(I)}(D)$ for all n. Further, we show that the feed-forward rate distortion function $R(D)$ is equal to $R^{(I)}(D)$. We also provide two ways to calculate numerically the value $R_n(D)$: using a BA-type algorithm and using a GP form.

We now state our main theorems.

Theorem 1 (Achievability of $R_n(D)$): For a discrete, stationary, ergodic source, and for any distortion D, any n, and delay $s = 1$, $R_n(D)$ is an achievable rate.

Theorem 2 (Feed-Forward Rate Distortion): For any distortion D, the operational rate distortion function $R(D)$ is equal to the mathematical expression $R^{(I)}(D)$ given in (5).

Theorem 3 (Algorithm for Calculating $R_n(D)$): For a fixed source distribution $p(x^n)$, there exists an alternating minimization procedure in order to compute $R_n(D)$ as in (6).

Theorem 4 (GP Form of $R_n(D)$): The nth order rate distortion function $R_n(D)$ is equal to the solution of the following GP standard form:

$$\max_{\lambda,\gamma(x^n),\{p'(x_i|x^{i-1},\hat{x}^i)\}_{i=1}^n} \frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n)\log\gamma(x^n)\right)$$

subject to the constraints

$$\begin{aligned}&\log(p(x^n)) + \log(\gamma(x^n)) - \lambda d(x^n, \hat{x}^n) \\ &\quad - \sum_{i=1}^{n} \log p'(x_i|x^{i-1}, \hat{x}^i) \le 0, \quad \forall\, x^n, \hat{x}^n \\ &\sum_{x_i} p'(x_i|x^{i-1}, \hat{x}^i) = 1, \quad \forall\, i, \forall\, x^{i-1}, \hat{x}^{i-1} \\ &\lambda \ge 0.\end{aligned}$$

Proofs of Theorems 1 and 2 are given in Sections III and IV, respectively. The algorithm in Theorem 3 is described in Section V and proved in Section VI and the proof for Theorem 4 is found in Section VII.

III. Achievability Proof (Theorem 1)

In this section, we follow the proof from [2] and show that if the source is stationary and ergodic, then $R_n(D)$, as given in (6), is achievable for any n. The main modification to Gallager's work made in this paper is the codebook construction. While in Gallager's book a codeword is a single stream of symbols, here we construct a code tree that depends on the delayed input. Hence, the proof of the probability of error changes as well.

The proof is developed as follows. First, we assume that the source is ergodic in blocks of length n, and show achievability. A source that is ergodic in blocks is one that, by looking at each n letters as a single letter from a super alphabet, we obtain an ergodic super source (presented in [2, Ch. 9.8]). Next, for the general ergodic sources, we follow a result given in [2] about ergodic modes, as explained further on. The distortion is assumed to be normalized, finite, and bounded as in (3). An example for such a distortion can be found in [5] and in Section VIII in an example called the stock market.

Theorem 5: Consider a discrete stationary source that is ergodic in blocks of length n. For any distortion D, such that $R_n(D) < \infty$ and $\delta > 0$, and for any L sufficiently large, there exists a codebook of trees $\mathcal{T}_C$ of length L with $|\mathcal{T}_C| \le 2^{L(R_n(D)+\delta)}$ code trees for which the average distortion per letter satisfies $\mathbb{E}\left[d(X^L, \hat{X}^L)\right] \le D + \delta$.

Proof: Let $p(\hat{x}^n|x^n)$ be the transition probability that achieves the minimum $R_n(D)$ and let $p(\hat{x}^n\|x^{n-1})$ be the causal conditioning probability that corresponds to $p(x^n)p(\hat{x}^n|x^n)$.

1) *Code design:* For any L, consider the *ensemble of codes* $\mathcal{T}_C$ with $|\mathcal{T}_C| = \lfloor 2^{L(R_n(D)+\delta)} \rfloor$ code trees of length L, where each code tree $\tau^L \in \mathcal{T}_C$ is a concatenation of L/n subcode trees of length n. Each subcode tree is generated independently according to $p(\hat{x}^n\|x^{n-1})$, as in Fig. 2.
2) *Encoder:* The encoder assigns a code tree $\tau^L \in \mathcal{T}_C$ for every x^L such that $d(x^L, \hat{x}^L(\tau^L, x^{L-1}))$ is minimal. The sequence $\hat{x}^L(\tau^L, x^{L-1})$ is determined by walking on tree τ^L and following the branch x^{L-1}.
3) *Decoder:* At time i, the decoder possesses the index of the tree τ^L and causal information of the source x^{i-1}, and returns the symbol $\hat{x}_i(\tau^L, x^{i-1})$ that it produces.

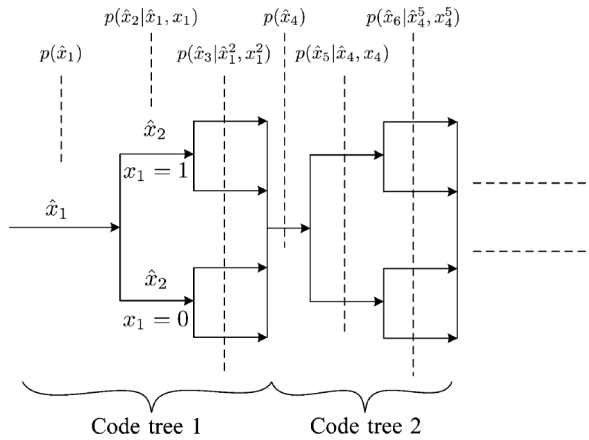

Fig. 2. Concatenation of two code trees for a binary alphabet, each of length $n = 3$. The upper branches are for $x_i = 1$, and the lower branches are for $x_i = 0$.

Let us define the *test channel ensemble* as the conditional probability

$$p_L(\hat{x}^L|x^L) = \prod_{i=0}^{L/n-1} p(\hat{x}_{ni+1}^{ni+n}|x_{ni+1}^{ni+n}) \tag{7}$$

where each $p(\hat{x}_{ni+1}^{ni+n}|x_{ni+1}^{ni+n})$ achieves $R_n(D)$. We also define the causal conditional probability

$$p_L(\hat{x}^L\|x^{L-1}) = \prod_{i=0}^{L/n-1} p(\hat{x}_{ni+1}^{ni+n}\|x_{ni+1}^{ni+n-1})$$

where the distribution is according to

$$\begin{aligned}P_{\hat{X}_{ni+1}^{ni+n}|X_{ni+1}^{ni+n}}(\hat{x}^n|x^n) &= P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) \\ P_{\hat{X}_{ni+1}^{ni+n}\|X_{ni+1}^{ni+n-1}}(\hat{x}^n\|x^{n-1}) &= P_{\hat{X}^n\|X^{n-1}}(\hat{x}^n\|x^{n-1}).\end{aligned}$$

Moreover, we define for every code tree τ^L of length L the measure

$$I_n(\tau^L \to x^L) = \log \frac{p_L(\hat{x}^L|x^L)}{p_L(\hat{x}^L||x^{L-1})} \tag{8}$$

where $\hat{x}^L = \hat{x}^L(\tau^L, x^{L-1})$. Note that $I_n(\tau^L \to x^L)$ is not the directed information between the sequences $\hat{x}^L$, x^L, but simply a measure between a source sequence x^L and the output $\hat{x}^L$ of the test channel $p_L(\hat{x}^L|x^L)$, as defined in (7).

Let $\mathcal{T}$ be the set of all code trees of length L, and consider the following set:

$$\begin{aligned}\mathcal{A} = \{&\tau^L \in \mathcal{T}, x^L \in \mathcal{X}^L : \text{either} \\ &I_n(\tau^L \to x^L) > L(R_n(D) + \delta/2) \quad \text{or} \\ &d(x^L, \hat{x}^L(\tau^L, x^{L-1})) > L(D + \delta/2)\}\end{aligned} \tag{9}$$

and let $p_t(\mathcal{A})$ be the probability of the set $\mathcal{A}$ over the test channel ensemble, as defined previously, constructed as in (7).

Let us use the notation

$$\hat{x}^L(\mathcal{T}_C, x^{L-1}) = \hat{x}^L\left(\arg\min_{\tau^L \in \mathcal{T}_C} d\big(x^L, \hat{x}^L(\tau^L, x^{L-1})\big), x^{L-1}\right)$$

where $\mathcal{T}_C$ is the ensemble of code trees, as described in the coding scheme. Now, let $p_c(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD)$ be the probability over the ensemble of codes $\mathcal{T}_C$ and source sequences such that the distortion exceeds LD. We wish to give an upper bound to the probability $p_c(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD)$; for this, we use the following lemma.

Lemma 1: For a given source, $\{X_i\}_{i\geq 1}$, and test channel, we have the following inequality:

$$\begin{aligned} &p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD\right)\\ &\quad\leq p_t(\mathcal{A})+\exp\{-|\mathcal{T}_C|2^{-LR_n(D)}\} \end{aligned} \tag{10}$$

where the set $\mathcal{A}$ is described in (9).

Proof: We first write $p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD\right)$ as

$$\begin{aligned} &p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD\right)\\ &=\sum_{x^L\in\mathcal{X}^L}p(x^L)p_c\big(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD|X^L=x^L\big)\,. \end{aligned}$$

For every x^L, let us define the set $\mathcal{A}_{x^L}$ as the set of all code trees $\tau^L\in\mathcal{T}$ for which $(\tau^L,x^L)\in\mathcal{A}$, i.e.,

$$\begin{aligned} \mathcal{A}_{x^L}=\{&\tau^L\in\mathcal{T}:\text{either}\\ &I_n(\tau^L\to x^L)>L(R_n(D)+\delta/2)\quad\text{or}\\ &d(x^L,\hat{x}^L(\tau^L,x^{L-1}))>L(D+\delta/2)\}. \end{aligned} \tag{11}$$

We observe that $d(x^L,\hat{x}^L(\mathcal{T}_C,x^{L-1}))>LD$ for a given x^L only if $d(x^L,\hat{x}^L(\tau^L,x^{L-1}))>LD$ for every $\tau^L\in\mathcal{T}_C$. Thus, $d(x^L,\hat{x}^L(\mathcal{T}_C,x^{L-1}))>LD$ only if $\tau^L\in\mathcal{A}_{x^L}$ for every $\tau^L\in\mathcal{T}_C$. Since τ^L are independently chosen

$$\begin{aligned} &p_c\Big(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD|X^L=x^L\Big)\\ &\qquad\leq\left(p_t(\mathcal{A}_{x^L})\right)^{|\mathcal{T}_C|}\\ &\qquad=\left(1-p_t(\mathcal{A}^c_{x^L})\right)^{|\mathcal{T}_C|} \end{aligned}$$

where $\mathcal{A}^c_{x^L}$ is the complement set of $\mathcal{A}_{x^L}$. We note that the probability of tree τ^L being in $\mathcal{A}^c_{x^L}$ depends only on the branch associated with x^L. In other words, if a tree $\tau^L\in\mathcal{A}^c_{x^L}$, then all other trees with the same branch associated with x^L is in $\mathcal{A}^c_{x^L}$ as well; the same goes for $\mathcal{A}_{x^L}$. Hence, we can divide the set of all code trees $\mathcal{T}$ into disjoint subsets, $B_{x^L,\hat{x}^L}$, that have the same branch associated with x^{L-1}, i.e.,

$$B_{x^L,\hat{x}^L}=\{\tau^L\in\mathcal{T}:\tau^L(x^{L-1})=\hat{x}^L\}$$

where $\tau^L(x^{L-1})$ is a walk on tree τ^L over the branch x^{L-1}. Clearly, the probability of each subset, $B_{x^L,\hat{x}^L}$, is

$$p_t(B_{x^L,\hat{x}^L})=p_L(\hat{x}^L\|x^{L-1})$$

since the left-hand side is a summation of the probabilities of all trees with the same branch associated with x^L, and we are left with the probability of that one branch.

Now, for every $\tau^L\in B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}$ and due to the definition of $\mathcal{A}^c_{x^L}$, we have

$$I_n(\tau^L\to x^L)=\log\frac{p_L(\hat{x}^L|x^L)}{p_L(\hat{x}^L||x^{L-1})}\leq LR_n(D).$$

Therefore

$$p_L(\hat{x}^L\|x^{L-1})\geq p_L(\hat{x}^L|x^L)2^{-LR_n(D)} \tag{12}$$

and we obtain that

$$\begin{aligned} &p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD|X^L=x^L\right)\\ &\quad\leq\left(1-p_t(\mathcal{A}^c_{x^L})\right)^{|\mathcal{T}_C|}\\ &\quad=\left(1-\sum_{B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p_t(B_{x^L,\hat{x}^L})\right)^{|\mathcal{T}_C|}\\ &\quad=\left(1-\sum_{\hat{x^L}:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p_L(\hat{x}^L\|x^{L-1})\right)^{|\mathcal{T}_C|}\\ &\quad\overset{(a)}{\leq}\left(1-2^{-LR_n(D)}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p_L(\hat{x}^L|x^L)\right)^{|\mathcal{T}_C|} \end{aligned}$$

where (a) follows the inequality in (12).

We now use the inequality $(1-ab)^k\leq 1-a+\exp\{-bk\}$, which is true for all $0\leq a\leq 1$ [2, eq. (9.3.22) and (9.3.23)]. Hence, taking $a=\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p_L(\hat{x}^L|x^L)$, $b=2^{-LR_n(D)}$, we find

$$\begin{aligned} &p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD|X^L=x^L\right)\\ &\quad\leq 1-\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p_L(\hat{x}^L|x^L)\\ &\qquad+\exp\{-|\mathcal{T}_C|2^{-LR_n(D)}\}. \end{aligned} \tag{13}$$

By taking a sum over x^L, we remain with

$$\begin{aligned} &p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD\right)\\ &\quad=\sum_{x^L}p(x^L)p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD|X^L=x^L\right)\\ &\quad\overset{(a)}{\leq}\sum_{x^L}p(x^L)\Bigg(1-\\ &\qquad\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p_L(\hat{x}^L|x^L)+\exp\{-|\mathcal{T}_C|2^{-LR_n(D)}\}\Bigg)\\ &\quad=1-\sum_{x^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}p(x^L,\hat{x}^L)\\ &\qquad+\exp\{-|\mathcal{T}_C|2^{-LR_n(D)}\} \end{aligned} \tag{14}$$

where (a) is due to (13). Note that

$$\begin{aligned}
&\sum_{x^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}} p(x^L,\hat{x}^L)\\
&\quad=\sum_{x^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}\sum_{\tau^L\in\mathcal{T}} p(x^L,\hat{x}^L,\tau^L)\\
&\quad\geq\sum_{x^L}\sum_{\hat{x}^L:B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}\sum_{\tau^L\in B_{x^L,\hat{x}^L}} p(x^L,\hat{x}^L,\tau^L)\\
&\quad\overset{(a)}{=}\sum_{x^L}\sum_{B_{x^L,\hat{x}^L}\subset\mathcal{A}^c_{x^L}}\sum_{\tau^L\in B_{x^L,\hat{x}^L}} p(x^L,\tau^L)\\
&\quad=\sum_{x^L}\sum_{\tau^L\in\mathcal{A}^c_{x^L}} p(x^L,\tau^L)\\
&\quad=p_t(\mathcal{A}^c)
\end{aligned}$$

where (a) follows the fact that if $\tau^L\in B_{x^L,\hat{x}^L}$, then $\hat{x}^L$ is determined by the tree τ^L and the branch x^L. Now, continuing from (14), we obtain

$$\begin{aligned}
&p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>LD\right)\\
&\quad\leq 1-p_t(\mathcal{A}^c)+\exp\{-|\mathcal{T}_C|2^{-LR_n(D)}\}\\
&\quad=p_t(\mathcal{A})+\exp\{-|\mathcal{T}_C|2^{-LR_n(D)}\}. \qquad (15)
\end{aligned}$$

■

We now use the result in (15) in order to complete the proof of the theorem. Furthermore, we can see that the average distortion of the code satisfies

$$\begin{aligned}
\mathbb{E}\left[d(X^L,\hat{X}^L)\right]&\leq(D+\delta/2)+\sup_{x^L,\hat{x}^L} d(x^L,\hat{x}^L)\\
&\quad\times p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>L(D+\delta/2)\right).
\end{aligned}$$

This arises, as in [2, Th. 9.3.1], from upper bounding the distortion by $D+\delta/2$ when the distortion $d(x^L,\hat{x}^L)\leq D+\delta/2$, and by

$$\sup_{x^L,\hat{x}^L} d(x^L,\hat{x}^L)$$

otherwise. By choosing $|\mathcal{T}_C|=\lfloor 2^{L(R_n(D)+\delta)}\rfloor$, the last term in (15) goes to zero with increasing L. Furthermore, the first term is bounded by

$$\begin{aligned}
p_t(\mathcal{A})\leq\ &p_t\left\{x^L\in\mathcal{X}^L,\ \tau^L\in\mathcal{T}:\right.\\
&\qquad\left.I_n(\tau^L\to x^L)>L(R_n(D)+\delta/2)\right\}\\
&+p_t\{x^L\in\mathcal{X}^L,\tau^L\in\mathcal{T}:\\
&\qquad d(x^L,\hat{x}^L(\tau^L,x^{L-1}))>L(D+\delta/2)\}. \qquad (16)
\end{aligned}$$

Note that

$$\begin{aligned}
&p_t\left(I_n(\tau^L\to x^L)>L(\frac{1}{n}R_n(D)+\delta/2)\right)\\
&=p_t\left(\frac{1}{L}\sum_{i=1}^{L/n-1}\log\frac{p(\hat{x}_{ni+1}^{ni+n}|x_{ni+1}^{ni+n})}{p(\hat{x}_{ni+1}^{ni+n}||x_{ni+1}^{ni+n-1})}>R_n(D)+\delta/2\right).
\end{aligned}$$

As assumed, the source is ergodic in blocks of length n. Furthermore, the test channel is defined to be memoryless for blocks of length n, and hence, the joint process is ergodic in blocks of length n. Thus, with probability 1

$$\begin{aligned}
\frac{1}{n}\lim_{L\to\infty}\frac{1}{L/n}\sum_{i=0}^{L/n-1}\log\frac{p(\hat{x}_{ni+1}^{ni+n}|x_{ni+1}^{ni+n})}{p(\hat{x}_{ni+1}^{ni+n}||x_{ni+1}^{ni+n-1})}&\\
=\frac{1}{n}\mathbb{E}\left[\log\frac{p(\hat{x}^n|x^n)}{p(\hat{x}^n\|x^{n-1})}\right]&\\
=R_n(D).&
\end{aligned}$$

Therefore, the probability of the first term in (16) goes to zero as L goes to infinity and the same goes for the second term due to the definition of the distortion. In order to finish the proof, and due to the fact that p_c goes to zero with increasing L and the fact that the distortion is finite, we can choose L large enough such that

$$\begin{aligned}
&p_c\left(d(X^L,\hat{x}^L(\mathcal{T}_C,X^{L-1}))>L(D+\delta/2)\right)\cdot\sup_{x^L,\hat{x}^L} d(x^L,\hat{x}^L)\\
&\quad\leq\delta/2.
\end{aligned}$$

In this case, we obtain $D_L\leq D+\delta$, and hence, the rate $R_n(D)$ is achievable for sources that are ergodic in blocks of length n. ■

Much like in Gallager's proof for the case where there is no feed forward, we note that not all ergodic sources are also ergodic in blocks and we need to address these cases as well. For that purpose, we need [2, Lemma 9.8.2] for ergodic sources. We recall that a discrete stationary source is ergodic if and only if every invariant set of sequences under a shift operator T is of probability 1 or 0. In [2, Ch. 9.8], the author looks at the operator T^n, i.e., a shift of n places, and considers an invariant set S_0, $p(S_0)>0$, with respect to T^n. In [2, Lemma 9.8.2], it is stated that one can separate the source S to n' invariant subsets, $\{S_i=T^i(S_0)\}_{i=0}^{n'-1}$, $p(S_i)=\frac{1}{n'}$, with regard to T^n such that n' divides n and the sets S_i, S_j are disjoint (except, perhaps, for an intersection of zero probability). These subsets are called *ergodic modes*, due to the fact that each invariant subset of them under the operator T^n is of probability 0 or $\frac{1}{n'}$. In other words, conditional on an ergodic mode S_i, each invariant subset of it with respect to T^n is of probability 0 or 1.

Recall that by definition

$$R_n(D)=\frac{1}{n}I_n(\hat{X}^n\to X^n)$$

where the right-hand side is the average directed information between the source and the reconstruction, determined according to $p(\hat{x}^n|x^n)$, which achieves $R_n(D)$. Let $I_n(\hat{X}^n\to X^n|i)$ be the average directed information between a source sequence from the ith ergodic mode and the ensemble of codes, using the probability $p(\hat{x}^n|x^n)$ that achieves $R_n(D)$. Note that the directed information can be written as

$$\begin{aligned}
I_n(\hat{X}^n\to X^n)&=\sum_{x^n,\hat{x}^n}p(x^n)p(\hat{x}^n|x^n)\log\frac{p(\hat{x}^n|x^n)}{p(\hat{x}^n\|x^{n-1})}\\
&=\sum_{x^n,\hat{x}^n}p(x^n)p(\hat{x}^n|x^n)\log\frac{p(\hat{x}^n|x^n)p(x^n)}{p(\hat{x}^n\|x^{n-1})p(x^n)}\\
&=D\left(p(x^n)p(\hat{x}^n|x^n)||p(\hat{x}^n||x^{n-1})p(x^n)\right)
\end{aligned}$$

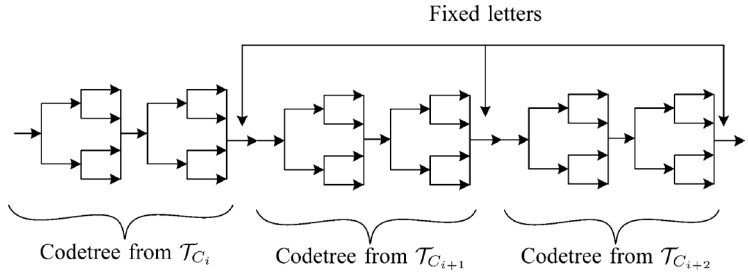

Fig. 3. Code tree from the ith codebook, $n = n' = 3$, $L = 6$.

which is convex over the input probability $p(x^n)$. Thus

$$I_n(\hat{X}^n \to X^n) \geq \frac{1}{n'} \sum_{i=0}^{n'-1} I_n(\hat{X}^n \to X^n|i). \qquad (17)$$

We now present a few observations.

1) First, we observe that $\frac{1}{n} I_n(\hat{X}^n \to X^n|i)$ is an upper bound to the nth order rate distortion function conditional on the ith ergodic mode. Hence, from Theorem 5, we know that there exists a codebook $\mathcal{T}_{C_i}$ with $|\mathcal{T}_{C_i}| = \lfloor 2^{L(\frac{1}{n} I_n(\hat{X}^n \to X^n|i)+\delta)} \rfloor$ code trees of length L such that the average distortion constraint holds.
2) Second, if a codebook $\mathcal{T}_{C_i}$ satisfies the distortion constraint, conditional on the ergodic mode S_i, then it has the same effect when conditional on the ergodic mode $T(S_{i-1})$.

Observation 2) means that we can use the codebook $\mathcal{T}_{C_{i-1}}$ to encode not only a source sequence from S_{i-i}, but also a shift of the source sequence in S_{i-1} with $\mathcal{T}_{C_i}$, since $T(S_{i-1}) \subseteq S_i$.

Using the aforementioned observations, we can now prove Theorem 1, i.e., the achievability of $R_n(D)$, where the source is ergodic and stationary. An equivalent version of Theorem 1 is the following: let $R_n(D)$ be the nth order rate distortion function for a discrete, stationary, and ergodic source. For any D such that $R_n(D) < \infty$ and $\delta > 0$ and any L sufficiently large, there exists a codebook of trees $\mathcal{T}_C$ of length L with $|\mathcal{T}_C| \leq 2^{L(R_n(D)+\delta)}$ code trees for which the average distortion per letter satisfies $\mathbb{E}\left[d(X^n, \hat{X}^n\right] \leq D + \delta$.

Proof of Theorem 1: Let $p(\hat{x}^n|x^n)$ be the transition probability that achieves $R_n(D)$ and let $p(\hat{x}^n \| x^{n-1})$ be the causal conditioning probability that corresponds to $p(x^n)p(\hat{x}^n|x^n)$.

1) *Code design:* For any L and any ergodic mode S_i, $0 \leq i \leq n'$, construct an ensemble of codes $\mathcal{T}_{C_i}$ with $|\mathcal{T}_{C_i}| = \lfloor 2^{L(\frac{1}{n} I_n(\hat{X}^n \to X^n|i)+\delta)} \rfloor$ "little" code trees of length L, where each "little" code tree is generated according to $p(\hat{x}^L \| x^{L-1})$, as in Fig. 2 in Theorem 5 above. Now, for every $0 \leq i \leq n'$, the ith codebook is an ensemble of "big" code trees. These are a concatenation of n' "little" code trees, starting from one in $\mathcal{T}_{C_i}$, and followed by one from $\mathcal{T}_{C_{i+1}}$ through to one from $\mathcal{T}_{C_{n'+i-1}}$, where the index is calculated modiolus n'. In the example of a "big" code tree in Fig. 3, we see additional letters at the end of each "little" code tree, i.e., in positions $L+1,\ 2(L+1), \ldots, n'(L+1)$, which are fixed. The purpose of the fixed letters is to shift the sequence and encode it with a code tree from the sequential codebook. Note that the overall length of a code tree sums up to $L' = Ln' + n'$.
2) *Encoder:* For every i, the encoder assigns for every source sequence $x^{L'} \in S_i$, a code tree $\tau^{L'}$ from the ith codebook such that $d(x^{L'}, \hat{x}^{L'}(\tau^{L'}, x^{L'-1}))$ is minimal. The sequence $\hat{x}^{L'}(\tau^{L'}, x^{L'-1})$ is determined by walking on tree $\tau^{L'}$ and following the branch $x^{L'-1}$.
3) *Decoder:* The decoder receives a tree $\tau^{L'}$ and causal information of $x^{L'}$ and returns the sequence $\hat{x}^{L'}$ that it produces.

Since the distortion constraint for every ergodic mode is satisfied, due to Theorem 5, the overall distortion is satisfied as well. The additional fixed letters are of unknown distortion but, due to the fact that the distortion is bounded, their contribution is negligible for large values of L. Moreover, note that for every i, the ith codebook is of the same size. Thus, the overall size of the codebook is

$$\begin{aligned}
|\mathcal{T}_C| &= n' \prod_{i=0}^{n'-1} |\mathcal{T}_{C_i}| \\
&\overset{(a)}{\leq} n' \prod_{i=0}^{n'-1} 2^{L(\frac{1}{n} I_n(\hat{X}^n \to X^n|i)+\delta)} \\
&= 2^{L(\frac{1}{n}\sum_{i=0}^{n'-1} I_n(\hat{X}^n \to X^n|i)+n'\delta+\frac{\log(n')}{L})} \\
&\overset{(b)}{\leq} 2^{L(\frac{n'}{n} I_n(\hat{X}^n \to X^n)+n'\delta+\frac{\log(n')}{L})} \\
&\overset{(c)}{=} 2^{Ln'(R_n(D)+\delta+\frac{\log(n')}{Ln'})} \\
&\overset{(d)}{=} 2^{(Ln'+n')(R_n(D)+\delta+\frac{\log(n')}{Ln'})}
\end{aligned}$$

where (a) is due to the fact that the codebook $\mathcal{T}_{C_i}$ is bounded by $2^{L(\frac{1}{n} I_n(\hat{X}^n \to X^n|i)+\delta)}$, as shown in Theorem 5, (b) follows (17), (c) is due to the fact that we use a conditional probability $p(\hat{x}^n|x^n)$ that achieves $R_n(D)$, and (d) is a simple algebraic manipulation. Recall that $L' = Ln' + n'$, so that by letting $\delta' = \delta + \frac{\log(n')}{Ln'}$, we conclude that $R_n(D)$ is an achievable rate for the general ergodic source, as required. ■

IV. Proof That $R(D) = R^{(I)}(D)$ (Theorem 2)

In this section, we show that the operational description of the rate distortion with feed forward is equal to the mathematical one given in (18). This will be done by, first, showing that the mathematical expression $R^{(I)}(D)$ is achievable and, second, showing that it is a lower bound to the rate distortion function. We recall that $R^{(I)}(D)$ is the solution to

$$\lim_{n\to\infty} \frac{1}{n} \min_{p(\hat{x}^n|x^n):\mathbb{E}\left[d(X^n,\hat{X}^n)\right]\leq D} I(\hat{X}^n \to X^n). \qquad (18)$$

To show that $R^{(I)}(D)$ is achievable, we first need to show that the limit of the sequence $\{R_n(D)\}$ exists. For this purpose, we use the following lemma.

Lemma 2: The sequence $R_n(D)$, as defined in (6), is subadditive, and thus

$$\inf_n R_n(D) = \lim_{n\to\infty} R_n(D).$$

Note that a sequence $\{a_n\}$ is called subadditive if for all m, l

$$(m+l)a_{m+l} \le ma_m + la_l.$$

The proof for Lemma 2 is given in Appendix A.

We now state a lemma for the achievability of $R^{(I)}(D)$.

Lemma 3 (Achievability of $R^{(I)}(D)$): The mathematical expression for the feed-forward rate distortion,$R^{(I)}(D)$ is achievable and thus upper bounds $R(D)$.

Proof: We showed in Theorem 1 that for any n, $R_n(D)$ is achievable. Further, in Lemma 2, we show that the limit exists and is equal to the infimum and hence is achievable too. Therefore, we conclude that the mathematical expression $R^{(I)}(D)$ is achievable and forms an upper bound to the operational description $R(D)$. ■

To show that $R^{(I)}(D)$ is a lower bound to the rate distortion function, we provide the following lemma.

Lemma 4 (Converse): The mathematical expression $R^{(I)}(D)$ is a lower bound to the operational rate distortion function.

For the completeness of this paper, we provide the proof of Lemma 4 in Appendix B. However, similar proof was presented by Venkataramanan and Pradhan in [4]. Their expressions involved the limit in probability of the entropy and directed information, as described in Section I.

Proof of Theorem 2: Combining Lemmas 3 and 4 provides us with the proof for our fundamental theorem, stated in Section II, viz., the operational rate distortion function $R(D)$ is equal to the mathematical one $R^{(I)}(D)$. ■

V. Extension of the Regular BA Algorithm for Rate Distortion With Feed Forward

In this section, we describe an algorithm for calculating $R_n(D)$, where

$$R_n(D) = \frac{1}{n} \min_{r(\hat{x}^n|x^n):\mathbb{E}\left[d(X^n,\hat{X}^n)\right]\le D} I(\hat{X}^n \to X^n) \quad (19)$$

using the alternating minimization procedure. This method was first used by Blahut [17] and Arimoto [18] to obtain a numerical solution for the i.i.d. source rate distortion and for the memoryless channel capacity.

Before we describe the algorithm, let us denote by $r = r(\hat{x}^n|x^n)$, $q = q(\hat{x}^n||x^{n-1})$ the PMFs that are participating in the minimization. Further, let us consider the double optimization problem given by

$$R_n(D) = \frac{1}{n}\left[-\lambda D + \min_{r,q} K(r,q)\right] \quad (20)$$

where

$$K(r,q) = I_{FF}(r,q) + \lambda \mathbb{E}_r\left[d(X^n,\hat{X}^n)\right]$$

and $I_{FF}(r,q)$ is the directed information given by

$$I_{FF}(r,q) = I(\hat{X}^n \to X^n) = \sum_{\hat{x}^n,x^n} p(x^n) r(\hat{x}^n|x^n) \log \frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n||x^{n-1})}. \quad (21)$$

In Section VI, we show that the double optimization problem given in (20) is equal to the one given in (19). Equations (20) and (21) allow us to apply the alternating minimization procedure.

A. Description of the Algorithm

In Algorithm 1, we present the steps required to minimize the directed information where the input PMF $p(x^n)$ is fixed.

Algorithm 1 Iterative algorithm for calculating $R_n(D)$, where $p(x^n)$ is fixed.

(a) Fix a value of $\lambda \ge 0$ that determines a point on the $R_n(D)$ curve.

(b) Start with a random, causally conditioned point $q^0(\hat{x}^n||x^{n-1})$. Usually, we start from a uniform distribution, i.e., $q^0(\hat{x}^n||x^{n-1}) = 2^{-n}$ for every $(x^n,\hat{x}^n)$.

(c) Set $k = 1$.

(d) Compute $r^k(\hat{x}^n|x^n)$ using the formula

$$r^k(\hat{x}^n|x^n) = \frac{q^{k-1}(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n} q^{k-1}(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}.$$

(e) Calculate the joint probability

$$p(x^n,\hat{x}^n) = p(x^n) r^k(\hat{x}^n|x^n)$$

and deduce the causal conditioned PMF $q^k(\hat{x}^n||x^{n-1})$ as in (1).

(f) Calculate the parameter

$$c^k_{\hat{x}^n,x^{n-1}} = \frac{q^k(\hat{x}^n||x^{n-1})}{q^{k-1}(\hat{x}^n||x^{n-1})}.$$

(g) Calculate

$$F = \log \max_{\hat{x}^n,x^{n-1}} c^k_{\hat{x}^n,x^{n-1}} - \sum_{x^n,\hat{x}^n} p(x^n) r^k(\hat{x}^n|x^n) \log c^k_{\hat{x}^n,x^{n-1}}.$$

(h) If $F \ge \epsilon$, set $k := k+1$, and return to (d).

(i) The rate distortion function, with distortion

$$D_k = \sum_{\hat{x}^n,x^n} p(x^n) r^k(\hat{x}^n|x^n) d(x^n,\hat{x}^n) \quad (22)$$

is

$$R^k_n(D_k) = \frac{1}{n} \sum_{x^n,\hat{x}^n} p(x^n) r^k(\hat{x}^n|x^n) \log \frac{r^k(\hat{x}^n|x^n)}{q^k(\hat{x}^n||x^{n-1})}.$$

The parameter λ is used in the Lagrangian with which we optimize the directed information. The value of D_k, and hence

TABLE I
MEMORY AND OPERATIONS NEEDED BY THE EXTENDED BA ALGORITHM FOR SOURCE CODING WITH FEED FORWARD

	Operation	Memory
$\min_{p(\hat{x}^n\|x^n):\mathbb{E}[d(X^n,\hat{X}^n)]\le D}\left(\frac{1}{n}I(\hat{X}^n;X^n)\right)$, regular BA algorithm	$O((\|\mathcal{X}\|\|\hat{\mathcal{X}}\|)^n)$	$(\|\mathcal{X}\|\|\hat{\mathcal{X}}\|)^n+\|\mathcal{X}\|^n+\|\hat{\mathcal{X}}\|^n$
$\min_{p(\hat{x}^n\|x^n):\mathbb{E}[d(X^n,\hat{X}^n)]\le D}\left(\frac{1}{n}I(\hat{X}^n\to X^n)\right)$, Algorithm 1	$O((\|\mathcal{X}\|\|\hat{\mathcal{X}}\|)^n)$	$2(\|\mathcal{X}\|\|\hat{\mathcal{X}}\|)^n+\|\mathcal{X}\|^n$

$R_n(D_k)$, depends on λ; thus, choosing λ appropriately sweeps out the $R_n(D_k)$ curve. The algorithm stops when $F<\epsilon$. In Appendix C, we provide upper and lower bounds that are used to show that if $F<\epsilon$, we ensure that $|R_n^k(D_k)-R_n(D_k)|<\epsilon$.

Now, let us present a special case and a few extensions for Algorithm 1.

1) *Regular BA algorithm*, i.e., *the delay* $s=n$. For delay $s=n$, the algorithm suggested here agrees with the regular BA algorithm, where instead of step (d) we have

$$r^k(\hat{x}^n|x^n)=\frac{q^{k-1}(\hat{x}^n)2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n}q^{k-1}(\hat{x}^n)2^{-\lambda d(x^n,\hat{x}^n)}}$$

and in step (e), $q^k(\hat{x}^n)$ corresponds to the joint probability $p(x^n)r^k(\hat{x}^n|x^n)$ as well. Moreover, the expression for $c^k_{\hat{x}^n,x^{n-1}}$ is reduced to

$$c^k_{\hat{x}^n}=\frac{q^k(\hat{x}^n)}{q^{k-1}(\hat{x}^n)}$$

and the termination of the algorithm in step (g) is defined by

$$F=\log\max_{\hat{x}^n}c^k_{\hat{x}^n}-\sum_{x^n,\hat{x}^n}p(x^n)r^k(\hat{x}^n|x^n)\log c^k_{\hat{x}^n}\le\epsilon$$

as in the regular BA algorithm [17].

2) *Function of the feed forward with general delay* s. We present a generalization of the algorithm, where the feed forward is a deterministic function of the source with some delay s, $z^{i-s}=f(x^{i-s})$. In that case, step (d) is replaced by

$$r^k(\hat{x}^n|x^n)=\frac{q^{k-1}(\hat{x}^n||z^{n-s})2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n}q^{k-1}(\hat{x}^n||z^{n-s})2^{-\lambda d(x^n,\hat{x}^n)}}$$

and in step (e), we have

$$q^k(\hat{x}^n\|z^{n-s})=\prod_{i=1}^n p(\hat{x}_i|\hat{x}^{i-1},z^{i-s})$$

where we calculate $p(\hat{x}_i|\hat{x}^{i-1},z^{i-s})$ from the joint distribution $p(x^n,\hat{x}^n)=p(x^n)r^k(\hat{x}^n|x^n)$. The algorithm is terminated in the same way, where

$$c^k_{\hat{x}^n,z^{n-s}}=\frac{q^k(\hat{x}^n\|z^{n-s})}{q^{k-1}(\hat{x}^n\|z^{n-s})}.$$

B. Complexity and Memory Needed

The computation complexity and the memory needed for the aforementioned algorithm are presented in Table I. The memory that is needed for the algorithm is only due to the PMFs q, r, and p. The first two need a memory of $(|\mathcal{X}||\hat{\mathcal{X}}|)^n$ and p needs $|\mathcal{X}|^n$. As for calculation complexity, we see that in step (d), we need for every $x^n,\hat{x}^n$, $O(|\mathcal{X}|^n)$ summation and multiplication operations. As for step (e), after calculating the joint distribution, we need $O(n(|\mathcal{X}||\hat{\mathcal{X}}|)^n)$ summations. Altogether, we need about $O((|\mathcal{X}||\hat{\mathcal{X}}|)^n)$ operations.

VI. DERIVATION OF ALGORITHM 1

In this section, we first derive the alternating minimization procedure and provide the settings of our optimization problem. Then, we prove its convergence to the global minimum given by

$$R_n(D)=\frac{1}{n}\min_{r(\hat{x}^n\|x^{n-1}):\mathbb{E}[d(X^n,\hat{X}^n)]\le D}I(\hat{X}^n\to X^n).$$

This is done by reforming the formula for $R_n(D)$, so it can be solved using the Lagrange multipliers method and the Karush–Kuhn–Tucker conditions. In the second part, we present a sequence of lemmas that conclude with the proof of Theorem 4. The outline of the second part is further detailed in the following.

Throughout this section, note that the input probability $p(x^n)$ is fixed in all minimization calculations. Further, we denote by $I_{FF}(r,q)$ the directed information, given by

$$I_{FF}(r,q)=\sum_{\hat{x}^n,x^n}p(x^n)r(\hat{x}^n|x^n)\log\frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n\|x^{n-1})}.$$

The alternating maximization procedure is described in [12] by two maximization functions operating on the two-variable objective $f(u_1,u_2)$, where $u_1\in\mathcal{A}_1$, $u_2\in\mathcal{A}_2$. The first is $c_1(u_2)\in\mathcal{A}_1$, which is the one that achieves $\sup_{u_1\in\mathcal{A}_1}f(u_1,u_2)$, and the second is $c_2(u_1)\in\mathcal{A}_2$, which is the point that achieves $\sup_{u_2\in\mathcal{A}_2}f(u_1,u_2)$. Although in this paper we wish to solve a minimization problem, its negative can be used in the alternating maximization procedure. For that purpose, assume that $c_2(u_1)\in\mathcal{A}_2$ and $c_1(u_2)\in\mathcal{A}_1$ for all $u_1\in\mathcal{A}_1$, $u_2\in\mathcal{A}_2$. Let us define an iteration as the following equation:

$$(u_1^k,u_2^k)=\left(c_1(u_2^{k-1}),c_2(c_1(u_2^{k-1}))\right)$$

and in each iteration, we consider the value $f^k=f(u_1^k,u_2^k)$. We now state the alternating maximization procedure lemma.

Lemma 5 (see [19, Lemmas 9.4 and 9.5]) "Convergence of the alternating maximization procedure": Let $f(u_1, u_2)$ be a real, concave, bounded-from-above function that is continuous and has continuous partial derivatives and let the sets $\mathcal{A}_1$ and $\mathcal{A}_2$, over which we maximize, be convex. Under these conditions, $\lim_{k\to\infty} f^k = f^*$, where f^* is the solution to the optimization problem.

The rate distortion function with feed forward can be, as in [17], carried out parametrically in terms of a parameter λ, which is introduced as a Lagrange multiplier. The motivation for this change is to replace the problem into one that can be solved using the alternating minimization procedure. In Appendix C, we show that the parameter λ defines the slope of the curve $R_n(D)$ at the point it parameterizes, and the slope is given by $\frac{-\lambda}{n}$. We now write the following parametric expression for $R_n(D)$:

$$R_n(D) = \frac{1}{n} \min_{r(\hat{x}^n|x^n)} \Big[I(\hat{X}^n \to X^n) + \lambda \left(\mathbb{E}_r \left[d(X^n, \hat{X}^n) \right] - D \right) \Big] \tag{23}$$

where D is the distortion at the point $r^*(\hat{x}^n|x^n)$ that achieves $R_n(D)$. Here, the value of D is not an input to the minimization, but is determined by the parameter λ.

Note that the directed information is a function of the joint distribution $p(x^n)r(\hat{x}^n|x^n)$. Since the source distribution is given, the directed information I_{FF} is determined by $r = r(\hat{x}^n|x^n)$ alone. Let us define by $q = q(\hat{x}^n\|x^{n-1})$ the causal conditioning probability. Now, let us define the functional

$$K(r, q) = I_{FF}(r, q) + \lambda \mathbb{E}_r \left[d(X^n, \hat{X}^n) \right]. \tag{24}$$

Note that now, the minimization is over all possible r, and thus, the functional $K(r, q)$ can be solved using the alternating minimization procedure.

From (23) and (24), we can see that $R_n(D)$ can be written as

$$\frac{1}{n} \left[-\lambda D + \min_r K(r, q) \right] \tag{25}$$

where $q(\hat{x}^n\|x^{n-1})$ corresponds to the joint distribution $p(x^n)r(\hat{x}^n|x^n)$ and D is the distortion at the point $r^*(\hat{x}^n|x^n)$ that achieves $R_n(D)$.

In the rest of this section, we show that we can use the alternating minimization procedure for computing $R_n(D)$, as given in (25). For this purpose, we present several lemmas that assist in proving our main theory. In Lemma 6, we show that the expression we minimize satisfies the conditions in Lemma 5. In Lemma 7, we show that we are allowed to minimize the functional $K(r, q)$ over $r(\hat{x}^n|x^n)$ and $q(\hat{x}^n\|x^{n-1})$ together, rather than over $r(\hat{x}^n|x^n)$ alone, and thus use the alternating minimization procedure to achieve the optimum value. Lemma 8 is a supplementary claim that helps us to prove Lemma 7 and in which we find an expression for $q(\hat{x}^n\|x^{n-1})$ that minimizes the functional $K(r, q)$, where $r(\hat{x}^n|x^n)$ is fixed. In Lemma 9, we find an explicit expression for $r(\hat{x}^n|x^n)$ that minimizes the functional $K(r, q)$, where $q(\hat{x}^n\|x^{n-1})$ is fixed. Theorem 3 combines all lemmas to show that the alternating minimization procedure, as described in Algorithm 1, converges. We conclude with a supplementary claim about the upper and lower bounds on the feed-forward rate distortion, and then prove that the stopping condition described in Algorithm 1 ensures that the error $|R_n^k(D) - R_n(D)| < \epsilon$.

Lemma 6: For a fixed input PMF $p(x^n)$, the functional $K(r, q)$ given in (24) is convex in $\{r, q\}$, continuous and with continuous partial derivatives. Moreover, the sets of the causal regular conditioned PMF r and the causal conditioned PMF q, over which we optimize, are convex.

Proof: Since the functional $K(r, q)$ consists of a linear (and thus convex) expression in r, i.e., $\mathbb{E}_r \left[d(X^n, \hat{X}^n) \right]$, we only need to verify that the directed information is convex. We first write the directed information in the following form:

$$\begin{aligned} I(\hat{X}^n \to X^n) &= -\sum_{\hat{x}^n, x^n} p(x^n, \hat{x}^n) \log \frac{p(x^n)}{p(x^n\|\hat{x}^n)} \\ &= -\sum_{\hat{x}^n, x^n} p(x^n, \hat{x}^n) \log \frac{p(x^n) q(\hat{x}^n\|x^{n-1})}{p(x^n\|\hat{x}^n) q(\hat{x}^n\|x^{n-1})} \\ &= -\sum_{\hat{x}^n, x^n} p(x^n, \hat{x}^n) \log \frac{q(\hat{x}^n\|x^{n-1})}{p(x^n, \hat{x}^n)/p(x^n)} \\ &= -\sum_{\hat{x}^n, x^n} p(x^n) r(\hat{x}^n|x^n) \log \frac{q(\hat{x}^n||x^{n-1})}{r(\hat{x}^n|x^n)} \\ &= I_{FF}(r, q). \end{aligned}$$

This form is the negative of a concave function, as proven in [12, Lemma 2]. Furthermore, in the same lemma, we show that the directed information is continuous with continuous partial derivatives; similar proof applies here. It is also simple to verify that both of the sets that we minimize over are convex, i.e., sets $\mathcal{A}_1,\ \mathcal{A}_2$, where

$$\begin{aligned} \mathcal{A}_1 &= \{r(\hat{x}^n|x^n) : r(\hat{x}^n|x^n) > 0 \\ &\qquad \text{is a regular conditioned PMF}\} \\ \mathcal{A}_2 &= \{q(\hat{x}^n\|x^{n-1}) : q(\hat{x}^n\|x^{n-1}) \\ &\qquad \text{is a causally conditioned PMF}\}. \end{aligned} \tag{26}$$

■

Recall that in order to use the alternating minimization procedure, we minimize over $\{r(\hat{x}^n|x^n),\ q(\hat{x}^n||x^{n-1})\}$ instead of over $r(\hat{x}^n|x^n)$ alone, and thus need the following lemma.

Lemma 7: For any discrete random variables, $X^n,\ \hat{X}^n$, $R_n(D)$, as defined in (25), is equal to the following optimization problem:

$$\frac{1}{n} \left[-\lambda D + \min_{r,q} K(r, q) \right]$$

where D is the distortion at the point $r^*(\hat{x}^n|x^n)$ that achieves $R_n(D)$.

To prove this lemma, we note that $\mathbb{E}_r \left[d(X^n, \hat{X}^n) \right]$, which does not contain the variable q, is part of the functional $K(r, q)$. Hence, it suffices to show that

$$\min_{\mathbf{r}} \frac{1}{n} I(\hat{X}^n \to X^n) = \min_{\mathbf{q}} \min_{\mathbf{r}} \frac{1}{n} I(\hat{X}^n \to X^n). \tag{27}$$

The proof is given after the following supplementary claim, in which we calculate the specific $q(\hat{x}^n\|x^{n-1})$ that minimizes the directed information when $r(\hat{x}^n|x^n)$ is fixed.

Lemma 8: For fixed $r(\hat{x}^n|x^n)$, there exists a unique $c_2(r)$ that achieves $\min_{q(\hat{x}^n\|x^{n-1})} I(\hat{X}^n \to X^n)$, and is given by

$$q^*(\hat{x}^n\|x^{n-1}) = \frac{p(x^n)r(\hat{x}^n|x^n)}{p(x^n||\hat{x}^n)} \tag{28}$$

where $p(x^n\|\hat{x}^n)$ is calculated using the joint distribution $p(x^n)r(\hat{x}^n|x^n)$.

Proof for Lemma 8: In order to prove this lemma, we show that the expression for $I_{FF}(r,q) - I_{FF}(r,q^*) > 0$ for any $q \neq q^*$. Consider the following chain of inequalities:

$$\begin{aligned}
&I_{FF}(r,q) - I_{FF}(r,q^*)\\
&= \sum_{x^n,\hat{x}^n} p(x^n)r(\hat{x}^n|x^n)\log\frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n\|x^{n-1})}\\
&\quad - \sum_{x^n,\hat{x}^n} p(x^n)r(\hat{x}^n|x^n)\log\frac{r(\hat{x}^n|x^n)}{q^*(\hat{x}^n\|x^{n-1})}\\
&= \sum_{x^n,\hat{x}^n} p(x^n)r(\hat{x}^n|x^n)\log\frac{q^*(\hat{x}^n||x^{n-1})}{q(\hat{x}^n||x^{n-1})}\\
&= \sum_{x^n,\hat{x}^n} p(x^n||\hat{x}^n)q^*(\hat{x}^n||x^{n-1})\log\frac{p(x^n||\hat{x}^n)q^*(\hat{x}^n||x^{n-1})}{p(x^n||\hat{x}^n)q(\hat{x}^n||x^{n-1})}\\
&= D\left(p(x^n||\hat{x}^n)q^*(\hat{x}^n||x^{n-1}) \parallel p(x^n||\hat{x}^n)q(\hat{x}^n||x^{n-1})\right)\\
&\overset{(a)}{\geq} 0
\end{aligned}$$

where (a) follows from the nonnegativity of the divergence. By the properties of the divergence function, equality holds if and only if the joint PMFs are the same, i.e., $q = q^*$. ∎

Proof of Lemma 7: The PMF that minimizes the directed information is the one that corresponds to the joint distribution $r(\hat{x}^n|x^n)p(x^n)$; thus, (27) holds and the functional $K(r,q)$ can be minimized over both r and q in combined. ∎

In the following lemma, we derive an explicit expression for $r(\hat{x}^n|x^n)$ that achieves $R_n(D)$, where $q(\hat{x}^n\|x^{n-1})$ is fixed.

Lemma 9: For fixed $q(\hat{x}^n\|x^{n-1})$, there exists $c_1(q)$ that achieves $R_n(D)$, and is given by

$$r(\hat{x}^n|x^n) = \frac{q(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n} q(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}.$$

Proof: Following [13, Ch. 5.5.3], since we are solving a convex optimization problem, we can apply the KKT conditions with the constraints $\sum_{\hat{x}^n} r(\hat{x}^n|x^n) = 1$, and set up the functional

$$\begin{aligned}
J &= \sum_{x^n,\hat{x}^n} p(x^n)r(\hat{x}^n|x^n)\log\frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n\|x^{n-1})}\\
&\quad + \lambda\left(\sum_{x^n,\hat{x}^n} p(x^n)r(\hat{x}^n|x^n)d(x^n,\hat{x}^n) - D\right)\\
&\quad + \sum_{x^n}\nu(x^n)\sum_{\hat{x}^n} r(\hat{x}^n|x^n).
\end{aligned}$$

Solving $\frac{\partial J}{\partial r(\hat{x}^n|x^n)} = 0$ yields the expression for $r(\hat{x}^n|x^n)$ as

$$r(\hat{x}^n|x^n) = \frac{q(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n} q(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}. \tag{29}$$

∎

Another lemma that is required is the one that states that the algorithm, when it converges, remains fixed on its variables. We already know that the variable q that optimizes the directed information is unique; we have to show that within the algorithm the variable r is unique as well. This is needed to ensure that we achieve a unique point when using the algorithm in order to show that the algorithm indeed converges. Let r^k be the PMF r that is obtained by the kth iteration using Algorithm 1.

Lemma 10: By following Algorithm 1 we obtain the sequence $\{r^k\}$ converges to the PMF r^* that achieves $R_n(D)$.

Proof: The uniqueness is proven in a similar way to a proof given by Blahut in [17, Th. 6] and we follow it with appropriate modifications. We recall that in the kth iteration

$$\begin{aligned}
K(r^k,q^k) &= I_{FF}(r^k,q^k) + \lambda\mathbb{E}_{r^k}\left[d(X^n,\hat{X}^n)\right]\\
&= \sum_{x^n,\hat{x}^n} p(x^n)r^k(\hat{x}^n|x^n)\log\frac{r^k(\hat{x}^n|x^n)}{q^k(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}.
\end{aligned}$$

Further, from [17, Th. 6], we can see that

$$\begin{aligned}
&K(r^{k+1},q^{k+1})\\
&= -\sum_{x^n,\hat{x}^n} p(x^n)r^k(\hat{x}^n|x^n)\log\left(\sum_{\hat{x}^n} q^k(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\right)\\
&\quad + \sum_{x^n,\hat{x}^n} p(x^n)r^{k+1}(\hat{x}^n|x^n)\log\frac{q^k(\hat{x}^n\|x^{n-1})}{q^{k+1}(\hat{x}^n\|x^{n-1})}.
\end{aligned}$$

Hence

$$\begin{aligned}
&K(r^k,q^k) - K(r^{k+1},q^{k+1})\\
&= \sum_{x^n,\hat{x}^n}\Bigg(p(x^n)r^k(\hat{x}^n|x^n)\cdot\\
&\quad \log\frac{r^k(\hat{x}^n|x^n)\sum_{\hat{x}^n} q^k(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{q^k(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}\Bigg)\\
&\quad + \sum_{x^n,\hat{x}^n} p(x^n)r^{k+1}(\hat{x}^n|x^n)\log\frac{q^{k+1}(\hat{x}^n||x^{n-1})}{q^k(\hat{x}^n||x^{n-1})}\\
&\overset{(a)}{\geq} \sum_{x^n,\hat{x}^n} p(x^n)r^k(\hat{x}^n|x^n)\cdot\\
&\quad \left(1 - \frac{q^k(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{r^k(\hat{x}^n|x^n)\sum_{\hat{x}^n} q^k(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}\right)\\
&\quad + \sum_{x^n,\hat{x}^n} p(x^n)r^{k+1}(\hat{x}^n|x^n)\left(1 - \frac{q^k(\hat{x}^n\|x^{n-1})}{q^{k+1}(\hat{x}^n\|x^{n-1})}\right)\\
&\overset{(b)}{=} \sum_{x^n,\hat{x}^n} p(x^n)r^k(\hat{x}^n|x^n)\left(1 - \frac{r^{k+1}(\hat{x}^n|x^n)}{r^k(\hat{x}^n|x^n)}\right)\\
&\quad + \sum_{x^n,\hat{x}^n} p(x^n||\hat{x}^n)q^{k+1}(\hat{x}^n||x^{n-1})\left(1 - \frac{q^k(\hat{x}^n||x^{n-1})}{q^{k+1}(\hat{x}^n||x^{n-1})}\right)\\
&= 0 + 0
\end{aligned}$$

where (a) follows from the inequality $\log(y) \geq 1 - \frac{1}{y}$ and (b) follows from (29), where $q = q^k$ and $r = r^{k+1}$. Note that

we have strict inequality unless $q^k = q^{k+1}$ and $r^k = r^{k+1}$. Thus, $K(r^k, q^k)$ is nonincreasing and is strictly decreasing unless the distribution stabilizes, and hence, the uniqueness of the optimum parameter r^* emerges. ■

Now, we can prove Theorem 3 as stated in Section II.

Proof of Theorem 3: First, we have to show the existence of a double minimization problem, i.e., an equivalent problem where we minimize over two variables instead of only one; this was shown in Lemma 7. Now, in order for the alternating minimization procedure to work on this optimization problem, we need to show that the conditions given in Lemma 5 are satisfied for the functional $K(r, q)$; this was shown in Lemma 6. The steps described in Algorithm 1 are proved in Lemmas 8 and 9, thus giving us an algorithm to compute $R_n(D)$, where the minimization is evaluated according to the parameter λ. ■

Our last step in proving the convergence of Algorithm 1 is to show why the stopping condition ensures a small error. For this reason, we state a lemma introducing the existence of bounds to the rate distortion with feed-forward function and then conclude that the stopping condition does ensure a small error in the algorithm, i.e., $|R_n^k(D_k) - R_n(D_k)| < \epsilon$, where $R_n^k(D_k)$ is the upper bound in the kth iteration, and

$$D_k = \mathbb{E}_{r^k}\left[d(X^n, \hat{X}^n)\right]. \tag{30}$$

For this purpose, we define the following expressions that takes part in each iteration:

$$\begin{aligned} c^k_{\hat{x}^n, x^{n-1}} &= \frac{q^k(\hat{x}^n \| x^{n-1})}{q^{k-1}(\hat{x}^n \| x^{n-1})} \\ \gamma^k(x^n) &= \left(\sum_{\hat{x}^n} q^{k-1}(\hat{x}^n \| x^{n-1}) 2^{-\lambda d(x^n, \hat{x}^n)}\right)^{-1} \end{aligned} \tag{31}$$

and the upper and lower bound

$$\begin{aligned} I_U^k(D_k) &= \frac{1}{n}\Bigg(-\lambda D + \sum_{x^n} p(x^n) \log \gamma^k(x^n) \\ &\quad - \sum_{x^n, \hat{x}^n} p(x^n) r^k(\hat{x}^n | x^n) \log c^k_{\hat{x}^n, x^{n-1}} \Bigg) \\ I_L^k(D_k) &= \frac{1}{n}\Bigg(-\lambda D + \sum_{x^n} p(x^n) \log \gamma^k(x^n) \\ &\quad - \log \max_{\hat{x}^n, x^{n-1}} c^k_{\hat{x}^n, x^{n-1}} \Bigg). \end{aligned} \tag{32}$$

Note that $R_n^k(D_k) = I_U^k(D_k)$.

Lemma 11: Let the parameter $\lambda \geq 0$ be given, and let $c^k_{\hat{x}^n, x^{n-1}}$, $\gamma^k(x^n)$ be as in (31) in the kth iteration of Algorithm 1. Then, if D_k is as in (30)

$$I_L^k(D_k) \leq R_n(D_k) \leq I_U^k(D_k).$$

The proof for Lemma 11 is given in Appendix C.

From Lemma 11, we can deduce the claim regarding the stopping condition. Let the error be denoted as $\epsilon_k = |R_n^k(D) - R_n(D)|$.

Corollary 1: The error denoted as ϵ_k satisfies

$$\epsilon_k \leq \log \max_{\hat{x}^n, x^{n-1}} c^k_{\hat{x}^n, x^{n-1}} - \sum_{x^n, \hat{x}^n} p(x^n) r^k(\hat{x}^n | x^n) \log c^k_{\hat{x}^n, x^{n-1}}$$

where $c^k_{\hat{x}^n, x^{n-1}}$ is defined in the kth iteration by (31).

Proof: The proof follows (32), in which the upper bound and lower bound differ only in their last term. Thus, if the RHS is bounded by ϵ, then $\epsilon_k \leq \epsilon$. ■

VII. GP Form to $R_n(D)$ (Theorem 4)

In this section, we show that the nth order rate distortion function with feed forward, $R_n(D)$, can be given as a maximization problem, written in a standard form of GP. For this purpose, we first state the following theorem, of which Theorem 4 is a direct consequence of it.

Define the following optimization problem.

Theorem 6: The nth order rate distortion function $R_n(D)$ is the solution of the following maximization problem:

$$\max_{\lambda \geq 0, \gamma(x^n)} \frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n) \log \gamma(x^n)\right) \tag{33}$$

where, for some causal conditioned probability $p'(x^n \| \hat{x}^n)$, $\gamma(x^n)$ satisfies the inequality constraint

$$p(x^n)\gamma(x^n) 2^{-\lambda d(x^n, \hat{x}^n)} \leq p'(x^n \| \hat{x}^n). \tag{34}$$

In Appendix D we provide two proofs for Theorem 6; the first is similar to Berger's proof in [20] for the regular rate distortion function based on the inequality $\log(y) \geq 1 - \frac{1}{y}$ and the second uses the Lagrange duality, as presented in [13] and [21], that transforms a minimization problem to a maximization one. The motivation for providing both proofs is to give, on the one hand, a simple proof that holds only to the rate distortion problem and on the other hand, a more general (but involved) one that is based on the well-known KKT conditions (or Lagrange duality).

Appendix D also demonstrate the connection between the rate distortion function and the parameter λ, which states that the slope of $R_n(D)$ at point D is $-\frac{\lambda}{n}$.

Proof of Theorem 4: Considering the aforementioned theorem, our interest now is to adjust the constraints in order to obtain a GP form. We note that the optimization problem in (33) does not change if we maximize over $p'(x^n \| \hat{x}^n)$ as well and the constraint in (34) is no longer for some p', i.e., $R_n(D)$ is the solution to

$$\max_{\lambda \geq 0, \gamma(x^n), p'(x^n \| \hat{x}^n)} \frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n) \log \gamma(x^n)\right) \tag{35}$$

where $\gamma(x^n)$, $p'(x^n \| \hat{x}^n)$ satisfy the inequality constraint

$$p(x^n)\gamma(x^n) 2^{-\lambda d(x^n, \hat{x}^n)} \leq p'(x^n \| \hat{x}^n). \tag{36}$$

The aforementioned statement is true since, on the one hand, the maximization in (33) increases upon maximizing over another variable, $p'(x^n\|\hat{x}^n)$, as in (35); on the other hand, the variable $\gamma^*(x^n)$, $p'^*(x^n\|\hat{x}^n)$ that achieves (35) satisfies constraint (34) in Theorem 6. Hence, the maximization problem in (35) cannot be greater than the one in (33).

To obtain a GP standard form, we transform the constraint in (36), such that

$$p(x^n)\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)}p'(x^n\|\hat{x}^n)^{-1} \leq 1.$$

Taking the log of both sides, we obtain

$$\log(p(x^n)) + \log(\gamma(x^n)) - \lambda d(x^n,\hat{x}^n) - \sum_{i=1}^{n} \log p'(x_i|x^{i-1},\hat{x}^i) \leq 0.$$

Note that maximizing over $p'(x^n\|\hat{x}^n)$ is the same as maximizing over its products $\{p'(x_i|x^{i-1},\hat{x}^i)\}_{i=1}^n$[10, Lemma 3]. Therefore, we can conclude that the rate distortion with feed forward $R_n(D)$ is the solution of the following GP maximization form:

$$\max_{\lambda,\gamma(x^n),\{p'(x_i|x^{i-1},\hat{x}^i)\}_{i=1}^n} \frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n)\log\gamma(x^n)\right)$$

subject to

$$\log(p(x^n)) + \log(\gamma(x^n)) - \lambda d(x^n,\hat{x}^n) - \sum_{i=1}^{n} \log p'(x_i|x^{i-1},\hat{x}^i) \leq 0, \quad \forall\, x^n,\hat{x}^n$$
$$\sum_{x_i} p'(x_i|x^{i-1},\hat{x}^i) = 1, \quad \forall\, i, \forall\, x^{i-1},\hat{x}^{i-1}$$
$$\lambda \geq 0.$$

Hence, we obtain a standard form of GP. This GP problem can be solved using standard convex optimization tools. ■

VIII. Numerical Examples

In this section, we present several examples for the rate distortion source coding with feed forward. First, by using Algorithm 1, we demonstrate for a specific example that feed forward does not decrease the rate distortion function where the source is memoryless (i.i.d.), as shown in [3]. Then, we provide two explicit examples for a Markovian source: one where the distortion is a single letter and one with a general distortion function, as presented in [5]. GP is also used to verify our results.

In all of the examples, we run Algorithm 1 with various values of λ and thus construct the graph of $R_n(D)$ using interpolations. Alternatively, one can use the GP form and find, for every distortion D given as input, the rate R.

A. Memoryless (i.i.d.) Source

Analogous to the memoryless channel, it was shown by Weissman and Merhav [3] that for an i.i.d. source feed forward does not decrease the rate distortion function. In this example, the source is distributed $X \sim B(\frac{1}{2})$, and the distortion function is a single letter, i.e.,

$$d(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i,\hat{x}_i).$$

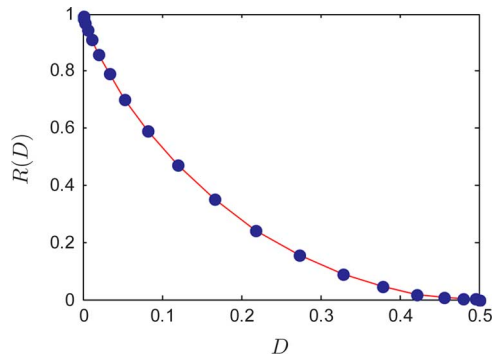

Fig. 4. Rate distortion function for a binary source and feed forward with delay 1. The circles represent the performance of Algorithm 1. The line is the plot of (37).

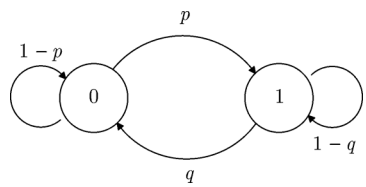

Fig. 5. Symmetrical Markov chain.

Running our algorithm with delay $s = 1$ and block length $n = 5$, we would expect to obtain the same result as with no feed forward at all (as shown in [22, ch. 10.3.1]), which is given by

$$R(D) = \begin{cases} H_b(p) - H_b(D), & 0 \leq D \leq \min\{p, 1-p\} \\ 0, & D \geq \min\{p, 1-p\}. \end{cases} \quad (37)$$

Note that $H_b(p)$, $H_b(D)$ are the binary entropies with parameters p, D, respectively. Indeed, the aforementioned function and the performance of Algorithm 1 coincide, as illustrated in Fig. 4. Note that the joint distribution $p(x^n)r(\hat{x}^n|x^n)$ is the same as the one that achieves the analytical calculation in which $p(x_i) = 0.5$ and $X \oplus \hat{X} \sim B(D)$. For $D = 0.2$ and $n = 3$, solving the geometrical programming form using a MATLAB code produces the rate $R = 0.278072$, which is close to $R(0.2)$ using (37). The value of λ turns out to be 6, which means that the slope at point $(R = 0.278072, D = 0.2)$ is -2.

In the following example, we present the performance of Algorithm 1 for a Markov source and a single-letter distortion.

B. Markov Source and Single-Letter Distortion

The Markov source is presented in Fig. 5. This model was solved by Weissman and Merhav in [3] for the symmetrical case $p = q$. We extend this model for the case of general transition probabilities p and q. The analytical solution for this example is detailed in Appendix E; there, we show that, for any n, the nth rate distortion function is equal to

$$\frac{1}{n}H_b(\pi) + \frac{n-1}{n}\left(\pi_1 H_b(p) + \pi_2 H_b(q)\right) - H_b(D). \quad (38)$$

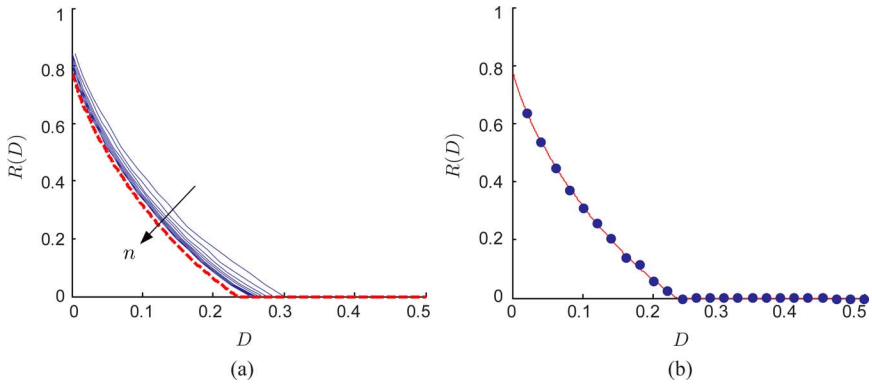

Fig. 6. $R(D)$ for the Markov source example and feed forward with delay 1. (a) Graph of $R_n(D)$; the arrow marks the way $R_n(D)$ responds to increasing n. The dashed line is the analytical calculation (b) Graph of $12D_{12}(R) - 11D_{11}(R)$. The circles represent the performance of Algorithm 1.

By taking n to infinity, we have

$$R(D) = \pi_1 H_b(p) + \pi_2 H_b(q) - H_b(D)$$

where $\pi = [\pi_1, \pi_2]$ is the stationary distribution of the source. Fig. 6 is displayed at the bottom of the page. In Fig. 6(a), we present the graphs of $R_n(D)$ for $n = 1$ up to $n = 12$, where $p = 0.3$, $q = 0.2$, and X_0 has the stationary distribution $[0.4, 0.6]$. It is evident that $R_n(D)$ decreases as n increases and converges to the analytical calculation.

In [12, Lemma 6], we provided another estimator for the limit of the nth directed information. This was the directed information rate, which served as an estimator for the feedback channel capacities. There, we show that if the limit exists, then

$$\lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n) = \lim_{n\to\infty} \left(I(X^n \to Y^n) - I(X^{n-1} \to Y^{n-1}) \right).$$

Hence, we can estimate $R(D)$, which is the limit of the nth directed information, as

$$R(D) = \lim_{n\to\infty} (nR_n(D) - (n-1)R_{n-1}(D)).$$

This estimator turns out to perform better than $R_n(D)$, due to the subtraction that eliminates any bias that might exist in $R_n(D)$. This is applied in two ways: either when the rate value is fixed or when the distortion value is fixed. In both cases, we first have to fix an axes vector and interpolate the other vector with respect to the fixed one; then, we can calculate the differences between the interpolated vectors.

In Fig. 6(b), we present this estimator only for $n = 12$ where the vector of the distortion is interpolated, i.e., $12D_{12}(R) - 11D_{11}(R)$. We can see that this estimation is much more accurate than the one in Fig. 6(a).

This is a good opportunity to present the performance of the upper and lower bounds of a specific rate distortion pair (R, D) and the geometrical programming solution to this problem, shown in Fig. 7. We ran Algorithm 1 for the specific parameters $\lambda = 9.216$, $n = 3$ that correspond to the rate distortion pair $(R = 0.35884, D = 0.10627)$ at slope $-\frac{9.216}{3} \approx -3$, shown in Fig. 7(a). Fig. 7(b) shows calculated $R_3(D)$ using GP and Algorithm 1 compared to the theoretical one in (38).

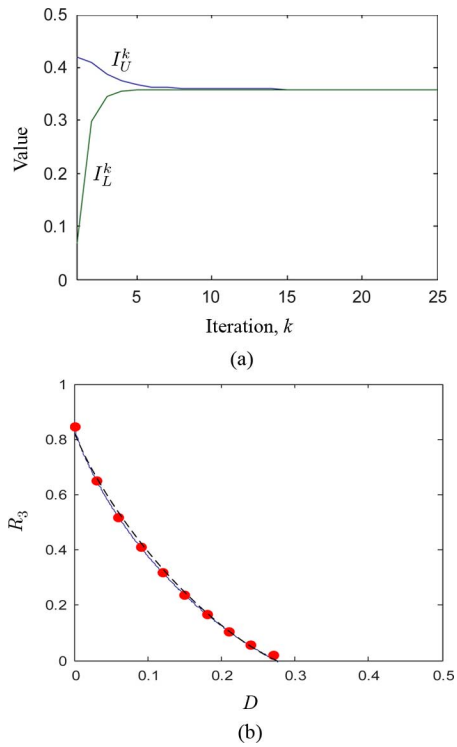

Fig. 7. Bounds for $R_3(D)$ and performance of GP and BA-type algorithm for $R_3(D)$. (a) Graph of the upper and lower bounds as a function of the iteration for $n = 3$ and $\lambda = 9.216$, as given in (32). (b) Graph of calculated $R_3(D)$. The solid line is $R_3(D)$ as in (38); the circles represent the performance of the GP and the dashed line is the BA-type algorithm result.

C. Stock Market Example. Markov Source and General Distortion

The stock market example, in which we wish to observe the behavior of a particular stock over an N-day period, was introduced and solved in [5]. Assume the stock can take $k+1$ values, $0 \le i \le k$, and is modeled as a $k+1$ state Markov chain. On a given day i, the probability for the stock value to increase by 1 is p_i, to decrease by 1 is q_i, and to remain the same is $1 - p_i - q_i$. When the stock value is in state 0, the value cannot decrease. Similarly, when in state k, the value cannot increase. If an investor would like to be forewarned whenever the stock value drops, he is advised with a binary decision $\hat{X}_n$. $\hat{X}_n = 1$ if the

TABLE II
DISTORTION $e(\hat{x}_i, x_{i-1}, x_i), j \in \{0, 1, \ldots, k\}$

	(x_{i-1}, x_i)		
	$j, j+1$	j, j	$j, j-1$
$\hat{x}_i = 0$	0	0	1
$\hat{x}_i = 1$	1	1	0

value drops from day $n-1$ to day n, and $\hat{X}_n = 0$ otherwise. The distortion is modulated in the following form:

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} e(\hat{x}_i x_{i-1}, x_i)$$

where $e(.,.,.)$ is given in Table II. It was shown in [5] that the rate distortion function of a general Markov chain source with k states is given by

$$R(D) = \sum_{i=1}^{k-1} \pi_i \left(H(p_i, q_i, 1-p_i-q_i) - H_b(\epsilon) \right) + \pi_k \left(H_b(q_k) - H_b(\epsilon) \right)$$

where $\pi = [\pi_0, \pi_1, \ldots, \pi_k]$ is the stationary distribution of the Markov chain and $\epsilon = \frac{D}{1-\pi_0}$.

In our special case, we have $k = 2$, i.e., two states for the Markov chain, and transition probabilities $p_i = 0.3,\ q_i = 0.2$, as illustrated in Fig. 5. The stationary distribution of such a source is $\pi = [0.4, 0.6]$, and we obtain

$$R(D) = \pi_1 \left(H_b(q) - H_b(\epsilon) \right) = 0.6(H_b(0.2) - H_b(\frac{D}{0.6})).$$

Since the rate cannot be less than zero, and is a descending function of the distortion, the rate distortion function is as earlier when $H_b(0.2) \geq H_b(\frac{D}{0.6})$, i.e., when $D \leq 0.12$, and thus, we obtain

$$R(D) = \begin{cases} 0.6(H_b(0.2) - H_b(\frac{D}{0.6})), & D \leq 0.12 \\ 0, & \text{otherwise.} \end{cases} \quad (39)$$

In Fig. 8(a), we present the graphs of $R_n(D)$ for $n = 1$ up to $n = 12$ with the distortion described previously and where X_0 has the stationary distribution $[0.4, 0.6]$. We can see that $R_n(D)$ decreases as n increases, as expected, and converges to the analytical calculation. In Fig. 8(b), we present the directed information rate estimator for $n = 12$, where the vector of the distortion is interpolated, i.e., $12D_{12}(R) - 11D_{11}(R)$. We can see that this estimator is much more accurate than the one in Fig. 8(a).

D. Effects of the Delay on $R_n(D)$

In this example, we use the Markov source (see Fig. 5) example with a single-letter distortion. We run Algorithm 1 with delays $s \in \{1, 2, .., 10\}$ and block length $n = 10$, where X_0 has a stationary distribution. We expect the rate distortion function to increase with the delay s. This is expected because as the delay s increases, the value of the directed information increases as well. Due to the fact that for $s \in \{3, 4, \ldots, 10\}$ all graphs are close together, we present $R_n(D)$ only for $s = 1, 2, 10$ and the results are shown in Fig. 9.

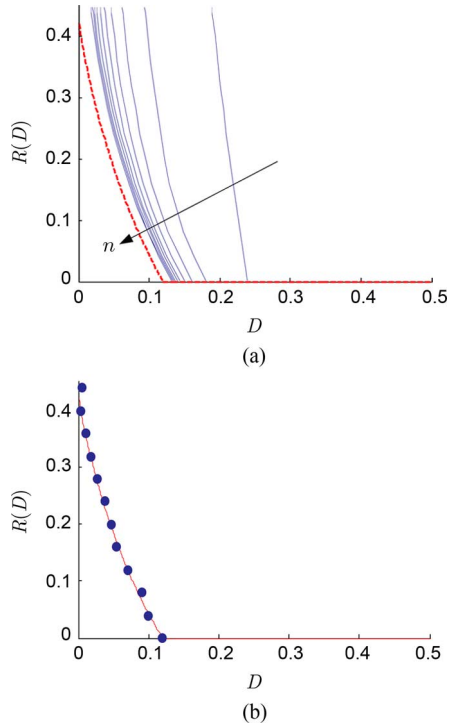

Fig. 8. $R(D)$ for the stock market example with feed forward and delay 1 (a) Graph of $R_n(D)$; the arrow marks the way $R_n(D)$ responds to increasing n. The dashed line is the analytical calculation. (b) Graph of $12D_{12}(R) - 11D_{11}(R)$. The circles represent the performance of Algorithm 1.

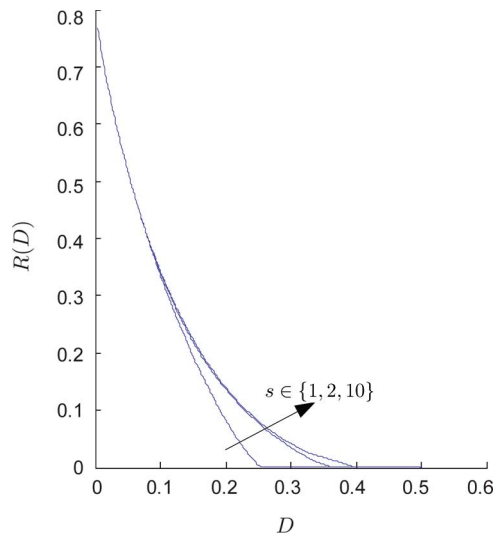

Fig. 9. $R_{10}(D)$ for a Markov source as a function of the delay.

IX. CONCLUSION

In this paper, we considered the rate distortion problem of discrete-time, ergodic, and stationary sources with feed forward at the receiver. We first derived a sequence of achievable rates, $\{R_n(D)\}_{n \geq 1}$, that converge to the feed-forward rate distortion. By showing that the sequence is subadditive, we proved that the limit of $R_n(D)$ exists and is equal to the feed-forward rate distortion. We provided an algorithm for calculating $R_n(D)$ using the alternating minimization procedure, presented a dual form for the optimization of $R_n(D)$, and transformed it into a GP maximization problem.

APPENDIX A
PROOF OF LEMMA 2

We start by showing that the sequence $\{R_n(D)\}$ is subadditive; the methodology is similar to Gallager's proof in [2, Th. 9.8.1] for the case of no feed forward. Then, by showing that the sequence $R_n(D)$ is subadditive, following [2, Lemma 4A.2] results with the proof, i.e.,

$$\lim_n R_n(D) = \inf_n R_n(D).$$

To commence, we recall that a sequence $\{a_n\}$ is called subadditive if for all m, l

$$(m+l)a_{m+l} \leq ma_m + la_l.$$

Let l and n be arbitrary positive integers and, for a given D, let $p_n(\hat{x}^n|x^n)$ and $p_l(\hat{x}^l|x^l)$ be the conditional PMFs that achieve the minimum of the directed information with block lengths of n and l, i.e., that achieve $R_n(D)$ and $R_l(D)$, respectively. Suppose we transmit $m = n + l$ samples as follows; the first n samples are transmitted using p_n and the subsequent l samples are transmitted using p_l. Hence, the overall conditional PMF is

$$p_{n+l}(\hat{x}^{n+l}|x^{n+l}) = p_n(\hat{x}^n|x^n)p_l(\hat{x}_{n+1}^{n+l}|x_{n+1}^{n+l}).$$

Note that since

$$p(x^m)p(\hat{x}^m|x^m) = p(x^m||\hat{x}^m)p(\hat{x}^m||x^{m-1})$$

we can write

$$\begin{aligned} I(\hat{X}^m \to X^m) &= H(X^m) - H(X^m\|\hat{X}^m) \\ &= H(\hat{X}^m\|X^{m-1}) - H(\hat{X}^m|X^m). \end{aligned}$$

Further, from the construction of the conditional overall PMF p_{n+l}, its clear that

$$H(\hat{X}^{n+l}|X^{n+l}) = H(\hat{X}^n|X^n) + H(\hat{X}_{n+1}^{n+l}|X_{n+1}^{n+l}).$$

Furthermore

$$\begin{aligned} &H(\hat{X}^m\|X^{m-1}) \\ &\quad= \sum_{i=1}^{n+l} H(\hat{X}_i|\hat{X}^{i-1}, X^{i-1}) \\ &\quad= H(\hat{X}^n||X^{n-1}) + \sum_{i=n+1}^{n+l} H(\hat{X}_i|\hat{X}^{i-1}, X^{i-1}) \\ &\quad\leq H(\hat{X}^n\|X^{n-1}) + \sum_{i=n+1}^{n+l} H(\hat{X}_i|\hat{X}_{n+1}^{i-1}, X_{n+1}^{i-1}) \\ &\quad= H(\hat{X}^n||X^{n-1}) + H(\hat{X}_{n+1}^{n+l}||X_{n+1}^{n+l-1}). \end{aligned}$$

Thus, it follows that

$$\begin{aligned} I(\hat{X}^{n+l} \to X^{n+l}) \leq\ & I(\hat{X}^n \to X^n) \\ &+ I(\hat{X}_{n+1}^{n+l} \to X_{n+1}^{n+l}). \end{aligned} \tag{40}$$

Since the source is stationary, we can start the input block at any given time index. Thus, the PMFs p_n and p_l achieve $nR_n(D) + lR_l(D)$ on the right-hand side of (40), while the left-hand side is greater than $(n+l)R_{n+l}(D)$ since we attempt to minimize the expression to achieve the rate distortion function. Hence, we obtain

$$(n+l)R_{n+l}(D) \leq nR_n(D) + lR_l(D).$$

Using [2, Lemma 4A.2] for subadditive sequences, we obtain

$$\inf_n R_n(D) = \lim_{n\to\infty} R_n(D).$$

■

APPENDIX B
PROOF OF LEMMA 4

In this appendix, we prove Lemma 4, which provides that the following mathematical expression for the feed-forward rate distortion $R^{(I)}(D)$ is equal to

$$\lim_{n\to\infty} \frac{1}{n} \min_{p(\hat{x}^n|x^n):\mathbb{E}\left[d(X^n,\hat{X}^n)\right]\leq D} I(\hat{X}^n \to X^n) \tag{41}$$

is a lower bound to the operational definition $R(D)$.

Proof: Consider any $(n, 2^{nR}, D)$ rate distortion with feed-forward code defined by the mappings $f,\ \{g_i\}_{i=1}^n$, as given in (4) in Section II, and distortion constraint $\mathbb{E}\left[d(X^n, \hat{X}^n)\right] \leq D + \epsilon_n$, where $\epsilon_n \to 0$ as n goes to infinity. Let the message sent be a random variable $T = f(X^n)$, and assume that the distortion constraint is satisfied. Then, we have the following chain of inequalities:

$$\begin{aligned} nR &\overset{(a)}{\geq} H(T) \\ &\geq I(X^n; T) \\ &\overset{(b)}{=} \sum_{i=1}^n I(X_i; T|X^{i-1}) \\ &= \sum_{i=1}^n \left(H(X_i|X^{i-1}) - H(X_i|X^{i-1}, T)\right) \\ &\overset{(c)}{=} \sum_{i=1}^n \left(H(X_i|X^{i-1}) - H(X_i|X^{i-1}, T, \hat{X}^i)\right) \\ &\overset{(d)}{\geq} \sum_{i=1}^n \left(H(X_i|X^{i-1}) - H(X_i|X^{i-1}, \hat{X}^i)\right) \\ &= \sum_{i=1}^n I(X_i; \hat{X}^i|X^{i-1}) \\ &\overset{(e)}{=} I(\hat{X}^n \to X^n) \end{aligned}$$

where (a) follows from the fact that the alphabet of T is nR, (b) follows from the chain rule for mutual information, (c) is due to the fact that given X^{i-1}, T, we know $\hat{X}^i$, and (d) is since conditioning reduces the entropy. Step (e) follows the chain rule for directed information. Taking n to infinity, we obtain $R \geq R^{(I)}(D)$ and the distortion constraint satisfies

$$\limsup_{n\to\infty} \mathbb{E}\left[d(X^n, \hat{X}^n)\right] \leq D.$$

■

APPENDIX C
PROOF OF LEMMA 11

In this appendix, we prove the existence of a sequence of upper and lower bounds to $R_n(D)$, the rate distortion function with feed forward. These bounds correspond to an iteration in Algorithm 1 and both converge to $R_n(D)$. To this end, we present and prove a few supplementary claims that assist in achieving our main goal. Theorem 6 provides an alternating form (Lagrange dual form) of an optimization problem achieving $R_n(D)$ that is proved in Appendix D. In Lemma C1, we show that in each iteration, we can obtain measures that satisfy the constraint in Theorem 6 to form a lower bound and that the bound is tight and achieved as the upper bound converges. We also provide a proof for the existence of an upper bound in each iteration.

Before we begin, we recall that a step in Algorithm 1 is defined by the following equality:

$$r^k(\hat{x}^n|x^n) = \frac{q^{k-1}(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}'^n} q^{k-1}(\hat{x}'^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}'^n)}}. \quad (42)$$

We shall use this equality throughout the proof.

As mentioned, we use Theorem 6 that provides us with the following alternating optimization problem:

$$\max_{\lambda\geq 0,\gamma(x^n)} \frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n)\log\gamma(x^n)\right) \quad (43)$$

where $\gamma(x^n)$ satisfies the inequality constraint

$$p(x^n)\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)} \leq p'(x^n\|\hat{x}^n) \quad (44)$$

for some causal conditioned probability $p'(x^n\|\hat{x}^n)$.

We now show that in each iteration in Algorithm 1, choosing $\gamma(x^n)$ appropriately forms a lower bound for $R_n(D)$. In the kth iteration of Algorithm 1, let

$$\gamma'^k(x^n) = \left(\sum_{\hat{x}^n} q^{k-1}(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\right)^{-1} \quad (45)$$

$$c^k_{\hat{x}^n,x^{n-1}} = \frac{q^k(\hat{x}^n\|x^{n-1})}{q^{k-1}(\hat{x}^n\|x^{n-1})} \quad (46)$$

and define

$$\gamma^k(x^n) = \frac{\gamma'^k(x^n)}{\max_{\hat{x}^n,x^{n-1}} c^k_{\hat{x}^n,x^{n-1}}}. \quad (47)$$

Lemma C1: With the aforementioned quantities, the constraint in (44) is satisfied and forms a tight lower bound given by

$$R_n(D) \geq \frac{1}{n}\Bigg(-\lambda D + \sum_{x^n} p(x^n)\log\gamma^k(x^n) - \log\max_{\hat{x}^n,x^{n-1}} c^k_{\hat{x}^n,x^{n-1}}\Bigg).$$

Proof: Let us fix the parameter $\gamma'^k(x^n)$, as in (45). Hence

$$\begin{aligned} p(x^n)\gamma'^k(x^n)2^{-\lambda d(x^n,\hat{x}^n)} &= p(x^n)\frac{2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n} q^{k-1}(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}} \\ &\overset{(a)}{=} \frac{p(x^n)r^k(\hat{x}^n|x^n)}{q^{k-1}(\hat{x}^n||x^{n-1})} \\ &\overset{(b)}{=} \frac{p'(x^n||\hat{x}^n)q^k(\hat{x}^n||x^{n-1})}{q^{k-1}(\hat{x}^n||x^{n-1})} \\ &\leq p'(x^n||\hat{x}^n)\max_{\hat{x}^n,x^{n-1}}\frac{q^k(\hat{x}^n||x^{n-1})}{q^{k-1}(\hat{x}^n||x^{n-1})} \end{aligned}$$

where (a) follows from the definition of a step in Algorithm 1 and is given previously in (42), and (b) follows the chain rule of causal conditioning and

$$p'(x^n\|\hat{x}^n) = \frac{p(x^n)r^k(\hat{x}^n|x^n)}{q^k(\hat{x}^n||x^{n-1})}$$

is a causal conditioned PMF. Hence, combined with (47), we obtain

$$\begin{aligned} p(x^n)\gamma^k(x^n)2^{-\lambda d(x^n,\hat{x}^n)} &= \frac{p(x^n)\gamma'(x^n)2^{-\lambda d(x^n,\hat{x}^n)}}{\max_{\hat{x}^n,x^{n-1}} c^k_{\hat{x}^n,x^{n-1}}} \\ &\leq p'(x^n\|\hat{x}^n). \end{aligned}$$

Thus, we can use Theorem 6 and obtain a lower bound for $R_n(D)$, i.e.,

$$\begin{aligned} R_n(D) &\geq \frac{1}{n}\left[-\lambda D + \sum_{x^n} p(x^n)\log\gamma^k(x^n)\right] \\ &= \frac{1}{n}\Bigg[-\lambda D + \sum_{x^n} p(x^n)\log\gamma'^k_{x^n} \\ &\qquad - \sum_{x^n} p(x^n)\log\left(\max_{\hat{x}^n,x^{n-1}} c^k_{\hat{x}^n,x^{n-1}}\right)\Bigg] \\ &= \frac{1}{n}\Bigg[-\lambda D + \sum_{x^n} p(x^n)\log\gamma'^k(x^n) \\ &\qquad - \log\left(\max_{\hat{x}^n,x^{n-1}} c^k_{\hat{x}^n,x^{n-1}}\right)\Bigg]. \end{aligned} \quad (48)$$

To complete the proof of this lemma, we are left to show that as k increases, i.e., the upper bound converges to $R_n(D)$, the lower bound is tight. On this matter, we note that the PMFs that achieve the optimum value $q^*,\ r^*$ are unique (This is shown in Lemma 10 in Section VI). Thus, it is clear that

$$c^*_{\hat{x}^n,x^{n-1}} = \frac{q^*(\hat{x}^n\|x^{n-1})}{q^*(\hat{x}^n\|x^{n-1})} = 1, \quad (49)$$

and

$$\begin{aligned} \gamma^k(x^n) &= \gamma'^k(x^n) \\ &= \left(\sum_{\hat{x}^n} q^*(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\right)^{-1}. \end{aligned} \quad (50)$$

Placing (50) and (49) in (48), as shown in Theorem 6, achieves equality instead of the chain of inequalities given. Thus, $R_n(D)$

is, in fact, the solution to the optimization problem given in (43) and we have demonstrated the existence of the lower bound. ■

Lemma C2: In the kth iteration in Algorithm 1, the upper bound to the rate distortion is given by

$$R_n(D_k) \leq \frac{1}{n}\Bigg(-\lambda D_k + \sum_{x^n} p(x^n)\log\gamma^k(x^n) - \sum_{x^n} p(x^n) r^k(\hat{x}^n|x^n)\log c^k_{\hat{x}^n,x^{n-1}}\Bigg)$$

where $D_k = \mathbb{E}_{r^k}\left[d(X^n,\hat{X}^n)\right]$.

Proof: Note that if $r^k(\hat{x}^n,x^n)$ produces a distortion D, then

$$\begin{aligned}
&nR_n(D)\\
&\leq I_{FF}(r^k,q^k)\\
&= \sum_{x^n,\hat{x}^n} p(x^n) r^k(\hat{x}^n|x^n)\log\frac{r^k(\hat{x}^n|x^n)}{q^k(\hat{x}^n\|x^{n-1})}\\
&\overset{(a)}{=} \sum_{x^n,\hat{x}^n} p(x^n) r^k(\hat{x}^n|x^n)\cdot\\
&\log\frac{q^{k-1}(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{q^k(\hat{x}^n||x^{n-1})\sum_{\hat{x}'^n} q^{k-1}(\hat{x}'^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}'^n)}}\\
&= -\lambda\mathbb{E}_{r^k}\left[d(X^n,\hat{X}^n)\right]\\
&- \sum_{x^n} p(x^n)\log\sum_{\hat{x}'^n} q^{k-1}(\hat{x}'^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}'^n)}\\
&- \sum_{x^n,\hat{x}^n} p(x^n) r^k(\hat{x}^n|x^n)\log\frac{q^k(\hat{x}^n\|x^{n-1})}{q^{k-1}(\hat{x}^n\|x^{n-1})}\\
&\overset{(b)}{=} -\lambda D_k + \sum_{x^n} p(x^n)\log\gamma^k(x^n)\\
&- \sum_{x^n,\hat{x}^n} p(x^n) r^k(\hat{x}^n|x^n)\log c^k_{\hat{x}^n,x^{n-1}} \qquad (51)
\end{aligned}$$

where (a) follows from the definition of a step in Algorithm 1 and is given previously in (42), and (b) follows from the definition of $\gamma^k(x^n)$, $c^k_{\hat{x}^n,x^{n-1}}$. Hence, we have formed an upper bound to the rate distortion, as in the lemma. Note that the only inequality is in the first line of the chain and is due to the fact that $I_{FF}(r^k,q^k) \geq \min_{r,q} I_{FF}(r,q)$. However, upon convergence, this inequality is tight. ■

We can now conclude our main objective in this appendix.

Proof of Lemma 11: Proving this lemma requires us to present upper and lower bounds that converge to $R_n(D)$. Lemma C1 provides us with a lower bound and its tightness, whereas Lemma C2 provides us with a tight upper bound. ■

Appendix D
Proof of Theorem 6

In this appendix, we provide a proof for Theorem 6. We recall that Theorem 6 states that the rate distortion function is the solution to the following optimization problem:

$$\max_{\lambda\geq 0,\gamma(x^n)} \frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n)\log\gamma(x^n)\right) \qquad (52)$$

where, for some causal conditioned probability $p'(x^n\|\hat{x}^n)$, $\gamma(x^n)$ satisfies the inequality constraint

$$p(x^n)\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)} \leq p'(x^n\|\hat{x}^n). \qquad (53)$$

In the rest of this appendix, we provide two proofs for this theorem, as mentioned in Section VII. We also provide the connection between the curve of $R_n(D)$ and the parameter λ; this is embodied in Lemma D1.

Before we begin, we refer the reader to Section V. There, we can see that a step in Algorithm 1 is defined by the following equality:

$$r^k(\hat{x}^n|x^n) = \frac{q^{k-1}(\hat{x}^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}'^n} q^{k-1}(\hat{x}'^n||x^{n-1})2^{-\lambda d(x^n,\hat{x}'^n)}} \qquad (54)$$

where r^k is a conditional probability converges, using Algorithm 1, to $R_n(D)$. This equality is the outcome of differentiating the Lagrangian when $q(\hat{x}^n\|x^{n-1})$ is fixed, as given in Section VI. We shall use this equality throughout the proof. Further, we write the directed information as $I_{FF}(r,q)$ for short

$$I_{FF}(r,q) = \sum_{x^n,\hat{x}^n} p(x^n) r(\hat{x}^n|x^n)\log\frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n\|x^{n-1})}$$

where q is the causal conditional probability taking part in the algorithm.

As mentioned, the first proof follows the one in [20].

Proof of Theorem 6: First, we show that for every $r(\hat{x}^n|x^n)$ for which the distortion constraint is satisfied, the following chain of inequalities holds:

$$\begin{aligned}
&I_{FF}(r,q) + \lambda D - \sum_{x^n} p(x^n)\log\gamma(x^n)\\
&\overset{(a)}{\geq} I_{FF}(r,q) + \lambda\mathbb{E}_{r(\hat{x}^n|x^n)}\left[d(X^n,\hat{X}^n)\right]\\
&\quad - \sum_{x^n} p(x^n)\log\gamma(x^n)\\
&= \sum_{x^n,\hat{x}^n} p(x^n) r(\hat{x}^n|x^n)\log\frac{r(\hat{x}^n|x^n)2^{\lambda d(x^n,\hat{x}^n)}}{q(\hat{x}^n||x^{n-1})\gamma(x^n)}\\
&\overset{(b)}{\geq} \sum_{x^n,\hat{x}^n} p(x^n) r(\hat{x}^n|x^n)\left(1 - \frac{q(\hat{x}^n\|x^{n-1})\gamma(x^n)}{r(\hat{x}^n|x^n)2^{\lambda d(x^n,\hat{x}^n)}}\right)\\
&= 1 - \sum_{x^n,\hat{x}^n} q(\hat{x}^n||x^{n-1})p(x^n)\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)}\\
&\overset{(c)}{\geq} 1 - \sum_{x^n,\hat{x}^n} q(\hat{x}^n||x^{n-1})p'(x^n||\hat{x}^n)\\
&\overset{(d)}{=} 0
\end{aligned}$$

where (a) follows from the fact that the distortion D exceeds $\mathbb{E}_{r(\hat{x}^n|x^n)}\left[d(X^n,\hat{X}^n)\right]$ for every $r(\hat{x}^n|x^n)$, as has been assumed, (b) follows from the inequality $\log\frac{1}{y} \geq 1 - \frac{1}{y}$, (c) is due to the constraint in (53), and (d) follows from the fact that $q(\hat{x}^n\|x^{n-1})p'(x^n\|\hat{x}^n)$ is equal to some joint distribution $p(x^n,\hat{x}^n)$ [6]. Since the chain of inequalities is true for every $r(\hat{x}^n|x^n)$, we can choose the one that achieves $R_n(D)$ and then divide by n to obtain the inequality in (52) in our Theorem.

To complete the proof of Theorem 6, we need to show that equality holds in the chain of inequalities above for some $\gamma(x^n)$ that satisfies the constraint. If so, let us denote by $r^*(\hat{x}^n|x^n)$ the conditional PMF that achieves $R_n(D)$. Further, we denote by $q^*(\hat{x}^n\|x^{n-1})$ the corresponding causal conditioned PMF. Now, consider the following chain of equalities:

$$\begin{aligned} nR_n(D) &= \sum_{x^n,\hat{x}^n} p(x^n) r^*(\hat{x}^n|x^n) \log \frac{r^*(\hat{x}^n|x^n)}{q^*(\hat{x}^n\|x^{n-1})} \\ &\overset{(a)}{=} \sum_{x^n,\hat{x}^n} p(x^n) r^*(\hat{x}^n|x^n) \cdot \\ &\quad \log \frac{2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}'^n} q^*(\hat{x}'^n||x^{n-1}) 2^{-\lambda d(x^n,\hat{x}'^n)}} \\ &\overset{(b)}{=} -\lambda \mathbb{E}_{r^k(\hat{x}^n|x^n)}\left[d(X^n,\hat{X}^n)\right] + \sum_{x^n} p(x^n) \log \gamma(x^n) \\ &= -\lambda D + \sum_{x^n} p(x^n) \log \gamma(x^n) \end{aligned}$$

where (a) is due to a step in the algorithm given by (54) and by the uniqueness of $r^*(\hat{x}^n|x^n)$ in the algorithm (again, we refer the reader to Lemma 10 in Section VI, where this is shown) and (b) follows the expression for $\gamma(x^n)$ given by

$$\gamma(x^n) = \left(\sum_{\hat{x}'^n} q^*(\hat{x}'^n\|x^{n-1}) 2^{-\lambda d(x^n,\hat{x}'^n)} \right). \tag{55}$$

Therefore, we are left to verify that the $\gamma(x^n)$ above satisfies the constraint

$$\begin{aligned} &p(x^n)\gamma(x^n) 2^{-\lambda d(x^n,\hat{x}^n)} \\ &\quad = p(x^n) \frac{2^{-\lambda d(x^n,\hat{x}^n)}}{\sum_{\hat{x}^n} q^*(\hat{x}^n\|x^{n-1}) 2^{-\lambda d(x^n,\hat{x}^n)}} \\ &\quad \overset{(a)}{=} \frac{p(x^n) r^*(\hat{x}^n|x^n)}{q^*(\hat{x}^n||x^{n-1})} \\ &\quad = \frac{p(x^n,\hat{x}^n)}{q^*(\hat{x}^n||x^{n-1})} \\ &\quad \overset{(b)}{=} p'(x^n||\hat{x}^n) \end{aligned}$$

where (a) follows from (54) and (b) is due to the causal conditioning chain rule. Hence, we showed that $R_n(D)$ is the solution to the optimization problem given in (52). ■

We also present an alternative proof for Theorem 6, this using the Lagrange duality, as in [13] and [21].

Alternative Proof for Theorem 6: Recall that $R_n(D)$ is the result of

$$\min_{r(\hat{x}^n|x^n)} \sum_{\hat{x}^n,x^n} p(x^n) r(\hat{x}^n|x^n) \log \frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n||x^{n-1})}$$

where $q(\hat{x}^n\|x^{n-1})$ is defined by $p(x^n)r(\hat{x}^n|x^n)$, subject to the following conditions:

$$\begin{aligned} &\sum_{x^n,\hat{x}^n} p(x^n) r(\hat{x}^n|x^n) d(x^n,\hat{x}^n) \le D \\ &\forall\, x^n : \sum_{\hat{x}^n} r(\hat{x}^n|x^n) = 1 \\ &\forall\, x^n,\hat{x}^n : r(\hat{x}^n|x^n) \ge 0. \end{aligned}$$

Let us define the Lagrangian as

$$\begin{aligned} J(r,\lambda,\gamma,\mu) &= \sum_{x^n,\hat{x}^n} p(x^n) r(\hat{x}^n|x^n) \log \frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n\|x^{n-1})} \\ &\quad + \lambda \left(\sum_{x^n,\hat{x}^n} p(x^n) r(\hat{x}^n|x^n) d(x^n,\hat{x}^n) - D \right) \\ &\quad + \sum_{x^n} \gamma(x^n) \left(\sum_{\hat{x}^n} r(\hat{x}^n|x^n) - 1 \right) \\ &\quad - \sum_{x^n,\hat{x}^n} \mu(x^n,\hat{x}^n) r(\hat{x}^n|x^n) \end{aligned}$$

where $\mu(x^n,\hat{x}^n) \ge 0$ for all $x^n,\hat{x}^n$. Differentiating the Lagrangian, $J(r,\lambda,\gamma,\mu)$, over the variable $r(\hat{x}^n|x^n)$, we obtain

$$\begin{aligned} \frac{\partial J}{\partial r(\hat{x}^n|x^n)} &= p(x^n) \log \frac{r(\hat{x}^n|x^n)}{q(\hat{x}^n\|x^{n-1})} \\ &\quad + \lambda p(x^n) d(x^n,\hat{x}^n) + \gamma(x^n) - \mu(x^n,\hat{x}^n). \end{aligned}$$

Solving the equation $\frac{\partial J}{\partial r(\hat{x}^n|x^n)} = 0$, in order to find the optimum value yields the following expression:

$$r(\hat{x}^n|x^n) = q(\hat{x}^n||x^{n-1}) \gamma'(x^n) 2^{\frac{\mu(x^n,\hat{x}^n)}{p(x^n)} - \lambda d(x^n,\hat{x}^n)} \tag{56}$$

where $\gamma'(x^n) = 2^{-\frac{\gamma(x^n)}{p(x^n)}}$. Multiplying both sides by $\frac{p(x^n)}{q(\hat{x}^n\|x^{n-1})}$, we are left with the constraint

$$\begin{aligned} p(x^n\|\hat{x}^n) &= p(x^n)\gamma'(x^n) 2^{\frac{\mu(x^n,\hat{x}^n)}{p(x^n)} - \lambda d(x^n,\hat{x}^n)} \\ &\ge p(x^n)\gamma'(x^n) 2^{-\lambda d(x^n,\hat{x}^n)} \end{aligned} \tag{57}$$

where $p(x^n\|\hat{x}^n)$ is induced by $r(\hat{x}^n|x^n)p(x^n)$.

From [13, Ch. 5.1.3], we know that $g(\lambda,\gamma,\mu) = J(r^*,\lambda,\gamma,\mu)$ is a lower bound to $R_n(D)$. Substituting the minimizer $r(\hat{x}^n|x^n)$ using (56) and the condition given by (57) into J, we obtain the Lagrange dual function given in (58) shown at the bottom of the page. By making the constraints explicit, and since the minimization problem is convex, we obtain the Lagrange dual problem, i.e., $R_n(D)$ is the solution to

$$\max_{\gamma(x^n),\lambda} \frac{1}{n} \left(-\lambda D + \sum_{x^n} p(x^n) \log \gamma(x^n) \right) \tag{59}$$

$$g(\lambda,\gamma') = \begin{cases} -\lambda D + \sum_{x^n} p(x^n) \log \gamma'(x^n), & p(x^n)\gamma'(x^n) 2^{-\lambda d(x^n,\hat{x}^n)} \le p(x^n\|\hat{x}^n) \\ -\infty, & \text{otherwise} \end{cases} \tag{58}$$

subject to

$$\forall\, x^n, \hat{x}^n : p(x^n)\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)} \le p(x^n\|\hat{x}^n)$$
$$\lambda \ge 0$$

for the $p(x^n\|\hat{x}^n)$ that is induced by $r(\hat{x}^n|x^n)p(x^n)$ and $r(\hat{x}^n|x^n)$ is the optimal PMF.

We use the notion of an *optimal* PMF if it achieves the optimal value. For example, the PMF $r(\hat{x}^n|x^n)$ achieves the minimum of the directed information, given the distortion constraint is optimal. We say that the PMF $p(x^n\|\hat{x}^n)$ is optimal if it is induced by the optimal $r(\hat{x}^n|x^n)$. Another example is the maximization problem in (59). We say that $\lambda, \gamma(x^n)$ are optimal if they achieve the maximum value. Therefore, $p(x^n\|\hat{x}^n)$ is optimal as well if it satisfies (57).

Now, we wish to transform the constraint to

$$\forall\, x^n, \hat{x}^n : p(x^n)\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)} \le p'(x^n\|\hat{x}^n) \tag{60}$$

for some $p'(x^n\|\hat{x}^n)$. First, note that we always achieve equality in (60) since we can increase the value of $\gamma(x^n)$ and thus increase the objective. This, combined with the fact that for $r(\hat{x}^n|x^n) > 0$, $\mu(x^n, \hat{x}^n)$ must be zero, we have equality in (57) as well (if $r(\hat{x}^n|x^n) = 0$, then $q(\hat{x}^n\|x^{n-1}) = 0$ and (56) holds too). Now, let us assume that the maximum in (59) with the constraint in (60) is achieved at a *nonoptimal* $p'(x^n\|\hat{x}^n)$, i.e., one that is not achieved using the optimal $\lambda, \gamma(x^n)$. Thus, the value obtained in (59) is larger than the value achieved by $p(x^n\|\hat{x}^n)$, i.e., $R_n(D)$ (since the maximization includes $p(x^n\|\hat{x}^n)$). However, from the Lagrange duality, it should be a lower bound to $R_n(D)$, thus contradicting the fact that the maximum is achieved at a nonoptimal $p'(x^n\|\hat{x}^n)$. ■

Note that we can construct the optimal PMF $r(\hat{x}^n|x^n)$ from the solution to the maximization problem presented here. Consider the parameters $\lambda,\ \gamma(x^n)$, that achieve (59), and calculate $p(x^n\|\hat{x}^n)$ according to (57). The calculation of $r(\hat{x}^n|x^n)$ is done recursively on $r(\hat{x}^i|x^i)$. For $i = 1$, calculate $r(\hat{x}^1|x^1)$ using

$$r(\hat{x}^1|x^1) = \frac{p(x^1||\hat{x}^1)}{p(x^1)}\sum_{x_1} p(x^1)r(\hat{x}^1|x^1).$$

Further, calculate $q(\hat{x}_1)$ using

$$q(\hat{x}_1) = \sum_{x_1} p(x^1)r(\hat{x}^1|x^1).$$

Now, once we have $r(\hat{x}^j|x^j),\ q(\hat{x}_j|\hat{x}^{j-1}x^{j-1})$ for every $j < i$, calculate $r(\hat{x}^i|x^i)$ using

$$r(\hat{x}^i|x^i) = \frac{p(x^i||\hat{x}^i)}{p(x^i)}\left[\prod_{j=1}^{i-1} q(\hat{x}_j|\hat{x}^{j-1}x^{j-1})\right]\cdot \frac{\sum_{x_i} p(x^i)r(\hat{x}^i|x^i)}{p(x^{i-1})r(\hat{x}^{i-1}|x^{i-1})}$$

and then

$$q(\hat{x}_i|\hat{x}^{i-1}x^{i-1}) = \frac{\sum_{x_i} p(x^i)r(\hat{x}^i|x^i)}{p(x^{i-1})r(\hat{x}^{i-1}|x^{i-1})}.$$

Continue this until $i = n$, and obtain our optimal $r(\hat{x}^n|x^n)$.

Another lemma that we wish to provide is the connection between the curve of $R_n(D)$ and the parameter λ. This lemma is similar to the one given by Berger in [20, Th. 2.5.1] for the case of no feed forward. Consider the expression for $R_n(D)$ given by

$$\frac{1}{n}\left(-\lambda D + \sum_{x^n} p(x^n)\log\gamma(x^n)\right)$$

where $\gamma(x^n)$ and λ are the variables that maximize (59). Recall that $\gamma(x^n)$ is of the form

$$\gamma(x^n) = \left(\sum_{\hat{x}^n} q^*(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\right)^{-1}.$$

Lemma D1: The slope at distortion D is $R_n'(D) = -\frac{\lambda}{n}$.

Proof: The proof is given simply by differentiating the expression for $R_n(D)$

$$\begin{aligned}\frac{dR_n}{dD} &= \frac{\partial R_n}{\partial D} + \frac{\partial R_n}{\partial\lambda}\frac{d\lambda}{dD} + \sum_{x^n}\frac{\partial R_n}{\partial\gamma(x^n)}\frac{d\gamma(x^n)}{dD}\\ &= \frac{1}{n}\left[-\lambda - D\frac{d\lambda}{dD} + \sum_{x^n}\frac{p(x^n)}{\gamma(x^n)}\frac{d\gamma(x^n)}{dD}\right]\\ &= -\frac{\lambda}{n} + \frac{1}{n}\left[-D + \sum_{x^n}\frac{p(x^n)}{\gamma(x^n)}\frac{d\gamma(x^n)}{d\lambda}\right]\frac{d\lambda}{dD}.\end{aligned}$$

Now, consider the following expression:

$$F = \sum_{x^n,\hat{x}^n} p(x^n)q^*(\hat{x}^n\|x^{n-1})\gamma(x^n)2^{-\lambda d(x^n,\hat{x}^n)}.$$

Using the $\gamma(x^n)$ given previously, we have $F = 1$, and thus, $\frac{\partial F}{\partial\lambda} = 0$. However

$$\begin{aligned}\frac{\partial F}{\partial\lambda} &= \sum_{x^n,\hat{x}^n}\left[\frac{d\gamma(x^n)}{d\lambda} - d(x^n,\hat{x}^n)\gamma(x^n)\right]\cdot\\ &\quad p(x^n)q^*(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\\ &= \sum_{x^n}\frac{d\gamma(x^n)}{d\lambda}p(x^n)\sum_{\hat{x}^n} q^*(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\\ &\quad - \sum_{x^n,\hat{x}^n} p(x^n)q^*(\hat{x}^n\|x^{n-1})2^{-\lambda d(x^n,\hat{x}^n)}\gamma(x^n)d(x^n,\hat{x}^n)\\ &= \sum_{x^n}\frac{d\gamma(x^n)}{d\lambda}\frac{p(x^n)}{\gamma(x^n)} - \sum_{x^n,\hat{x}^n} p(x^n)r^*(\hat{x}^n|x^n)d(x^n,\hat{x}^n)\\ &= \sum_{x^n}\frac{d\gamma(x^n)}{d\lambda}\frac{p(x^n)}{\gamma(x^n)} - D\\ &= 0.\end{aligned}$$

Hence, we can conclude that

$$\begin{aligned}\frac{dR_n}{dD} &= -\frac{\lambda}{n} + \frac{1}{n}\left[-D + \sum_{x^n}\frac{p(x^n)}{\gamma(x^n)}\frac{d\gamma(x^n)}{d\lambda}\right]\frac{d\lambda}{dD}\\ &= -\frac{\lambda}{n}.\end{aligned}$$

■

APPENDIX E
SOLUTION TO $R(D)$ FOR AN ASYMMETRICAL MARKOV SCHEME

The Markov source is presented in Fig. 5 above. We can describe the process $\{X_i\}$ using the equation

$$\begin{aligned} X_i &= X_{i-1}W_1 + (1 - X_{i-1})W_2 \\ &= (X_{i-1}(W_1 \oplus W_2)) \oplus W_2 \end{aligned}$$

where $W_1 \sim B(q)$, $W_2 \sim B(p)$. This allows us to evaluate $H(X_n|X_{n-1})$

$$\begin{aligned} H(X_n|X_{n-1}) &= H((X_{n-1}(W_1 \oplus W_2)) \oplus W_2|X_{n-1}) \\ &= p(x_{n-1} = 1)H(W_1 \oplus W_2 \oplus W_2) \\ &\quad + p(x_{n-1} = 0)H(W_2) \\ &= \pi_1 H(W_1) + \pi_2 H(W_2) \end{aligned}$$

where π is the stationary distribution of the source. Now, to find the rate distortion of this model, we start with the converse

$$\begin{aligned} \frac{1}{n} I(\hat{X}^n \to X^n) &= H(X^n) - H(X^n \| \hat{X}^n) \\ &= \frac{1}{n} H(X_1) + \frac{n-1}{n} H(X_n|X_{n-1}) \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} H(X_i|X^{i-1}, \hat{X}^i) \\ &\overset{(a)}{\geq} \frac{1}{n} H_b(\pi) + \frac{n-1}{n} H(X_n|X_{n-1}) \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} H(X_i|\hat{X}_i) \\ &\overset{(b)}{\geq} \frac{1}{n} H_b(\pi) + \frac{n-1}{n} H(X_n|X_{n-1}) - H_b(D) \\ &= \frac{1}{n} H_b(\pi) + \frac{n-1}{n} (\pi_1 H_b(p) + \pi_2 H_b(q)) \\ &\quad - H(D) \end{aligned}$$

where (a) follows from the fact that conditioning reduces entropy and (b) follows from the fact that $P(X_i \neq \hat{X}_i) \leq D$ and that $H_b(D)$ increases with D for $D \leq \frac{1}{2}$.

However, this lower bound can be achieved by letting X_i depend on $\hat{X}_i$ and X_{i-1}, as in Fig. 10, where $p_1,\ p_2$ must hold for the following equations:

$$\begin{aligned} p_1 D + (1 - p_1)(1 - D) &= 1 - p \\ p_2 D + (1 - p_2)(1 - D) &= 1 - q \end{aligned}$$

i.e.,

$$\begin{aligned} p_1 &= \frac{D - p}{2D - 1} \\ p_2 &= \frac{D - q}{2D - 1}. \end{aligned}$$

Note that under this construction, the source X^n is still Markovian. Further, from Fig. 10, we can see that $X_{i-1} - \hat{X}_i - X_i$ forms a Markov chain and $H(X_i|\hat{X}_i) = H_b(D)$. Thus, we obtain equality in (a) and (b) in the aforementioned chain of inequalities and hence show that

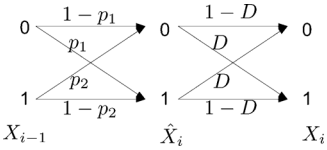

Fig. 10. Distribution of X_i given X_{i-1} and $\hat{X}_i$.

$$R_n(D) = \frac{1}{n} H_b(\pi) + \frac{n-1}{n} (\pi_1 H_b(p) + \pi_2 H_b(q)) - H_b(D).$$

By taking n to infinity, we obtain

$$R(D) = \pi_1 H_b(p) + \pi_2 H_b(q) - H_b(D).$$

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and anonymous referees for their comments, which significantly helped to improve the content and the organization of the paper. The authors gratefully acknowledge Yossef Steinberg for very helpful discussions.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
[2] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[3] T. Weissman and N. Merhav, "On competitive prediction and its relation to rate-distortion theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3185–3194, Dec. 2003.
[4] R. Venkataramanan and S. S. Pradhan, "Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 2154–2179, Jun. 2007.
[5] R. Venkataramanan and S. S. Pradhan, "On computing the feedback capacity of channels and the feed-forward rate-distortion function of sources," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 1889–1896, Jul. 2010.
[6] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Appl.*, Nov. 1990, pp. 303–305.
[7] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, Jan. 2003.
[8] H. Marko, "The bidirectional communication theory—a generalization of information theory," *IEEE Trans. Commun.*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.
[9] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, Jan. 2009.
[10] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, Feb. 2009.
[11] Y. H. Kim, "Feedback capacity of stationary Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 57–85, Jan. 2010.
[12] I. Naiss and H. H. Permuter, "Extension of the Blahut–Arimoto algorithm for maximizing directed information," IEEE Trans. Inf. Theory 2010, CoRR, abs/1012.5071.
[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.
[14] M. Grant and S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, Version 1.21 Apr. 2011 [Online]. Available: http://cvxr.com

[15] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Budapest, Hungary: Akademiai Kiado, 1986.
[16] I. Csiszár and G. Tusnady, "Information geometry and alternating minimization procedures," *Statist. Decis., Suppl.*, no. 1, pp. 205–237, 1984.
[17] R. Blahut, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.
[18] S. Arimoto, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.
[19] R. W. Yeung, *Information Theory and Network Coding*. New York: Springer-Verlag, 2008.
[20] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[21] M. Chiang and S. P. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Trans. Inf. Theory*, vol. 50, no. 2, pp. 245–258, Feb. 2004.
[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New-York: Wiley, 2006.

Iddo Naiss received his B.Sc. (*summa cum laude*) degrees in Electrical and Computer Engineering and in Mathematics from the Ben-Gurion University, Israel, in 2010. He also received his M.Sc. degree on the fast track program for honor students in Electrical and Computer Engineering from the Ben-Gurion University, Israel, in 2012. Currently Iddo is working with Samsung-Israel on coding for flash memories.

Haim H. Permuter (M'08) received his B.Sc. (*summa cum laude*) and M.Sc. (*summa cum laude*) degree in Electrical and Computer Engineering from the Ben-Gurion University, Israel, in 1997 and 2003, respectively, and Ph.D. degrees in Electrical Engineering from Stanford University, California in 2008.

Between 1997 and 2004, he was an officer at a research and development unit of the Israeli Defense Forces. He is currently a senior lecturer at Ben-Gurion university.

Dr. Permuter is a recipient of the Fullbright Fellowship, the Stanford Graduate Fellowship (SGF), Allon Fellowship, Marie Curie Reintegration fellowship, and the 2009 U.S.–Israel Binational Science Foundation Bergmann Memorial Award.